The Analysis of Response Patterns on IRT Ability Estimation Methods in Computerized Adaptive Test

Deng-Jyi Chen, Ah-Fur Lai and Chia-Chi Mao
Department of Computer Science, National Chiao Tung University, Taiwan
djchen@csie.nctu.edu.tw

Abstract

The main purpose of the study was to compare four ability estimation engines (which are OWEN, EAP, MLE, and WLE) regarding the convergent state and divergent state, and dynamic behaviors under different types of test-taker's response pattern in CAT.

1. Introduction

Due to the rapid progress of information technology, the computerized adaptive test (CAT) has been applied in many areas of test. The Graduate Record Examination (GRE) is a typical one. An appropriate ability estimation engine is a very important component in an efficient and accurate CAT system. When estimated by different ability estimation methods, different response patterns, especially extreme ones, will reveal some problems such as divergence or slow convergence. The ability estimation engine affects not only the outcome of the test but also its testing efficiency. As a whole, MLE yields less bias than Bayesian methods (such as OWEN, EAP) especially at ability extremes, but OWEN and EAP yield lower standard error as well as greater administrative efficiency than MLE [1][2][3]. In other words, different ability estimator possesses distinctive advantages and disadvantages on efficiency and accuracy. General speaking, a CAT system is embedded with one or two ability estimation engines, so it becomes more difficult to accomplish the simulation comparison among different estimators. The study have developed and implemented four ability estimator programs (as afore mentioned) to be used as the IRT ability estimation engines for a CAT system. Then different response patterns were fed to the IRT ability estimation engines in evaluating and comparing the convergent state under various response patterns, and in investigating the dynamic behavior under no test length limitation.

2. Response Patterns for Comparison

To simulate the authentic testing situation in the CAT system, this study generated 50 response patterns. These patterns include the general and extreme response patterns, can be categorized into 5 types and they are shown in Table 1 and as follows:

Type 1: all "1" bits: 111……1111 (abbrev. 1^{160})
Type 2: all "0" bits: 000……0000 (abbrev. 0^{160})
Type 3: "0" bits shifted right into all "1"s: 0…01111…1 et al. (abbrev. 01^{159} ~$0^{15}1^{145}$)
Type 4: "1" bits shifted right into all "0"s: 1…1000….0 et al. (abbrev. 10^{159} ~$1^{19}0^{141}$)
Type 5: consecutive multiple bits toggling.

Table 1. Measured proficiency and average testing length of four ability estimator

