Analysis of the performance tests (Objective Structured Clinical Examination) by using Item Response Theory (IRT) and Neural Test Theory (NTT)

M Miyamoto, Y Mori, A Miyazaki, T Kubota, H Yoneda, K Suzuki
Education Center, Osaka Medical College, JAPAN

The Generalizability Theory has showed that the variance of correlations between examinees and stations is so larger than the inter-rator variance. This means that scenario or station specificity is very important in OSCE. It will be very useful if we can quantify the scenario specificity in the terms of the assessment items (checklist) used in each scenario.

Method

The Graduation OSCEs of Osaka Medical College for 6th year students have been performed from 2006 through 2009.

In the case of abdomen station, we have two scenarios of "abdominal pain" and "abdominal distension". The number of examinees for "abdominal pain" was a total of 272, 68 in 2006, 72 in 2007, 68 in 2008 and 64 in 2009. The number of examinees for "abdominal distension" was a total of 256, 62 in 2006, 70 in 2007 and 64 in 2008, and 60 in 2009.

We have used the same scenarios over the past 4 years. Each scenario for "abdominal pain" and "abdominal distension" contained 45 items including 20 for medical interview (shared) and 25 for physical examination (partially shared among both scenarios). Each student's performance was judged by two assessors. A student can get a point for each item just in the case that both rators admit.

In order to overview the properties of all the items, we used NTT. NTT is a nonparametric test theory that uses a mechanism based on the self-organizing map (SOM).

On the other hand, the Item Response Theory (IRT) needs uni-dimensionality strictly to analyze the items. It results that the IRT could be applied only to the selected assessment items. Uni-dimensionality was checked by MPLUS using Categorical Factor Analysis. We used 2-parameter-model IRT calculated by M-PLUS with the weighted least square method. The values of these difficulty and discrimination parameters were calculated for the selected items.

The examinees' abilities were calculated by BILOG-MG with Bayes (EAP)

method.

Results

1) The assessment items used in each scenario and their correct answer rates were shown in Table1 (Assessment items for Medical Interview) and 2 (Assessment items for Medical Examination Techniques).

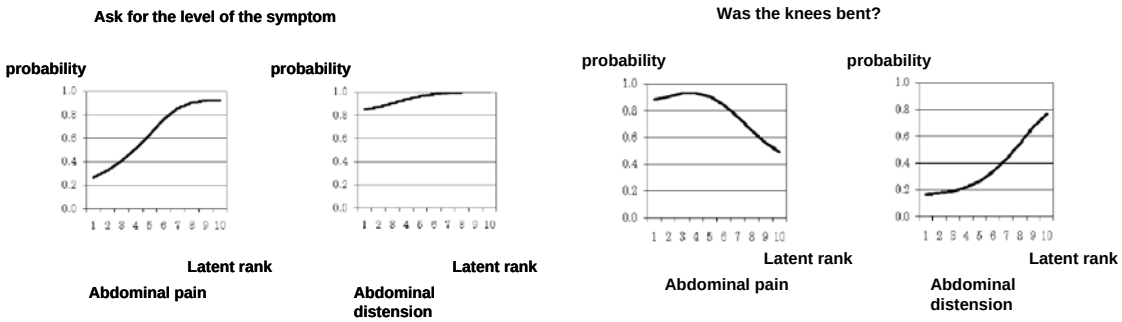
Table1　Assessment items for Medical Interview

