Empirical vs. Expected IRT-Based Reliability Estimation in Computerized Multistage Testing (MST)

Yanwei Zhang
Krista Breithaupt
Aster Tessema
David Chuah

AICPA

Paper Presented at the Annual Conference of the National Council of Measurement in Education, San Francisco, CA
April 2006

Abstract

Two IRT-based procedures to estimate test reliability for a certification exam that used both adaptive (via a MST model) and non-adaptive design were considered in this study. Both procedures rely on calibrated item parameters to estimate error variance. In terms of score variance, one procedure (Method 1) uses the empirical ability distribution from a particular sample of examinees, and the other procedure (Method 2) assumes a normal distribution of ability and is sample-free. Due to the problem of sampling restriction in adaptive tests, Method 1 was modified (Method 1 extension) to "beef up" the sample and estimate reliability for each testlet in a MST panel before aggregating the estimates into an overall estimate for a test form or route. Overall, results imply that Method 1 and Method 2 tend to produce similar results for both adaptive and non-adaptive tests on the panel level and the test section level. Method 1 should not be applied to individual test forms in adaptive tests by a MST design. In the latter case, the modified or extended procedure can be used to alleviate the problem of restricted sample. The algorithms of the discussed procedures can be implemented in common statistical programming language such as SAS and SPSS as flexible alternatives to the theoretical and empirical reliability estimates computed, for example, by the BILOG-MG software.

Empirical vs. Expected IRT-Based Reliability Estimation
in Computerized Multistage Testing (MST)

Yanwei Zhang, Krista Breithaupt, Aster Tessema, David Chuah
AICPA

Introduction

As a central concept in testing, reliability refers to the degree to which test scores are free from errors (Crocker & Algina, 1986). In the classical test theory (CTT), reliability can be defined precisely as the ratio of the true score variance to the observed score variance, and equivalently, as one minus the ratio of error variance to observed score variance (Feldt & Brennan, 1993):

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2},$$

where $\rho_{XX'}$ is the reliability, σ_X^2 is the observed score variance, σ_T^2 is the true score variance, and σ_E^2 is the error score variance.

A closely related concept is the standard error of measurement (SEM), defined as the standard deviation of the error score component (i.e. σ_E). SEM can be estimated with the standard deviation of the observed test scores and the reliability, as in

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}.$$

The classical reliability and SEM estimates are aggregates or averages across the score scale. In the modern test theories such as IRT, Test Information Function (TIF) and conditional SEM, the latter expressed as the inverse of the square root of TIF, replaced the classical concepts, and they can be estimated at any particular level on the score scale to represent the measurement precision (Hambleton, Swaminathan, & Rogers, 1991).

With IRT scoring, the reliability estimation that relies on aggregation or generalization provides less information than TIF or CSEM. However, for high-stake testing programs scored by IRT, some reliability index as a general characteristic of a test form or one of its components is always desirable (AERA, APA, & NCME, 2002). Because reliability is such an established criterion for test quality, newer and modified aggregate estimates are needed for IRT-based test models. At the same time, performance tests involving multi-dimensional score models are gaining popularity, often associated with mixed item types and/or scoring methods. For such tests the correlation between different test components has to be adjusted by a single reliability estimate from each component.

The purpose of this study is to compare two potentially appropriate algorithms for aggregate reliability estimate using the results from a computerized national certification exam that includes both multiple-choice questions (MCQ) and performance tasks, and to provide some guidelines for their application in large-scale adaptive testing programs.

Data Source

The testing program providing the context for this study is a high-stake computerized certification exam that has four independent subjects or sections. Three of the sections use an adaptive, multistage testing model (MST) for the multiple-choice items. The basic structure of MST is a panel of multiple stages sequentially administered to examinees. At the first stage, all examinees take a common set of items (a module) of medium difficulty. Depending on their performance at the first stage, the examinees are asked to take another module, either of high or medium difficulty, at the second stage. Then their performance in the first two stages will determine which module, high or medium in difficulty, they are to take at the third stage. Each panel will then have up to four different routes or test forms of MCQ items that will match the ability level of the examinees (see Figure 1 for a diagram of the panel). For a given panel, two routes are primary that most examinees will take: Medium-Medium-Medium (MMM) and Medium-Hard-Hard (MHH). The rest of examinees will be placed on one of two secondary routes, Medium-Medium-Hard (MMH) and Medium-Hard-Medium (MHH). In addition to multiple-choice items, the examinees of the three sections are required to answer a common set of performance tasks called simulations including written communication questions. The fourth section currently has only multiple-choice items and uses a non-adaptive design by which all examinees of a panel take the same three modules without being routed. Put differently, this section applies a 1-1-1 model.

