ED 465 764 TM 033 903

AUTHOR	Hau, Kit-Tai; Wen, Jian-Bing; Chang, Hua-Hua
TITLE	Optimum Number of Strata in the a-Stratified Computerized Adaptive Testing Design.
PUB DATE	2002-04-00
NOTE	22p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Ability; *Adaptive Testing; *Computer Assisted Testing; Estimation (Mathematics); *Selection; Simulation; *Test Items
IDENTIFIERS	*Item Discrimination (Tests); Item Exposure (Tests); Test Security

ABSTRACT

In the a-stratified method, a popular and efficient item exposure control strategy proposed by H. Chang (H. Chang and Z. Ying, 1999; K. Hau and H. Chang, 2001) for computerized adaptive testing (CAT), the item pool and item selection process has usually been divided into four strata and the corresponding four stages. In a series of simulation studies, researchers examined the optimum number of strata by systematically varying the number of strata, pool size (200, 400, and 800 items), item characteristics (0., 0.5 correlation between difficulty and discrimination), and item selection method (largest information, matching estimated ability with difficulty). Results show that quite independent of the item pool size and the correlation between item discrimination and difficulty, ability estimation deteriorated while the number of over- and under-exposed items decreased with an increase in stratum number. There is a diminishing return in that dividing the pool into too many strata can also be problematic because when the stratum is too small, there are not any items of close difficulty for each particular examinee. The results are in general agreement with the speculation that too few and too many strata may not provide optimum efficiency and balanced item pool utilization. It is shown that the ideal and optimum number of strata to be used in each specific application depend on the item pool structure, test length, and other testing conditions. (Contains 8 figures, 24 tables, and 8 references.) (Author/SLD)

Optimum Number of Strata in the a-Stratified Computerized Adaptive Testing Design

Kit-Tai Hau, The Chinese University of Hong Kong
Jian-Bing Wen, The Chinese University of Hong Kong
Hua-Hua Chang, University of Texas, Austin

Note. Paper presented at the American Educational Research Association Conference, New Orleans, April 1-5, 2002. Please address all correspondence to Professor Kit-Tai Hau, Faculty of Education, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (email: KTHAU@CUHK.EDU.HK).

Abstract

In the a-stratified method, a popular and efficient item exposure control strategy proposed by Chang (Chang & Ying, 1999; Hau & Chang, 2001) for computerized adaptive testing (CAT), the item pool and item selection process has been usually divided into four strata and the corresponding four stages. In a series of simulated studies, we examined the optimum number of strata by systematically varying the number of strata, pool size (200, 400, 800 items), item characteristics (0, .5 correlation between difficulty and discrimination), and item selection method (largest information, matching estimated ability with difficulty). Results showed that quite independent of the item pool size and the correlation between item discrimination and difficulty, ability estimation deteriorated while the number of over- and under-exposed items decreased with an increase in stratum number. But there is a diminishing return in that dividing the pool into too many strata would also be problematic because when the stratum was too small, there would not be any item of close enough difficulty for each particular examinee. The results are in general agreement with our speculation that too few and too many strata may not provide the optimum efficiency and balanced item pool utilization. It is shown that the ideal and optimum number of strata to be used in each specific application depend on the item pool structure, test length, and other testing conditions.

With the advancement in computer technology and respective psychometric theories, computerized adaptive testing (CAT) has moved from pure research to large scale implementation during the early 1990s. In the a-stratified method, a popular and efficient item exposure control strategy proposed by Chang (Chang & Ying, 1999; Hau & Chang, 2001), the item pool and item selection process has been usually divided into four strata and the corresponding four stages. In this study, the optimum number of stages and strata with respective to item pool and testing characteristics was explored.

In CAT, tailoring items to test-takers' ability through the selection of appropriate items would be desirable because an examinee is measured most effectively when the items are neither too difficult nor too easy. The logic behind the most prevalent item selection strategy can be mathematically derived (Hau & Chang, 2001). In item selection, aside from non-statistical considerations such as content balancing, the most common strategy in the last three decades has been the maximization of item information. Specifically, an item will be selected if it has the maximum information at the currently estimated θ level, which is calculated from the examinee's available responses at that instant (see also other alternatives, e.g., Chang & Ying, 1996; Owen, 1975).

Item information has been typically defined as Fisher information that varies as a function of the test-taker's ability θ. Consider the simple case when all items follow $c \equiv 0$ (i.e., a two parameter model). Then, Fisher information increases monotonically with a, items with high a's will be preferentially selected (e.g., see Hau & Chang, 2001).

Test Security, Exposure Control and a-Stratified Design

Test security has been a serious problem in CAT. In contrast to a paper-and-pencil test where examinees are tested with an identical set of items at the same time, in a CAT examinees are tested individually or in small groups with items being reused for examinees at different sessions. Understandably, test security becomes a problem because examinees can remember and share the item content with others. To avoid item content leakage, it is therefore important to control the frequency with which an item is administered to test-takers. In other words, monitoring items' exposure rate to prevent overexposure is necessary to enhance test security.

Remedies to restrain the over-exposure of high discrimination items have been proposed by McBride & Martin (1983), Sympson and Hetter (1985), Stocking and Lewis (1995), Davey & Parshall (1995), Thomasson (1995), and others. This issue has drawn particularly great attention from researchers when CAT is implemented in high stake tests like TOEFL and ASVAB-CAT. Working with a totally different item selection philosophy in that a proactive mechanism should be devised to equalize the exposure of high and low discrimination items, Chang (see review, Chang & Ying, 1999) demonstrated the benefit of using their multi-stage a-stratified design.

Essentially in the a-stratified method, the item pool is divided into several strata in an ascending order of their discrimination parameter (for details see Chang & Ying, 1999 or Hau

& Chang, 2001). The corresponding CAT is also divided into the same number of stages. Within each stage of testing, items with difficulty closest to the estimated ability are selected from the corresponding pool stratum. Thus, in actual operation, items with smaller a-parameters are selected first from the strata with less discriminating items, while larger a-parameter items are left for latter stages. Since the estimates of examinee's ability are not close to the true value during early stages, the use of high a-parameter items do not necessarily imply a greater precision in ability estimation. Actually simulation studies showed that this a-stratified method can equalize item exposure without damaging ability estimation efficiency and accuracy (Chang & Ying, 1999).

If test security is the only concern, then all examinees should be given a random sample of items from the pool. The random selection tends to approximately equalize the exposure rates of all items in the pool and consequently will help to minimize the item overlap among examinees. On the other hand, if efficiency in ability estimation is the only concern, then according to Fisher information criterion, the high discrimination items should be used instead. The efficiency gain will be at the expense of the unbalanced item usage and the greater cost in item replenishment. In other words, if the total budget in test maintenance is kept constant, apparently there is a tradeoff between test security and efficiency. If both factors are important as in a high stakes examination, then the testing agency has no choice but to spend more money on test development and maintenance, which subsequently results in a many folds increase in the examination fee. Despite the seeming incompatibility between test security and efficiency, the above tradeoff may be avoidable if a method can be found that has a balanced item usage yet maintains efficiency.

The a-stratified strategy has at least three potential advantages. Firstly, it may provide an efficiency in ability estimation comparable to the traditional maximum information approach. Secondly, it automatically leads to a more even item exposure rate control. The major cause for unevenly distributed item exposure and subsequent security problems is that large *a* items are more likely to be selected than the small *a* ones. In the a stratified method, exposure rates will become more evenly distributed because proportionally equal numbers of items are chosen from strata of high, medium and low *a* parameters. Thirdly, in comparison to maximum information integrated with Sympson and Hetter Method, the stratified method is simpler to implement (see Hau & Chang, 2001).

