5000 System Panels' MCQs – what have we learnt?

Curriculum Retreat 2009
Faculty of Medicine, CUHK

Background and goals of the present study

Background:

- There appears to be more failures in PHOM panel-end MCQ tests when compared with those of other system panels.

Goals:

- Whether there are differences in the psychometric properties of MCQ used in SA of different system panels?
- What is/are the indicator(s) of a MCQ test (with a pass mark of 50) that would predict the presence of failures?

Scope of the analyses

- System panel summative assessments (Year 1 to 3) since the implementation of the new medical curriculum (2001 to 2009) were included in the analysis. Data from skill panel SA (COSK, LLSK. PCLM) and PHES were not included.
- There were 154 SA involving 6475 items.

Year	Panels	No. of SA	No. of items
1 *(2001-2009)*	PCAR1, PFOS1, PGIN1, PHOM1, PHUS1, PMUS1	50	2377
2 *(2002-2009)*	PCAR2, PGIN2, PHAE2, PHOM2, PHUS2, PMUS2, PMDT2, PNEU2	56	2426
3 *(2003-2009)*	PCAR3, PGIN3, PHAE3, PHOM3, PMUS3, PMDT3, PNEU2, PREP3	48	1672

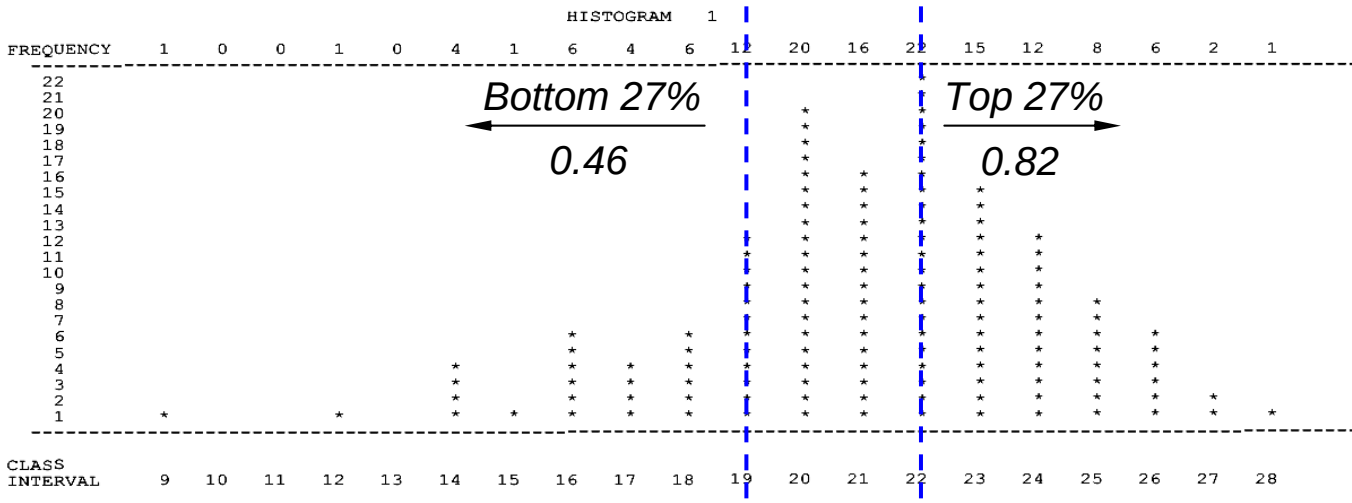
- Psychometric properties of MCQ (i.e. difficulty level, discriminating power), number of options picked and mean-2SD of students' score in each test.

Item analysis report provided by OES

```
****TEST STATISTICS****

NUMBER OF EXAMINEES          137

TOTAL SCORE:
                MEAN       21.04
            VARIANCE       10.09
  STANDARD DEVIATION        3.18
    KR-20 RELIABILITY      0.5247
S.E. OF MEASUREMENT        2.1895
```




```
ITEM   2: DIF=0.620, RPB=  0.380, CRPB=  0.228 (95% CON=  0.062, 0.381)
                    RBIS= 0.485, CRBIS= 0.290, IRI=0.185
   GROUP    N    INV    NF   OMIT      A       B       C*      D       E
   TOTAL  137     0     0     0     0.07    0.17    0.62    0.02    0.12
    HIGH   44     0           0     0.02    0.14    0.82    0.00    0.02
    MID    58     0           0     0.09    0.14    0.57    0.03    0.17
    LOW    35     0           0     0.11    0.26    0.46    0.03    0.14
   TEST SCORE MEAN %:          0       62      66      73      63      66
   DISCRIMINATING POWER        0    -0.09   -0.12    0.36   -0.03   -0.12
   STANDARD ERROR OF D.P.      0     0.06    0.09    0.11    0.03    0.06
```