The data used in this study were collected from one administrational window of the exam that spanned two months for continuous testing.

Methodology

Both methods to estimate reliability involve computing score variance based on IRT parameters. Method 1 applies the empirical distribution of the ability estimates while Method 2 assumes strict normal distribution of ability. Their algorithms are briefly described as follows:

<u>Method 1</u>

The basic conception of reliability is a (transformed) ratio between error variance and score variance. In this case it is

$$\hat{\rho} \cong 1 - \frac{\overline{S}^{2}_{\hat{\theta}|\theta}}{S^{2}_{\hat{\theta}}},$$

where $\overline{S}^{2}_{\hat{\theta}|\theta}$ is the averaged error variance across all the theta estimates of a particular sample, and $S^{2}_{\hat{\theta}}$ is simply the score variance of the theta estimates of the same sample. So

$$S^{2}_{\hat{\theta}} = \frac{\sum_{j=1}^{N} (\hat{\theta}_j - \overline{\hat{\theta}})^2}{N-1}$$

where *j* (from 1 to N) stands for an individual examinee. On the other hand,

$$\overline{S}^2_{\frac{\hat{\theta}}{\theta}} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sum_{i}^{K} I_i(\hat{\theta}_j)}$$

where j (from 1 to N) stands for an individual examinee, and i (from 1 to K) stands for an item the jth examinee answered for scoring purpose. Here $I_i(\hat{\theta}_j)$ is the information of the ith item given the theta estimate of the jth examinee, and $\sum_{i}^{K} I_i(\hat{\theta}_j)$ is then the test information of all K items (for a test form or a route of 75 items, for example) given the theta estimate of the jth examinee.

Once inversed, the result

$$\frac{1}{\sum_{i}^{K} I_i(\hat{\theta}_j)}$$

will be the error variance of the theta estimate of the jth examinee based on all K items he or she was administered. Finally, it gets pooled across the whole sample and then averaged, letting the number of examinees in each theta range take care of the density of the distribution.

Method 2

The reliability here is defined as the ratio of item true variance to the observed item variance, under the true-score model (Lord & Novick, 1968), which can be expressed as

$$\rho_{XX'} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2},$$

where σ_τ^2 is the expected true score variance and σ_e^2 is the expected error score variance.

The direct solution to both σ_τ^2 and σ_e^2 involves complex numerical integration. In this case, σ_τ^2 and σ_e^2 were estimated by the approximation methods (Dimitrov, 2003) outlined as follows.

For a test of *n* binary items, expected error variance σ_e^2 is a sum of expected item error variance $\sigma_{(e_i)}^2$ as in

$$\sigma_e^2 = \sum_{i=1}^{n} \sigma_{(e_i)}^2 .$$

The expected true score variance can be expressed as

$$\sigma_\tau^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{\left[\pi_i(1-\pi_i) - \sigma_{(e_i)}^2\right]\left[\pi_j(1-\pi_j) - \sigma_{(e_j)}^2\right]}.$$

In the above formulas, π_i (or π_j) is the expected item score. Let a_i, b_i, and c_i stand for the discrimination, difficulty, and guessing parameters in the 3PL IRT model respectively,

$$\pi_i = c_i + (1-c_i)\frac{1-erf(X_i)}{2},$$

where $X_i = a_i b_i / \sqrt{2(1+a_i^2)}$ and *erf* is a known mathematics function called the *error function*.