Response pattern			Average Measured proficiency				Average testing length			
Type	No	Formal expression	MLE	OWEN	WLE	EAP	MLE	OWEN	WLE	EAP
Type 3	1	01^{159}	1.140459	3	0.65578	2.813138737	7.9	21.8	12.3	20.5
	2	$0^{2}1^{158}$	0.270301	3	0.17866	2.799110503	11.7	42.3	13.5	32.8
	3	$0^{3}1^{157}$	-0.0219	3	-0.05338	2.776596592	15.4	70.8	17.3	58.3
	4	$0^{4}1^{156}$	-0.4623	2.99	-0.3995	2.754059963	17.5	116.9	18.6	75.9
	5	$0^{5}1^{155}$	-0.73248	2.985	-0.64608	2.741118327	19.1	151	19.5	94.7
	6	$0^{6}1^{154}$	-0.96511	2.365	-0.89257	2.169351832	20.5	142.6	20.2	130.4
	7	$0^{7}1^{153}$	-1.081	0.03	-1.08338	-2.740431952	27	75.4	20.1	6.9
	8	$0^{8}1^{152}$	-1.13705	-0.22	-1.12127	-2.737083179	29	58.8	21.4	7
	9	$0^{9}1^{151}$	-1.21308	-3	-1.2117	-2.738554443	31.7	7.5	22.3	7
	10	$0^{10}1^{150}$	-1.27638	-3	-1.29733	-2.741133975	35	8.2	24.9	6.9
	11	$0^{11}1^{149}$	-1.30806	-3	-1.34897	-2.739490783	37	8.2	28	7
	12	$0^{12}1^{148}$	-1.34547	-3	-1.3545	-2.73802124	40	8.1	25.5	7
	13	$0^{13}1^{147}$	-3	-3	-1.42882	-2.739972255	23	7.8	25.4	7
	14	$0^{14}1^{146}$	-3	-3	-1.50772	-2.737514816	23	7.7	26.8	6.9
	15	$0^{15}1^{145}$	-3	-3	-1.71596	-2.736416393	23	7.7	28	6.7
Type 4	16	10^{159}	-3	-3	-1.12311	-2.73912113	22.3	11	12.5	12.2
	17	$1^{2}0^{158}$	-3	-3	-1.02142	-2.73807901	24.3	13.3	7.5	14.9
	18	$1^{3}0^{157}$	-2.45084	-3	-1.09997	-2.737624389	23.6	15.4	9.8	17.2
	19	$1^{4}0^{156}$	-1.00719	-3	-0.3592	-2.735986373	12.3	18.1	12.5	18.6
	20	$1^{5}0^{155}$	0.626983	-3	0.08402	-2.736548792	8.4	20.2	13.9	20.3
	21	$1^{6}0^{154}$	-2.22801	-3	1.007075	-2.738975896	31.2	23.2	16.1	20.3
	22	$1^{7}0^{153}$	1.158391	-3	0.45178	-2.737104662	11	26.1	17	20.1
	23	$1^{8}0^{152}$	1.250299	-3	-0.13407	-2.738013422	12.1	27.1	14	20.9
	24	$1^{9}0^{151}$	1.326948	-1.8	0.22279	-2.73818342	13.3	23.9	14.7	20.9
	25	$1^{10}0^{150}$	1.233276	0.6	0.5616	1.758823586	15	17.6	15.1	12.6
	26	$1^{11}0^{149}$	2.680514	3	0.62286	2.886962787	14	9.9	18.6	10.6
	27	$1^{12}0^{148}$	3	3	0.84884	2.885447664	16	10.2	20	10.4
	28	$1^{13}0^{147}$	3	3	0.163478	2.88747604	17	9.9	21.8	10.7
	29	$1^{14}0^{146}$	3	3	1.18876	2.885531224	18	10.2	21.4	10.5
	30	$1^{15}0^{145}$	3	3	-0.42201	2.884750028	18	10.3	18	10.2
	31	$1^{16}0^{144}$	3	3	-0.27106	2.888223759	19	10.5	21.8	10.7
	32	$1^{17}0^{143}$	3	3	-2.08413	2.886549503	20	10.1	26	10.5
	33	$1^{18}0^{142}$	3	3	0.48196	2.888392655	21	10.3	25.2	10.9
	34	$1^{19}0^{141}$	3	3	0.22599	2.886079981	22	10.1	22	10.4
Type 5	35	$(10)^{80}$	-0.62862	-1.215	-0.37001	-1.341786832	8.6	130.3	10.2	160
	36	$(01)^{80}$	-2.36326	-1.41111	-0.63635	-1.775128409	11	132.2	9.6	160
	37	$(1100)^{40}$	-0.26732	-1.345	-0.18634	-1.363207141	9.5	160	11.5	160
	38	$(0011)^{40}$	-2.75744	-1.62	-0.97093	-1.650274271	16.1	160	13.1	160
	39	$(111000)^{26}1\,0$	-0.4272	-1.3	-0.24291	-1.527541047	14.5	160	11.8	160
	40	$(000111)^{26}0\,1$	-2.38905	-1.585	-1.15895	-1.802418673	34.5	160	11.2	160
	41	$(11110000)^{20}$	0.285764	-1.405	0.28377	-1.490668803	9.2	160	14	160
	42	$(00001111)^{20}$	-1.74592	-1.635	-1.16386	-2.010156756	16.2	160	10.8	160
	43	$(1111100000)^{16}$	0.634165	-1.455	0.15075	-1.539256226	8.4	160	16	160
	44	$(0000011111)^{16}$	-1.9401	-1.765	-0.75924	-1.766020416	16.2	144.5	12	160
	45	(111111000000) 1	0.726589	-1.51	0.62069	-1.537470228	16.3	160	19.2	160
	46	(000000111111) 0	-1.6942	-1.725	-1.34596	-1.912469359	13	144.6	14.6	160
	47	(11111110000000) 1 0	1.15212	-1.575	1.61745	-1.46637061	11	160	19.7	160
	48	(00000001111111) 0 1	-2.23212	-2.155	-1.52996	-2.739248982	15	98.7	16.1	6.8

3. Result and Discussion

The result of simulation is shown in Table 1. The right side of Table 1 shows the max test length

0-7695-2916-X/07 $25.00 © 2007 IEEE
IEEE COMPUTER SOCIETY

required for estimating ability when these four engines reach the convergence status, and the result indicated that the average demanded test lengths for OWEN and EAP are larger than those measured from MLE and WLE. The test length required for ability estimation in the simulated test indicates that the OWEN ability estimator converges slowly, or cannot converge even if all of the test items in the item pool are selected for testing, ranging from response pattern No. 2 to 8 and 35 to 48. For example, in reaching convergent state, response pattern No. 6 consumes 142.6 items estimated by the OWEN ability engine, but it consumes only 20.5, 20.2 as estimated by MLE and WLE respectively. For pattern No. 2 to 6 and 35 to 47, the EAP ability estimator encounters the same problem like OWEN.