0.82 - 0.46 = 0.36

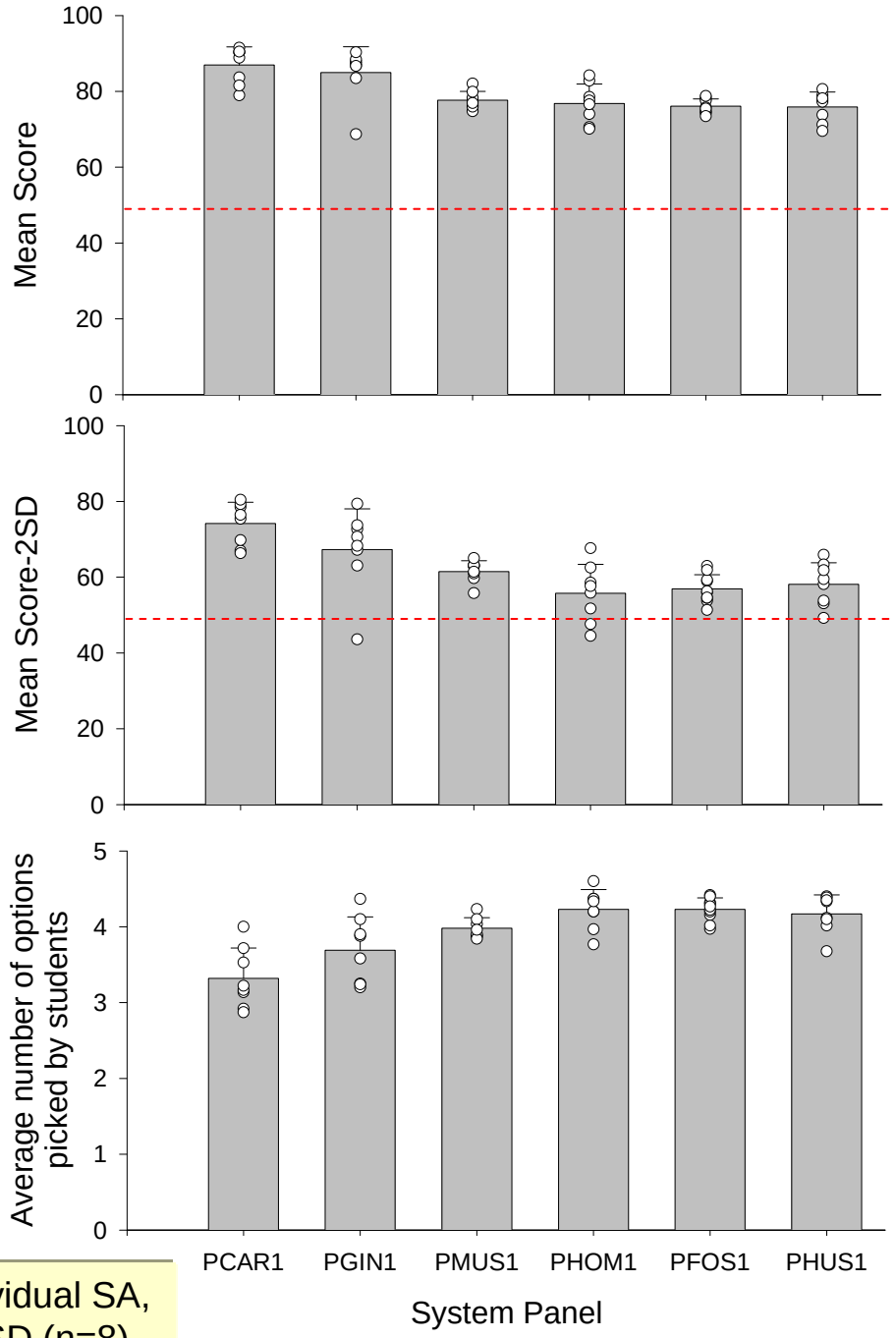
DP is defined as the difference between the proportions of the HIGH (top 27%) and LOW (bottom 27%) groups giving this response.

```
ITEM   4: DIF=1.000, RPB=  0.000, CRPB=  0.000 (95% CON= -0.174, 0.174)
                    RBIS= 0.000, CRBIS= 0.000, IRI=0.000
   GROUP    N    INV    NF   OMIT      C*
   TOTAL  137     0     0     0     1.00
    HIGH   44     0           0     1.00
    MID    58     0           0     1.00
    LOW    35     0           0     1.00
   TEST SCORE MEAN %:          0       81
   DISCRIMINATING POWER        0     0.00
   STANDARD ERROR OF D.P.      0     0.00
```

```
ITEM  16: DIF=0.934, RPB=  0.356, CRPB=  0.278 (95% CON=  0.116, 0.426)
                    RBIS= 0.691, CRBIS= 0.540, IRI=0.088
   GROUP    N    INV    NF   OMIT      C       D       E*
   TOTAL  137     0     0     0     0.02    0.04    0.93
    HIGH   44     0           0     0.00    0.00    1.00
    MID    58     0           0     0.02    0.03    0.95
    LOW    35     0           0     0.06    0.11    0.83
   TEST SCORE MEAN %:          0       60      54      71
   DISCRIMINATING POWER        0    -0.06   -0.11    0.17
   STANDARD ERROR OF D.P.      0     0.04    0.05    0.06
```