Optimum Number of Stratum

In most of the stratified designs (e.g., Chang & Ying, 1999; Hau & Chang, 2001), four strata have been used. However, there has not been any attempt to determine how the number of strata would affect the efficiency and item over-exposure. There can be two extremes in the number of strata. On one extreme, if only one stratum, instead of the usual four strata, is used, then all items will be in the same stratum. Within this stratum, items with difficulty nearest to the examinee's current estimated ability will be selected. The stratified design in that case will differ from the maximum information approach in that in the former design, the discrimination parameter has not been considered. Thus, such a stratified design with one stratum should have an efficiency lower than that of the maximum information approach. However, if the distribution of item difficulty matches that of the examinees, then item usage will be relatively balanced.

On the other hand, if the number of strata equals to the preset test length, then these strata and hence the items selected will be arranged strictly in the order of ascending discrimination items. That is, item selection will always start from the stratum with the lowest discrimination items and then the items selected will monotonically increase in discrimination. If there are insufficient items of diversified difficulties within each of these strata, then dividing the item pool into many strata may decrease the chance of getting an item close enough to the desired difficulty. In that case, efficiency in ability estimation will suffer, but the impact on item usage may be quite complicated depending on the original pool characteristics.

It can also be speculated that the overall testing performance depends on the number of strata and hence the size of items within each stratum. If there are many items of various levels of discrimination and difficulty within each stratum, then using many strata will lead to a relatively high efficiency, while perhaps at some degree of sacrifice of a more balanced item usage.

The present study will examine the above hypothesis as regards the optimum number of strata through simulation studies with item pool imitating operational conditions as well as other characteristics. The objective is to find the relationship between testing performance (efficiency and item pool usage) the stratification process (number of strata adopted).

Simulated Studies

In a series of simulated studies, we systematically varied the Number of Strata in the stratified approach under a 3 Pool Size (number of items in the pool, 3 levels) X 2 Item Characteristics X 2 Item Selection methods design.

Pool Size. Three item bank of different sizes were examined which contain 200 (small pool), 400 (medium pool) and 800 (large pool) items respectively.

Item Characteristics. Two item banks were purposely designed to examine how item characteristics might interact with the number of strata. The two-parameter logistic model is used in these two item banks. Both item banks contained items with a normal distribution of item difficulty matching students' ability distribution. The first set of items displayed a hypothetical situation in which item difficulty and discrimination were not correlated in the sense that within each ability range, there were items with various levels of discrimination (a = 0.4 to 2.0). On the other hand, the second set of items demonstrated a situation in which difficulty was moderated correlated with discrimination at .5. That means more difficult items were relatively more discriminating while easier items were relatively less discrimination.

Latent trait distribution. Five thousand θ values were generated from a standardized normal distribution $N(0,1)$.

Test algorithm. Two different test lengths, 24 and 48 items respectively, were used in simulations. The item pool was partitioned into 1, 2, 3, 4, 6, 8, 12, 24 strata when tests had 24 items, and 1, 2, 3, 4, 6, 8, 12, 16, 24, 48 strata when tests had 48 items. Testing was divided into respective stages parallel to each stratum. Two different item selection methods were used to select items in each test stage. In one, items which provided the most information to the current estimated ability level was selected; while in the other, items whose difficulty was closest to the estimated ability were selected. The maximum likelihood method was used to estimate the ability in simulations.

Evaluation Criterion. The different designs were compared in terms of the test information, error of ability estimation, item exposure and test overlap rate.

Test information can be taken as the index of test efficiency in fix-length CAT tests. The larger the amount of test information test provide, the more efficient the test algorithm is. Test information is the sum of all the Fisher item information in the test.

$$I(\theta)=\sum_{i=1}^{n} I_i(\theta)=\sum_{i=1}^{n}\frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}$$

Bias and mean squared error (MSE) are used to evaluate accuracy of ability estimate, which are respectively defined as:

$$\text{Bias}=\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i-\theta_i)$$

$$\text{MSE}=\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i-\theta_i)^2$$

where m is the number of simulated examinees and θ_i and $\hat{\theta}_i$ are the true and estimated ability of the *i*th examinee. The correlation of the θ_i and $\hat{\theta}_i$ is also calculated and taken as one index of the estimation accuracy.

For item exposure, the χ^2 statistics proposed by Chang & Ying (1999) is used to measure the skewness of item exposure rate distribution in variable length CAT.

$$\chi^2=\sum_{i=1}^{N}\frac{\left[A_i-\left(\sum_{i=1}^{N}A_i\Big/N\right)\right]^2}{\sum_{i=1}^{N}A_i\Big/N}$$

where N is the total number of items in the bank, A_i is the item exposure rate of the ith item in the bank. The smaller the χ^2 statistics, the closer to the uniform distribution the item exposure rate is. All item exposure rates are equal when χ^2 statistics is 0.

The test overlap rate is another parameter indicating the quality of different item selection design. It is defined as the expected number of common items encountered by two randomly selected examinees divided by the expected test length in variable-length CAT. There are C_M^2 pairs of tests among M examinees,

$$R_t = \frac{TO_{总} / C_M^2}{(\sum_{i=1}^{M} L_i)/M} = \frac{2TO_{总}}{(M-1)\sum_{i=1}^{M} L_i}.$$

The numbers of over- and under-exposed items are also used as additional information about the item pool usage in these methods.

Results and Discussion

The results of simulations can be seen from the general trends in Tables 1 to 24. All the methods being examined were generally satisfactory in ability estimation with average bias not larger than 0.01. The correlation between the true and estimated abilities was consistently above .97 for test length 24, and was larger than .98 when test length increased to 48. For all stages in the testing and in congruence with common sense, when the pool size increased, the test overlap rate would decrease accordingly. It is understandable because with greater number of items in the pool, the probability of an item being selected will be decreased in general which subsequently lead to a lowering of the test overlap rate (Chang & Ying, 1999).

Selecting items whose difficulty level is closest to the estimated ability level would lead to less efficient item pool usage when the number of strata increased. As test overlap rate increased, the chi square statistics became larger when the item pool and the testing were partitioned into more strata. This trend was also reflected by the increase in the number of over- and under-exposed items. When there was only one stratum, with items selected solely on item difficulty, the item pool usage would be most balanced. For testing with items being partitioned into more strata, it is quite difficult to find items to match examinees' estimate abilities. For simulations with the same item pool and the same number of strata, results showed that test length had a direct effect on chi-square – an indicator of skewness of item exposure, with skewness being increased with an increase in test length.

When items of maximum information were selected from a stratum, it is logical to expect that the larger the size of the stratum (i.e., the smaller the number the pool is being stratified), the greater the chance to find a suitable item of large information. So, the most informative item would be chosen if there is only one stratum. When the item pool was partitioned into more strata, test information would decrease and the estimation would become worse.

Quite independent of the item pool size and the correlation between *a*s and *b*s, the MSE of estimates increased and the correlation between estimates and true values decreased when the number of strata increased. That is, ability estimation deteriorated with increasing stratum number. However, in terms pool usage, the number of over- and under-exposed items decreased with an increase in stratum number. The test overlap rate and Chi square statistics would decrease accordingly. But there is a diminishing return in that dividing the pool into too many strata would also be problematic because when the stratum was too small, there would not be any item of close enough difficulty for each particular examinee.

The results are in general agreement with our speculation that too few and too many strata may not provide the optimum efficiency and balanced item pool utilization. It is shown that the ideal and optimum number of strata to be used in each specific application depend on the item pool structure, test length, and other testing conditions. The results also confirm that test efficiency and the balanced usage of items do not necessarily increase or decrease monotonically with the number of strata.