The expected item error variance $\sigma_{(e_i)}^2$ (or $\sigma_{(e_j)}^2$) with the 3PL IRT model can be approximated by

$$\sigma_{(e_i)}^2 = c_i + (1-c_i)(1-\pi_i) + (1-c_i)^2 m_i \exp\left[-0.5(b_i / d_i)^2\right],$$

in which m_i and d_i depend on the item discrimination, as in:

$$m_i = 0.2646 - 0.118a_i + 0.0187a_i^2$$
$$d_i = 0.7427 - 0.7081/a_i + 0.0074a_i^2$$

The outputs of the above method have an approximation error ranging from 0 to 0.005 in the absolute value (with a mean of 0.001 and a standard deviation of 0.001).

Comparison of Method 1 and Method 2

From the above descriptions, one can see that Method 1 is more analogous to the classical concept of reliability in the sense that the error variance and the observed score variance are both sample-driven. Method 2, on the other hand, is sample-free because it estimates through approximation the true score variance and the error score variance directly assuming normal distribution of ability. Once that assumption is made, the calculation in Method 2 is based solely on item parameters.

For the MST exam in question, the reliability estimate by Method 1 can be made at different sampling levels, i.e. by section, by panel, and by route (test form). When the total sample of a section is used, the estimate is aggregated upon all the items and examinees across all the panels. When the sample of a panel is used, the estimate is a generalization across the four possible routes or test forms in the panel. When the sample of a route within in a panel is used, the estimate is restricted to that particular test form. On the other hand, the reliability estimate by Method 2 can only be made at the level of route or test form, but the estimates can be aggregated to the level of panel or section.

The difference between the fourth section and the first three in the exam offers an opportunity to compare the two methods in both adaptive and non-adaptive situations. In both situations, because panels are randomly assigned to examinees, the ability distribution at the panel and section levels should be approximately equivalent. In the non-adaptive situation, the sampling for Method 1 is the same for the panel and the test form because they are identical. In the adaptive situation, however, the sampling distribution of ability will be more restricted on the level of individual test form because of the routing of examinees. In other words, the Method 1 reliability may be underestimated at this level, particularly for the two secondary routes or test forms with potentially very few examinees.

Given the differences between the methods, the reliability estimate by Method 2 is expected to be higher than the corresponding estimate by Method 1, depending on how the sampling distribution of ability deviates from the *N(0, 1)* distribution assumed by Method 1. In the non-adaptive situation, comparison can be made at each level of the test. In the adaptive situation, comparison can be made at the panel and section levels when the estimates by Method 2 are aggregated to those levels. On the test form or route level, the Method 1 estimate is expected to be lower due to the restricted range in the sampling distribution of examinee ability as well as its deviance from normality.

Extension of Method 1

To address the problem of underestimation of route-based reliability by Method 1 due to restriction of sampling distribution, this study also considered an extension of Method 1 by aggregating reliability estimated from the testlet at each stage.

The new approach can be illustrated with a particular route, say MHM. First, reliability of the medium testlet at Stage One can be estimated by Method 1 based on all examinees who take that testlet. Similarly, reliability can be estimated for the hard testlet at Stage Two and the medium testlet at Stage Three, in both cases based on all examinees who take the testlet. Since each testlet has the same number of items, the three estimates can be combined to generate an overall reliability estimate for three testlets, or for a route. Let ρ_{M1} be the reliability of the medium testlet at Stage One, ρ_{H2} be the reliability of the hard testlet at Stage Two, and ρ_{M3} be the reliability of the medium test at Stage Three, then the overall reliability of the MHM route (ρ_{MHM}) can be computed as

$$\rho_{MHM} = \frac{\rho_{M1} + \rho_{H2} + \rho_{M3}}{1 + (n-1)\overline{\rho}},$$

where n is the number of testlets or stages (n = 3), and $\overline{\rho}$ is the simple average of ρ_{M1}, ρ_{H2}, and ρ_{M3}.

This approach addressed the problem of sampling restriction of Method 1 for secondary routes, because reliability of each testlet involved is estimated from all examinees exposed to that testlet, either from a particular panel, or even across panels if a full sample is desired when testlets are used in multiple panels. The arithmetic average of estimates in the denominator makes sense in this case because each testlet has the same number of items. It is expected that the route-based reliability estimates by this approach will be higher than those by the original Method 1.

