A preliminary study of the use of Rasch analysis in a survey on teachers' and schools' readiness in adopting onscreen marking and post-assessment data analysis

2018

Background

1. The Hong Kong Examinations and Assessment Authority (HKEAA) is developing an onscreen marking (OSM) module for schools as part of an integrated post-assessment platform. Upon completion, teachers would be able to finish marking using tablet computers and item marks would be stored in the system for subsequent data analysis using the assessment quality-assurance platform (AQP).

2. This pilot study on teachers' and schools' readiness in adopting onscreen marking and post-assessment data analysis was conducted by a questionnaire survey administered to about 100 participants at the end of the 2018 Quality Assessment Management Accreditation Scheme (QAMAS) ceremony. The participants received some background information about OSM and AQP from several presentations and experience sharing sessions before they completed the questionnaires.

3. The purposes of this study were
   (a) To validate the data collected in this survey;
   (b) To identify teachers' and schools' biggest concerns in adopting OSM and AQP;
   (c) To assess teachers' and schools' readiness in adopting OSM and AQP;
   (d) To delineate possible target groups of teachers for promotion of OSM and AQP;
   (e) To detect possible problematic items that may require rewording or cancellation.

Methodology

4. In this study, it is assumed that readiness in adopting OSM and AQP is a latent variable that cannot be measured directly. Yet, by measuring teachers' responses to some observable variables, such as their use of tablet computers, or their schools' provision of assessment literacy training, if the data fits in a Rasch model, a picture of readiness in adopting OSM and AQP can be revealed.

5. Using Rasch model to validate questionnaires and thus to establish a common scale of

item difficulty and person ability is a common practice in the field of education, psychology, health and medical sciences, and many other areas and the methodologies were well documented by Mesbah *et. al.* (2002), Smith and Smith (2004, 2007), and Bond and Fox (2015).

6. To serve the above purpose, the items in the questionnaire in this pilot survey were designed to create a hierarchy for showing a construct map of readiness. Some items are supposed to be easier to endorse and they will be more likely to appear at the bottom of the scale (e.g. recognizing the idea of assessment for learning) while some other items are assumed to be more difficult to endorse (e.g. making assessment top priority in school management) and they will be more likely to show up at the top of the scale. This scale also reflects the underlying assumption of unidimensionality in Rasch measurement. That is readiness is the single and only latent variable to be studied in this survey.

7. The questionnaire contains 20 items in which 10 were about teachers' personal experience or perception and the other 10 were about their schools' assessment management. A 4-point Likert scale (from 1 strongly disagree to 4 strongly agree) was used to show respondents' degree of agreement to the statement of each item. The items were phrased using positive wordings in general. In addition, personal data about respondents' school type, position at school, experience in OSM and AQP were also collected. The English translation of items and item labels are listed in Table 1 and the original questionnaire in Chinese is attached in the ***Annex***.

Table 1 English translation of items and item labels

Item number	Item label	English translation
1.	P_TABLET	I used to work using tablet computers.
2.	P_NEWTECH	I am willing to learning new technologies.
3.	P_AFORL	I recognize the idea of assessment for learning.
4.	P_BETTERITEM	I think the exam questions I wrote before have rooms to improve.
5.	P_EFFICIENCY	I think OSM will increase efficiency in the long run.
6.	P_HEALTH	I think OSM brings no harm to health with rest.
7.	P_TRAINING	I give professional development courses in assessment first priority.
8.	P_SECURITY	I think there is a risk of bringing answer scripts home to mark.
9.	P_WORKLOAD	I think OSM will reduce my workload.
10.	P_REFLECT	I reflected on my teaching during marking.

11.	S_AFORL	My school recognizes the idea of assessment for learning.
12.	S_NEWTECH	My school supports the use of new technologies.
13.	S_DATADRIVEN	My school will use data to support decision making.
14.	S_BETTERITEM	My school management thinks the internal exam questions have rooms to improve.
15.	S_SECURITY	My school pays attention to exam paper security.
16.	S_MONEY	My school is able to allocate money to develop OSM.
17.	S_WORKLOAD	My school considers teachers' workload.
18.	S_PRIORITY	My school will give assessment top priority in its development plan.
19.	S_TRAINING	My school organizes assessment-related training in staff development days from time to time.
20.	S_HISTD	My school demands teachers to try their best in avoiding any mistake in assessment activities.

Data analysis

8. A total of 69 valid questionnaires were collected. The responses were analyzed using MINISTEP version 4.3.2. A rating scale model was adopted. Although it is impossible to ensure that the teachers treated the agree-disgree scale the same way to each item, a partial credit model was not used because the limited number of samples could not guarantee a stable estimate of different parameters in a more complicated form of model.

Results

Separation and reliability

9. The overall model statistics of the initial model was satisfactory. The real item separation was 3.47 and the item reliability was 0.92, both were above the respective recommended cut-off value of 3 and 0.9 (Linarce 2018). At the same time, the real (extreme and non-extreme) person separation of 2.47 and reliability of 0.86 were also found having values above the respective recommended levels of 2 and 0.8. These numbers imply that the person sample is large enough to confirm the item difficulty hierarchy of the instrument and the questionnaire is sensitive enough to distinguish more ready teachers from less ready teachers.

