A Study of Multiple Stage Adaptive Test Designs

Ronald Armstrong & Jennifer Edmonds

Rutgers Business School

Newark/New Brunswick

Rutgers University

Ronald D. Armstrong
Management Science & Information Systems
Rutgers University
180 University Avenue
Newark, NJ 07102-1895

mprescott@comcast.net

March 26, 2004

Paper presented at the annual meeting of NCME, San Diego, CA, April, 2004.

This research has been funded by a grant from the Law School Admission Council.

A Study of Multiple Stage Adaptive Test Designs

Abstract

This paper evaluates several multiple stage adaptive test (MST) designs under the criteria of (a) accuracy of test scores, (b) simplicity of the design to facilitate review, and (c) efficiency of item pool usage to reduce the cost of item development. A commercially available mixed integer programming package was used to assemble MSTs. Analytical techniques, based on Item Response Theory, were used to evaluate scoring accuracy of an MST. Results with an operational item pool were tabulated.

A Study of Multiple Stage Adaptive Test Designs

Introduction.

Traditional paper and pencil (P&P) tests are linear tests and the same items are administered to every examinee during an examination. In computer adaptive testing (CAT), examinees receive items that match their current ability estimate. The last decade has seen paper-and-pencil (P&P) tests being replaced by computerized adaptive tests (CATs) within many testing programs. A multi-stage computer adaptive test (MST) combines characteristics of both a standard CAT and P&P test because it adapts to the ability of the examinee like CAT and provides P&P benefits such as test specialist review, exposure of pre-selected items and parallel test forms. The advantages and disadvantages of the MST approach to testing can be found in Luecht and Nungester (1998), Luecht and Nungester (2000), Luecht and Burgin (2003), and Armstrong, Jones, Koppel and Pashley (to appear). A discussion of MST designs for credentialing exams is given by both Hambleton and Xing (2004), and Xing and Hambleton (to appear).

MST designs for an admission test are considered in this paper. Several MST designs are presented and each design is systematically evaluated. The evaluation considers the following criteria: simplicity of the design, item pool usage, and scoring reliability. The possibility of review by test specialists and the quantity of response data where many examinees have been administered identical item sequences are enhanced by a simple design. Also, a simple design generally requires fewer items for the MST. A valuable resource for any testing agency is its item pool. The design should effectively utilize the pool. The capability of the design to support multiple tests from the pool will be studied under different scenarios. The main reason to employ an adaptive test is to achieve better scoring accuracy than a P&P test with fewer items. Several measures of scoring accuracy will be used to compare the reliability of the tests created from the design.

Once designs have become operational at a testing agency, they may be difficult to change since standardization across administrations is desired. A small benefit achieved by one design over another can have a significant impact over years. The effort taken to thoroughly evaluate the designs is highly justified and important for the testing agency. This study presents the summary results necessary for management to make their decisions. A comprehensive study on the strengths and weaknesses of various designs will assist the administration at a testing agency when determining the appropriate design to implement in an operational setting.

The MST structure contains multiple stages and consists of bins in which testlets are placed. The bins at a given stage are arranged in levels corresponding to ability classifications. This paper presents methods to evaluate an MST design (MSTD). Table 1 shows the skeleton of a possible MSTD. A bin may contain more than one testlet when the MST is created. For example, two testlets would be included in Stage 1 and Stage 3 of an MST assembled with the design of Table 1. A shorthand notation for the design of Table 1 would be a 1-1-2-3-3 design, where the numbers indicate levels at the stage and the number is repeated for each testlet in the stage.

(insert Table 1 about here)

Within a stage, bins are indexed using the convention that the lower index describes a targeting for a lower ability group. For discussion purposes, assume that every testlet for this MSTD contains between 5 and 7 items, and there are between 35 and 37 items on each path. The first stage (bin 1) will contain two testlets (5-7 items each) designed for the complete ability range of the test-taking population. The test taker advances to one of the two bins in the third stage based on the number of correct responses in the first two testlets. A possible method for routing out of stage (*i*) is the following. Proceed to bin 2 if the total number of correct responses to the items in bins 1 is less than 7, and proceed to bin 3 otherwise. This study will use the number correct for routing as described in Armstrong and Roussos (2002). A possible sequence of bins an examinee may traverse is referred to as a *path*. The items on a path are predetermined and correspond to a linear test form.