Results

The methods described above were applied to two exam sections, one adaptive with the MST design, and one non-adaptive (linear). In both cases, only MCQ items were included for estimation.

Table 1 presents the overall reliability results by Method 1 (empirical) and Method 2 (expected) for both adaptive and non-adaptive tests. For Method 1, the estimate was based on all the examinees in a section. For Method 2, the estimate was the average from all panels (non-adaptive test) or from all routes of all panels (adaptive test). One can see that for both tests, Method 1 results were lower than Method 2 results, while the difference was larger in non-adaptive tests (0.04) than in adaptive tests (0.01).

Results on the panel level are listed in Table 2 for adaptive tests and in Table 3 for non-adaptive tests. Again, reliability estimates by Method 2 were the average of all four routes of a panel in adaptive test, and Method 1 used all examinees taking that panel. In non-adaptive tests, route and panel were equivalent for Method 2.

Table 2 indicates that the reliability estimates on the panel level were more or less the same for adaptive tests, where the difference ranged from 0 to 0.02 (results of Method

2 always higher when there was a difference). For non-adaptive tests, on the other hand, Table 3 clearly indicates much larger variance among the empirical estimates by Method 1 (ranging from 0.76 to 0.84). The expected results by Method 2 ranged from 0.83 to 0.87. At the same time, the difference between empirical and expected results also varied greatly from 0.02 (Panel 6) to 0.10 (Panel 34).

Before presenting results on the route level, the numbers of examinees by route and panel are shown in Table 4 for adaptive tests. It is clear that most examinees were placed on the MHH route, which was not surprising for a certification exam. The MMM route had significantly lower number of examinees than the MHH route, ranging from 26 to 59. The two secondary routes, MMH in particular, had very small number of examinees (4 on the MMH route for Panel 20).

The extension of Method 1 was applied to each route in adaptive tests. Table 5 lists the reliability estimates for each testlet across the three stages of a 1-2-2 MST design, based on all examinees in a panel who took the testlet. Table 5 indicates that the highest estimates were expected from the first medium testlet at Stage 1, and they became lower, for both medium and hard testlets, towards the end of testing.

The final results of Method 1 extension, after aggregating testlet reliability for a route, are listed in Table 6 (under M1+) along with those by original Method 1 and Method 2.

For the two primary routes (MMM and MHH), the results by Method 1 (empirical) were significantly lower than those by Method 2 (expected) by about 0.16. The empirical reliability estimates for the two secondary routes were even lower, sometimes outrageous. Method 1 was apparently inappropriate for the two secondary routes.

On the other hand, the extension of Method 1 produced results comparable to those of Method 2. One can see that estimates by M1+ were very consistent both across all routes and across all panels, as found in the results by Method 2. On average, the extension of Method 1 produced slightly lower estimates, by 0.03, than Method 2, but much higher than the estimates by the original Method 1, especially for the two secondary routes.

Discussion

This study considered two IRT-based procedures to estimate test reliability for a certification exam that used both adaptive (via a MST model) and non-adaptive design. Both procedures rely on calibrated item parameters to estimate error variance. In terms of score variance, one procedure (Method 1 in this study) uses the empirical ability distribution from a particular sample of examinees, and the other procedure (Method 2 in this study) assumes a normal distribution of ability and is sample-free. Due to the problem of sampling restriction in adaptive tests, Method 1 was modified (Method 1 extension in this study) to "beef up" the sample and estimate reliability for each testlet in a MST panel before aggregating the estimates into an overall estimate for a test form or route.

With Method 2 reliability can be directly estimated only on the level of individual route or test form within a panel, while Method 1 can estimate reliability based on the sample for a test form, a panel or a whole section. To get a general estimate for a whole panel or an exam section with numerous panels, the estimates from individual test forms by Method 2 were averaged to the panel level and to the section level.

On both the exam section level and the panel level, the results are that reliability estimates by Method 2 were always higher than those by Method 1. In this particular case, the Method 1 and Method 2 estimates from the adaptive tests are practically the same, while the results from the non-adaptive tests indicate greater discrepancy between the two methods that may necessitate further analysis.