Category structure

10. It is necessary to ensure the Likert form of category structure functioned properly in this survey. As shown in Table 2a, there are only very few responses in "Strongly Disagree". This category can be combined with the "Disagree" category to achieve a more meaningful rating scale.

Table 2a Summary of the 4-category structure model

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	5	0	1.84	-.36	1.90	2.35	NONE	(-4.33)	1 S DISAGREE
2	2	131	10	.48*	.60	.95	.93	-3.19	-1.85	2 DISAGREE
3	3	812	59	2.12	2.13	.95	.95	-.51	1.60	3 AGREE
4	4	423	31	3.97	3.95	.97	.98	3.69	(4.80)	4 S AGREE
MISSING		9	1	.46						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

CATEGORY LABEL	STRUCTURE MEASURE	S.E.	SCORE-TO-MEASURE AT CAT.	-ZONE-		50% CUM. PROBABLTY	COHERENCE M->C	C->M	ESTIM DISCR	
1	NONE		(-4.33)	-INF	-3.35					1 S DISAGREE
2	-3.19	.46	-1.85	-3.35	-.39	-3.25	67%	22%	.75	2 DISAGREE
3	-.51	.10	1.60	-.39	3.73	-.46	71%	87%	1.02	3 AGREE
4	3.69	.07	(4.80)	3.73	+INF	3.71	72%	56%	1.03	4 S AGREE

M->C = Does Measure imply Category?
C->M = Does Category imply Measure?

11. A new model was constructed after recoding "1 Strongly disagree" to "2 Disagree" and the summary of category structure of this new model is shown in Table 2b. A comparison between the two models shows that the 3-category structure model reported slight improvement in separation and reliability (Table 3). All subsequent analyses were based on the 3-category structure model.

Table 2b Summary of the 3-category structure model

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
2	2	136	10	-1.13	-1.15	1.04	1.03	NONE	(-3.25)	1 DISAGREE
3	3	812	59	.52	.53	.96	.95	-2.14	.00	3 AGREE
4	4	423	31	2.44	2.43	.98	.99	2.14	(3.25)	4 S AGREE
MISSING		9	1	-1.20						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

CATEGORY LABEL	STRUCTURE MEASURE	S.E.	SCORE-TO-MEASURE AT CAT.	-ZONE-		50% CUM. PROBABLTY	COHERENCE M->C	C->M	RMSR	ESTIM DISCR	
2	NONE		(-3.25)	-INF	-2.18		70%	22%	.8084		1 DISAGREE
3	-2.14	.10	.00	-2.18	2.18	-2.15	72%	88%	.3209	.99	3 AGREE
4	2.14	.07	(3.25)	2.18	+INF	2.15	73%	58%	.5409	1.01	4 S AGREE

M->C = Does Measure imply Category?
C->M = Does Category imply Measure?

Table 3 comparison of separation and reliability of the 4-category structure model and the 3-category structure model

Model	Item separation	Item reliability	Person separation	Person reliability
4-category structure	3.47	0.92	2.47	0.86
3-category structure	3.48	0.92	2.54	0.87

Item polarity and item fit statistics

12. Another criterion to check the validity of items is the point-correlation between the data code and person raw scores. Table 4 shows the correlation coefficients of the 20 items. All items had a correlation above +0.3 and there was no item with zero or negative values. According to Bond & Fox (2015) the positive PTMEASUR-AL correlation value represents that the item is carefully constructed to measure what it supposes to measure. In other words, all items in this study were making contribution to the measurement of the readiness scale.

Table 4 Item statistics

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
8	207	69	.93	.25	1.57	3.07	1.62	3.05	A .42	.56	53.6	69.7	P_SECURITY
1	215	69	.42	.25	1.47	2.59	1.47	2.39	B .52	.56	56.5	70.0	P_TABLET
20	241	69	-1.32	.27	1.28	1.62	1.33	1.43	C .30	.55	63.8	72.7	S_HISTD
9	190	69	2.02	.25	1.24	1.46	1.30	1.64	D .63	.57	65.2	70.1	P_WORKLOAD
19	201	68	1.16	.25	1.24	1.42	1.27	1.46	E .47	.56	64.7	70.0	S_TRAINING
16	199	66	.94	.26	1.19	1.12	1.18	1.02	F .63	.55	63.6	70.0	S_MONEY
15	240	69	-1.25	.27	1.01	.09	1.05	.29	G .46	.56	71.0	72.5	S_SECURITY
18	200	68	1.22	.25	1.02	.18	1.05	.35	H .46	.56	75.0	69.9	S_PRIORITY
6	209	69	.81	.25	.96	-.17	.99	-.01	I .56	.56	71.0	69.6	P_HEALTH
17	212	68	.43	.25	.92	-.46	.86	-.78	J .72	.56	70.6	69.9	S_WORKLOAD
2	238	69	-1.11	.27	.83	-1.06	.91	-.36	j .55	.56	73.9	72.5	P_NEWTECH
5	213	67	.17	.26	.90	-.54	.88	-.66	i .67	.56	67.2	69.8	P_EFFICIENCY
10	231	69	-.63	.26	.85	-.90	.84	-.82	h .55	.56	73.9	71.7	P_REFLECT
7	213	69	.55	.25	.84	-.99	.82	-1.05	g .57	.56	76.8	69.7	P_TRAINING
14	225	68	-.42	.26	.83	-1.04	.81	-1.00	f .58	.56	82.4	71.2	S_BETTERITEM
12	236	69	-.97	.26	.79	-1.34	.73	-1.40	e .68	.56	84.1	72.3	S_NEWTECH
13	236	69	-.97	.26	.75	-1.63	.73	-1.44	d .65	.56	82.6	72.3	S_DATADRIVEN
3	235	69	-.90	.26	.73	-1.75	.72	-1.48	c .65	.56	81.2	72.1	P_AFORL
4	219	69	.17	.25	.72	-1.85	.67	-2.01	b .50	.56	81.2	70.1	P_BETTERITEM
11	240	69	-1.25	.27	.63	-2.59	.56	-2.40	a .70	.56	82.6	72.5	S_AFORL
MEAN	220.0	68.5	.00	.26	.99	-.1	.99	-.1			72.0	70.9	
P.SD	15.9	.8	.98	.01	.26	1.5	.28	1.5			8.7	1.2	

