Running Head: ITEM PARAMETER RECOVERY WITH ADAPTIVE TESTS

Item Parameter Recovery With Adaptive Tests

Ben-Roy Do

Siang Chee Chuah

Fritz Drasgow

Department of Psychology

University of Illinois at Urbana-Champaign

Correspondence concerning this article should be addressed to Ben-Roy Do, Department of Psychology, University of Illinois at Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820. Phone number (217) 333-9631. Fax Number (217) 244-5876. Electronic mail may be sent to benroydo@s.psych.uiuc.edu. We thank Krista Breithaupt for help in assembling CAST panels.

Abstract

Previous research regarding sample sizes requirements for item parameter estimation was based on unrestricted samples. However, in adaptive testing, range restriction becomes an important issue. Range restriction can not only have significant consequences for pretesting new items to be added into the operational item pool, but also influences item parameter recovery in operational computerized adaptive tests. A series of simulations were conducted using a multi-stage adaptive testing with 1-3-3 and 1-2-2 designs and item parameters from a licensing examination item pool. Results suggest that range restriction has such a debilitating effect that even with relatively large sample sizes, item parameter recovery is undermined. Implications and suggestions for future research are discussed.

Item Parameter Recovery With Adaptive Tests

Item parameter estimates obtained from responses to traditional paper-and-pencil exams may not be comparable to those estimated from responses to computerized adaptive test (CAT), as the conditions such as item ordering and test administration modes are different in CAT (Ban et al., 2001). But prior research has not emphasized the role of sample range restriction in CAT settings. Instead, previous research regarding sample sizes requirements for item parameter estimation has been based on unrestricted samples. However, in CAT, as not all participants route through all test items and testlets, and therefore sample range restriction becomes an issue. Sample range restriction could influence item estimation in operational computerized adaptive tests, as well as pre-testing new items to be added to the operational item pool.

Conventional test construction is subject to constraints, such as item contents, item overlap with other items, item features in relation to subsets of other items, and statistical properties from pretesting (Stocking & Swanson, 1993). In CAT, items maybe selected based on maximum likelihood item information at the examinee's current ability estimate. Thus, more accurate results across a broad range of ability levels may be obtained (Lord, 1977), and improved measurement efficiency (Hambleton & Swaminathan, 1985). But as examinees are routed through a limited items and testlets in CAT, the traditional way of estimating item parameter with unrestricted samples may not be possible and item parameter recovery could be subject to errors due to range restrictions.

We conducted a series of simulations using a multi-stage adaptive testing (MST) design, also known as computer adaptive sequential testing (CAST; Luecht & Nungester, 1998). In a MST, multiple blocks of items (testlets or modules) are assembled. These testlets were grouped together to form test administration units called panels. Each panel is divided into two or more stages, and each testlet in the panel is targeted to meet specific proficiency levels, such as easy, moderate, or hard difficulty levels.

There can be several designs within a panel in a MST. Figure 1 shows a MST 1-3-3 design where the panel consists of three stages. Here the first stage has one routing testlet, and the second and third stages have 3 testlets each. Therefore, there are a total of seven testlets in the MST 1-3-3 design. In the first stage, all examinees first go through the moderate difficult Testlet 1, and are then routed as shown by the arrows. Those who performed poorly on Testlet 1 are routed to the easier Testlet 2, those who did moderately well go to the moderate difficult Testlet 3, and those who performed well are routed to a more difficult Testlet 4. Similarly, in the third stage, people can be routed to Testlet 5, 6, or 7 according to the routes shown by the arrows. Note that examinees cannot be routed from Testlet 4 to Testlet 5 or from Testlet 2 to Testlet 7.

Insert Figure 1 around Here

Another possible assembly is a MST 1-2-2 design, shown in Figure 2. Here, after routing through the moderate difficult Testlet 1, there are only 2 different difficulty levels in the second and third stages. In this example, Testlet 2 is composed of moderately difficult items, and Testlet 3 items are more difficult. Again, in the third stage, people can be routed to Testlet 4 or 5 depending on their performance in the second stage. Although there are many other MST designs based on the number of stages in a panel and the levels of difficulty in a stage, the current study will focus on the MST 1-3-3 and 1-2-2 designs.