A 3-parameter IRT (Lord, 1980) model will be used. Target information functions and target characteristic curves are defined for each bin. The sum of the bin targets on a path yields the path targets. The MST assembly requires path information functions and path characteristic curves to be within a tolerance of the path targets. All other constraints (categorical constraints) are scaled versions of the constraints of a corresponding P&P test.

Experiment Design.

The MST assembly problem can be modeled as a mixed integer programming (MIP) problem (Nemhauser and Wolsey, 1988). All test assembly constraints are satisfied by each path of the MST. The constraints on the optimization problem are those usually associated with automated test assembly. They include limitations on the cognitive skills content, answer key count, topic, diversity and stimulus usage. The constraint examples can found in Boekkooi-Timminga, E. (1990), Theunissen (1985) and van der Linden (1998). A complete statement of the mixed integer programming model can be found in Armstrong and Edmonds (2003).

Item pool

This study used an operational item pool designed to support a paper-and-pencil test (LSAT). No attempt was made to provide an improved design for a pool to support an MST approach. Xing and Hambleton (to appear) note the importance of the item pool on the overall quality of the tests produced from it. A method to derive an item bank design is given by van der Linden, Veldkamp, and Reese (2000). While using an existing item pool yields important results, a different pool (even an artificially created pool) might be better for MST design review. Of course, any pool used for evaluation purposes must be a reasonable pool that could be supported by item writers, test specialists and/or item cloning. The use of an existing pool does allow a strong argument that the pool can be supported.

Bin percentiles

Modifying the bin percentiles affects both the bin targets and the routing rules. All the designs considered here have 100% of the population taking the first two testlets (first stage). The second stage always has two bins with a 50%-50% split between the bins. This means that the lowest 50% of all scores in the first stage go to the lower bin of stage 2 and the upper 50% of the scores go to the upper bin of stage 2. It is impossible to achieve exactly the split stated for the latter stages when the same routing rules are used for all examinees. For designs with four stages, the third stage will always have three bins with a 33%-33%-33% split. The final stage can have three types of bin partitioning rules: even splits of the test-taking population, 20% in each outer bin of the final stage, and 13% in each outer bin of the final stage. Examples of the bin partitioning rules are show in Tables 2 and 3 for both three and four stage MSTDs. By directing a smaller group of individuals to the outer bins, the intention was to provide additional scoring accuracy for the examinees with extremely high or low abilities.

(insert Tables 2 & 3 about here)

A design with four stages allows for further categorization of the test-takers. The 4-stage designs have eight possible paths for test-takers. Experience has indicated that more than four levels at the final stage provide a negligible increase in scoring accuracy and make the design too complex.

Parameters for bin target creation

Bin targets will be created using the method described in Armstrong and Roussos (2002). Armstrong and Roussos derive targets using an omniscient testing method which simulates the administration of a multiple linear test with knowledge of the test takers' true abilities; the test takers are drawn from standard normal distribution. The omniscient test assembly problem has the following objective function:

$$\text{Minimize } \sum_i \left(\alpha_1 \tilde{u}_i + \alpha_2 \bar{u}_i - \alpha_3 I_i(\theta) \right) x_i \, .$$

The binary variable x_i equals 1 if the i^{th} item is assigned to the test and equals 0 otherwise. The parameter $\tilde{u}_i$ is the random cost associated with each item, and this study sets α_1 equals 1. The parameter $\bar{u}_i$ is the exposure rate of the i^{th} item, which can be calculated as the total number of tests in which the item has appeared divided by the total number of tests administered. The value of $\alpha_2 > 0$ can be considered as a penalty coefficient for item usage. The third term, $-\alpha_3 I_i(\theta)$, represents the focus of standard CAT implementations where the objective is to maximize information; and $I_i(\theta)$ is the information associated with each item. The constraints in the optimization problem are the same as the constraints mentioned above, excluding the information and characteristic function constraints. This study used 15,000 examinees to stabilize the exposure rate and another 15,000 examinees to obtain the targets.