13. Table 4 also shows statistics about the item fit. With reference to recommendation by Wright and Linacre (1994), the reasonable range of item mean-square statistics for infit and outfit is between 0.6 and 1.4. In this 20-item model, three items were marginally out of such recommended range. They included two items with high fit MNSQ: P_SECURITY (personal-aspect: risk of bringing answer scripts home to mark) and P_Tablet (personal-aspect: used to work using tablet computers). According to Smith (1996), the misfit in these two items is likely to be caused by some noisy outliers. The third misfit item was S_AFORL (school-aspect: recognizing the concept of assessment for learning) which has an outfit mean-square value of 0.56, representing a possible overfitting problem. Nevertheless, the degree of deviation of fit statistics from recommended values was generally small and this should not adversely affect the validity of measurement.

Dimensionality

14. Further investigation of item characteristics was conducted through a Rasch principal components analysis of residuals method which aims at confirming that the items are all measuring only one latent construct or dimension. Tables 5 & 6 and Figure 1 show the

results of the analysis. Following the method recommended by Linacre (2018), the proportion of unexplained variance should be studied first and then followed by a search of patterns in item cluster graph and an examination of the disattenuated correlation between item clusters.

15. As reported in Table 5, almost 60% of the total raw variance (about 12 items) was not explained and the unexplained variance in the first contrast is only 8.4% of the total raw variance (fewer than 2 items). This implies that item by item variations rather than grouping of items (different dimension) are present.

16. As shown in Figure 1, three clusters of the items are found. Positive loadings are associated with 7 school-aspect items (cluster 1) and negative loadings are associated with 6 personal-aspect items (cluster 3) while the remaining 7 items with close to zero loadings form another cluster in between (cluster 2). An examination of the disattenuated correlation between each pair of these three clusters all reported positive correlations greater than 0.5 which implies that all three clusters are measuring the same underlying attribute. The above evidence altogether confirms that the assumption of dimensionality holds in this model. The model can be used to evaluate teachers' and schools' readiness in OSM and AQP.

Table 5 Table of standardized residual variance in eigenvalue (item information) units

Eigenvalue	Observed	Expected			
Total raw variance in observations =	33.6808	100.0%		100.0%	
Raw variance explained by measures =	13.6808	40.6%		40.3%	
Raw variance explained by persons =	6.5830	19.5%		19.4%	
Raw Variance explained by items =	7.0979	21.1%		20.9%	
Raw unexplained variance (total) =	20.0000	59.4%	100.0%	59.7%	
Unexplned variance in 1st contrast =	2.8142	8.4%	14.1%		
Unexplned variance in 2nd contrast =	2.1424	6.4%	10.7%		
Unexplned variance in 3rd contrast =	1.9479	5.8%	9.7%		
Unexplned variance in 4th contrast =	1.7586	5.2%	8.8%		
Unexplned variance in 5th contrast =	1.5258	4.5%	7.6%		