Insert Figure 2 around Here

Using a MST design allows test developers to have greater control over item presentation. Within each testlet, the content and quality of items can be examined by test developers to maintain robustness, control for order effects, and minimize context effects such as cross-information and unbalanced content (Wainer & Kiely, 1987). However, MST has its drawbacks. Kingsbury and Zara (1991) pointed out that using a MST adaptive testing rather than a fully adaptive testing reduces information by 30% to 50%, and thus more items may be needed to obtain a desired level of precision.

Test developers may wish to re-estimate item parameters from operational adaptive tests in at least two situations. First, as a testing program converts from paper-and-pencil administrations to computerized administrations, there may be a need to pretest large numbers of items. Sometimes, items are pretested using sample of volunteers who are administered items conventionally (i.e., with no adaptivity). Such examinees may be less motivated than individuals answering items under operational conditions. Thus, re-estimating item parameters from an operational administration may be desirable. Second, pretest items may be seeded into an operational adaptive test. However, the number of examinees who answer each item may be small, leading to concerns about the accuracy of item parameter estimates. Again, re-estimating item parameters from a hopefully large operational administration may be warranted.

As noted earlier, when item parameters are estimated from an operational MST, the range of ability is restricted to obtain item parameter estimates as accurate as those from a random sample of examinees. Consequently, we conducted a study to evaluate “how much larger?” in the context of MST.

Method

Item response generation

Item parameters were obtained from an operational licensing examination item pool. The three-parameter logistic (3PL) model was used to generate dichotomized item responses. In the 3PL model, the probability of an examine with a latent trait or ability level of θ answering the i^{th} item correctly can be expressed as

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}}, \qquad (1)$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the pseudo-guessing parameter, and D is a scaling factor set equal to 1.7 (Hulin et al., 1983). Ability levels were randomly drawn from a normal distribution with a mean of 0 and standard deviation of 1.

Sample sizes

Sample sizes of N = 300 and 1000 (denoted as 1k in tables) were used for simulated examinees taking a complete un-routed test. These two sample sizes were chosen following the suggestions of Hulin, Lissak, and Drasgow (1982) that to achieve highly accurate estimation, sample sizes of 1000 examinees with tests of 60 items may be required. These unrestricted samples served as the baseline for comparisons with restricted samples, where not all participants route through all test items and testlets.

To observe the effect of sample range restriction, restricted samples sizes of N = 1000, 2000, 3000, 4000, and 5000 were simulated. Note that a sample of 1000 in a MST will lead to samples of approximately 300 examinees being administered each Stage Two and Three testlets in the 1-3-3 design and approximately 500 examinees being administered each Stage Two and Three testlets in the 1-2-2 design.

Testlets and panels

In the Study 1 MST 1-3-3 design, two panels were assembled, each consisted of 7 testlets. Each testlet had 20 items. Testlets 1, 3, 6 were moderatel difficult testlets, Testlets 2 and 5 were the easier testlets, and Testlets 4 and 7 were the more difficult testlets. The means and standard deviations of the item parameters for the three difficulty levels are presented in Table 1. There were no significant differences between the two panels for the "a", "b", or "c" parameters.

Insert Table 1 around Here

In the Study 2 MST 1-2-2 design, two panels were assembled, each consisting of 5 testlets. Each testlet had 25 items. Testlets 1, 2, 4 were of moderate difficulty, and Testlets 3 and 5 were the more difficult testlets. The means and standard deviations of the two difficulty levels are presented in Table 2. Again, there were no significant differences between the two panels for the "a", "b", or "c" parameters.

Insert Table 2 around Here

It should be noted that the moderately difficult testlets in the MST 1-3-3 design were different from the MST 1-2-2 design. As there was no easy testlet in the MST 1-2-2 design, the "b" parameters of moderate difficulty testlets were between the easy and moderate testlets from the MST 1-3-3 design.

Evaluation of estimation

Recovery of item characteristic curves (ICC) was examined by computing the root mean squared error (RMSE) between the true ICC and the estimated ICC recovered for each item at 31 θ values from –3.0 to +3.0 chosen at equal intervals,

$$\text{RMSE}=\sqrt{\frac{1}{31}\sum_{j=1}^{31}[P_i(\theta_j)-\hat{P}_i(\theta_j)]^2} \qquad (2)$$

Hulin, Lissak, and Drasgow (1982) suggested that RMSEs of recovered ICCs provide the best overall index of item parameter estimation. The Average RMSE for each testlet was computed as the average of the item RMSEs.