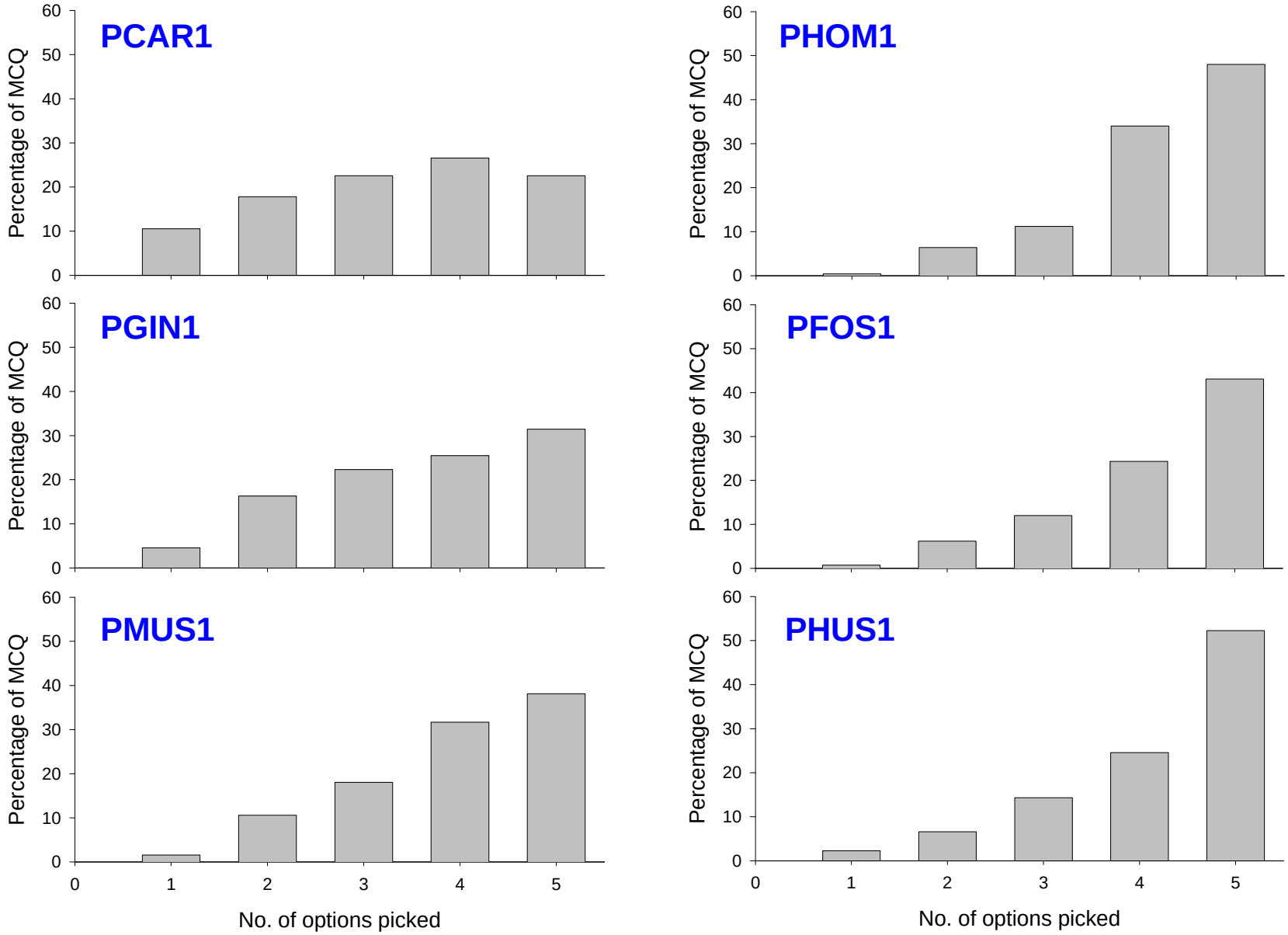
Medical Year 1

- Mean score is given by the average of the difficulty levels of MCQs used in the test.
- There are significant differences in the mean scores of SA (i.e. difficulty levels of items) offered by different system panels.
- Differences among panels become more evident when mean-2SD are compared, and more of these values are close to the 50% mark.
- Those SA with higher mean scores are made up of MCQs with fewer numbers of options picked by students.

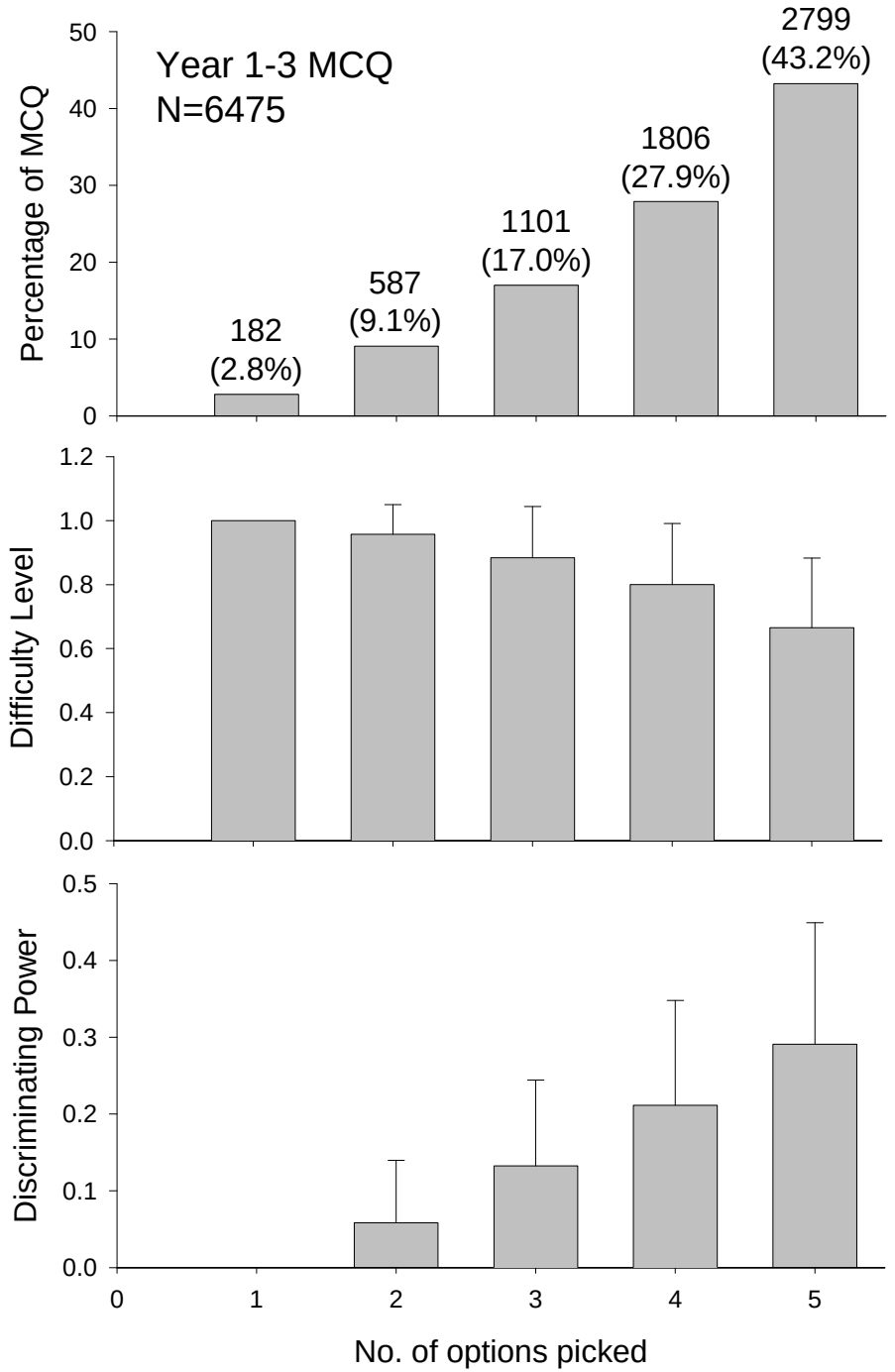


Open circle indicates the mean score of individual SA, and the bar and error bar show their mean±SD (n=8).

