The Impact of Statistical Constraints on Classification Accuracy for Multistage Tests

Ryoungsun Park

AICPA American Institute of CPAs

Abstract

This study examined classification testing (i.e., pass/fail) using various levels of test information functions, overall test difficulties, and cutoff scores based on a 1-2-2 multistage test (MST) design and simulation components. Automated test assemblies were performed using a linear programming model to ensure the desired test information functions. The results indicated that the correct classification rates can be improved by more than 2% by increasing the test information by 50%. The "diminishing return" of accuracy as a function of test information was also observed. The overall test difficulty also affected classification accuracy.

Theoretical Framework

The credentialing or licensing process involves high-stakes tests that are used to determine whether a particular candidate has the skills or knowledge to be certified or licensed (Bergstrom & Lunz, 1999). This type of examination is conveniently called classification testing because the main objective of the test is to classify an examinee into mutually exclusive categories (e.g., a category in which examinees can receive licenses). To achieve this objective, the test development, administration, scoring, and reporting of test scores to candidates are carefully governed by the certification and licensure board.

Computer-based tests (CBT) have garnered greater attention related to classification testing. For example, computer adaptive testing (CAT) administers an individual item based on the examinee's current ability estimate; thus, it is referred to as fully adaptive testing (van der Linden & Glas 2000). Meanwhile, multistage tests (MST), another CBT approach, select a collection of items (i.e., a module) to match the examinee's estimated ability (Patsula, 1999). Therefore, MST is considered to be a compromise between fully CAT and non-adaptive linear test forms (Jodoin, Zenisky, & Hambleton, 2006). The main benefit of MST compared to other testing approaches is that test developers can achieve quality assurance for various design aspects (e.g., content distribution of the test or the exposure control for an individual item) because test forms are constructed before the actual administration.

MST's individual test forms are called panels, and several panels can be constructed simultaneously. Within a panel, there are a multiple number of modules. A group of modules in a panel form a stage, and examinees are routed to one of the modules within a stage. The module to be administered within a stage is determined by the current ability estimate. The routes of modules that examinees take within a panel are called pathways (Luecht, 2000). In construing

MST, under the umbrella of item response theory (IRT), statistical constraints are specified as the target test information function (TIF). The TIF determines the amount of information in either module level or pathway level along the ability scale. The extensive test information is directly translated to the higher measurement accuracy because test information is inversely related to the standard error of measurement (Embretson & Reise, 2000).

Previous studies have examined the effect of TIFs on MST classification accuracy. Zenisky (2004) conducted a simulation study to determine the classification accuracy of MST relative to the different levels of TIFs using a dichotomously scored item pool. In addition, the impact of allocating more information in the first module was investigated. The results showed that accuracy increased as the TIFs increased. In addition, when a large amount of information was allocated for the first module, the performance was better in terms of classification accuracy. More recently, Kim, Chung, Dodd, and Park (2012) investigated the impact of TIFs of the first module in the context of classification testing using a mixed-format item pool. They controlled the information for the first module by increasing and decreasing the regular levels of TIFs by 50%. The simulation results revealed that correct classification rates for tests with 50% increased first module information were up to 1.44% greater compared to the MST with 50% decreased first module information.

Objectives

It is crucial to consider variations of amounts of TIFs across test assemblies in order to make more accurate decisions about examinees' abilities. A few MST studies have attempted to offer variations on levels of TIFs, but they only considered small variations in the TIF levels in their MST designs (e.g., Zenisky 2004, and Kim et al., 2012). In addition, most previous MST research considered up to three cutoff points based primarily on the normal distribution of

abilities (e.g., 30%, 50%, and 70% passing rates; Kim et al., 2012). Finally, many MST studies have considered only one difficulty condition. To fill these gaps, the current study aims to determine how much decision accuracy is influenced by providing relatively more variations in the TIF levels, numbers of cutoff scores, and the overall test difficulties.

Method

This study considered various MST design conditions to determine how pass/fail decisions were made. Seven levels of TIFs were constructed by varying levels (or heights) of TIFs, and six cutoff points (i.e., thetas of -0.5, -0.3, -0.1, 0.1, 0.3, and 0.5) were empirically chosen at various points along the ability scale based on the current data set. The study included three levels of overall test difficulties (i.e., easy, medium, and hard levels). The two major pathways in each test difficulty condition were peaked at the thetas of (-1.5 and -0.4), (-0.6 and 0.4), and (0.0 and 1.0) for easy, medium, and hard tests, respectively. Therefore, the simulation study included MST conditions of 7 (levels of TIFs) x 6 (cutoffs) x 3 (overall test difficulties). Thus, 126 conditions were evaluated for classification accuracy.