Figure 1 Standardized residual plot of contrast 1

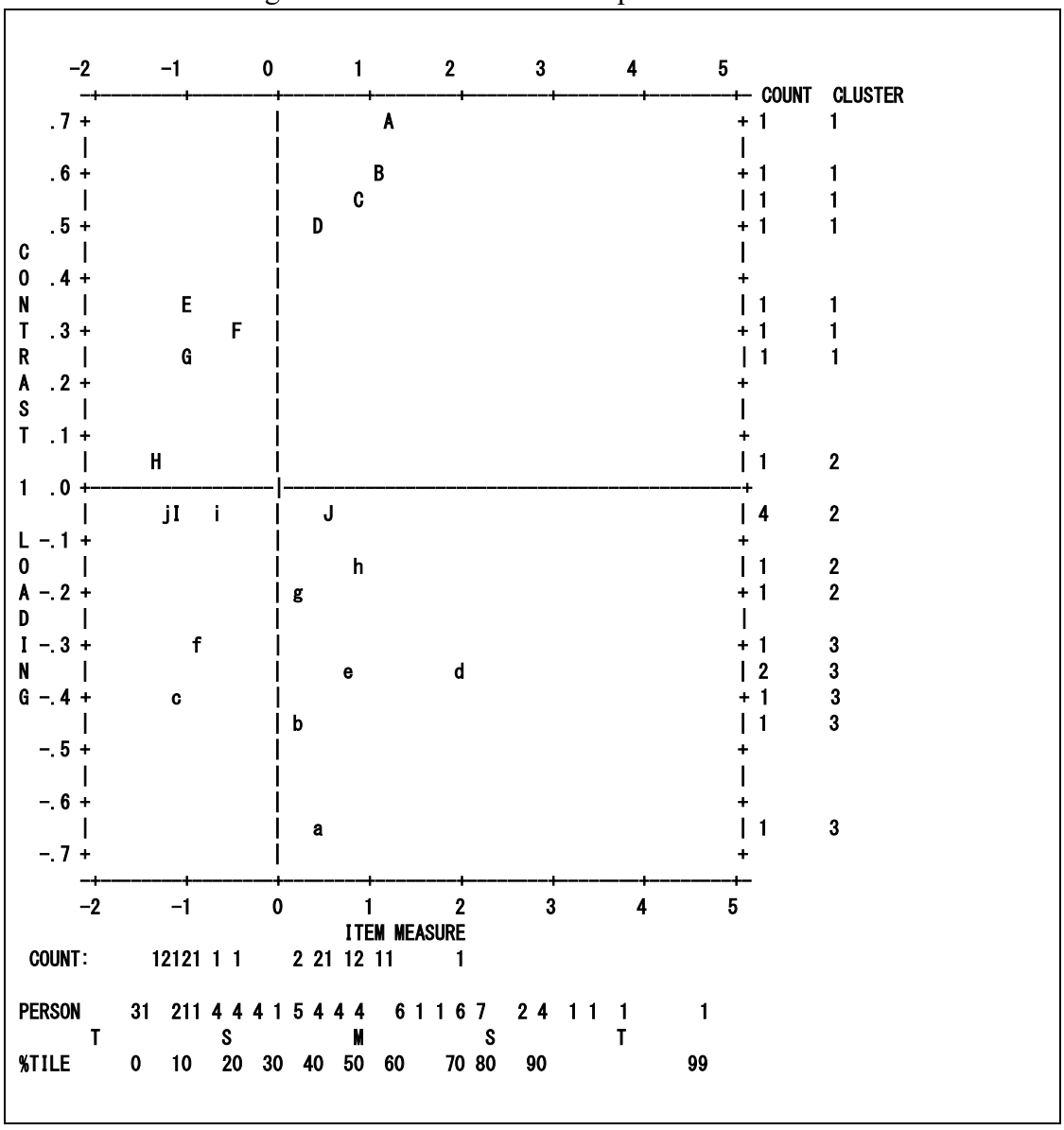


Table 6 Approximate relationships between the person measures

PCA Contrast	ITEM Clusters	Pearson Correlation	Disattenuated Correlation	Pearson+Extr Correlation	Disattenuated+Extr Correlation
1	1 - 3	0.4385	0.5598		
1	1 - 2	0.5393	0.7516		
1	2 - 3	0.6406	0.9006		

Person fit

17. At the same time, person fit statistics should be examined to ensure consistency of the person responses. Based on the simulations of Smith et al. (1998), a value of standardized mean-squared unweighted person fit value of 2.00 can be used as a cut score for flagging persons for misfit. Refer to Table 7, there 7 persons associated with high positive misfit and there are 8 persons associated with high negative misfit. In addition, there are another 3 persons with negative point-measure correlation. Although there were a number of misfit persons, such person fit issues are acceptable in this study because the purpose is to identify patterns rather than to classify each respondent.