When Method 1 was applied to individual test forms in adaptive tests, the impact of sampling restriction on score variance due to routing in a MST design caused the estimates to be much lower than those by Method 2 for the two primary routes. For the two secondary routes where the sampling problem was most severe, Method 1 was obviously not applicable.

The modified Method 1 estimated reliability for each testlet in a MST panel first. Here results indicate that the adaptive design still had an impact, although minor, because estimates from an earlier stage tended to be larger than a later stage. When testlet reliability estimates were aggregated for a test form or route, the new estimates were a lot more comparable with results by Method 2, and they were very consistent across routes and panels.

Overall, results imply that Method 1 and Method 2 tend to produce similar results for both adaptive and non-adaptive tests on the panel level and the test section level. Method 1 should not be applied to individual test forms in adaptive tests by a MST design. In the latter case, the modified or extended procedure can be used to alleviate the problem of restricted sample. The algorithms of the discussed procedures can be implemented in common statistical programming language such as SAS and SPSS as flexible alternatives to the theoretical and empirical reliability estimates computed by the BILOG-MG software (Scientific Software International, 2003).

It is beyond the scope of this study to provide theoretical justification for a choice between Method 1 (or its modification) and Method 2. When the two procedures produce same, or similar, results, they are practically equivalent, and either one should work, although Method 2 is easier to implement because no examinee data are required. In situations where the procedures produce significantly different results, the question to answer, again not by this study, is really whether reliability estimation in IRT scored tests, especially adaptive tests, should be sample-driven or sample-free.

References

AERA, APA, & NCME (2002). Standards for educational and psychological testing. Washington, DC: Authors.

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Holt, Rinehart and Winston.

Dimitrov, D. M. (2003). Marginal True-Score Measures and Reliability for Binary Items as a Function of Their IRT Parameters. *Applied Psychological Measurement*, 27(6), 440-458.

Feldt, L. S. & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.). Educational Measurement (3rd Edition). Phoenix, AZ: ACE/The Orys Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury, CA: Sage.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA:Addison-Wesley.

Scientific Software International (2003). *IRT from SSI.* (M. du Toit, Ed.). Lincolnwood, IL: Author.

Table 1: Overall Reliability Estimates for Both Adaptive and Non-adaptive Tests

Reliability	Method 1 (Empirical)	Method 2 (Expected)
Adaptive Test	0.90	0.91
Non-adaptive Test	0.81	0.85

Table 2: Reliability Estimates by Panels for Adaptive Tests (Number of Items = 75)

Panel ID	Method 1 (Empirical)	Method 2 (Expected)
1	0.89	0.91
2	0.88	0.91
3	0.90	0.91
4	0.89	0.91
5	0.90	0.90
6	0.89	0.91
7	0.89	0.91
8	0.88	0.91
9	0.88	0.91
10	0.91	0.91
11	0.90	0.91
12	0.91	0.91
13	0.90	0.91
14	0.89	0.91
15	0.90	0.91
16	0.90	0.91
17	0.89	0.90
18	0.90	0.91
19	0.89	0.91
20	0.90	0.91
21	0.91	0.91
22	0.90	0.91
23	0.91	0.91
24	0.90	0.91
25	0.90	0.91
26	0.91	0.91
27	0.89	0.91
28	0.90	0.91
29	0.90	0.91
30	0.90	0.90
31	0.90	0.91
32	0.91	0.91
33	0.89	0.91

Table 3: Reliability Estimates by Panels for Non-adaptive Tests

Panel ID	Method 1 (Empirical)	Method 2 (Expected)
1	0.80	0.85
2	0.83	0.86
3	0.81	0.86
4	0.83	0.86
5	0.81	0.84
6	0.83	0.85
7	0.84	0.87
8	0.78	0.84
9	0.80	0.85
10	0.79	0.85
11	0.79	0.85
12	0.83	0.86
13	0.84	0.86
14	0.76	0.84
15	0.82	0.85
16	0.83	0.87
17	0.83	0.87
18	0.79	0.83
19	0.82	0.85
20	0.81	0.86
21	0.79	0.84
22	0.82	0.87
23	0.81	0.86
24	0.82	0.86
25	0.84	0.86
26	0.82	0.86
27	0.83	0.86
28	0.82	0.85
29	0.79	0.85
30	0.83	0.86
31	0.80	0.85
32	0.81	0.85
33	0.80	0.84
34	0.75	0.85
35	0.81	0.84