In order to further survey the difference of dynamic process under four ability estimators, this study selects response pattern No. 2, 6, 35, 41 and 47, and investigates the behavior of estimated provisional ability. Fig. 1 portrays the dynamic behavior picture of response pattern No. 6. As can be seen, four ability estimators, and response pattern No. 2, 35, 41 and 47 have the same tendency as Fig. 1.

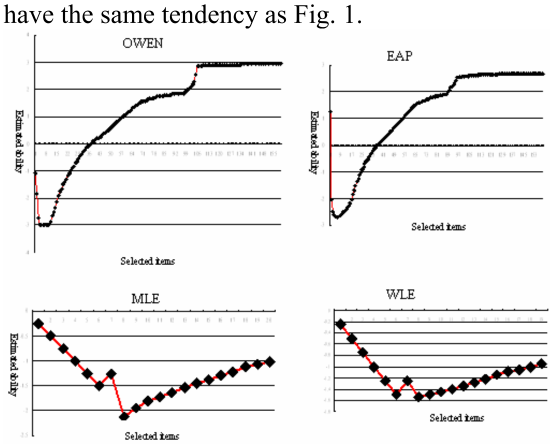

Fig. 1. The estimated ability dynamic behaviors of response pattern No. 6

But in pattern No. 2 to 6, MLE estimator has tendency of inward bias, i.e., this estimator will figure out higher proficiency value for low ability test takers and evaluate lower proficiency value for high ability ones as illustrated in Fig. 2. This result is the same as the research reported by [4]. As shown in Fig. 2, it also indicates that the estimated ability using MLE is significantly changeable for the same pattern, especially in pattern No. 2 to 6. In other words, MLE estimator will encounter the inward bias of ability estimated error.

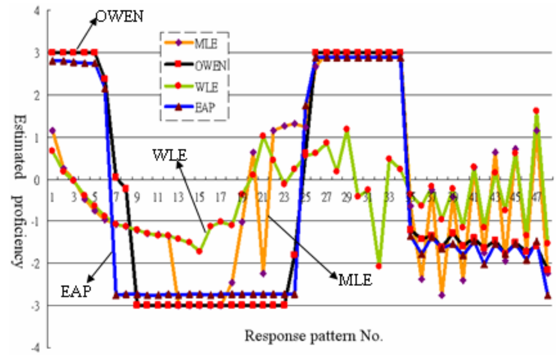

Fig. 2. Estimated proficiency of 48 response patterns by four ability estimators

4. Conclusion

This study designs 50 response patterns to simulate the convergence status of four ability estimation methods. The result indicated that the Bayesian ability methods (OWEN and EAP) would result in slow convergence or divergence even if running out of the item pool for some response patterns. On the contrary, MLE and WLE would typically produce convergent status for the same response patterns.

To overcome the issue of contradiction between ability estimation accuracy and testing efficiency caused by a specific response pattern for different ability estimation methods, it is suggested that a CAT system driven by a single ability estimation engine is to be transformed into the multiple CAT ability estimation engines (e.g. EAP+WLE) scheme. Whenever the system detects a specific response pattern to cause the default engine to reach the convergent state slowly, the mechanism of the multiple ability estimation engines would automatically switch to another appropriate engine and to continue re-estimating the examinee's proficiency again for the same response pattern.

5. References

[1] T. Wang, and W.P. Vispoel, "Properties of ability Estimation Methods in Computerized Adaptive Testing," *Jounal of Educational Measurement*, 35 (2), 109-135, 1998.

[2] T.A. Warm, "Weighted likelihood estimation of ability in item response theory," *Psychometrika*, pp. 54, 427-450, 1989.

[3] R.D. Bock, and R.J. Mislevy, "Adaptive EAP Estimation of Ability in a Microcomputer Environment," *Applied Psychological Measurement*, 6, 431-444, 1982.

[4] Q. Yi, T. Wang, and J.C. Ban, "Effects of scale transformation and test-termination rule on the precision of ability estimation in computer adaptive testing," *Journal of educational Measurements*, 38 (3), 267-292, 2001.

IEEE COMPUTER SOCIETY