Armstrong and Roussos adjust the information provided by an item with a penalty based on the exposure rate. The two parameters, α_2 and α_3, can be varied in the target creation process; a larger α_2 puts an emphasis on item pool utilization, and a larger α_3 puts an emphasis on item information. An increased emphasis on information will lead to a reduced utilization of the pool, and an increased emphasis on controlling exposure will reduce the test reliability. In this study, α_2 will be set at 75, a level identified by Armstrong and Roussos; α_3 will range from

20 to 40. As the information parameter is increased, the targets for a given MSTD also increase. Figure 1 shows the individual bin targets for the testlets in the final stage of the MSTD shown in Table 1.

(insert Figure 1 about here)

These individual targets for each bin increase as α_3 increases. The figure compares the targets when the information parameter is equal to 20 to the targets when the information parameter is equal to 40. The path targets are a sum of the testlet targets corresponding to that path, so the path targets also increase as α_3 increases. Although each incremental increase in α_3 leads to small increase in the corresponding targets, sizeable differences in results are seen across the range of α_3 values. For example, the number of non-overlapping MSTs assembled from the item pool drops from 21 to 19 as α_3 increases from 20 to 25; as α_3 is increased from 20 to 40, the number of MSTs decreases from 21 to 15 (see Table 5).

Number and position of bins

All designs have the number of bins at a stage non-decreasing. The MSTDs are distinguished according to the number of testlets appearing in the bins of a stage. If the stage has two testlets to a bin, the stage number will be repeated twice, if there is only one testlet per bin in a stage, the stage number will only appear once. For example, the MSTD discussed in Table 1 will be named 1-1-2-3-3. This name means that there are two testlets in each bin of stages *(i)*, and *(iii)*. Since the number of levels at a stage is non-decreasing, and all MSTDs begin with one level in the first stage, one can infer that there is one level in stage *(i)*, one level in stage *(ii)*, and three levels in stage *(iii)*. The following 3-stage designs will be evaluated: 1-1-2-2-3, 1-1-2-3-3, and 1-1-2-2-3-3. All of the 3-stage MSTDs had 4 paths, where an examinee reaching the upper (lower) level at stage 2 could not advance to the lowest (highest) level at stage 3. The 4-stage equivalents of these designs will also be evaluated; these include: 1-1-2-3-4, 1-1-2-3-3-4, and 1-1-2-2-3-4. All the 4-stage MSTDs had 8 paths.

Number of testlets per bin

The number of items per testlet will follow the current rules used for the P&P LSAT with set based items. The stimulus and associated items would be a testlet when set based items are used. The MSTs of this paper constrain the number of items in a testlet to be between 5 and 7. All designs begin with two testlets in the first stage. There will never be more than two testlets at a stage. The total number of testlets on any path will be limited to 6, and all bins at a stage will have the same number of testlets.

The above stipulations reduced the number of possible designs, but there still were too many design variations to perform a thorough evaluation of every design; in particular, the target possibilities were extremely large. The number of designs was logically reduced to a manageable number.

Results and Discussion.

This section presents computational results from assembling MSTs with discrete items from an operational item pool. There were 1336 items in the item pool. All solution times came from runs on a desktop personal computer with a Pentium 4 CPU, 3.06 GHz, 2.00GB of RAM, and Windows XP operating system. The item pool was saved in a Microsoft Access database. All MIP problems were solved with CPLEX (ILOG, 2000). Programs extracting data from the database, formulating the problems, and writing the assembled MSTs to the database were written in AMPL (Fourer, Gay, and Kernighan (2003)).

Simplicity of the Design

Simpler designs typically have fewer items per test form and fewer paths. An item can appear multiple times on an MST, but at most once on a path. The assembly model encourages multiple occurrences of an item on an MST by randomly choosing a value for each item from a uniform distribution. This value is then used as the objective coefficient for the MIP for assignments of the item to any bin. The random numbers are generated anew for each MST assembly. Table 4 shows the average number of items per MSTD with the changing information parameter.