Percentage distribution of Year 1 Summative MCQ based on the number of options picked by students

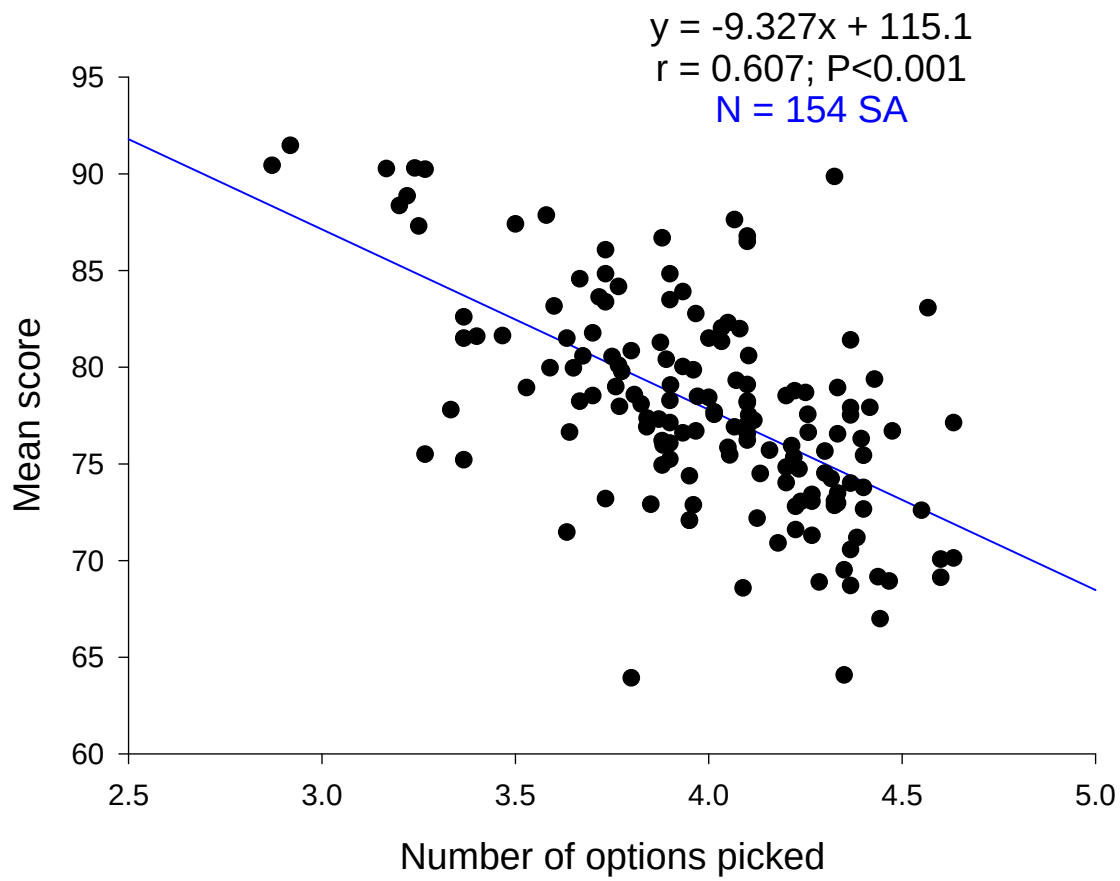


Number of usable options is related to difficulty level and discriminating power



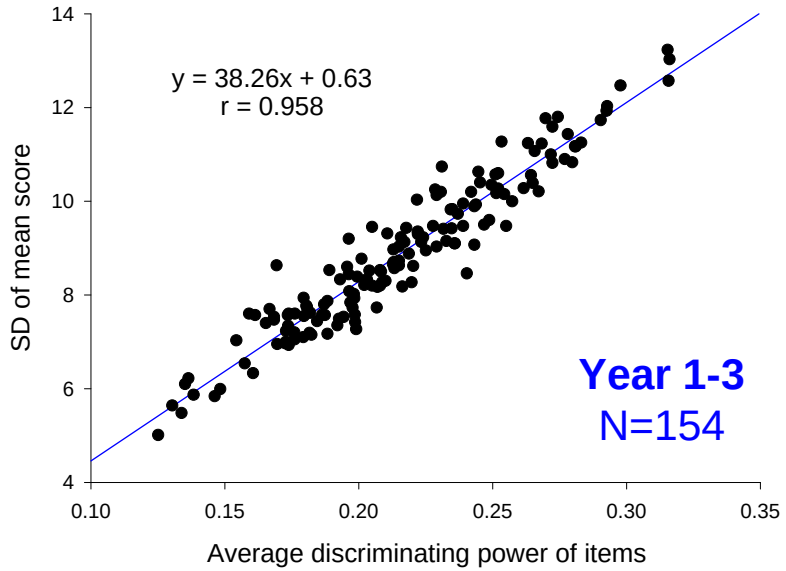
Data represents mean±SD with the number of items shown within brackets. Differences between groups are statistically significant (P<0.001), by Kruskal-Wallis one-way ANOVA on ranks.

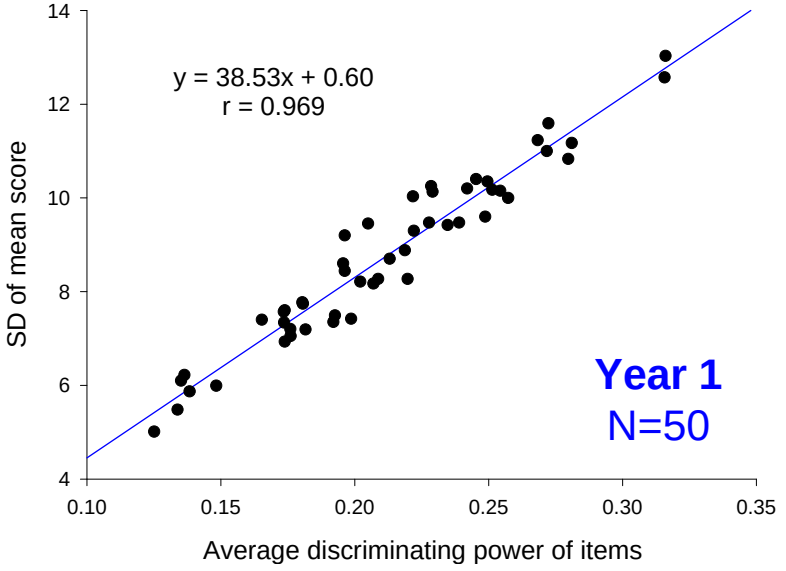
Average number of usable options in MCQ is inversely related to the mean score of the test