Table 7 Person statistics

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
16	69	20	1.99	.47	3.05	5.26	3.06	5.10	A .27	.45	35.0	68.9	116 1 3 2 1
40	66	20	1.33	.47	2.30	3.47	2.47	3.53	B .61	.43	45.0	67.9	140 2 2 2 2
52	73	20	2.92	.51	1.52	1.83	2.01	2.41	C-.18	.42	55.0	70.8	152 2 2 3 1
7	72	20	2.67	.49	1.99	3.12	1.82	2.28	D .29	.43	75.0	69.4	107 1 2 1 1
10	57	20	-.67	.47	1.88	2.24	1.95	2.25	E .76	.41	40.0	72.2	110 1 3 3 1
62	62	20	.45	.47	1.76	1.99	1.84	2.03	F .48	.41	50.0	72.2	162 2 2 1 2
61	66	20	1.33	.47	1.68	2.08	1.73	2.07	G-.08	.43	35.0	67.9	161 2 2 1 2
67	74	20	3.19	.53	1.44	1.51	1.66	1.52	H-.05	.41	55.0	73.1	167 1 1 2 1
48	70	20	2.21	.47	1.64	2.17	1.56	1.85	I .17	.45	55.0	68.7	148 2 3 2 1
60	69	20	1.99	.47	1.56	1.93	1.60	1.97	J .49	.45	50.0	68.9	160 2 3 2 2
68	60	20	.00	.48	1.56	1.50	1.59	1.52	K .38	.40	60.0	73.7	168 1 3 1 1
66	70	20	2.21	.47	1.55	1.93	1.56	1.86	L .51	.45	65.0	68.7	166 1 3 2 1
22	68	20	1.77	.47	1.41	1.46	1.49	1.63	M .24	.45	65.0	68.5	122 1 2 2 2
41	70	20	2.21	.47	1.41	1.51	1.28	1.05	N .60	.45	65.0	68.7	141 1 3 3 2
54	70	20	2.21	.47	1.35	1.30	1.26	.99	O .65	.45	65.0	68.7	154 2 2 1 2
15	67	20	1.55	.47	1.30	1.09	1.34	1.15	P-.09	.44	60.0	68.1	115 9 1 9 9
31	53	20	-1.56	.47	1.31	1.11	1.34	1.11	Q .60	.44	55.0	67.9	131 2 4 3 1
19	66	20	1.33	.47	1.26	.95	1.31	1.03	R .59	.43	65.0	67.9	119 2 2 1 2
44	69	20	1.99	.47	1.24	.94	1.29	1.08	S .09	.45	60.0	68.9	144 1 3 2 1
51	73	20	2.92	.51	1.07	.37	1.29	.89	T .27	.42	65.0	70.8	151 2 1 3 2
17	66	20	1.33	.47	1.15	.60	1.20	.71	U .38	.43	65.0	67.9	117 1 3 2 1
53	70	20	2.21	.47	1.09	.41	1.18	.73	V .25	.45	65.0	68.7	153 2 1 1 1
33	59	20	-.22	.48	1.14	.51	1.17	.56	W .57	.40	70.0	73.7	133 1 3 1 2
38	70	20	2.21	.47	1.15	.63	1.15	.63	X .56	.45	65.0	68.7	138 1 2 1 1
23	63	20	.68	.47	1.07	.33	1.14	.50	Y .14	.41	75.0	70.6	123 1 2 3 2
3	61	20	.23	.48	1.09	.37	1.08	.35	Z .61	.40	65.0	73.3	103 2 2 1 1
BETTER FITTING NOT SHOWN													
13	75	20	3.48	.55	.93	-.15	.74	-.44	.46	.38	80.0	76.8	113 1 2 3 1
24	55	20	-1.12	.47	.80	-.62	.81	-.51	.65	.43	80.0	69.2	124 1 2 3 1
65	70	20	2.21	.47	.78	-.83	.78	-.78	.55	.45	85.0	68.7	165 1 3 1 2
43	78	20	4.71	.77	.77	-.26	.42	-.47	.50	.27	90.0	90.0	143 1 1 3 1
50	53	18	-.27	.50	.77	-.58	.74	-.66	.65	.42	77.8	73.1	150 1 3 3 1
28	58	20	-.45	.48	.70	-.87	.66	-.96	z .45	.41	75.0	73.3	128 2 3 1 1
45	64	20	.90	.47	.68	-1.09	.63	-1.17	y .32	.42	75.0	68.6	145 2 2 1 2
58	64	20	.90	.47	.67	-1.12	.66	-1.04	x .30	.42	75.0	68.6	158 1 2 3 2
59	64	20	.90	.47	.66	-1.16	.65	-1.09	w .32	.42	75.0	68.6	159 1 1 1 1
8	57	20	-.67	.47	.65	-1.10	.62	-1.15	v .62	.41	80.0	72.2	108 1 2 3 1
30	76	20	3.81	.59	.65	-1.04	.43	-1.09	u .68	.36	85.0	80.9	130 1 2 1 1
35	72	20	2.67	.49	.65	-1.48	.56	-1.60	t .71	.43	75.0	69.4	135 1 2 2 1
47	61	20	.23	.48	.63	-1.14	.58	-1.29	s .71	.40	75.0	73.3	147 2 3 2 2
34	63	20	.68	.47	.62	-1.28	.60	-1.23	r .23	.41	85.0	70.6	134 1 2 3 1
37	66	20	1.33	.47	.60	-1.54	.57	-1.55	q .58	.43	85.0	67.9	137 1 1 3 1

69	55	20	-1.12	.47	.60	-1.43	.54	-1.56	p .53	.43	70.0	69.2	169 1 2 1 1
6	58	20	-.45	.48	.57	-1.40	.54	-1.47	o .58	.41	85.0	73.3	106 1 2 1 1
29	56	18	.45	.50	.54	-1.52	.55	-1.39	n .17	.41	88.9	71.5	129 1 3 1 1
11	62	20	.45	.47	.54	-1.59	.54	-1.46	m .12	.41	90.0	72.2	111 1 3 2 1
63	61	20	.23	.48	.49	-1.73	.49	-1.65	l-.22	.40	95.0	73.3	163 2 3 1 2
26	61	20	.23	.48	.47	-1.83	.44	-1.88	k .52	.40	85.0	73.3	126 2 2 3 2
12	58	20	-.45	.48	.46	-1.89	.43	-1.95	j .27	.41	85.0	73.3	112 1 2 1 1
21	61	20	.23	.48	.45	-1.96	.41	-2.07	i .56	.40	85.0	73.3	121 1 4 1 2
25	45	16	-1.03	.53	.45	-1.74	.40	-1.85	h .59	.42	87.5	72.5	125 2 3 1 2
18	64	20	.90	.47	.44	-2.23	.39	-2.29	g .59	.42	85.0	68.6	118 2 3 1 1
27	57	20	-.67	.47	.44	-2.06	.40	-2.12	f .48	.41	90.0	72.2	127 1 2 1 1
57	56	20	-.90	.47	.38	-2.52	.33	-2.57	e .68	.42	85.0	70.6	157 1 2 2 2
4	57	20	-.67	.47	.36	-2.52	.32	-2.59	d .58	.41	90.0	72.2	104 1 2 3 1
64	59	20	-.22	.48	.36	-2.36	.35	-2.34	c .13	.40	90.0	73.7	164 1 3 1 1
56	59	20	-.22	.48	.33	-2.59	.30	-2.64	b .22	.40	90.0	73.7	156 1 2 3 2
32	58	20	-.45	.48	.32	-2.67	.29	-2.72	a .48	.41	95.0	73.3	132 2 3 1 2

18. In short, based on the satisfactory separation and reliability statistics, a general high degree of item fit, an acceptable person fit, and a confirmation of model unidimensionality, the data collected in the survey basically fits in a Rasch model and therefore relationship between respondents (person) and factors affecting adopting OSM and AQP (items) can be established and analyzed.