Item Pool

The item pool was obtained from the actual operational licensure exam but a necessary modification was made to satisfy the required security while preserving statistical characteristics. For the current study, 3,046 dichotomously scored items were calibrated according to the three-parameter logistic (3PL) model. The pool includes six content domains and 84 sub-domains. Each content domain comprises 11.6%, 16.2%, 25.6%, 20.7%, 11.2%, and 14.7% of the item pool, respectively.

MST Assembly

MST construction was written in R (R development Core Team, 2013) for the current

study using the lpSolveAPI package (Konis, 2011).In order to control the peak of the test information, minimax LP models(van der Linden, 1987) were specified according to the target TIFs for each module construction. The LP model specifying statistical constraints is written as:

minimize y, (1)

subject to

$$\sum_{i=1}^{N} I_i(\theta_k)x_i - T(\theta_k) \leq y, \quad for\ all\ k, \tag{2}$$

$$T(\theta_k) - \sum_{i=1}^{N} I_i(\theta_k)x_i \leq y, \quad for\ all\ k, \tag{3}$$

$$\sum_{i=1}^{N} x_i = n, \tag{4}$$

$$y \geq 0, \tag{5}$$

and

$$x_i \in \{0,1\}, i = 1,...,I, \tag{6}$$

where x_i is the decision variable for test unit i; y is the real-valued absolute deviation from TIF; N is the number of items; $I_i(\theta_k)$ is the information for item i at θ_k; $T(\theta_k)$ is the TIF at θ_k; k is the number of theta points at the latent trait scale; and n is the number of items in a module.

In terms of the MST panel structure, a 1-2-2 MST panel structure (one panel) with four pathways was constructed, with 25 items included within each module (i.e., a total of 75 items per pathway).

Statistical Constraints

The statistical constraints were manipulated by the levels of TIFs at the difficulty for each module. A common practice to determine TIFs involves performing preliminary constructions (Luecht & Nungester, 1998; Zenisky, 2004). The average information of the preliminary

constructions is taken as a feasible TIF given the item pool. However, there is no formal proof that TIFs obtained in this method present the best possible TIFs. For this study, a large number of potential TIFs in the given item pool were considered as a variable of MST simulation. The largest TIFs can be obtained by choosing items of the greatest information given the location. Seven different height values were considered by adjusting the peak of TIF detrimentally. For example, the largest peak of TIF for pathways (e.g., medium-medium-medium pathway and medium-hard-hard pathway) was obtained from the preliminary assembly of 1-2-2 MST. When the largest peak was considered to have 100% of TIFs, the rest of the MST designs had 90%, 80%, 70%, 60%, 50%, and 40% of TIFs, respectively. In this way, measurement accuracy across the varied amount of information is highlighted and inferences could be drawn regarding the relative importance of the peak values. Furthermore, the content specifications require that each module include 4, 5, 4, 4, 4, and 4 items from each of six content domains, respectively, but no more than three items from the same sub-domain.

Data Generation and MST Simulation

Data generation and simulation routines were written in R for the current study. One hundred replications with 1,000 simulated examinees each were generated according to the standard normal distribution or uniform distribution (between theta=-4.0 and 4.0).The MST simulation was performed according the following steps. For each simulated examinee, all items within the first module were administered. After finishing entire items, ability estimation and routing decision were performed to assign an examinee to the remaining MST modules. To estimate ability, expected a posteriori (EAP) was used (e.g., Jodoin, 2003; Luecht et al., 2006; Hambleton & Xing, 2006). For the routing method, the cutoff theta points were used to assign each examinee to either the medium or hard module at the subsequent stages. For instance, if the

ability estimates of simulated examinees were greater than the cutoff theta, examinees were routed to the hard module. After all 75 items were administered, simulated examinees were classified into "pass" or "fail" categories by comparing their ability estimates to cutoff scores points.

Outcome Measure

All conditions were compared in terms of the following decision-making criteria averaged across 100 replications: (1) correct classification rate; (2) false positive error rate; and (3) false negative error rate.