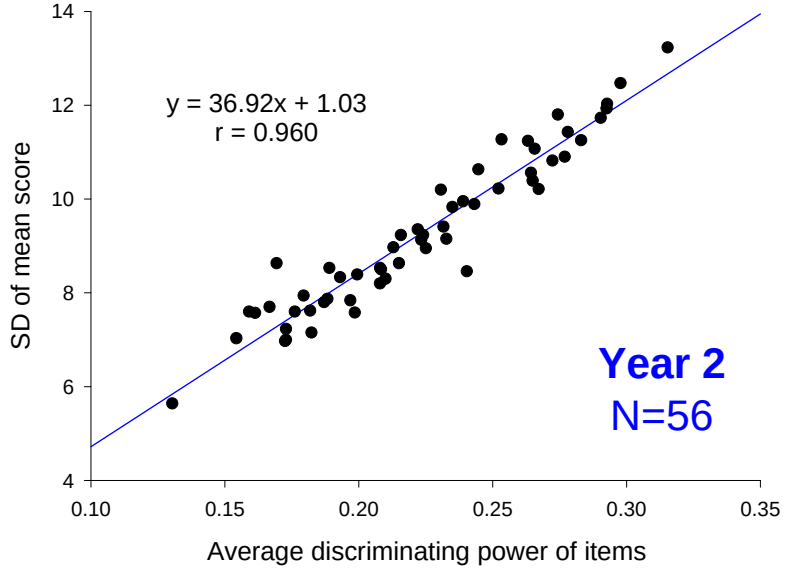


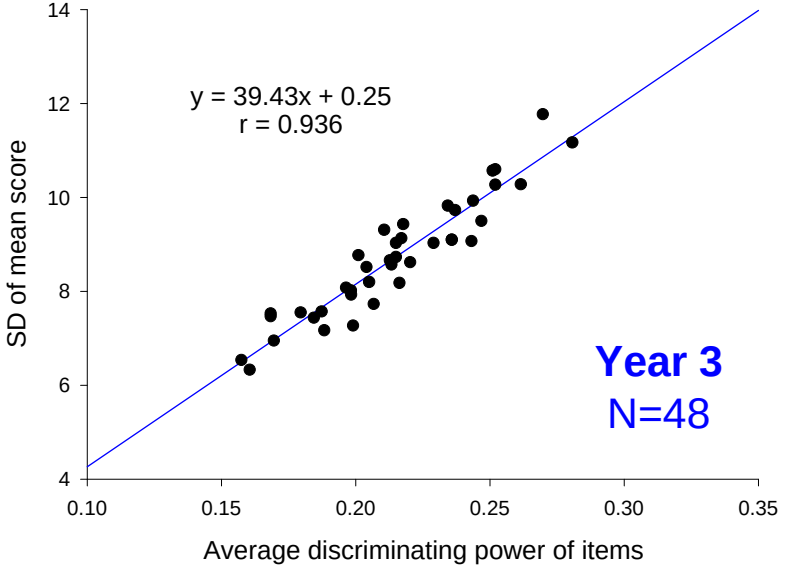
What determines SD of the mean score (i.e. spread of marks in a MCQ test)?

SD of the mean score shows a strong positive correlation to the discriminating power of MCQ.



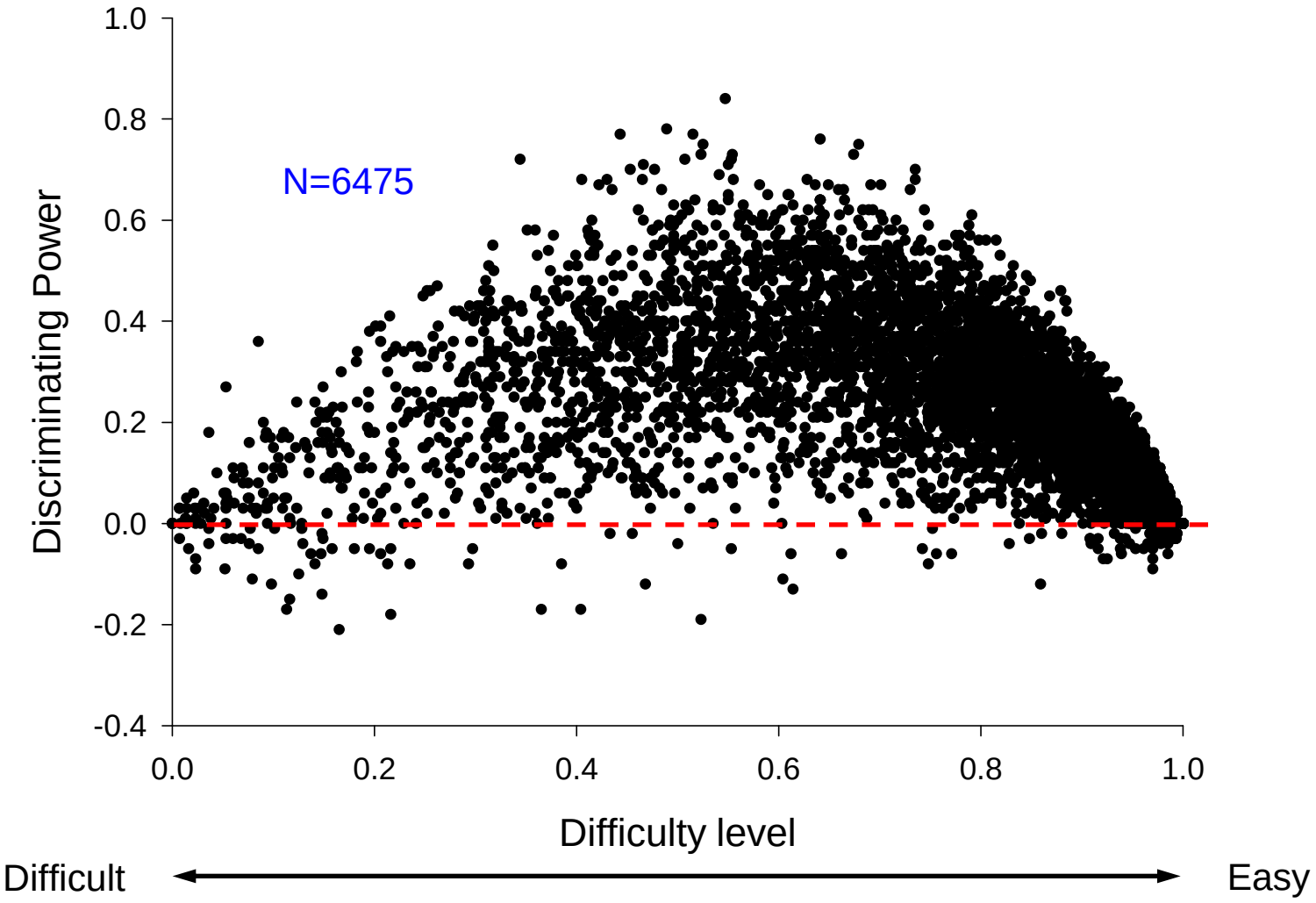






How is the discriminating power of an item related to its difficulty level?