An implication for item pool management is that in an operational CAT design, the optimum number of strata should be determined through simulation studies under conditions specifically chosen for that particular application. Furthermore, future research should be conducted in which the philosophy of using less discrimination items in the earlier stages of testing without can be implemented without physically partitioning and stratification of the item pools.

References

Chang, H. H. & Ying, Z. L. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23(3),* 211-222.

Davey, T., & Parshall, C. (1995 April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper Presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Hau, K. T., & Chang, H. H. (2001). Item Selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38, 249-266.*

McBride, J. R. & Martin, J. T. (1983*). Reliability and validity of adaptive ability tests in a military setting*. In D.J. Weiss (Ed.), *New horizons in testing* (p223-226). New York, Academic Press.

Owen, Z. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of American Statistical Association, 70*, 351-356.

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing*. Research Report 95-25. Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing*. Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thomasson, G. L. (1995, June). *New item exposure control algorithms for computerized adaptive testing*. Paper presented at the Annual Meeting of Psychometric Society, Minneapolis, MN.

Table 1 Indicators of Test performance at max length = 48, Pool Size = 200 item, R_{ab}=.5, selecting items with max information in each stratum

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0000	0.0226	0.9888	48	111	26.5915	0.3728	55.05
2	0.0063	0.0258	0.9876	37	107	24.3689	0.3616	49.65
3	0.0006	0.0265	0.9872	32	116	21.3675	0.3466	46.27
4	0.0032	0.0267	0.9870	23	117	16.7313	0.3235	46.36
6	0.0012	0.0300	0.9855	15	118	16.9837	0.3247	43.27
8	-0.0018	0.0374	0.9822	13	116	15.3636	0.3166	42.22
12	0.0021	0.0366	0.9824	13	116	14.1558	0.3106	39.07
16	0.0039	0.0397	0.9810	13	111	17.9920	0.3297	38.37
24	-0.0044	0.0429	0.9797	12	110	19.8045	0.3388	35.47
48	-0.0031	0.0579	0.9742	27	94	49.4643	0.4871	28.33

Table 2 Indicators of Test performance at max length = 48, Pool Size = 400 item, R_{ab}=.5, selecting items with max information in each stratum

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	-0.0006	0.0167	0.9916	205	114	69.7351	0.2941	67.16
2	0.0027	0.0202	0.9900	198	111	63.3637	0.2782	58.15
3	0.0040	0.0205	0.9897	182	114	56.2720	0.2605	53.68
4	0.0030	0.0208	0.9895	163	104	49.1741	0.2427	52.73
6	-0.0005	0.0229	0.9886	150	90	45.2822	0.2330	48.94
8	0.0051	0.0229	0.9886	127	78	34.1641	0.2052	49.22
12	0.0052	0.0244	0.9881	111	62	34.0843	0.2050	45.74
16	0.0014	0.0252	0.9877	86	51	28.2334	0.1904	45.83
24	0.0035	0.0298	0.9853	76	45	34.2584	0.2054	40.51
48	0.0012	0.0400	0.9808	95	67	45.6680	0.2340	32.35

Table 3 Indicators of Test performance at max length = 48, Pool Size = 800 item, R_{ab}=.5, selecting items with max information in each stratum

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0017	0.0147	0.9926	569	91	146.3511	0.2427	79.70
2	0.0002	0.0163	0.9919	554	99	127.7846	0.2195	65.73
3	0.0001	0.0187	0.9907	540	84	121.5655	0.2118	58.50
4	0.0050	0.0192	0.9904	523	73	106.6910	0.1932	56.69
6	0.0002	0.0214	0.9893	489	53	93.4599	0.1766	52.36
8	0.0001	0.0216	0.9893	476	42	76.3342	0.1552	52.48
12	0.0002	0.0230	0.9885	445	29	68.1021	0.1449	48.16
16	0.0036	0.0237	0.9883	422	26	61.7990	0.1370	48.91
24	0.0017	0.0272	0.9868	397	20	51.3116	0.1239	45.14
48	-0.0012	0.0328	0.9840	430	23	63.5775	0.1393	36.44

Table 4 Indicators of Test performance at max length = 48, Pool Size = 200 item, R_{ab}=.0, selecting items with max information in each stratum

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0020	0.0201	0.9900	48	102	29.3150	0.3864	56.70
2	-0.0003	0.0210	0.9896	35	115	23.0109	0.3549	51.17
3	-0.0007	0.0229	0.9887	23	118	17.3640	0.3266	48.74
4	-0.0001	0.0236	0.9885	17	127	13.6139	0.3079	48.38
6	-0.0019	0.0254	0.9875	12	130	12.6232	0.3029	45.09
8	-0.0040	0.0265	0.9870	9	127	11.9321	0.2995	45.07
12	-0.0011	0.0286	0.9858	8	118	12.7070	0.3033	40.51
16	-0.0013	0.0289	0.9856	4	119	14.9548	0.3146	40.42
24	-0.0012	0.0335	0.9838	7	103	21.2522	0.3461	36.06
48	0.0031	0.0402	0.9806	31	96	41.5227	0.4474	28.19

Table 5 Indicators of Test performance at max length = 48, Pool Size = 400 item, R_{ab}=.0, selecting items with max information in each stratum

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0010	0.0150	0.9925	200	115	66.9274	0.2871	70.92
2	-0.0011	0.0175	0.9912	181	104	59.0635	0.2675	61.79
3	0.0014	0.0185	0.9907	162	93	49.5359	0.2436	57.16
4	-0.0011	0.0195	0.9903	146	88	43.2696	0.2280	56.34
6	-0.0028	0.0206	0.9897	118	81	35.3348	0.2081	52.50
8	-0.0004	0.0206	0.9897	111	64	27.2541	0.1879	52.75
12	0.0008	0.0226	0.9887	72	47	22.2132	0.1753	48.73
16	-0.0005	0.0219	0.9891	61	27	19.5103	0.1686	49.21
24	-0.0002	0.0249	0.9876	62	48	23.8777	0.1795	42.96
48	-0.0020	0.0338	0.9836	102	63	42.3178	0.2256	34.06

Table 6 Indicators of Test performance at max length = 48, Pool Size = 800 item, R_{ab}= 0, selecting items with max information in each stratum

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	-0.0007	0.0124	0.9937	577	101	153.5595	0.2517	85.55
2	0.0002	0.0156	0.9922	550	90	127.5477	0.2192	70.17
3	0.0014	0.0169	0.9914	529	72	111.7488	0.1995	62.10
4	-0.0007	0.0175	0.9912	508	64	99.4887	0.1842	60.24
6	0.0050	0.0192	0.9904	472	46	84.4746	0.1654	54.76
8	0.0021	0.0191	0.9904	442	37	68.8343	0.1458	54.59
12	0.0024	0.0209	0.9895	416	21	54.8971	0.1284	50.10
16	0.0023	0.0204	0.9898	376	10	46.8826	0.1184	51.28
24	-0.0054	0.0235	0.9883	353	9	39.3870	0.1090	46.26
48	0.0004	0.0297	0.9854	385	12	48.3519	0.1202	36.18