Results

Figure 1 depicts the seven levels of TIFs for the two main pathways constructed for the easy difficult test. Two target peak locations for two major pathways occur at -1.6 and -0.4 on the theta scale, respectively. Pathway information (i.e., the accumulation of three modules' information) when theta was -1.6 was 22.18, 19.96, 17.22, 14.64, 12.63, 9.96, and 8.00 from 100% TIF to 40% TIF conditions, respectively. When the theta was -0.4, the information was 25.04, 24.03, 21.75, 19.16, 16.17, 13.17, and 9.62 for each TIF condition, respectively. Five quadrature points were specified as the target TIF during the ATA. Table 1 presents the correct classification rates (CCRs), false negative error rate (FNER), and false positive error rate (FPER) in percentages of MST simulation conditions with varied TIFs and cutoff scores for pathway-level information peaked at -1.6 and -0.4. The data indicated a positive relationship between TIF and CCR as CCR increases while TIF increases. The improvement of CCR due to the increase of TIF was large when TIFs were relatively small. For example, on average a 1.41% improvement of CCR occurred when TIF increased from 40% to 50%, but only a 0.26% increase was observed when TIF increased from 90% to 100%. Figure 2 shows the CCRs for different levels of TIFs

across the cutoff scores in the easy difficulty test. The classification decision accuracy decreased as the amount of information decreased from 100%. A noticeable decrease of accuracy was observed when TIF dropped to 60% or more. Within the cutoff conditions considered in this study, CCR decreased in general as the cutoff point increased. The majority of conditions resulted in CCR greater than 89. For instance, CCR ranged from 91.32 to 87.95 when the cutoff was 0.1.

[Insert Figures 1 and 2 about here]

[Insert Table 1 about here]

Figure 3 depicts the seven levels of TIFs for the two main pathways actually constructed for the medium difficulty test. The pathway information when the theta was -0.6 was 22.18, 19.96, 17.22, 14.64, 12.63, 9.96, and 8.00 from 100% TIF to 40% TIF conditions, respectively. When the theta was 0.4, pathway information was 25.04, 24.03, 21.75, 19.16, 16.17, 13.17, and 9.62 for each TIF condition, respectively. Table 2 presents the CCRs, FNERs, and FPERs in percentages of MST simulation conditions with varied TIFs and cutoff scores for the easy test. Figure 4 shows the CCRs for different levels of TIFs across cutoff scores with the easy test. CCRs were smallest around the cutoff of 0.0, but the accuracy increased as cutoffs shifted to both extreme ends. The CCR ranged from 92.57 to 88.55 when the cutoff was 0.1.

[Insert Figures 3 and 4 about here]

[Insert Table 2 about here]

Figure 5 depicts the seven levels of TIFs for the two main pathways actually constructed for the hard difficulty test. The pathway information when the theta was 0.0 was 20.81, 20.15, 18.38, 16.09,13.79, 11.65, and 9.21 from 100% TIF to 40% TIF conditions, respectively. When the theta was 1.0, pathway information was constructed to 19.41, 18.90, 17.61, 14.91, 12.92,

10.09, and 7.94 for each TIF condition, respectively. Table 3 presents the CCRs, FNERs, and FPERs in percentages of MST simulation conditions with varied TIFs and cutoff scores for the hard difficulty test. Figure 6 shows the CCRs for different levels of TIFs across cutoff scores with the hard difficulty test. CCR increased as the cutoff moved away in both directions from when the theta was 0.0, and CCRs ranged from 92.49 to 89.03 when the cutoff was 0.1.

[Insert Figures 5 and 6 about here]

[Insert Table 3 about here]

Discussion

Given the cutoff and test difficulty, the CCR decreased as the test information decreased. The amount of the CCR decrease was not uniform for all levels of TIFs as the rate of the drop increased as TIF decreased from 100%. A significant drop of CCR (i.e., more than a 2% drop) from its maximum value was manifested when TIF decreased below 60%.

An interaction was observed between the test difficulty and cutoff scores for CCR. For medium test difficulties, the smallest CCR was observed at a theta around 0.0. This is expected as most candidates are concentrated around a theta of 0.0 in the population distribution of the standard normal distribution. As the test difficulty shifted left on the theta scale (i.e., easy test difficulty), the minimum point of CCR shifted to the right from a theta of 0.0 on the theta scale (see Figure 2). For instance, the CCRs at the 0.5 cutoff were lower than the 0.1 cutoff for the easy test difficulty condition. Although fewer candidates were concentrated around the 0.5 cutoff, the low information of the easy difficulty test at the 0.5 theta actually increased the classification errors.