(insert Table 3 about here)

From this table, it can be observed that for the 3-stage test designs, the simplest design came from the first bin partitioning rule, even splits of the test-taking population. When the information parameter was greater than 35, however, the lowest number of items per test form came from the second bin partitioning rule, where the outer bins received the highest and lowest 20% of the population. Table 9 shows that for the 4-stage MSTDs, the fewest items per test form was also achieved with the even splits rule.

Item Pool Usage

Item pool usage can be described using the maximum number of tests assembled from the item pool. Table 5 shows the maximum number of tests assembled for the various test designs evaluated. The 5-testlet designs require fewer items per path than the 6-testlet design; thus, it is reasonable that these designs find more tests than the 6-testlet design.

(insert Table 5 about here)

Table 5 also shows that the most tests assembled come from designs using the first bin partitioning rule, with the population of test takers evenly split among the bins of the test design, and with the lowest level of the information parameter ($\alpha_3 = 20$). The assembly process found fewer tests when $\alpha_3 < 20$ than at higher levels of the information parameter. Also, the scoring reliability achieved for $\alpha_3 < 20$ was considered unsatisfactory. Thus, MSTDs with the information parameter less than 20 were not evaluated further.

Scoring Reliability

Scoring reliability is described using the conditional standard error of the scaled score and fidelity. Scaled scores were obtained by true score equating table and the scaled scores ranges from 20 to 80. These summary values were calculated assuming a stand normal distribution of the population ability.

(insert Tables 6 & 7)

The 1-1-2-2-3-3 design has the lowest overall SE out of the three 3-stage designs evaluated. Within that MSTD, for each bin partitioning rule, no significant improvement is seen with α_3>35. However, for the 5-stage designs, the 13% rule achieves the lowest standard error. Fidelity is the correlation between the scaled scores obtained from the number correct probability distribution conditioned on ability and the scaled score based on the true score derived from the assumed known ability. Numerical integration was used. The highest values for fidelity are seen with the 1-1-2-2-3-3 MSTD; and for the 5-testlet designs, the 13% rule achieves better results (higher fidelity values) than the other bin partitioning rules, but for the 6-testlet designs, the 33% or even splits rule outperforms the other bin partitioning rules in terms of maximizing fidelity.

(insert Table 8 about here)

(insert Table 9 about here)

Table 9 shows the maximum number of tests assembled from the item pool with a sequential assembly method where MSTs were assembled one at a time and all items previously used on MSTs were removed from the pool. Table 9 also gives the average number of items per test form, standard error, standard deviation of scaled test scores, and fidelity values from the 4-stage MSTDs with the information parameter set at 25. Surprisingly, the results were less favorable than for the 3-stage designs.

(insert Figure 2 about here)

The reason for this reduction in scoring accuracy comes from lower information targets. This result contradicts the generally accepted theory that additional adaptation improves scoring accuracy. We are currently looking closely at this paradox. Nevertheless, trial runs, to this point, show minimal improvement in going beyond three levels at the final stage. A major disadvantage of the 4-stage designs is that it is more complex, requiring more items and more paths to review. Also, the run time for the assembly models was longer than that for the 3-stage models. For a full comparison of solution time with respect to MSTD and information parameters, refer to Tables 10 & 11.

(insert Tables 10 and 11 about here)

Conclusions.

The implementation of the MST approach to CAT can take place in fairly straightforward manner. The items (testlets) are predetermined and simple routing rules based on number correct are specified. However, the various components of an MSTD make the process of selecting the “best” MSTD a complex problem. Within these MSTDs, the first bin partitioning rule, even splits of the test-taking population, is the most favorable in terms of item pool usage and simplicity of the test design. Also, it provides approximately the same exposure rate for each testlet at a given stage. However, the 13% rule was the most favorable in terms of maximizing scoring reliability. For example, conditional on $\theta = 3.0$, the conditional error of the scaled score was generally reduced by about 6.5% by using the 13% rule as opposed to the even splits rule.