Discussion

(a) To identify teachers' and schools' biggest concerns in adopting OSM and AQP

Item person map

19. The item-person map is the key feature in a Rasch analysis because it puts item difficulty and person ability on the same scale. The item-person map of this preliminary study was shown in Figure 2. On the right hand side of the figure, the items form a hierarchy according to their level of difficulty to be endorsed by the respondents. Easier items are found at the bottom and more difficult items are put to the top.

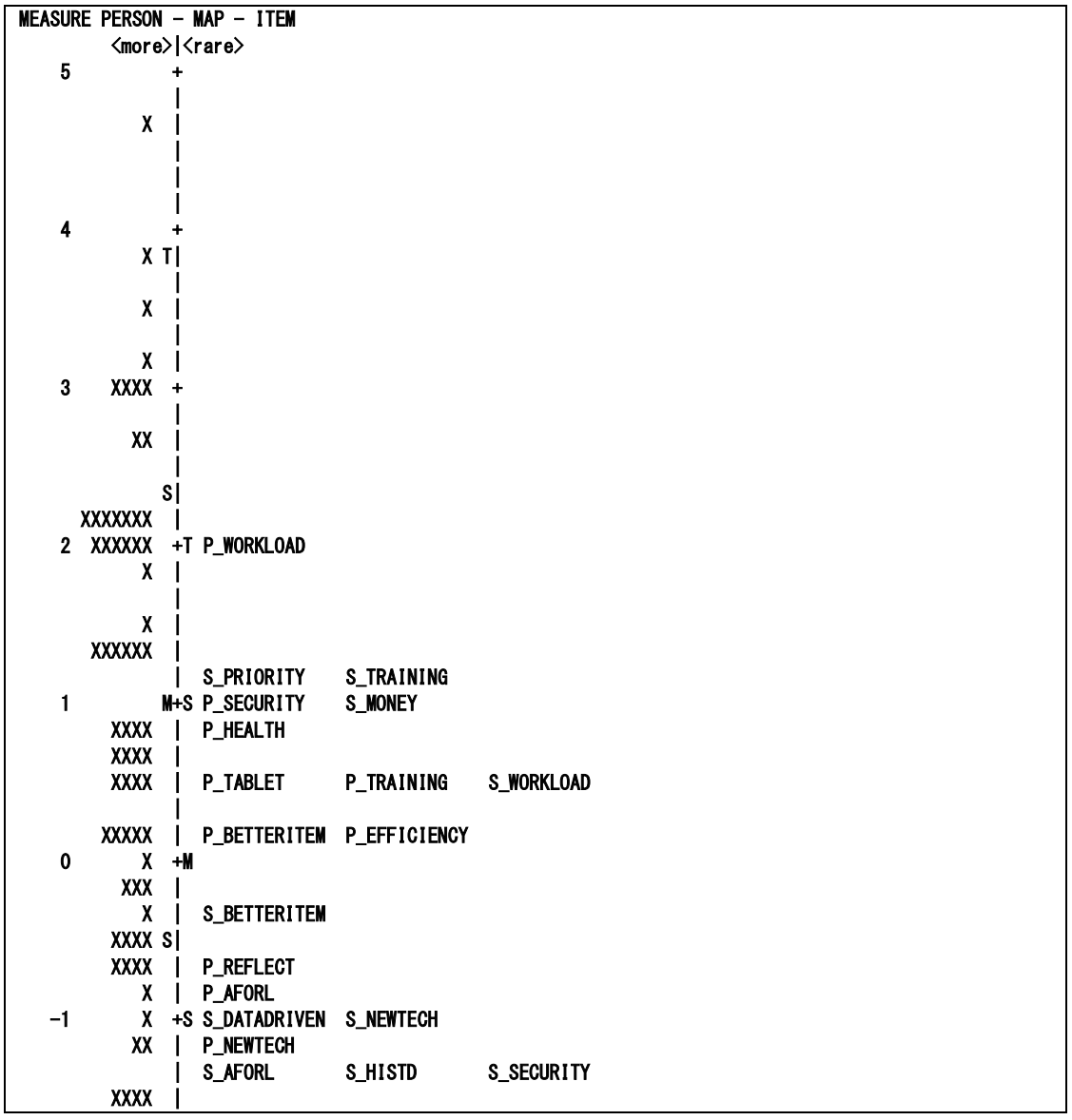
20. Figure 2 shows that the most difficult to endorse item is a personal-aspect item of workload, followed by three school-aspect factors of priority, training, and money. The second and third most difficult personal aspect items are security and health respectively.

21. For personal-aspect items, it was anticipated that workload and health concerns would be some difficult items. However, it is interesting to find out that not many teachers recognize the risks of taking answer scripts home to mark. To the other extreme, the

easiest items are teachers' willingness to use new technology, their recognition of the idea of assessment for learning, and their reflection on teaching and assessment design during marking. In other words, they do not resist new assessment technologies as long as the negative impacts can be eliminated or reduced.