The distance between the test difficulty and cutoff points affected CCRs. Greatest accuracy was achieved when the test difficulty was as close to the target cutoff as possible.

Given the -0.5 cutoff, for example, CCRs were largest when the test difficulty was easy (i.e., around a theta of -0.5). On the other hand, when the cutoff was 0.5, the hard difficulty test produced the greatest CCR. This implies that, in order to achieve a high classification accuracy, it is best to construct a test of high information on the cutoff score.

Target TIFs are difficult to assess before the construction for test developers. However, realistic TIFs might be assessed on the reference of the pool's average item information. Based on the average item information in the pool, the average information for a test of 75 items corresponded to the 50% TIF condition (i.e., pathway information of about 10.0). Given that, we can predict that pathway information should be above 10.0 and could be close to the 100% TIF condition depending on the number of panels under construction; if a small number of panels is required, a large value for the TIF is realizable. However, when the actual constructions require multiple panels, not all panels are expected to obtain the maximum test information (e.g., 100% TIF condition).

This study provided the simulation results of classification accuracy across levels of cutoffs, test difficulties, and the amount of TIFs. The evidence demonstrated that the amount of information as well as the location of test difficulty should be controlled as important factors that increase the measurement accuracy of the classification test. The increased test information produced an increase in accuracy. However, the rate of the accuracy increase decreased as the information increased. Therefore, test developers need to be aware of these "diminishing returns" between test accuracy and test information.

References

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure.*Innovations in computerized assessment*, 67-91.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* (Vol. 4). Psychology Press.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education*, *19*(3), 221-239.

Jodoin, M. G. (2003). *Psychometric properties of several computer-based test designs with ideal and constrained item pool.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Jodoin, M. G., Zenisky, A., & Hambleton, R.K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*, 203-220.

Kim, J., Chung, H., Dodd, B. G., & Park, R. (2012). Panel design variations in the multistage test using the mixed-format tests. *Educational and Psychological Measurement*, 72 (4), 574-588.

Konis, K (2011). *lpSolveAPI: R Interface for lpsolve version 5.5.2.0*. R package version 5.5.2.0-5.

Luecht, R. M. (2000). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education New Orleans, LA.

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

Luecht, R. M., Brumfield, T., &Breithaupt, K. (2006) A Testlet Assembly Design for Adaptive Multistage Tests. *Applied Measurement in Education,* 19(3), 189-202.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Issues in innovative item types. In *Practical considerations in computer-based testing*(pp. 70-91). Springer New York.

Patsula, L. N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0, URL http://www.R-project.org.

van der Linden, W. J. (1987). Automated test construction using minimax programming. In W.J. van der Linden (Ed.), *IRT-based test construction* (Research Report 87–2, chap. 3). Enschede: University of Twente, Department of Education.

van der Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: Theory and practice*. Kluwer Academic Pub.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