Simulation procedure

Item parameters were obtained from an operational licensing examination item pool. A computer program, 3PLGEN.exe (Stark, 2000), was used to sample theta values and create simulated responses for unrestricted samples. Another computer program, ADAPTT2.exe (Stark & Chuah, 2000), was used to simulate responses in a MST setting. These output from these programs consisted of simulated binary item responses (1 = correct, 0 = incorrect), which were subsequently used to obtain item parameters estimates with the in BILOG computer program (Mislevy & Bock, 1991). A third computer program, ICCDIF.exe (Chuah & Do, 2003) was written to calculate the mean square error (MSE) difference between the two ICCs (from true and estimated parameters). Then, RMSEs were obtained and the average RMSE for each testlet was computed. For each panel and MST design, three runs were simulated and averaged to minimize fluctuations due to sampling.

Results

Study 1: The MST 1-3-3 design

Table 3 shows the average RMSE for each testlet for the unrestricted samples (*N* = 300 and 1000, the later denoted as 1k) and the restricted samples (*N* = 1000, 2000, 3000, 4000, 5000). The letters “a” and “b” denotes the two panels. Recall that testlets consisted of 20 items, so that items 1-20 refer to Testlet 1, items 21-40 refer to Testlet 2, and so forth.

--

Insert Table 3 around Here

--

First, consider effects of routing within restricted samples for each sample size (*N* = 1000 to 5000). In this case, the average RMSE is substantially smaller in Testlet 1. However, two factors cause Testlet 1 to have smaller RMSEs: there is a wide range of ability for simulated examinees and the entire sample was administered this testlet.

Second, examinating the effect of sample size effect for the restricted samples from 1000 to 5000, it is apparent that the average RMSE becomes smaller as *N* increases.

Third, comparing the two panels shows few differences.

It is most interesting to compare the average RMSE for unrestricted (300 and 1k baselines) versus restricted (1000 to 5000 levels) samples. Some interesting interactions are apparent. For example, in the restricted sample of 1000, although the average RMSE is generally close to *N* =300 baseline as expected, the error is about 1.86 times larger in Testlet 6. In the restricted sample of 3000, although the average RMSE is usually close to 1k baseline, it is 2.65 times larger in Testlet 6. Furthermore, in the case of the restricted sample of 5000, the average RMSE is still 2.09 times larger than the 1k baseline for Testlet 6.

It is obvious that range restriction has a debilitating effect. Even with relatively large restricted samples, item parameter recovery can be undermined in the MST 1-3-3 design. Therefore, to compare unrestricted baselines with range restricted samples, it is necessary to have a large sample sizes to reduce RMSEs.

Study 2: The MST 1-2-2 design

Table 4 shows the average RMSE for each testlet for the unrestricted samples (*N* = 300 and 1000, the later denoted as 1k) and restricted samples (*N* = 1000, 2000, 3000, 4000, 5000). The letters "a" and "b" denote the two comparable panels. Testlets were presented as sets of 25 items.

Insert Table 4 around Here

First, consider the effect of routing for each sample size (*N* = 1000 to 5000). In this case, although the average RMSE still becomes larger in later stages, the difference was not as large as in the MST 1-3-3 design. The difference between Testlet 1 and Testlet 4 (Stage 3 moderate difficulty items) is smaller than .01 (2nd decimal difference).

Second, the sample size effect across restricted samples from 1000 to 5000 was compared. As expected, when the sample size gets larger, the average RMSE becomes smaller. The difference between Testlet 1 and Testlet 4 is less than .017 for samples of 1000 and .011 for samples of 5000, a great improvement over the MST 1-3-3- design.

Third, a comparison across two panels shows that the two panels performed similarly.

Finally, it is also interesting to see that the range restriction effect is less influential in the MST 1-2-2 design. In the restricted sample of 1000, the average RMSE is comparable to the 300 baseline, and there was not a huge increase in error as in the MST 1-3-3 design. In the restricted sample of 3000, the average RMSE is 1.20 times larger in Testlet 5 (Stage 3 hard items). Furthermore, in the case of restricted sample size of 5000, the average RMSE is only 1.02 times larger than the 1k baseline in Testlet 5.

Although the effect of range restriction is smaller in the MST 1-2-2 design, item parameter recovery was still undermined. Therefore, when one wishes to compare unrestricted baselines with range restricted samples in a MST situation, it is desirable to have a large sample size to reduce the RMSE.