Discriminating power is less for an item which is either too difficult or too easy.



For items that few students can get the “right” answer, are they really difficult?

Difficult items could have their stems/options poorly written, and therefore they are confusing to students. Or it is possible that a wrong key was used.

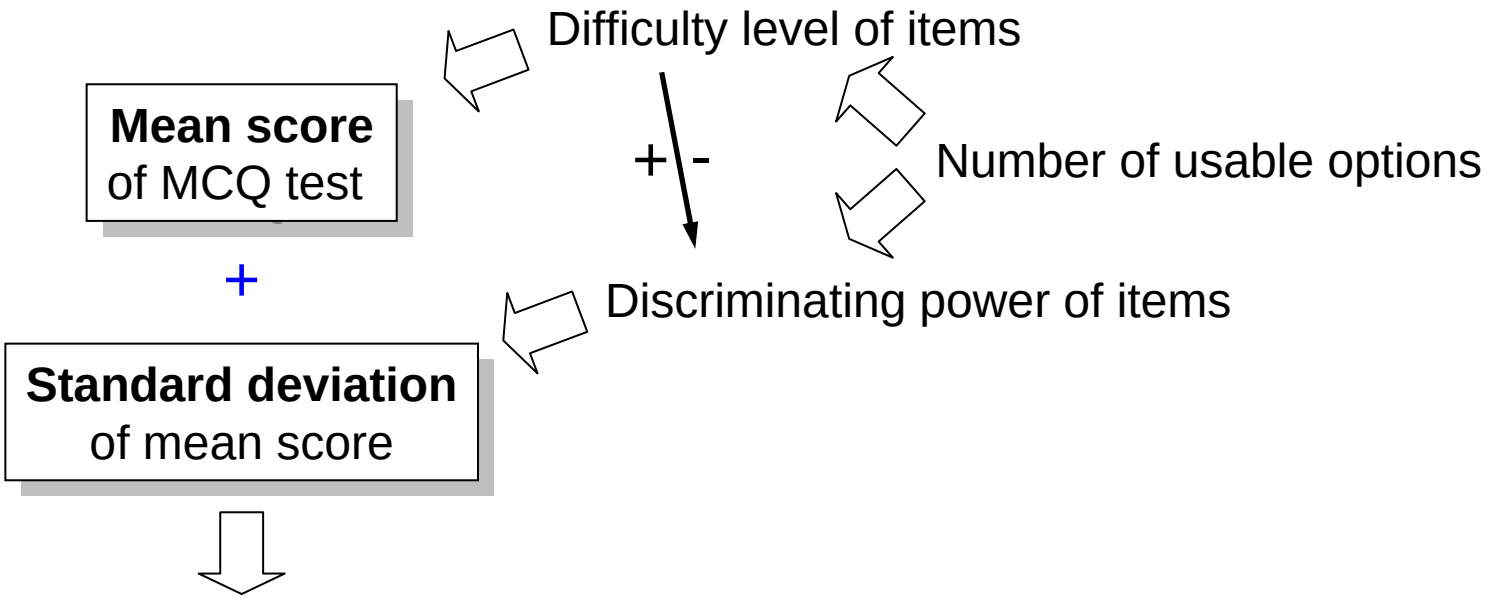
```
ITEM  21: DIF=0.180, RPB=  0.035, CRPB= -0.010 (95% CON= -0.176, 0.157)
                     RBIS= 0.052, CRBIS=-0.015, IRI=0.014
    GROUP    N    INV    NF   OMIT      A       B       C*      D       E
    TOTAL   139     0     0     0     0.39    0.16    0.18    0.23    0.04
     HIGH    36     0                 0.53    0.08    0.14    0.22    0.03
     MID     66     0                 0.35    0.20    0.20    0.24    0.02
     LOW     37     0                 0.32    0.16    0.19    0.22    0.11
    TEST SCORE MEAN %:                  71      66      70      69      61
    DISCRIMINATING POWER              0.20   -0.08   -0.05    0.01   -0.08
    STANDARD ERROR OF D.P.            0.12    0.08    0.09    0.10    0.06
```

```
ITEM   9: DIF=0.000, RPB=  0.000, CRPB=  0.000 (95% CON= -0.171, 0.171)
                     RBIS= 0.000, CRBIS= 0.000, IRI=0.000
    GROUP    N    INV    NF   OMIT      D
    TOTAL   132     0     0     0     1.00
     HIGH    27     0                 1.00
     MID     67     0                 1.00
     LOW     38     0                 1.00
    TEST SCORE MEAN %:                  78
    DISCRIMINATING POWER              0.00
    STANDARD ERROR OF D.P.            0.00
```

How to determine pass/fail?

- Professor C. B. Hazlett has addressed the issue of standard setting in Curriculum Retreats held in 2004 and 2007.
- What we are using for our SA is an “outdated” approach of setting an absolute and fixed pass/fail point: “a particular score or a % that has been determined prior to administering the test is set as the pass mark (e.g. 50%)”.
- Another approach would be norm-referencing of setting a relative and not fixed pass/fail point: “students are compared with each other and those who fail are “X” SDs below the mean performance of all candidates”.
- More “modern” practice would be to use test-centered or examinee-centered approach (or a combination of both) along with the judgment of “subject matter experts” factored into the method of choice.