Table 4: Number of Examinees by Routes in Adaptive Tests

Panel	MMM	MHH	MMH	MHM
1	43	280	8	12
2	40	275	14	14
3	40	237	15	14
4	43	246	15	18
5	45	232	24	17
6	40	238	12	28
7	37	241	17	17
8	42	274	11	19
9	30	241	12	26
10	39	244	9	7
11	49	224	21	12
12	41	271	8	34
13	38	263	13	20
14	37	261	6	26
15	50	256	9	20
16	49	291	14	18
17	42	230	12	20
18	36	280	14	26
19	33	228	11	23
20	50	260	4	18
21	41	245	7	14
22	39	266	5	14
23	63	241	8	16
24	55	224	17	9
25	26	263	17	16
26	45	240	13	20
27	45	235	15	12
28	43	239	22	13
29	49	233	11	12
30	48	238	19	26
31	32	255	13	21
32	42	251	5	20
33	59	271	13	33

Table 5: Reliability Estimates by Testlets at Each Stage in Adaptive Tests

Panel	Stage 1	Stage 2		Stage 3	
	Medium	Medium	Hard	Medium	Hard
1	0.81	0.66	0.65	0.63	0.63
2	0.80	0.70	0.65	0.67	0.59
3	0.83	0.74	0.66	0.66	0.62
4	0.81	0.73	0.66	0.67	0.63
5	0.83	0.74	0.65	0.67	0.63
6	0.82	0.66	0.65	0.63	0.57
7	0.81	0.69	0.63	0.61	0.65
8	0.82	0.69	0.66	0.61	0.59
9	0.81	0.69	0.66	0.62	0.61
10	0.83	0.74	0.67	0.67	0.67
11	0.83	0.64	0.65	0.55	0.63
12	0.83	0.70	0.69	0.62	0.61
13	0.83	0.68	0.69	0.66	0.63
14	0.83	0.67	0.65	0.64	0.60
15	0.83	0.66	0.67	0.60	0.57
16	0.83	0.72	0.65	0.67	0.63
17	0.80	0.68	0.65	0.62	0.63
18	0.83	0.74	0.63	0.67	0.62
19	0.81	0.66	0.65	0.61	0.63
20	0.82	0.66	0.66	0.66	0.61
21	0.83	0.69	0.66	0.67	0.63
22	0.83	0.67	0.66	0.67	0.59
23	0.83	0.69	0.65	0.70	0.59
24	0.83	0.69	0.69	0.63	0.61
25	0.83	0.74	0.72	0.55	0.67
26	0.82	0.70	0.66	0.66	0.65
27	0.81	0.64	0.65	0.63	0.63
28	0.81	0.73	0.65	0.60	0.63
29	0.81	0.70	0.65	0.61	0.59
30	0.83	0.74	0.69	0.64	0.61
31	0.83	0.67	0.66	0.61	0.65
32	0.83	0.72	0.65	0.67	0.63
33	0.81	0.69	0.63	0.63	0.57
Ave.	**0.82**	**0.69**	**0.66**	**0.64**	**0.62**