The 5-testlet designs provided less scoring reliability than the 6-testlet designs. The 5-testlet designs, however, found, on average, three more tests per assembly with a given MST design. This was not a surprise since the number of items on each path was about six more when the additional testlet was included. The real issue was the magnitude of the differences and how the agency views the trade-offs. The 4-stage was inferior to the 3-stage design in terms of simplicity and item pool usage. The experiments, to date, indicate little improvement in scoring reliability by adding a fourth stage. Further computational investigation is ongoing with the consideration of additional designs. In particular, designs with three levels at the second stage will be considered along with additional operational item pools.

The conclusions derived from this study are highly dependent on the item pool used to assemble the MSTs. As an agency moves from a linear testing approach to an adaptive one, a consideration for the make-up of the item pool should be considered. Here, an operational pool constructed to provide linear test forms was used. If an MST approach is implemented at an agency, the characteristics of items used on test forms will change from those used on linear tests. It is important to find an item pool design (make-up or blue print) that can be supported and will produce “good” tests under the criteria outlined in this paper.

Various issues need to be considered before a final multi-stage adaptive test design is selected. The trade-offs between scoring accuracy, cost and practicality of item development, and ease of form review have to be weighed. Once a design is chosen and becomes operational, it is difficult to change. The intent of the study reported here is to provide management information to determine an acceptable design that can be functional for several years.

References

1. Armstrong, R.D. & Edmonds, J.J. (2003). The assembly of multiple stage adaptive tests with discrete items. Newtown, PA: Law School Admission Council Report.

2. Armstrong, R. D., Jones, D. H., Pashley, P. & Koppel, N. (to appear) Computerized adaptive testing with multiple form structures. *Applied Psychological Measurement.*

3. Armstrong, R.D. & Roussos, L. (2002). A method to determine targets for multiple-form structures. Newtown, PA: Law School Admission Council Report.

4. Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item pools. *Journal of Educational Statistics*, *15*, 129-145.

5. Bradlow, E., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

6. Hambleton , R. & Xing, D. (2004). Computer-based test designs with optimal and non-optimal tests for making pass-fail decisions. Research Report, University of Massachusetts, Amherst, MA.

7. Fourer, R., Gay, D. & Kernighan, B. (2003). *AMPL: A modeling language for mathematical programming*. Brooks/Cole-Thompson Learning.

8. ILOG (2002). *CPLEX 8.0 user's manual.* Incline Village NV.

9. Kingsbury, G. & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359-375.

10. Lee, G. (2000). A comparison of methods of estimating conditional standard errors of measurement for testlet-based test scores using simulation techniques. *Journal of Educational Measurement*, *36*, 91-112.

11. Lee, G. , Dunbar, S. & Frisbie, D. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. *Educational and Psychological Measurement*, *61*, 958-975.

12. Lord, F. (1971). A theoretical study of two stage testing. *Psychometrika*, *36*, 227-242.

13. Lord, F. (1980). A*pplications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

14. Luecht, R. M. & Burgin, W. (2003). Test information targeting strategies for adaptive multistage testing designs. Paper presented at the 2003 Annual Meeting of NCME, Chicago IL.

15. Luecht, R. M. & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229-247.

16. Luecht, R. M. & Nungester, R. J. (2000). Computer-adaptive sequential testing. . In W. van der Linden & C. A. Glas (eds.), *Computerized adaptive testing: Theory and practice* (pp. 117-128). Boston, Kluwer.

17. Nemhauser, G., & Wolsey, L. (1988). *Integer and combinatorial optimization.* New York: John Wiley & Sons.

18. Theunissen, T.J.J.M. (1985) Binary programming and test design. *Psychometrika*, *50*, 411-420.

19. van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195-211.

20. van der Linden, W.J., Veldkamp, B.P. & Reese, L.M. (2000). An integer programming approach to item bank design. *Applied Psychological Measurement*, 24, 139-150.