Table 7 Indicators of Test performance at max length = 48, Pool Size = 200 item, R_{ab}=.5, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0069	0.0280	0.9862	1	137	4.4029	0.2618	42.52
2	0.0013	0.0290	0.9858	1	109	8.9089	0.2843	42.38
3	0.0040	0.0285	0.9860	4	109	10.9096	0.2943	41.86
4	0.0010	0.0308	0.9850	2	109	10.6869	0.2932	41.82
6	0.0054	0.0295	0.9854	5	111	13.5917	0.3078	40.86
8	-0.0019	0.0307	0.9853	5	103	15.8884	0.3192	39.28
12	0.0023	0.0320	0.9843	10	110	16.8275	0.3239	38.22
16	0.0052	0.0343	0.9835	13	103	20.9611	0.3446	37.06
24	0.0082	0.0363	0.9828	15	106	18.7229	0.3334	35.79
48	0.0065	0.0392	0.9814	20	101	24.8545	0.3641	32.37

Table 8 Indicators of Test performance at max length = 48, Pool Size = 400 item, R_{ab}=.5, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0083	0.0261	0.9870	11	14	5.6486	0.1339	43.51
2	0.0013	0.0247	0.9878	36	55	15.6183	0.1588	45.22
3	0.0039	0.0240	0.9881	48	51	18.2597	0.1654	44.64
4	0.0039	0.0239	0.9881	63	62	19.7809	0.1693	45.75
6	0.0054	0.0252	0.9873	78	58	21.2723	0.1730	43.77
8	0.0042	0.0247	0.9877	82	58	22.4248	0.1759	45.30
12	0.0013	0.0261	0.9870	79	62	23.6313	0.1789	42.66
16	0.0004	0.0282	0.9860	85	62	28.8186	0.1918	43.58
24	-0.0017	0.0288	0.9858	98	70	33.6032	0.2038	38.96
48	-0.0028	0.0418	0.9805	124	72	54.1303	0.2551	31.93

Table 9 Indicators of Test performance at max length = 48, Pool Size = 800 item, R_{ab}=.5, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0058	0.0270	0.9866	310	0	12.5232	0.0755	42.58
2	0.0049	0.0245	0.9878	413	7	23.7616	0.0895	44.95
3	0.0030	0.0238	0.9880	429	21	28.6336	0.0956	45.57
4	0.0024	0.0234	0.9883	426	25	34.5217	0.1030	45.54
6	-0.0016	0.0225	0.9887	419	22	31.4544	0.0991	45.80
8	0.0022	0.0235	0.9883	431	24	35.8152	0.1046	45.77
12	-0.0015	0.0231	0.9885	444	27	38.2821	0.1077	45.15
16	0.0042	0.0242	0.9880	450	24	39.5071	0.1092	44.92
24	0.0026	0.0245	0.9879	455	22	43.1798	0.1138	44.01
48	-0.0014	0.0278	0.9864	458	32	50.3252	0.1227	41.89

Table 10 Indicators of Test performance at max length = 48, Pool Size = 200 item, R_{ab}= 0, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0005	0.0240	0.9881	0	155	3.6245	0.2579	44.15
2	0.0022	0.0254	0.9874	0	141	4.9502	0.2646	44.20
3	0.0038	0.0244	0.9878	2	135	6.7112	0.2734	44.27
4	0.0008	0.0250	0.9879	0	138	6.9610	0.2746	43.92
6	0.0033	0.0258	0.9871	2	118	10.1230	0.2904	42.39
8	0.0020	0.0263	0.9871	1	116	11.9262	0.2994	41.54
12	-0.0005	0.0274	0.9865	4	109	14.0770	0.3102	39.97
16	0.0013	0.0285	0.9859	6	108	16.6571	0.3231	39.15
24	-0.0001	0.0300	0.9853	6	102	19.0438	0.3350	37.39
48	0.0019	0.0333	0.9837	17	96	25.9798	0.3697	33.19

Table 11 Indicators of Test performance at max length = 48, Pool Size = 400 item, R_{ab}= 0, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	-0.0006	0.0236	0.9882	14	18	6.4502	0.1359	45.65
2	0.0020	0.0225	0.9887	27	27	9.4102	0.1433	48.32
3	0.0002	0.0222	0.9889	27	27	9.2526	0.1429	47.96
4	-0.0016	0.0220	0.9890	29	32	10.8693	0.1470	49.20
6	-0.0004	0.0222	0.9889	34	43	11.8546	0.1494	47.19
8	-0.0017	0.0228	0.9886	55	49	15.6763	0.1590	48.69
12	0.0003	0.0232	0.9884	68	56	18.4289	0.1659	45.47
16	-0.0023	0.0220	0.9891	74	56	23.2075	0.1778	46.72
24	0.0010	0.0257	0.9872	94	58	28.4660	0.1910	41.29
48	-0.0025	0.0360	0.9826	120	73	49.4220	0.2434	33.55

Table 12 Indicators of Test performance at max length = 48, Pool Size = 800 item, R_{ab}= 0, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	-0.0007	0.0260	0.9869	325	1	12.6819	0.0757	40.61
2	-0.0009	0.0238	0.9881	322	2	12.6850	0.0757	44.67
3	-0.0004	0.0225	0.9887	346	2	13.3311	0.0765	45.78
4	-0.0020	0.0223	0.9888	330	1	12.9378	0.0760	46.35
6	-0.0018	0.0229	0.9886	331	1	12.4443	0.0754	46.70
8	0.0012	0.0224	0.9887	358	1	13.7734	0.0770	46.57
12	0.0007	0.0237	0.9882	376	3	17.7725	0.0820	46.07
16	-0.0026	0.0234	0.9884	381	3	17.4435	0.0816	45.75
24	-0.0006	0.0230	0.9886	398	8	23.4815	0.0892	44.68
48	0.0005	0.0248	0.9875	410	15	31.3449	0.0990	42.39

Table 13 Indicators of Test performance at max length = 24, Pool Size = 200 item, R_{ab}= .50, selecting items with maximum information

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0054	0.0370	0.9822	101	50	33.2450	0.2860	32.86
2	0.0040	0.0423	0.9799	93	53	29.5935	0.2678	27.55
3	0.0015	0.0465	0.9778	90	55	26.1380	0.2505	25.17
4	0.0055	0.0495	0.9764	75	46	21.8199	0.2289	24.07
6	0.0091	0.0491	0.9760	68	35	20.3323	0.2215	22.58
8	0.0047	0.0534	0.9742	57	31	16.5650	0.2026	21.39
12	0.0032	0.0587	0.9726	47	31	13.3237	0.1864	20.13
24	0.0020	0.0651	0.9687	38	35	11.6573	0.1781	18.36

Table 14 Indicators of Test performance at max length = 24, Pool Size = 400 item, R_{ab}= .50, selecting items with maximum information

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0024	0.0288	0.9858	289	47	74.9121	0.2471	38.88
2	0.0033	0.0351	0.9827	281	44	67.2981	0.2280	31.44
3	0.0038	0.0394	0.9808	271	34	61.2960	0.2130	28.32
4	0.0034	0.0402	0.9801	261	42	51.6573	0.1889	26.79
6	0.0045	0.0449	0.9780	244	19	43.7907	0.1693	24.84
8	0.0034	0.0465	0.9773	225	13	35.2784	0.1480	23.85
12	0.0007	0.0506	0.9752	205	11	30.7097	0.1366	22.55
24	0.0036	0.0565	0.9725	194	11	22.0036	0.1148	20.61

Table 15 Indicators of Test performance at max length = 24, Pool Size = 800 item, R_{ab}= .50, selecting items with maximum information

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0059	0.0260	0.9871	675	41	147.9329	0.2147	43.58
2	0.0041	0.0341	0.9832	666	35	127.5425	0.1892	33.80
3	0.0028	0.0363	0.9822	654	29	117.7298	0.1770	29.79
4	0.0033	0.0398	0.9806	643	23	96.3508	0.1502	27.78
6	0.0033	0.0428	0.9790	632	10	82.8697	0.1334	25.78
8	0.0040	0.0423	0.9794	612	7	64.3495	0.1102	24.86
12	0.0077	0.0487	0.9763	594	6	52.9836	0.0960	23.38
24	0.0038	0.0515	0.9750	647	7	33.2624	0.0714	21.87