22. For school-aspect items, all three most difficult items are real practices reflecting how schools value the importance of assessment activities. They were expected to be difficult items when the questionnaire was design. The easiest items are schools' pursuit of high assessment standards, their recognition of the idea of assessment for learning, and their serious attitude towards exam security. In other words, there seems to be a gap between schools' ideals and actions when they manage assessment activities.

Figure 2 Item-person map



(b) To assess teachers' and schools' readiness in adopting OSM and AQP

23. Comparing the distribution of respondents' ability level and that for item difficulty, it was found that the mean ability level is higher than the mean item difficulty. However, it did not imply that many teachers' had a high degree of readiness to adopt OSM and AQP. The mean ability level is at the difficulty level of school-aspect item of money. In other word, about one half of the respondents would find the concerns described in the top three personal-aspect and school-aspect items challenging enough to hinder them from trying to adopt OSM and AQP. Yet, there were still 17 teachers (top 25%) whose ability was found to be above the level of the most difficult item.

(c) To delineate possible target groups of teachers for promotion of OSM and AQP

24. A follow-up analysis to identify the demographic characteristics of the top 25% group was performed to specify some features of such potential targets for promotion of OSM and AQP. The results are shown in Table 8. Three rows of numbers are presented in Table 8. They represent the counts for three different groups: the top 25% group (first row); the top 50% group (second row); and the bottom 50% (third row).

Table 8 Demographic information of respondent by their ability level
top 25% group (first row),top 50% group (second row), bottom 50% group (third row)

School type	Position	AQP Experience	OSM Experience
Primary **12 of 44** 23 of 44 21 of 44	Principals **4 of 7** 7 of 7 0 of 7	Nil **6 of 30** 13 of 30 17 of 30	Yes **12 of 42** 21 of 42 21 of 42
Secondary **5 of 23** 11 of 23 12 of 23	Management teachers **9 of 33** 17 of 33 16 of 33	<2 years **4 of 14** 10 of 14 4 of 14	No **5 of 25** 13 of 25 12 of 25
Others	Subject teacher	>=2 years	

0 of 0	**4 of 26** 11 of 26 15 of 26	**7 of 23** 11 of 23 12 of 23	

25. It was found that among the top 25% of the respondents with the highest level of readiness, 12 (about two-third) are primary school teachers and only 5 (about one-third) are secondary school teachers. This proportion is almost the same as that for all respondents (44 to 23).

26. Principals who know school policies better were found more ready to adopt OSM and AQP. All principals have above mean person ability and 4 of them are in the top 25% group. While principals are most ready to adopt OSM and AQP, they do not need to do marking and data analysis themselves. Refer to the item-person map in Figure 2, "Schools concern about teachers' workload" was found indeed an above mean difficult item. In contrast, only 4 subject teachers among the 26 in total are in the top 25% group and fewer than half of the 26 have above mean ability. For teachers in the management, about half are above mean and the other half are below.

27. Finally, for the top 25% group, previous experience in AQP and OSM do help increase their readiness in adopting AQP and OSM. However, this pattern was not observed for the top 50% group.

(d) To detect possible problematic items that may require rewording or cancellation

Local independence

28. One of the assumptions of Rasch measurement is local independence. This means that there should not be any correlation between two items after the effect of the underlying trait is conditioned out. In other words, the correlation of residuals should be zero theoretically. The items should only be correlated through the latent trait that the instrument is measuring (Lord and Novick, 1968).

29. A check for local dependence of items using Yen's Q_3 statistics is one the common method in Rasch analysis. The Q_3 statistics is based on a Pearson correlation coefficient between the residuals of a pair of items after partialling out the measured construct (Yen, 1984, 1993). A positive correlation in Q_3 statistics between a pair of items may indicate possible local item dependency. As show in Table 9, there are five pairs of items

showing an over 0.2 positive correlation coefficient.

30. As recommended by Lincare (2018), caution is needed for a correlation around 0.7 and a correlation of 0.4 should be considered low dependency. In this pilot study, there are no pairs of items with correlation greater than 0.7. The pair of items with the largest correlation of 0.54 seems to show certain duplicated features of each other. A teacher who used to work using tablet computers is more likely to be willing to learning new technology. One may argue that new technology is the set while tablet is a subset of new tech and therefore the two items can be combined into one. However, it was also true that the new technology item tried to measure teachers' attitude but the tablet item intended to capture their action or practice, so two separate items were needed.