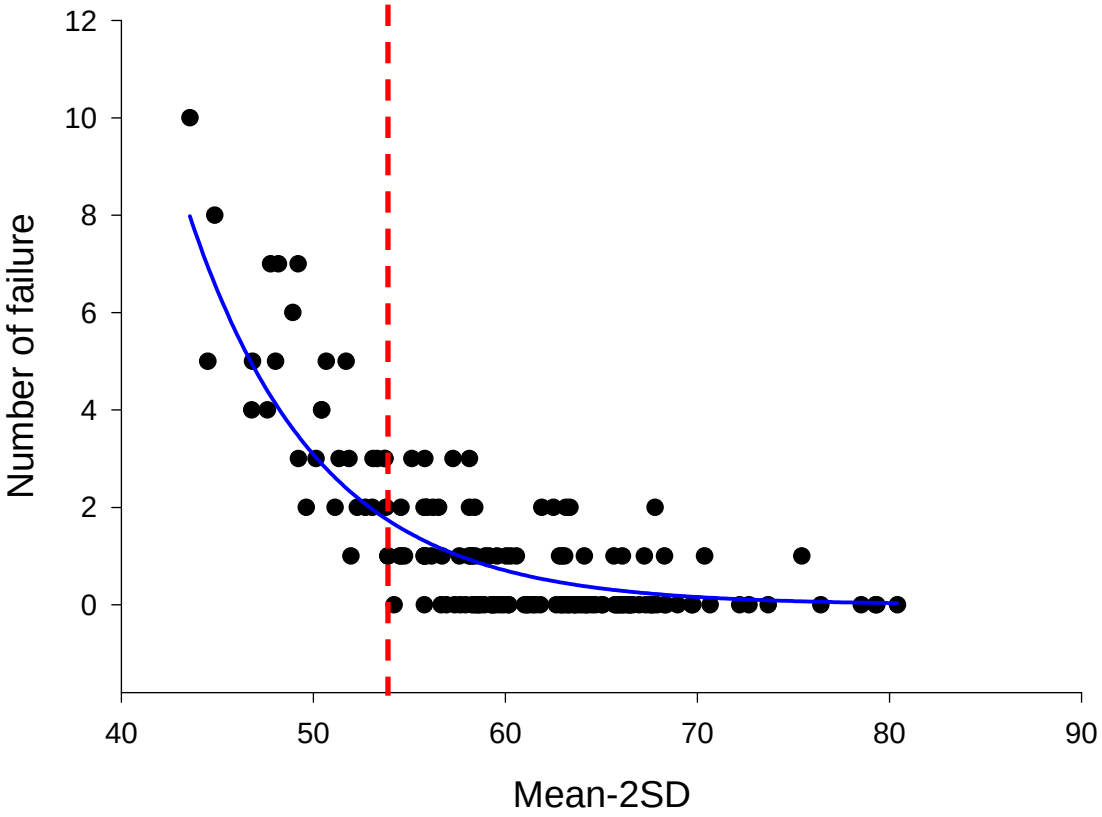
What have I learnt?



Norm-referencing based on mean-2SD

More likely to identify outliers who might score below 50, therefore requiring to take the supplementary exam.

How is the number of failures related to mean-2SD score?



The mean-2SD score approaching or falling below 50 represents the indicator that predicts the presence of failures in a MCQ test.

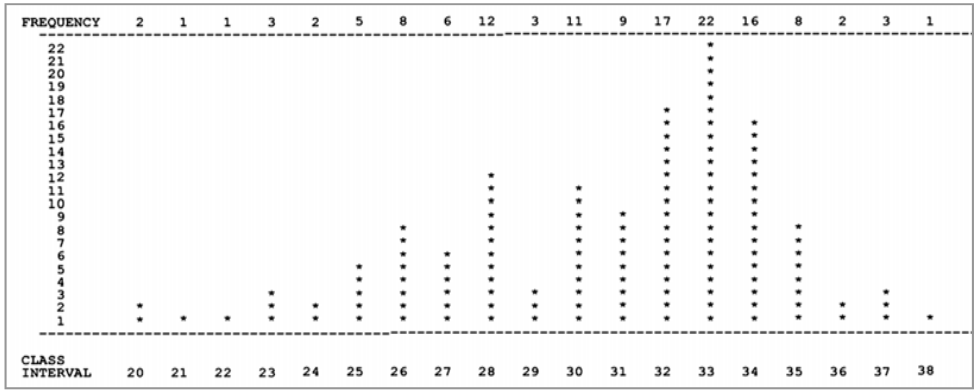
What we could learn?

- Keep track of the psychometric properties of each item, and the number of options picked by students, when being used over the years in different SA.

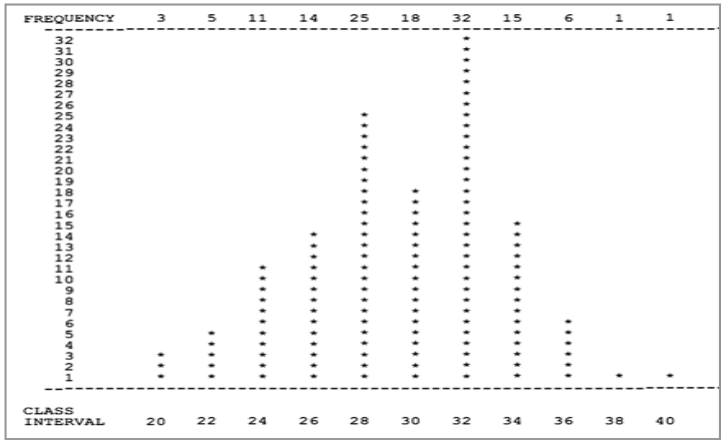
Year used	Difficulty Level	Discrimination Index					Selection percentage				
		A	B	C	D	E	A	B	C	D	E
03-04	0.682	-0.11	**0.46**	-0.09	-0.10	-0.17	0.08	**0.68**	0.06	0.06	0.11
05-06	0.655	-0.13	**0.52**	-0.18	-0.09	-0.12	0.10	**0.65**	0.14	0.06	0.05
07-08	0.766	0.02	**0.45**	-0.15	-0.18	-0.15	0.02	**0.77**	0.09	0.08	0.05
08-09	0.611	-0.15	**0.50**	-0.23	-0.07	-0.06	0.13	**0.61**	0.15	0.06	0.06

- Weed out those items that all students can pick the right answer (Diff level = 1).
- Improve on the writing of MCQ to eliminate confusion and to improve on the quality of the distracters (i.e. maximize the number of options that may be picked by students).
- Keep those items that test on important concepts/knowledge that students should know, even though they are not too difficult.
- Avoid over-testing students with difficult items that are beyond the level required or cover unimportant concepts/knowledge.