21. Wainer, H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-201.

22. Xing, D. & Hambleton, R. (in press). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*.

Stage	*(i)*	*(ii)*	*(iii)*
Percentile	*10-14 items*	*5-7 items*	*10-14 items*
[67,100]			Bin 6
[50,100]		Bin 3	
[0,100],[33,67]	Bin 1		Bin 5
[0,50]		Bin 2	
[0,33]			Bin 4

Table 1. An MST design with 3 stages and 6 bins is given. Each bin in the MST depicted is targeted for a particular population percentile range as indicated in the left margin. The range on the number of items assigned to each bin is given in the column header. This is a set based design with 5 to 7 items from each set.

	Stage (iii)		
Bin partitioning rule	**33%**	**20%**	**13%**
	[67,100]	[80,100]	[87,100]
	[33,67]	[20,80]	[13,87]
	[0,33]	[0,20]	[0,13]

Table 2. A display is given for the three bin partitioning rules for 3-stage MSTDs.

	Stage (iv)		
Bin partitioning rule	**25%**	**20%**	**13%**
	[75,100]	[80,100]	[87,100]
	[50,75]	[50,80]	[50,87]
	[25,50]	[20,50]	[13,50]
	[0,25]	[0,20]	[0,13]

Table 3. . A display is given for the four bin partitioning rules for 4-stage MSTDs.

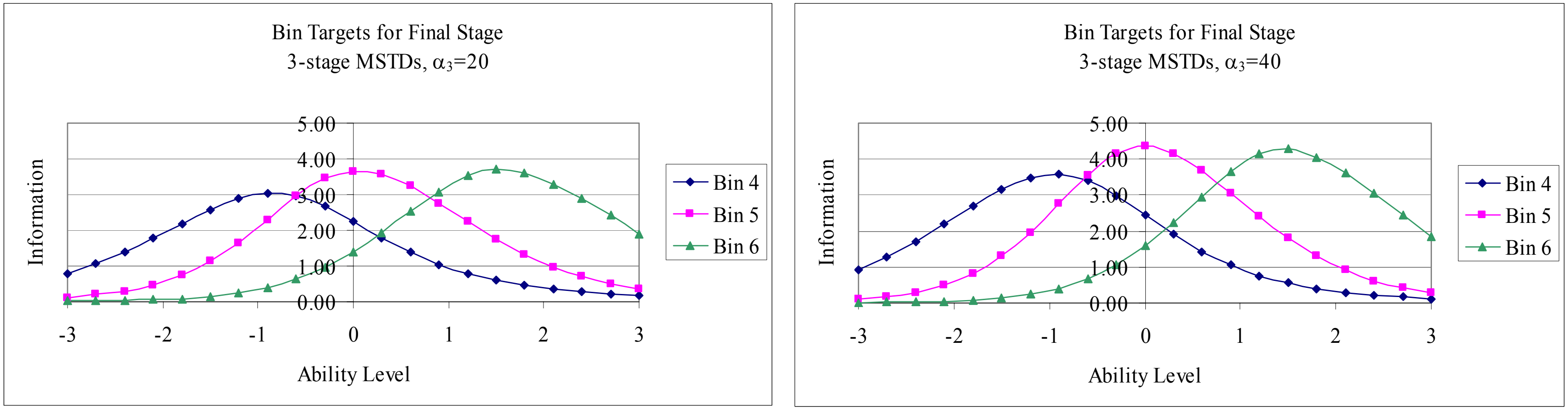

Figure 1. Bin Information Curves for the bins of the final stage of the 1-1-2-2-3-3 MSTD. In the figure on the left, the information parameter α_3=20; in the figure on the right, the information parameter α_3=40.