Table 9 Largest observation residual correlations (Q_3 statistics)

CORRELATION	ENTRY NUMBER	ITEM	ENTRY NUMBER	ITEM
.54	1	P_TABLET	2	P_NEWTECH
.40	18	S_PRIORITY	19	S_TRAINING
.40	15	S_SECURITY	20	S_HISTD
.37	16	S_MONEY	17	S_WORKLOAD
.35	16	S_MONEY	18	S_PRIORITY
.30	1	P_TABLET	5	P_EFFICIENCY
.27	12	S_NEWTECH	13	S_DATADRIVEN
-.47	6	P_HEALTH	19	S_TRAINING
-.40	1	P_TABLET	14	S_BETTERITEM
-.40	1	P_TABLET	18	S_PRIORITY
-.36	1	P_TABLET	13	S_DATADRIVEN
-.34	4	P_BETTERITEM	17	S_WORKLOAD
-.34	2	P_NEWTECH	14	S_BETTERITEM
-.33	7	P_TRAINING	12	S_NEWTECH
-.33	5	P_EFFICIENCY	20	S_HISTD
-.30	9	P_WORKLOAD	17	S_WORKLOAD
-.29	3	P_AFORL	18	S_PRIORITY
-.29	16	S_MONEY	20	S_HISTD
-.27	5	P_EFFICIENCY	13	S_DATADRIVEN
-.27	6	P_HEALTH	17	S_WORKLOAD

Limitations of the study

31. Above all, one of the major limitations of this preliminary survey was its small sample size. Among the respondents, about two-third of them came from primary schools, and over half of them were involved in school management. In terms of AQP and OSM experience, again over half of them had used AQP before and about two-third of them had experience in using OSM systems. This unbalanced proportion of teacher combination might also impact on the results of the survey.

32. Another problem of this study was related to whether to provide the respondents information about OSM and AQP before they took the questionnaire survey. As OSM and AQP were new to some teachers, it would be meaningless if these teachers answered the questions based on their imagination. As a compromise, we provided the respondents both the pros and cons of adopting OSM and AQP in the experience sharing session, although the benefits were highlighted more and the drawbacks were discussed in lesser details.

Conclusions

33. Based on the above findings of this pilot study, it was concluded that:
    (a) Using Rash analysis, it was confirmed that the data collected in this survey was basically valid and reliable;
    (b) Personal-aspect concerns of increasing workload and possible health hazards were found to be some factors that might possibly hinder teachers from adopting OSM and AQP;
    (c) School-aspect concerns that may hinder them from adopting OSM and AQP are their incapability in allocating money and prioritizing assessment in their school development plans;
    (d) Respondents' readiness in adopting OSM and AQP was found to be generally high with half of them having readiness above all items except personal concern of increased workload;
    (e) Principals were found to be more ready to adopt OSM and AQP than teachers;
    (f) There was no evidence to show the need of item revision and the instrument could be used for more a comprehensive survey.

Reference:

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Linacre, J.M. (2018). *A user's guide to WINSTEPS MINISTEPS Rasch Model computer programs*.

Lord, F.M. and Novick, M.R. (1968) Statistical Theories of Mental Test Scores. Addison-Wesley, Menlo Park.

Mesbah, M., Cole, B.F., Lee, M.L.T. (Eds.) (2002). *Statistical Methods for Quality of Life Studies: Design, Measurement and Analysis*. Kluwer Academic, Boston.

Smith, E.V., & Smith, R.M. (Eds.) (2004). *Introduction to Rasch Measurement: Theory, Models, and Applications*. Maple Grove, MN: JAM Press.

Smith, E.V., & Smith, R.M. (Eds.) (2007). *Rasch Measurement: Advanced and Specialized Applications*. Maple Grove, MN: JAM Press.

Smith, R. M. (1996) Polytomous Mean-Square Fit Statistics. *Rasch Measurement Transactions,* 10:3, 516-517.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8,370-371.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.

Annex

綜合試後電子平台
意見問卷調查

請就閣下和　貴校的情況，對以下各項的認同程度表達意見。請圈選適合的數字。

	個人層面	相當不同意	不同意	同意	相當同意
1.	我習慣使用平板電腦	1	2	3	4
2.	我願意學習新科技	1	2	3	4
3.	我認同評核促進學習的理念	1	2	3	4
4.	我認為自己過往設計的測考題目有改善空間	1	2	3	4
5.	我認為電子評卷長遠能增加工作效率	1	2	3	4
6.	我認為配合適當休息，電子評卷無損身體健康	1	2	3	4
7.	我會優先考慮進修有關評核的課程	1	2	3	4
8.	我認為攜帶試卷回家批改存在相當風險	1	2	3	4
9.	我認為電子評卷可以減少我的工作量	1	2	3	4
10.	我在過往評卷的過程中會反思自己的教學	1	2	3	4
	學校層面	相當不同意	不同意	同意	相當同意
11.	校方認同評核促進學習的理念	1	2	3	4
12.	校方支持使用新科技	1	2	3	4
13.	校方會運用數據支持決策	1	2	3	4
14.	校方認為校內測考題目有改善空間	1	2	3	4
15.	校方關注試卷的保安	1	2	3	4
16.	校方能調撥資源發展電子評卷	1	2	3	4
17.	校方關顧教師的工作量	1	2	3	4
18.	校方會把評估列作優先發展項目	1	2	3	4
19.	校方不時在教師發展日安排有關評估的培訓	1	2	3	4
20.	校方要求教師在測考工作上盡量避免錯失	1	2	3	4

其他意見　　如空間不足，可利用問卷背面發表意見

個人資料　　請圈選適合的數字

任職學校類型	1. 小學　2. 中學　3. 其他教育機構
職位	1. 校長　2. 副校長或課程/教務/評估主任 3. 科任老師　4. 其他，請註明：_____
使用 AQP 的經驗	1. 沒有　2. 少於兩年　3. 兩年或以上
使用電子評卷的經驗	1. 沒有　2. 有 (包括擔任 HKDSE/TSA 閱卷員)

～謝謝您的寶貴意見～