Discussion

Minimum sample size

How do we tell what is an adequate item parameter recovery? How large should the sample size be for estimating an MST? In a non-adaptive testing situation, extending Lord's (1968) suggestion, Hulin, Lissak, and Drasgow (1982) suggested that a sample size of 1000 for the 3PL model appears sufficient for accurate recovery of ICCs. Although the estimation of item parameters improves as the sample size gets larger, the testlets in later stages do not perform as well as in the routing Testlet 1. Range restriction has an impact on item parameter recovery. In the MST 1-3-3 design, even with a tripled sample size to 3000, item parameters estimation is still not comparable to the unrestricted sample of 1000. Although the average RMSE becomes smaller in the restricted sample of 5000, they are still rather large in Testlets 6 and 7. It seems that a restricted sample of 5000 is not as good as the unrestricted sample of 1000.

However, the story is different in the MST 1-2-2 design. In the restricted sample of 3000, the average RMSE is comparable to the unrestricted sample of 1000. When the restricted sample size increases, the results are not only closer to the baseline, but in some cases, better than the unrestricted samples. It seems that in the MST 1-2-2 design, a restricted sample of 2000 is nearly as good as an unrestricted sample size of 1000.

Larger average RMSE in later stages

It should be noted that the average RMSEs are larger in later stages, particularly the moderate or hard testlets in the third stage. This is expected because the Stage Three subsample should be more restricted than the Stage Two subsamples. In the MST 1-3-3 design, this effect is particularly evident in Testlet 6. However, in the MST 1-2-2 design, the testlets that have the largest average RMSE are Testlet 3 in Panel A and Testlet 5 in Panel B. In Panel B, initially Testlet 3 had larger errors than Testlet 5. But as the sample size increased, the RMSE in Testlet 5 became larger than Testlet 3. It seems that misclassification somehow helps to alleviate sample range restriction, but the effect is limited.

MST Designs

As there are many MST designs, which design works better than others would be a topic worth future research. But no matter which design one chooses, range restriction remains an issue, especially for MST designs with more testlets and stages. Considering the effects of sample range restriction, the MST 1-2-2 design produces smaller errors than the MST 1-3-3 design when comparing restricted versus unrestricted samples. To have results comparable to the *N* = 1000 unrestricted sample, samples more than 5 times larger for the MST 1-3-3 design, or 3 times larger for the MST 1-2-2 design may be required. Although other forms of MST designs might be more comparable to the unrestricted baselines, the MST 1-2-2 design is the simplest of them all. As more complex MST designs might further restrict the sample and produce less accurate estimation, the MST 1-2-2 design is probably the one that has the smallest average RMSE.

Conclusion

In an adaptive testing, sample range restriction has such a debilitating effect that even with relatively large sample sizes, item parameter recovery is undermined. Although average RMSE decreases as sample sizes increase, it is still worse than with unrestricted range samples. Range restriction hurts item parameter estimation, especially for the later stage testlets. It is therefore desirable to have a large sample size, say 3000 in the MST 1-2-2 design and 5000 in the MST 1-3-3 design, and a simpler structure such as MST 1-2-2 design to reduce the effect of sample range restriction.

References

Ban, J.-C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item – Calibration/Scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191-212.

Chuah, S. C., & Do, B.-R. (2003). ICCDIF.exe. Computer program available upon request.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer/Nijhoff.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones Irwin.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied psychological measurement, 6*, 249-260.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied measurement in education, 4*, 241-261.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillside, NJ: Lawrence Erlbaum Associates.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229-249.

Mislevy, R. J., & Bock, R. D. (1991). *BILOG users' guide*. Chicago: Scientific Software.

Stark, S. (2000). 3PLGEN.exe. Computer program available upon request.

Stark, S., & Chuah, S. C. (2000). ADAPTT2.exe. Computer program available upon request.

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

Table 1.

Means and standard deviations of item parameters in the MST 1-3-3 design

		Item Parameters					
	Difficulty	*a*		*b*		*c*	
Panel	*Level*	*M*	*SD*	*M*	*SD*	*M*	*SD*
A	easy	.66	.23	-1.25	.58	.12	.09
	moderate	.80	.33	-.39	.89	.17	.10
	hard	.81	.18	.68	.34	.23	.10
B	easy	.57	.14	-1.29	.57	.10	.07
	moderate	.79	.34	-.13	.82	.17	.11
	hard	.84	.27	.79	.31	.22	.11

Table 2.