		"abdominal pain"		"abdominal distension"	
Inteview process		average	SD	average	SD
1)	Greeting	0.981	0.138	0.934	0.249
2)	Selfintroduction	0.986	0.120	0.990	0.101
3)	Confirmation of name	0.976	0.154	0.974	0.158
4)	Listening to pt's complaints at opening	0.966	0.181	0.974	0.158
5)	Appropriate, polite wording	0.971	0.168	0.944	0.231
6)	Appropriate eye contact	0.947	0.224	0.959	0.198
7)	Appropriate posture and attitude	0.971	0.168	0.969	0.173
8)	Sympathetic words and attitude	0.760	0.428	0.878	0.329
9)	Appropriate summary and confirmation	0.933	0.251	0.918	0.275
10)	Ask whether anything has been missheard or for any questions at the end.	0.827	0.379	0.765	0.425
Interview information					
11)	Ask for the location of the symptom	0.981	0.138	0.964	0.186
12)	Ask for the properties of the symptom	0.957	0.204	0.990	0.101
13)	Ask for continuations of the symptom	0.841	0.366	0.903	0.297
14)	Ask for the level of the symptom	0.654	0.477	0.954	0.210
15)	Ask for the progress of the symptom	1.000	0.000	0.985	0.123
16)	Ask for situations when the symptom occurs	0.856	0.352	0.765	0.425
17)	Ask for factors in which it betters or worsens	0.716	0.452	0.796	0.404
18)	Ask for accompanying symptoms	0.981	0.138	0.939	0.240
19)	Ask the patient's response to the symptom	0.928	0.259	0.816	0.388
20)	Ask the condition, and symptoms of the whole body.	0.889	0.314	0.872	0.334
21)	Ask for the past history	0.798	0.402	0.837	0.371
22)	Ask for the family history	0.952	0.214	0.974	0.158
23)	Ask for extra information	0.904	0.296	0.786	0.411
24)	Ask for the social situation	0.851	0.357	0.806	0.396
25)	Ask for the pt's interpretation model	0.875	0.332	0.745	0.437
	Rating scale	4.740	0.652	4.816	0.904

Table1 2 Assessment items for Medical Examination Techniques

Scenario 1	"abdominal pain"			Scenario2	"abdominal distension"		
Consideration to patient		average	SD			average	SD
1)	Ask for physical exam and get agreement	0.913	0.282	1)	Same as scenario 1-1)	0.918	0.275
2)	Imposing voice in each examination	1.000	0.000	2)	Same as scenario 1-2)	0.985	0.123
3)	Comprehensible imposing of voice	0.981	0.138	3)	Same as scenario 1-3)	0.985	0.123
4)	Consideration to pain	0.990	0.098	4)	Same as scenario 1-4)	0.980	0.142
5)	Warming of the stethoscope	0.942	0.234	5)	Same as scenario 1-5)	0.964	0.186
Examination techniques							
6)	Was the abdomen exposed enough?	0.986	0.120	6)	Same as scenario 1-6)	0.918	0.275
7)	Was sexuality considered?	0.976	0.154	7)	Same as scenario 1-7)	0.969	0.173
8)	Examine in order of inspection → auscultating → percussion →palpation?	0.938	0.243	8)	Same as scenario 1-8)	0.898	0.303
＜Inspection＞				＜Inspection＞			
9)	Inspection of important symtoms and expression of the findings	0.928	0.259	9)	Same as scenario 1-9)	0.811	0.392
＜Auscultation＞				＜Auscultation＞			
10)	Auscultate the bowel sounds and express the findings	0.904	0.296	10)	Same as scenario 1-10)	0.954	0.210
11)	Examine the pitching sounds properly and express the findings	0.861	0.347	11)	Same as scenario 1-11)	0.878	0.329
＜Percussion＞				＜Percussion＞			
12)	Accurate percussion	0.962	0.193	12)	Same as scenario 1-12)	0.974	0.158
13)	Proper percussion on the the whole abdomen and confirmation of the painful area	0.861	0.347	13)	Same as scenario 1-13)	0.888	0.316
				14)	Was the peritoneal fluid confirmed by accurate technique and the finding expressed?	0.408	0.493
				15)	Were the upper and lower borders of the liver dullness confirmed by accurate percussion?	0.847	0.361
				16)	Was spleen enlargement confirmed by accurate percussion?	0.689	0.464
＜Palpitation＞				＜Palpitation＞			
14)	Were the lower limbs bent	0.774	0.419	17)	Same as scenario 1-14)	0.801	0.400
15)	Proper superficial palpitation on the whole abdomen first	0.947	0.224	18)	Same as scenario 1-15)	0.964	0.186
16)	Was the deep palpation performed correctly next?	0.938	0.243	19)	Same as scenario 1-16)	0.939	0.240
17)	Was the painful part (in the right low quadrant) palpated at the end?	0.803	0.399	20)	Was the liver confirmed by accurate palpation?	0.776	0.418
18)	Was the pain pressure point confirmed?	0.745	0.437				
19)	Was rebound tenderness confirmed?	0.750	0.434				
20)	Was muscular guarding confirmed?	0.620	0.487				
Rating scale		4.808	0.793	Rating scale		4.699	0.953