Table 16 Indicators of Test performance at max length = 24, Pool Size = 200 item, R_{ab}= .5, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0156	0.0605	0.9711	17	15	4.5894	0.1427	19.97
2	0.0032	0.0595	0.9720	23	33	9.0781	0.1652	20.45
3	0.0038	0.0574	0.9729	28	30	9.8611	0.1691	20.66
4	0.0115	0.0585	0.9728	25	34	10.3243	0.1714	20.62
6	0.0088	0.0569	0.9726	29	39	12.1688	0.1806	20.27
8	0.0045	0.0579	0.9723	45	32	13.3385	0.1865	19.74
12	0.0033	0.0629	0.9703	47	34	15.2270	0.1959	19.11
24	0.0007	0.0663	0.9682	58	39	15.6672	0.1981	17.92

Table 17 Indicators of Test performance at max length = 24, Pool Size = 400 item, R_{ab}= .5, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0091	0.0587	0.9711	170	0	7.1052	0.0776	19.78
2	0.0020	0.0521	0.9746	208	6	12.5680	0.0912	20.79
3	0.0016	0.0532	0.9741	212	8	14.9510	0.0972	21.03
4	-0.0031	0.0508	0.9751	213	9	15.8635	0.0995	21.25
6	0.0043	0.0516	0.9750	219	9	17.3199	0.1031	21.04
8	0.0002	0.0539	0.9741	207	10	17.7527	0.1042	21.14
12	0.0063	0.0536	0.9741	215	8	17.7671	0.1042	20.76
24	0.0025	0.0570	0.9722	225	11	20.2718	0.1105	19.87

Table 18 Indicators of Test performance at max length = 24, Pool Size = 800 item, R_{ab}= .5, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0173	0.0620	0.9702	664	0	11.6322	0.0443	19.35
2	0.0072	0.0533	0.9739	673	0	17.0832	0.0512	21.07
3	0.0048	0.0506	0.9754	667	0	19.0893	0.0537	21.47
4	0.0090	0.0525	0.9747	669	3	23.1211	0.0587	21.55
6	0.0043	0.0509	0.9751	670	2	22.7154	0.0582	21.67
8	0.0014	0.0491	0.9760	666	1	23.6849	0.0594	21.78
12	0.0041	0.0515	0.9750	669	3	26.6440	0.0631	21.49
24	-0.0021	0.0536	0.9743	663	5	30.8575	0.0684	20.94

Table 19 Indicators of Test performance at max length = 24, Pool Size = 200 item, R_{ab}= 0, selecting items with maximum item information in each stratum

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	-0.0009	0.0312	0.9846	102	50	35.4999	0.2973	35.43
2	0.0020	0.0367	0.9821	92	50	31.9279	0.2794	29.43
3	0.0059	0.0399	0.9804	73	47	23.7744	0.2387	26.85
4	0.0015	0.0425	0.9793	69	45	20.7337	0.2235	25.51
6	-0.0037	0.0447	0.9779	58	34	16.5787	0.2027	23.78
8	0.0001	0.0479	0.9767	50	30	13.7023	0.1883	22.72
12	-0.0020	0.0500	0.9758	33	24	10.8214	0.1739	21.46
24	0.0056	0.0591	0.9717	40	27	12.6740	0.1832	19.14

Table 20 Indicators of Test performance at max length = 24, Pool Size = 400 item, R_{ab}= 0, selecting items with maximum item information in each stratum

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	-0.0010	0.0257	0.9872	286	48	73.7171	0.2441	41.01
2	-0.0018	0.0332	0.9835	276	41	64.5355	0.2211	33.38
3	0.0011	0.0347	0.9829	270	36	56.7150	0.2016	30.08
4	-0.0007	0.0368	0.9816	256	29	47.1555	0.1777	28.51
6	0.0051	0.0405	0.9801	230	17	37.4599	0.1534	26.39
8	-0.0047	0.0421	0.9791	215	11	29.6782	0.1340	25.35
12	-0.0008	0.0453	0.9778	192	9	21.9747	0.1147	23.81
24	-0.0011	0.0493	0.9759	175	6	14.3786	0.0957	21.50

Table 21 Indicators of Test performance at max length = 24, Pool Size = 800 item, R_{ab}= 0, selecting items with maximum item information in each stratum

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0007	0.0225	0.9888	675	43	150.0131	0.2173	48.71
2	0.0001	0.0281	0.9860	664	37	123.9967	0.1848	36.83
3	0.0020	0.0333	0.9825	658	33	109.8222	0.1671	32.06
4	-0.0001	0.0368	0.9818	640	19	96.4262	0.1503	29.70
6	-0.0026	0.0403	0.9800	631	12	78.3909	0.1278	27.00
8	-0.0004	0.0416	0.9794	612	7	64.8405	0.1109	25.60
12	-0.0060	0.0450	0.9781	597	6	46.3135	0.0877	24.06
24	0.0026	0.0481	0.9764	650	5	23.8534	0.0596	21.94

Table 22 Indicators of Test performance at max length = 24, Pool Size = 200 item, R_{ab}= 0, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	-0.0018	0.0511	0.9748	8	12	4.0251	0.1399	20.68
2	0.0016	0.0487	0.9762	10	15	4.8375	0.1440	21.74
3	0.0008	0.0507	0.9755	23	17	6.5475	0.1525	21.74
4	0.0015	0.0500	0.9757	23	21	6.9296	0.1544	21.66
6	-0.0006	0.0534	0.9742	24	24	9.4783	0.1672	20.97
8	0.0005	0.0521	0.9745	31	30	10.3007	0.1713	20.70
12	0.0014	0.0581	0.9718	41	34	12.6108	0.1829	19.77
24	0.0031	0.0595	0.9718	52	36	16.7403	0.2035	18.56