Means and standard deviations of item parameters in the MST 1-2-2 design

		Item Parameters					
	Difficulty	*a*		*b*		*c*	
Panel	*Level*	*M*	*SD*	*M*	*SD*	*M*	*SD*
A	moderate	.66	.21	-.59	.64	.22	.09
	hard	.73	.23	.79	.79	.23	.10
B	moderate	.65	.21	-.58	.75	.21	.08
	hard	.74	.24	.88	.53	.22	.09

Table 3.

Average root mean squared errors for the MST 1-3-3 design

		Unrestricted Baselines				*Restricted Sample Sizes*									
Testlet	*Items*	*300a*	*300b*	*1ka*	*1kb*	*1000a*	*1000b*	*2000a*	*2000b*	*3000a*	*3000b*	*4000a*	*4000b*	*5000a*	*5000b*
1	1-20	.040	.040	.023	.025	.023	.022	.019	.017	.014	.014	.013	.013	.010	.011
2	21-40	.042	.041	.026	.024	.037	.040	.029	.036	.023	.030	.021	.029	.021	.028
3	41-60	.040	.040	.025	.027	.037	.045	.035	.041	.034	.036	.030	.037	.027	.037
4	61-80	.043	.047	.026	.026	.042	.059	.033	.052	.029	.050	.032	.048	.032	.047
5	81-100	.047	.045	.023	.022	.045	.041	.036	.032	.025	.026	.024	.023	.022	.023
6	101-120	.055	.058	.027	.032	.100	.110	.078	.093	.074	.081	.070	.075	.055	.068
7	121-140	.053	.054	.033	.029	.061	.052	.061	.044	.050	.044	.053	.044	.051	.041

Table 4.

Average root mean squared errors of in the MST 1-2-2 design

		Unrestricted Baselines				*Restricted Sample Sizes*									
Testlet	*Items*	*300a*	*300b*	*1ka*	*1kb*	*1000a*	*1000b*	*2000a*	*2000b*	*3000a*	*3000b*	*4000a*	*4000b*	*5000a*	*5000b*
1	1-25	.045	.046	.026	.026	.031	.025	.021	.017	.015	.017	.014	.014	.013	.012
2	26-50	.046	.048	.027	.027	.036	.041	.029	.031	.022	.029	.021	.024	.019	.021
3	51-75	.051	.052	.031	.029	.053	.050	.047	.040	.042	.035	.038	.030	.035	.029
4	76-100	.054	.048	.028	.030	.047	.038	.031	.028	.027	.022	.021	.023	.024	.021
5	101-125	.049	.053	.028	.030	.049	.048	.038	.034	.036	.034	.028	.031	.028	.032

Figure 1.