2) NTT was very useful to overview each features of the checklist items. We presented two examples here. In "abdominal pain", the item of "Ask for the level of the symptom" showed good discriminancy at moderate difficulty level. However, the item of "Was the knees bent?" showed that the students with high ability could not do well paradoxically.

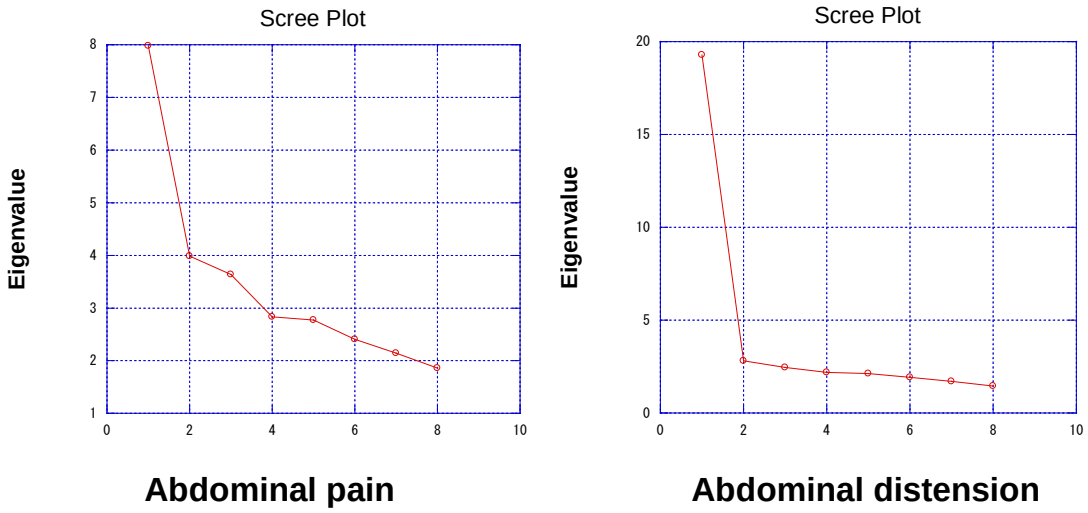


3) Eigen-values and factor loading values were calculated using tetrachoric correlation matrix by the Categorical Factor Analysis.

The 11 items and the 23 items were chosen from "abdominal pain" and "abdominal distension". The 11 items for "abdominal pain" and the 23

items for “abdominal distension” of which factor loading values exceeded 0.3, were chosen according to Lord’s criteria.

The first Eigen-values were much larger than those of the second group. The IRT-available items were mostly different between “abdominal pain” and “abdominal distension”, and these IRT-available items were specific to the scenarios.



4)The values of the difficulty and discrimination parameters in IRT model.