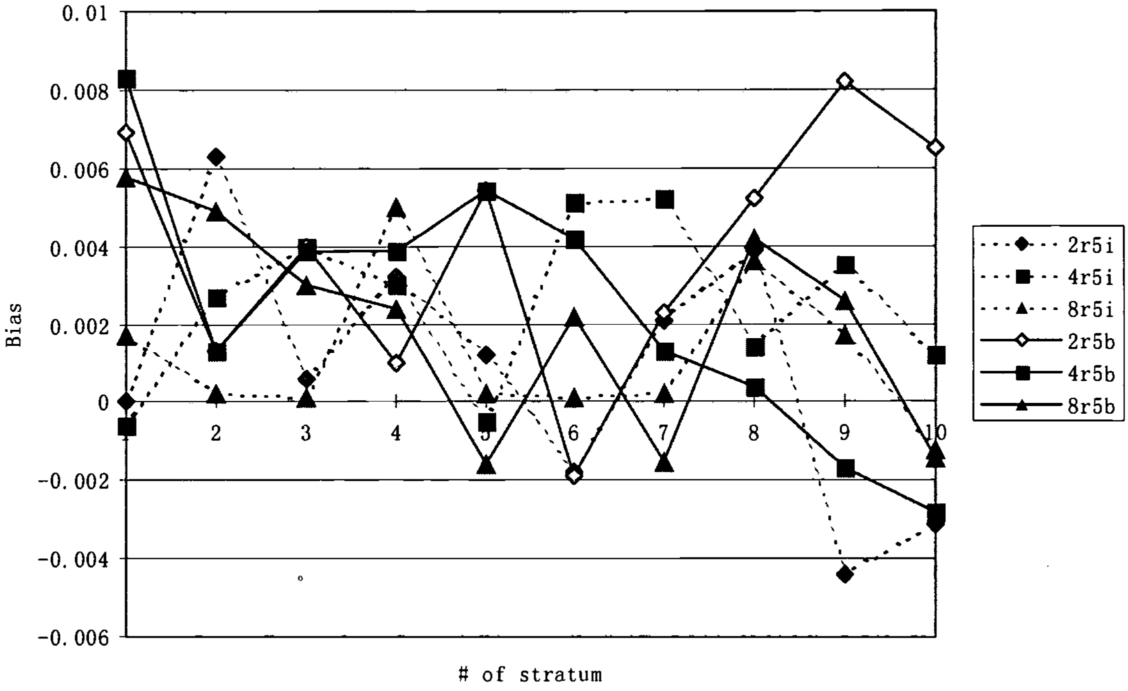
Table 23 Indicators of Test performance at max length = 24, Pool Size = 400 item, R_{ab}= 0, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0010	0.0530	0.9742	175	2	7.7041	0.0791	20.17
2	0.0048	0.0482	0.9767	185	4	9.3019	0.0831	22.08
3	0.0019	0.0460	0.9774	185	3	9.1280	0.0826	22.62
4	0.0034	0.0464	0.9771	191	5	10.2211	0.0854	22.75
6	0.0065	0.0485	0.9765	207	3	11.7357	0.0891	22.64
8	0.0012	0.0462	0.9771	212	6	14.3997	0.0958	22.64
12	-0.0026	0.0499	0.9756	201	5	14.7793	0.0967	22.04
24	0.0036	0.0511	0.9750	207	6	17.2195	0.1028	20.71

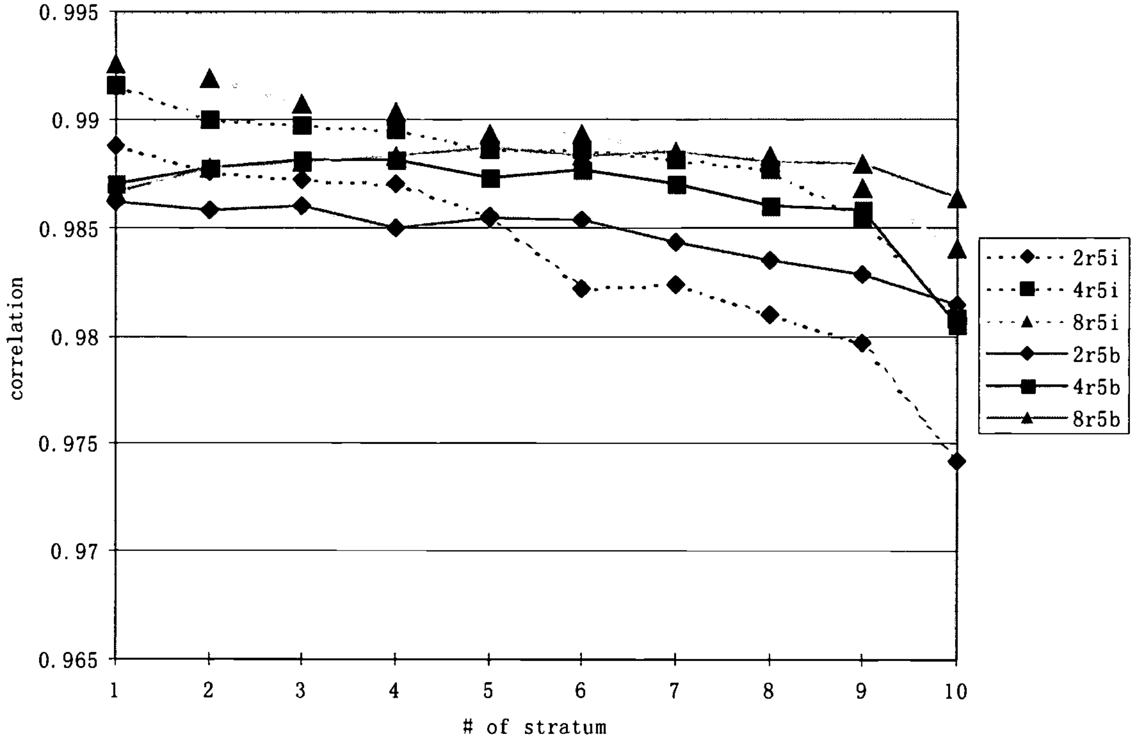
Table 24 Indicators of Test performance at max length = 24, Pool Size = 800 item, R_{ab}= 0, selecting items matching item difficulty

Stage number	bias	MSE	R	Under-exposed <=0.05	Over-exposed >=0.20	Chi^2	Test overlap	Test Info
1	0.0044	0.0608	0.9702	676	0	11.8420	0.0446	18.02
2	0.0042	0.0551	0.9733	695	0	10.9582	0.0435	20.37
3	0.0060	0.0488	0.9758	685	0	11.4178	0.0441	21.10
4	0.0024	0.0505	0.9752	682	0	11.0138	0.0436	21.55
6	0.0010	0.0465	0.9771	687	0	11.2557	0.0439	21.84
8	0.0044	0.0470	0.9767	683	0	11.2328	0.0438	21.84
12	0.0009	0.0481	0.9763	672	0	13.4224	0.0466	21.65
24	0.0019	0.0524	0.9746	672	0	17.5653	0.0518	20.96