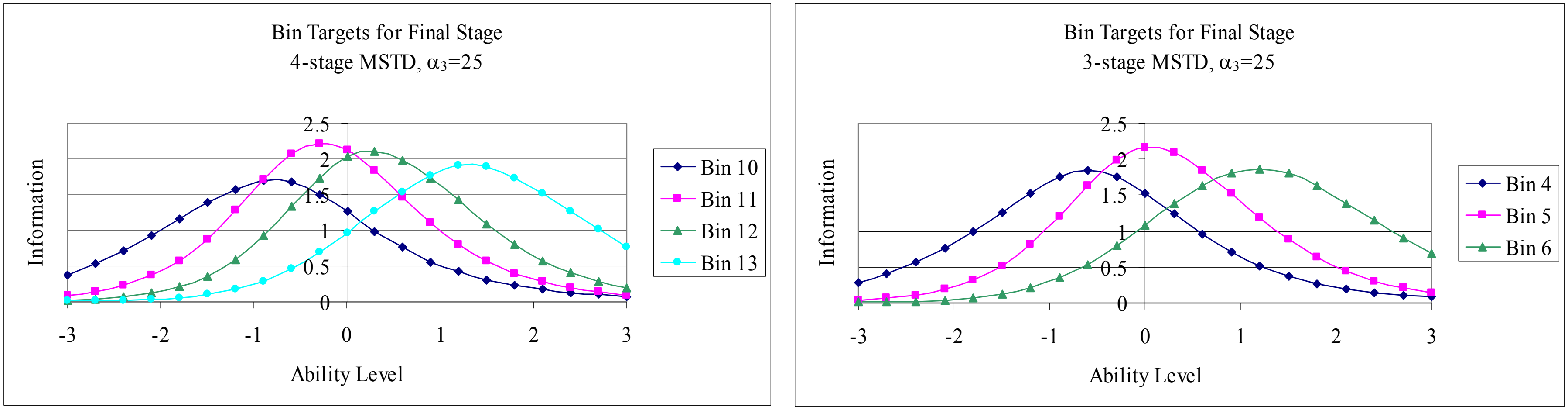

Figure 2. Bin Information Curves for the bins of the final stage of the 1-1-2-3-4 MSTD (on the left) and the 1-1-2-2-3 MSTD (on the right). The even split bin partitioning rule used for both designs.

	5-testlet MSTDs											**6-testlet MSTDs**				
	1-1-2-2-3					**1-1-2-3-3**						**1-1-2-2-3-3**				
33%	37.4	44.3	44.9	46.4	51.2	40.1	47.6	48.1	50.2	50.9		52.3	54.3	56.2	56.9	57.8
20%	47.2	45.0	50.6	48.2	46.1	46.8	50.1	51.2	49.4	46.7		52.2	56.5	58.4	56.5	51.8
13%	45.3	46.3	48.0	50.0	54.3	49.7	52.1	52.4	54.4	55.1		55.5	59.2	60.9	60.0	63.1
α_3	**20**	**25**	**30**	**35**	**40**	**20**	**25**	**30**	**35**	**40**		**20**	**25**	**30**	**35**	**40**

Table 4. The average number of distinct items per test from each MSTD with varying information parameters is given. The results come from an average across six replications.

	5-testlet MSTDs											**6-testlet MSTDs**				
	1-1-2-2-3					**1-1-2-3-3**						**1-1-2-2-3-3**				
33%	25	21	19	18	17	23	19	18	16	15		21	19	17	16	15
20%	22	20	19	17	16	20	18	16	16	15		20	17	16	16	15
13%	22	20	18	16	15	18	17	15	14	13		17	16	15	14	13
α_3	**20**	**25**	**30**	**35**	**40**	**20**	**25**	**30**	**35**	**40**		**20**	**25**	**30**	**35**	**40**

Table 5. Maximum number of tests assembled after six replications from each test design.

	5-testlet MSTDs											**6-testlet MSTDs**				
	1-1-2-2-3					**1-1-2-3-3**						**1-1-2-2-3-3**				
33%	4.66	4.29	4.11	4.05	4.01	4.40	4.20	4.08	3.99	3.91		3.81	3.66	3.55	3.46	3.42
20%	4.44	4.26	4.10	4.01	4.26	4.34	4.15	4.03	4.04	4.11		3.94	3.62	3.53	3.46	3.58
13%	4.41	4.25	4.12	4.01	3.99	4.30	4.17	4.05	3.96	3.87		3.78	3.66	3.55	3.43	3.38
α_3	**20**	**25**	**30**	**35**	**40**	**20**	**25**	**30**	**35**	**40**		**20**	**25**	**30**	**35**	**40**

Table 6. Conditional standard error from each test design is shown. The results come from an average of all *n* tests assembled from a given design (shown in Table 5).