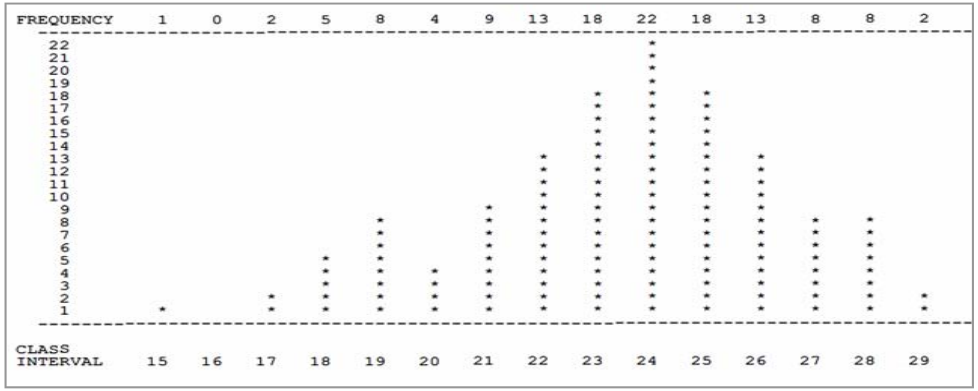
Results of PHOM Panel SA in 2008-09



PHOM1: 40 MCQ
Highest: 38
Lowest: 20
Mean: 76.6
Mean-2SD: 57.7

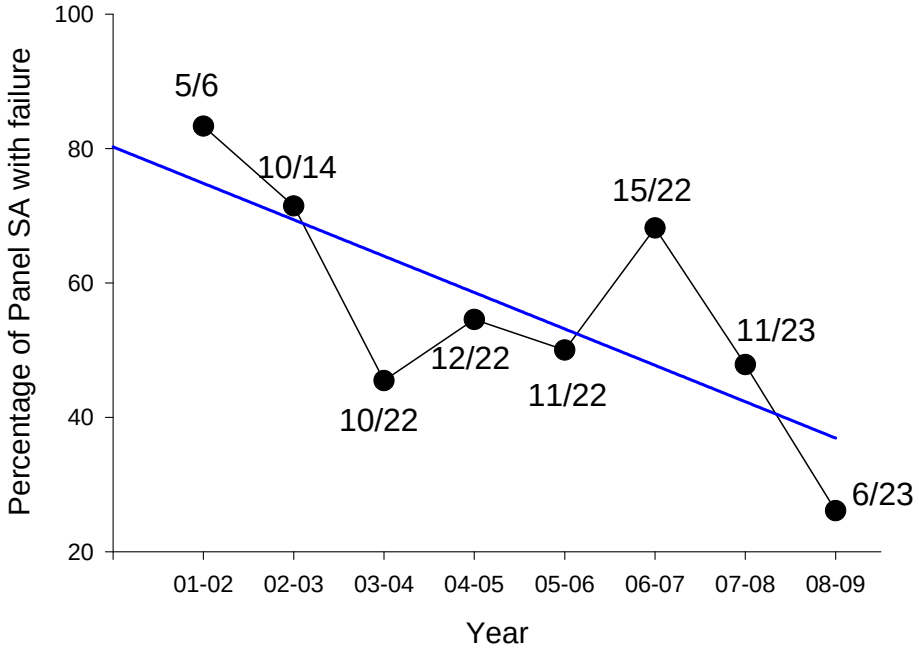


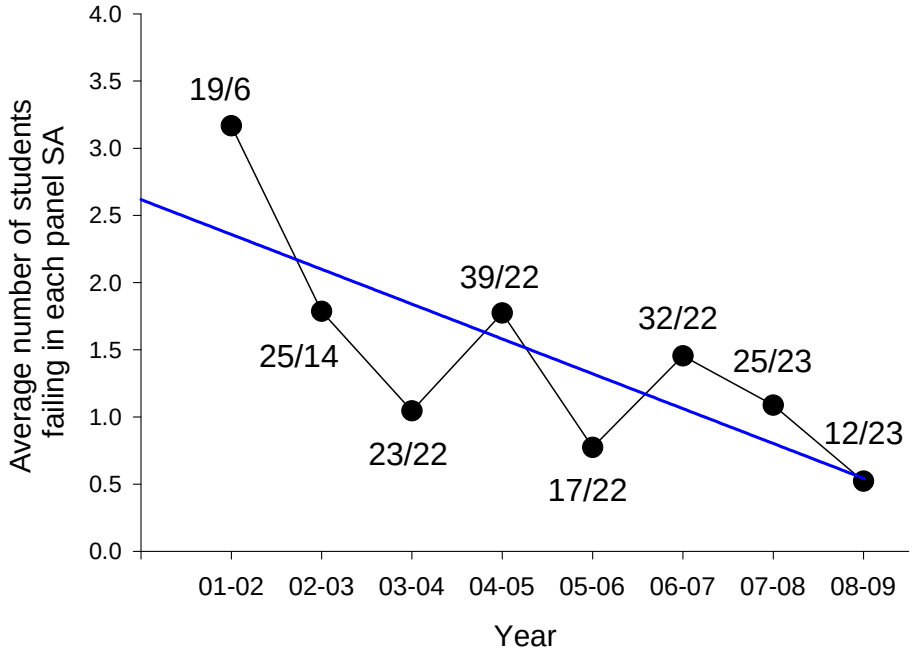
PHOM2: 40 MCQ
Highest: 39
Lowest: 19
Mean: 72.6
Mean-2SD: 52.7



PHOM3: 30 MCQ
Highest: 29
Lowest: 15
Mean: 78.2
Mean-2SD: 59.3

Trends in the number of failures in Panel SA





Over the years, there appears to be a downward trend in the number of panel SA having failures, and the number of students failing in SA could also be on the decline.

Acknowledgement

- Thanks to Prof CB Hazlett and Dr SPY Yip for allowing me to access the item analysis data of Panel SA.
- Thanks to Prof CB Hazlett for his comments.
- Thanks to Prof SM Kumta for allowing me to present the study here.
- Thanks to Miss Diana Kwan for providing well organized records of the item analysis.