MST 1-3-3 design

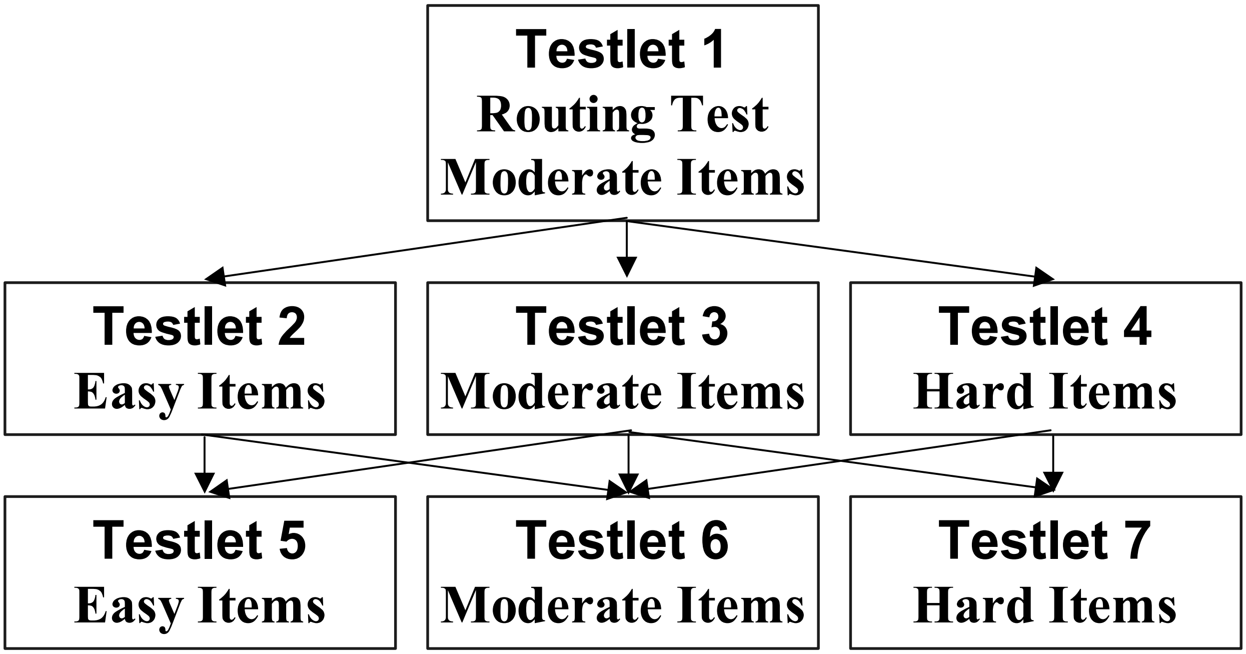
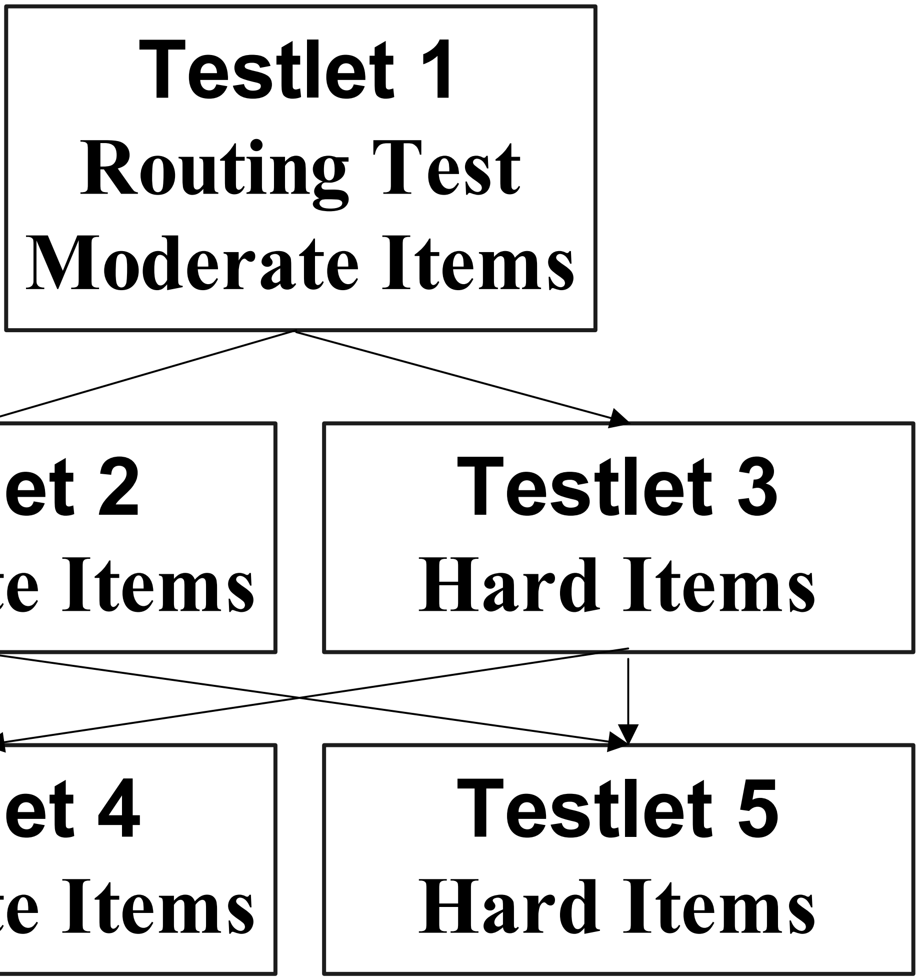


Figure 2.

MST 1-2-2 design


Testlet 1
Routing Test
Moderate Items
Testlet 2
Moderate Items
Testlet 3
Hard Items
Testlet 4
Moderate Items
Testlet 5
Hard Items