	5-testlet MSTDs											**6-testlet MSTDs**				
	1-1-2-2-3					**1-1-2-3-3**						**1-1-2-2-3-3**				
33%	10.96	10.70	10.64	10.62	10.61	10.70	10.64	10.60	10.55	10.53		10.41	10.36	10.33	10.31	10.31
20%	10.74	10.65	10.59	10.57	10.65	10.62	10.54	10.51	10.53	10.66		10.43	10.31	10.28	10.31	10.36
13%	10.67	10.65	10.58	10.54	10.52	10.55	10.51	10.46	10.42	10.36		10.33	10.28	10.27	10.24	10.18
α_3	**20**	**25**	**30**	**35**	**40**	**20**	**25**	**30**	**35**	**40**		**20**	**25**	**30**	**35**	**40**

Table 7. Standard Deviation of scaled scores from each test design is given. The results come from an average of all *n* tests assembled from a given design (shown in Table 5).

	5-testlet MSTDs											6-testlet MSTDs				
	1-1-2-2-3					1-1-2-3-3						1-1-2-2-3-3				
33%	0.904	0.915	0.922	0.924	0.925	0.911	0.918	0.922	0.925	0.928		0.930	0.935	0.938	0.941	0.943
20%	0.910	0.916	0.921	0.924	0.916	0.912	0.918	0.923	0.923	0.922		0.925	0.936	0.939	0.941	0.938
13%	0.910	0.916	0.920	0.924	0.925	0.912	0.917	0.921	0.924	0.927		0.930	0.934	0.938	0.942	0.943
α_3	**20**	**25**	**30**	**35**	**40**	**20**	**25**	**30**	**35**	**40**		**20**	**25**	**30**	**35**	**40**

Table 8. The average Fidelity from each test design is displayed. The results come from an average of all *n* tests assembled from a given design (shown in Table 5).

	5-testlet MSTD						6-testlet MSTDs									
	1-1-2-3-4						1-1-2-2-3-4					1-1-2-3-3-4				
	No. of tests	*Items per test*	*SE*	*stddev*	*Fidelity*		*No. of tests*	*Items per test*	*SE*	*stddev*	*Fidelity*	*No. of tests*	*Items per test*	*SE*	*stddev*	*Fidelity*
25%	17	49.8	4.70	10.91	0.902		18	56.2	4.05	10.56	0.923	15	52.6	4.23	10.67	0.917
20%	16	57.0	4.86	10.95	0.895		17	56.8	3.99	10.52	0.925	15	58.6	4.12	10.56	0.920
13%	14	58.4	4.90	10.92	0.893		17	57.5	4.05	10.53	0.922	14	60.2	4.15	10.55	0.919

Table 9. Analytical summary results from 4-stage MSTDs where α_3=25.

	5-testlet MSTDs											6-testlet MSTDs				
	1-1-2-2-3					1-1-2-3-3						1-1-2-2-3-3				
33%	24.0	72.3	109.7	90.8	147.5	101.0	136.3	60.6	156.0	120.1		7.5	8.2	6.9	15.7	21.6
20%	79.2	119.7	75.8	121.0	110.6	114.1	111.3	117.5	125.4	136.3		12.8	6.3	18.0	6.2	7.1
13%	117.4	100.0	96.9	95.2	155.0	28.8	52.3	100.3	75.2	220.8		34.0	17.1	27.3	17.4	31.3
α_3	**20**	**25**	**30**	**35**	**40**	**20**	**25**	**30**	**35**	**40**		**20**	**25**	**30**	**35**	**40**

Table 10. Solution times in seconds from 3-stage MSTDs with varying information parameters. The results come from an average across six replications.

	5-testlet MSTD		6-testlet MSTDs	
	1-1-2-3-4		1-1-2-2-3-4	1-1-2-3-3-4
25%	42.9		56.9	231.1
20%	284.2		43.8	389.1
13%	711.0		83.3	394.6

Table 11. Solution times in seconds from 4-stage MSTDs where α_3=25. The results come from an average across six replications.