Table 6: Reliability Estimates by Route for Adaptive Tests by Different Methods

Panel	MMM			MHH			MMH			MHM		
	M1	M2	M1+	M1	M2	M1+	M1	M2	M1+	M1	M2	M1+
1	0.78	0.91	0.88	0.77	0.91	0.87	0.12	0.91	0.87	0.40	0.91	0.87
2	0.79	0.91	0.89	0.74	0.90	0.86	-0.64	0.91	0.87	-0.41	0.90	0.88
3	0.73	0.91	0.90	0.77	0.91	0.88	-0.50	0.91	0.89	-0.49	0.91	0.89
4	0.76	0.91	0.89	0.76	0.91	0.88	0.53	0.91	0.89	-0.35	0.91	0.88
5	0.78	0.90	0.90	0.76	0.90	0.88	0.46	0.90	0.89	-0.58	0.90	0.88
6	0.75	0.91	0.88	0.70	0.91	0.86	0.23	0.91	0.87	-0.25	0.91	0.88
7	0.65	0.91	0.88	0.75	0.91	0.88	0.08	0.90	0.88	-0.54	0.91	0.87
8	0.68	0.91	0.88	0.74	0.91	0.87	-3.25	0.91	0.88	-0.27	0.91	0.87
9	0.69	0.90	0.88	0.75	0.91	0.87	0.60	0.91	0.88	-0.39	0.91	0.87
10	0.80	0.91	0.90	0.77	0.91	0.89	0.41	0.91	0.90	-0.19	0.91	0.89
11	0.69	0.91	0.86	0.77	0.91	0.88	0.29	0.91	0.87	-2.08	0.91	0.86
12	0.78	0.91	0.88	0.79	0.90	0.88	0.24	0.90	0.88	-0.09	0.91	0.88
13	0.79	0.91	0.89	0.79	0.91	0.88	0.55	0.91	0.88	-1.00	0.91	0.89
14	0.69	0.90	0.88	0.76	0.91	0.87	-1.54	0.91	0.88	0.07	0.91	0.88
15	0.74	0.91	0.87	0.77	0.91	0.87	0.17	0.91	0.87	0.65	0.90	0.88
16	0.77	0.91	0.89	0.77	0.91	0.88	0.60	0.91	0.89	0.30	0.91	0.88
17	0.69	0.91	0.88	0.77	0.90	0.87	-0.33	0.90	0.88	-0.87	0.91	0.87
18	0.78	0.91	0.90	0.74	0.91	0.87	-0.41	0.91	0.89	-0.13	0.91	0.88
19	0.80	0.91	0.87	0.75	0.91	0.87	-0.12	0.91	0.87	-0.10	0.91	0.87
20	0.80	0.91	0.88	0.73	0.91	0.87	0.07	0.91	0.87	-0.73	0.90	0.88
21	0.80	0.91	0.89	0.81	0.91	0.88	-1.80	0.91	0.88	0.25	0.91	0.89
22	0.76	0.91	0.89	0.77	0.91	0.87	-0.32	0.91	0.87	0.07	0.91	0.88
23	0.81	0.91	0.89	0.75	0.92	0.87	0.17	0.91	0.88	0.16	0.91	0.89
24	0.77	0.91	0.88	0.77	0.92	0.88	0.08	0.92	0.88	-0.67	0.91	0.89
25	0.67	0.91	0.88	0.81	0.90	0.90	0.40	0.91	0.90	0.14	0.91	0.87
26	0.74	0.91	0.89	0.78	0.91	0.88	-0.11	0.91	0.89	-0.54	0.91	0.88
27	0.71	0.91	0.87	0.75	0.91	0.88	-0.06	0.91	0.87	0.35	0.91	0.87
28	0.74	0.91	0.88	0.77	0.91	0.87	0.51	0.91	0.89	-0.30	0.91	0.87
29	0.77	0.90	0.88	0.75	0.91	0.87	0.54	0.91	0.88	0.39	0.90	0.87
30	0.77	0.90	0.89	0.75	0.90	0.88	0.54	0.91	0.89	0.32	0.90	0.89
31	0.78	0.91	0.88	0.80	0.91	0.88	0.27	0.91	0.88	-0.44	0.91	0.88
32	0.78	0.91	0.90	0.75	0.91	0.88	0.03	0.91	0.89	-0.03	0.91	0.88
33	0.77	0.91	0.88	0.70	0.91	0.86	0.25	0.90	0.87	-0.17	0.91	0.87
Ave.	**0.75**	**0.91**	**0.88**	**0.76**	**0.91**	**0.88**	**-0.06**	**0.91**	**0.88**	**-0.23**	**0.91**	**0.88**

Figure 1: Diagram of an MST model with 1-2-2 design