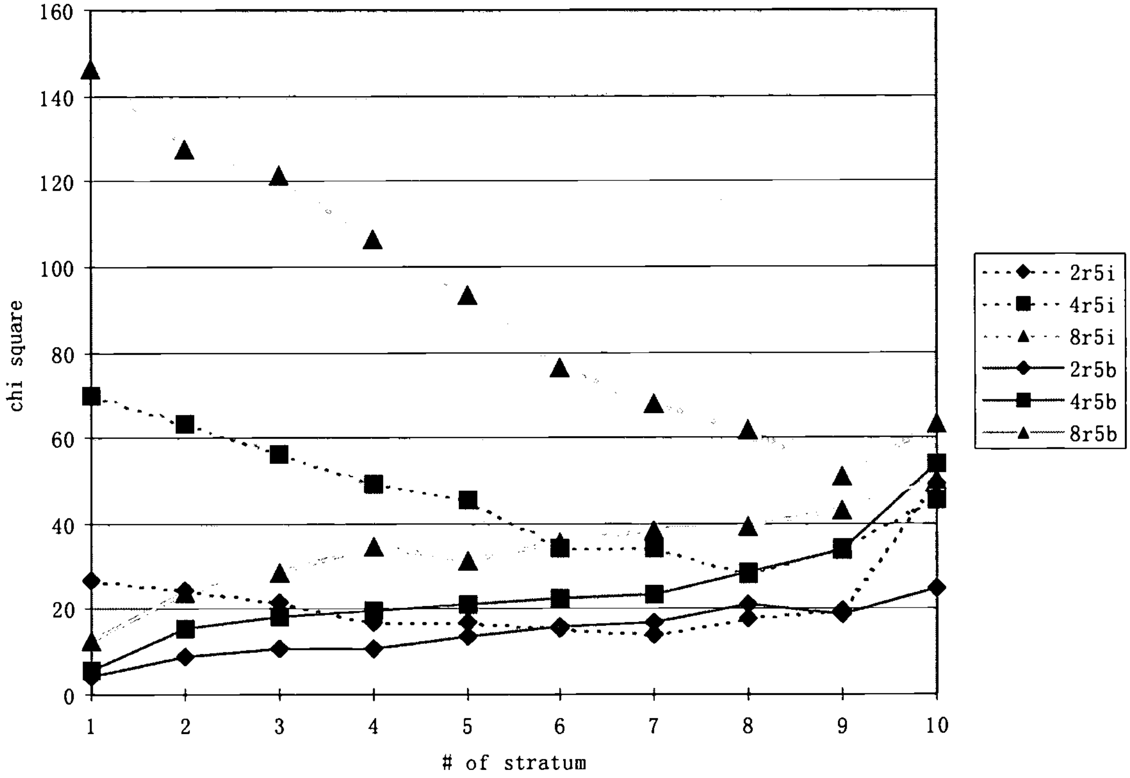
48 items



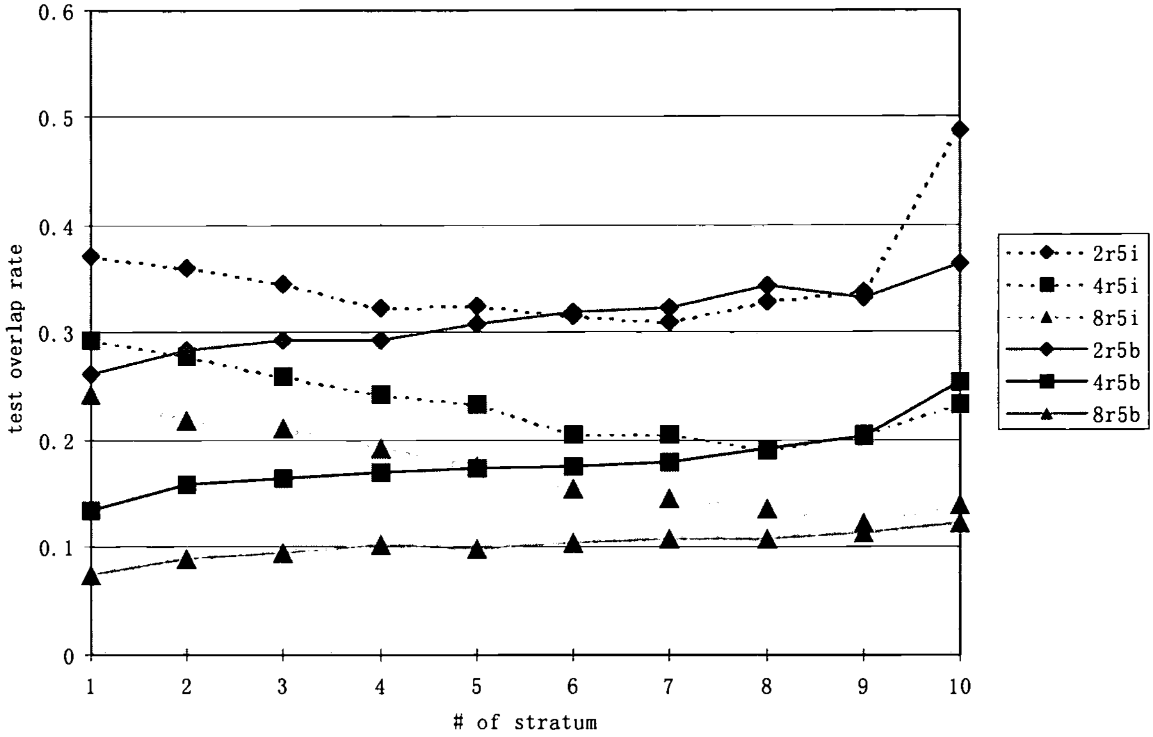
48 items



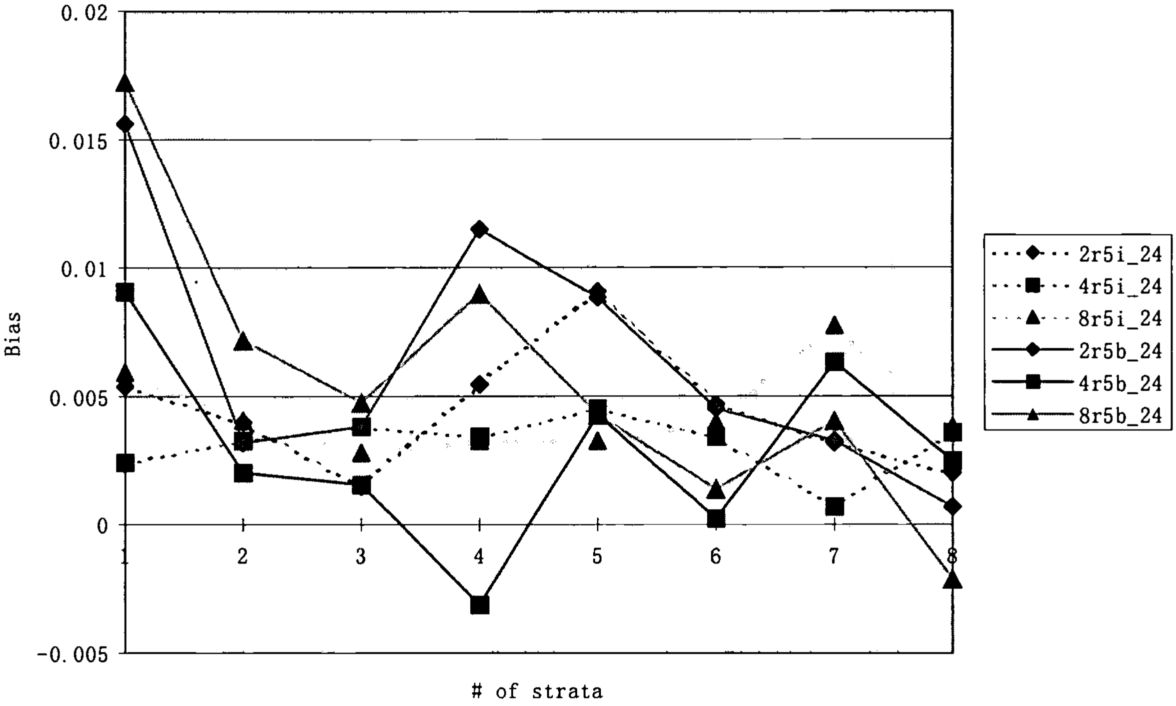
48 items



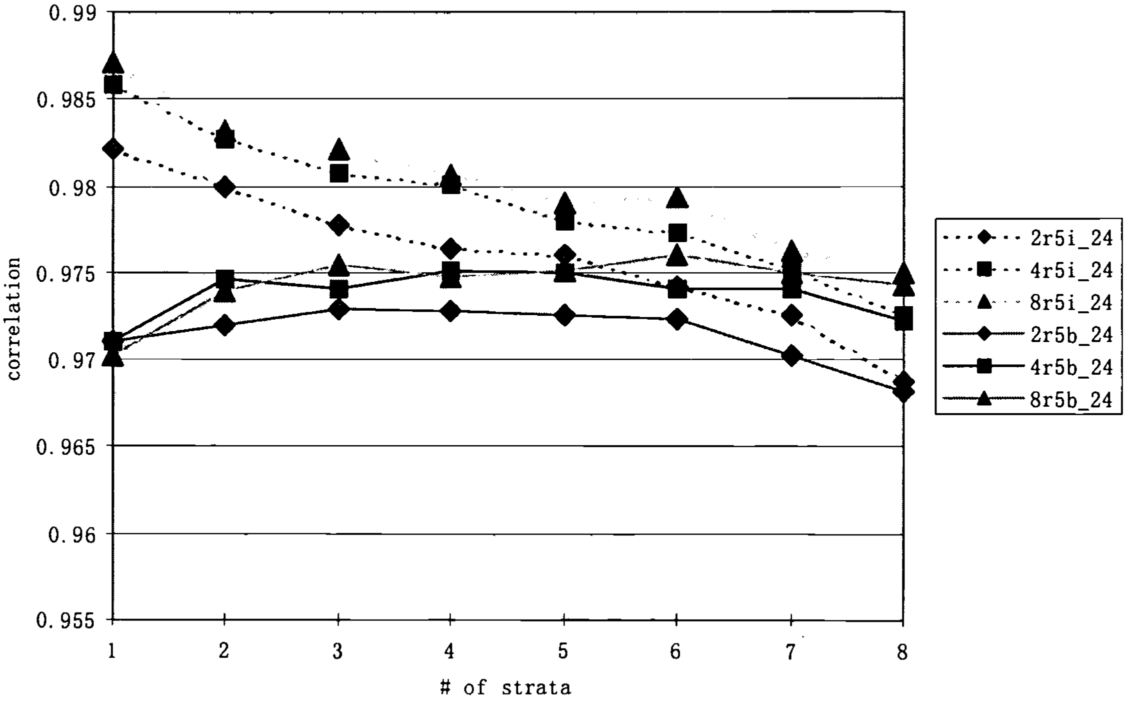
48 items



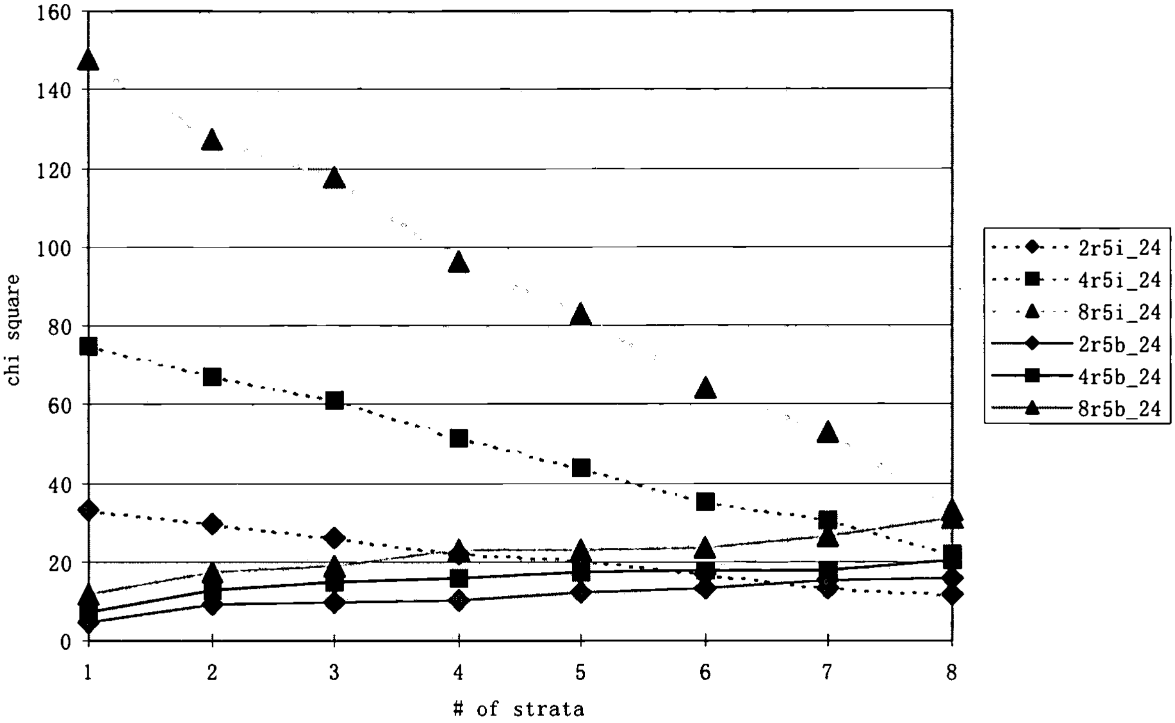
24 items



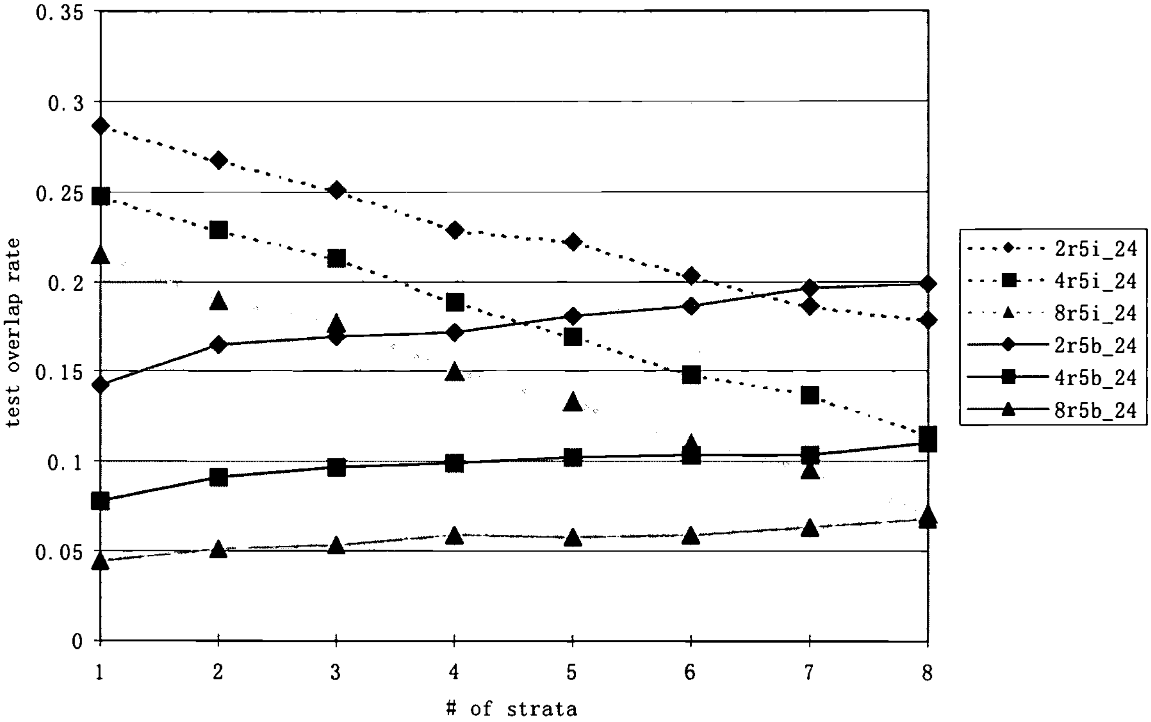
24 items



24 items



24 items



TM033903

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: optimum number of strata in the α-stratified computerized adaptive testing design	
Author(s): Kit-Tai Hau, Hua-Hua chang, Jian-Bing Wen	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
☑	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:	Printed Name/Position/Title: department chair. educational Psychology dept. HAU, KIT-TAI. professor	
Organization/Address: Faculty of Education, The chinese University of Hongkong. Shatin. N.T. Hong kong	Telephone: (852) 2609 6944	Fax: (852) 2603 6129
	E-mail Address: KTHAU@CUHK.Edu.HK	Date: 10/April/2002

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:	
ERIC Clearinghouse on Assessment and Evaluation **1129 Shriver Laboratory (Bldg 075)** **College Park, Maryland 20742**	**Telephone: 301-405-7449** **Toll Free: 800-464-3742** **Fax: 301-405-8134** **ericae@ericae.net** **http://ericae.net**