Table 3 "Estimated and P value of Discrimination and Difficulty for "abdominal pain" and "abdominal distension"

	Discrimination			Two-Tailed	Difficulty			Two-Tailed
Scenario1"abdominal pain"	Estimate	S.E..	Est./S.E.	P-Value	Estimate	S.E..	Est./S.E.	P-Value
Interview process 6	1.395	0.534	2.611	0.009	-1.99	0.326	-6.101	0
Interview process 7	0.901	0.353	2.554	0.011	-2.835	0.674	-4.204	0
Interview information 4	0.526	0.14	3.763	0	-0.85	0.257	-3.306	0.001
Interview information 14	0.405	0.157	2.584	0.01	-4.197	1.381	-3.038	0.002
Examination technique 1	1.079	0.261	4.139	0	-2.98	0.434	-6.86	0
Examination technique 4	0.784	0.28	2.795	0.005	-2.367	0.579	-4.088	0
Examination technique 7	0.858	0.339	2.526	0.012	-2.717	0.703	-3.865	0
Examination technique 10	1.89	0.869	2.175	0.03	-1.83	0.253	-7.242	0
Examination technique 11	1.372	0.665	2.061	0.039	-1.899	0.372	-5.108	0
Examination technique 13	0.721	0.174	4.143	0	-1.128	0.238	-4.736	0
Examination technique 14	0.387	0.14	2.775	0.006	-1.868	0.637	-2.931	0.003
average	0.938				-2.242			
standard deviation	0.465				0.929			
	Discrimination			Two-Tailed	Difficulty			Two-Tailed
Scenario2"abdominal distension"	Estimate	S.E..	Est./S.E.	P-Value	Estimate	S.E..	Est./S.E.	P-Value
Interview process 1	1.231	0.401	3.066	0.002	-1.937	0.294	-6.585	0
Interview process 8	0.648	0.164	3.955	0	-2.139	0.433	-4.937	0
Interview process 9	0.564	0.266	2.125	0.034	-2.837	1.039	-2.73	0.006
Interview process 10	0.471	0.11	4.295	0	-1.698	0.391	-4.345	0
Interview information 3	0.884	0.193	4.578	0	-1.961	0.296	-6.628	0
Interview information 4	1.447	0.685	2.113	0.035	-2.049	0.345	-5.944	0
Interview information 6	0.852	0.125	6.791	0	-1.115	0.176	-6.336	0
Interview information 7	0.747	0.131	5.72	0	-1.382	0.23	-6.014	0
Interview information 8	0.858	0.344	2.496	0.013	-2.372	0.585	-4.051	0
Interview information 10	0.461	0.167	2.766	0.006	-2.72	0.855	-3.182	0.001
Interview information 11	0.333	0.15	2.222	0.026	-3.105	1.298	-2.392	0.017
Consideration to patient 3	0.663	0.241	2.747	0.006	-2.523	0.67	-3.763	0
Examination technique 3	0.785	0.211	3.718	0	-2.057	0.386	-5.325	0
Examination technique 4	0.329	0.141	2.327	0.02	-2.824	1.139	-2.478	0.013
Examination technique 5	1.314	0.597	2.201	0.028	-2.118	0.381	-5.558	0
Examination technique 6	0.502	0.18	2.791	0.005	-2.593	0.782	-3.314	0.001
Examination technique 8	0.828	0.176	4.704	0	-1.904	0.295	-6.452	0
Examination technique 9	0.505	0.108	4.672	0	0.515	0.224	2.302	0.021
Examination technique 10	0.703	0.14	5.025	0	-1.779	0.297	-5.986	0
Examination technique 11	0.475	0.1	4.755	0	-1.148	0.286	-4.018	0
Examination technique 12	0.524	0.126	4.146	0	-1.82	0.404	-4.501	0
Examination technique 14	0.716	0.306	2.338	0.019	-2.654	0.783	-3.388	0.001
Examination technique 15	0.378	0.118	3.197	0.001	-2.14	0.646	-3.315	0.001
average	0.705				-2.016			
standard deviation	0.301				0.762			

Conclusions

1) NTT could overview the features of all the items.
2) IRT-available items were specific to each scenario.
3) We could judge the features and the qualities of the scenarios including the checklist assessment items by IRT and NTT.
4) The values of difficulty and discrimination, and the examinee-ability in OSCE, can be measured and common-scaled, then can be equated horizontally and vertically.

Reference

Shojima, K. (2007) Neural test theory. Proceedings of the International Meeting of the Psychometric Society 2007, Tokyo, Japan. p.160.