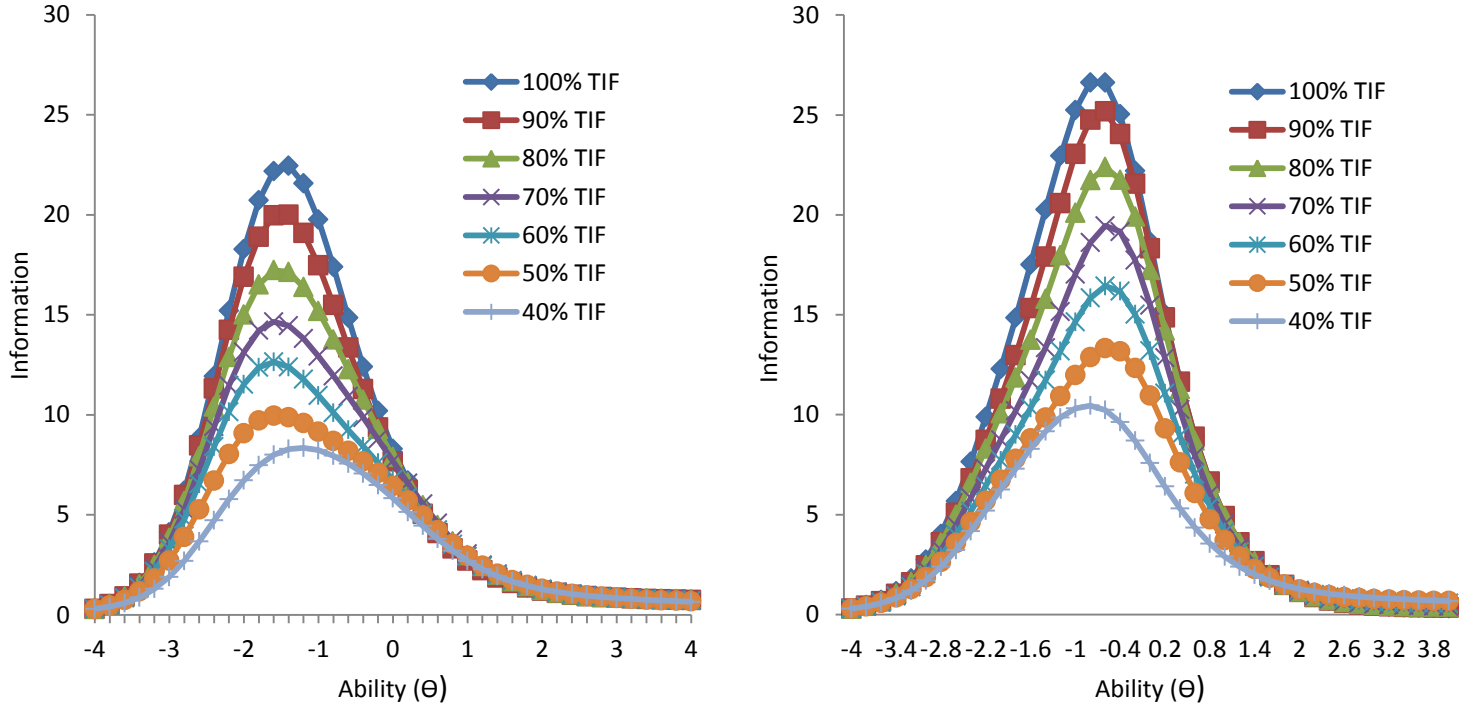

Figure 1. Various levels of test information functions for medium–medium–medium (left) pathway and medium–hard–hard pathway (right) with the easy difficulty test.
Note. TIF= test information function.

Table 1: *Comparisons for the Classification Error and Accuracy Rates of MST simulation with Varied Levels of TIFs, Cutoffs for the easy difficulty test (Averaged Across 100 Replications)*

	CO = -0.5			CO = -0.3			CO = -0.1			CO = 0.1			CO = 0.3			CO = 0.5		
L-TIF	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER
100%	93.53	4.2	2.26	92.74	3.84	3.42	91.66	3.44	4.9	91.32	4.68	4	90.44	4.99	4.57	89.94	4.98	5.08
90%	93.35	4.26	2.38	92.55	4.02	3.42	91.64	3.64	4.73	91.06	4.84	4.1	90.19	5.17	4.65	89.94	5.23	4.82
80%	93.21	4.23	2.56	92.27	4.22	3.51	91.3	3.98	4.72	90.71	5.07	4.22	90.21	5.27	4.52	89.96	5.14	4.9
70%	92.72	4.28	3	91.91	4.29	3.8	91.2	4.07	4.73	90.64	5.03	4.33	89.99	5.32	4.7	89.92	4.97	5.1
60%	92.3	4.26	3.43	91.11	4.76	4.14	90.55	4.61	4.84	89.89	5.36	4.75	89.38	5.69	4.93	89.28	5.48	5.24
50%	91.7	4.32	3.98	90.61	4.86	4.53	89.77	4.92	5.3	89.36	5.4	5.24	88.85	5.69	5.46	88.72	5.61	5.67
40%	91.08	4.06	4.85	89.83	4.69	5.48	88.85	4.88	6.26	87.95	5.43	6.62	87.39	5.78	6.83	87.21	6.01	6.78

Note. Each of the 100 replications contained 1,000 observations. CO= cutoff; D = ability distribution; L-TIF = levels of test information function; CCR = correct classification rate; FNER = false negative error rate; FPER = false positive error rate.

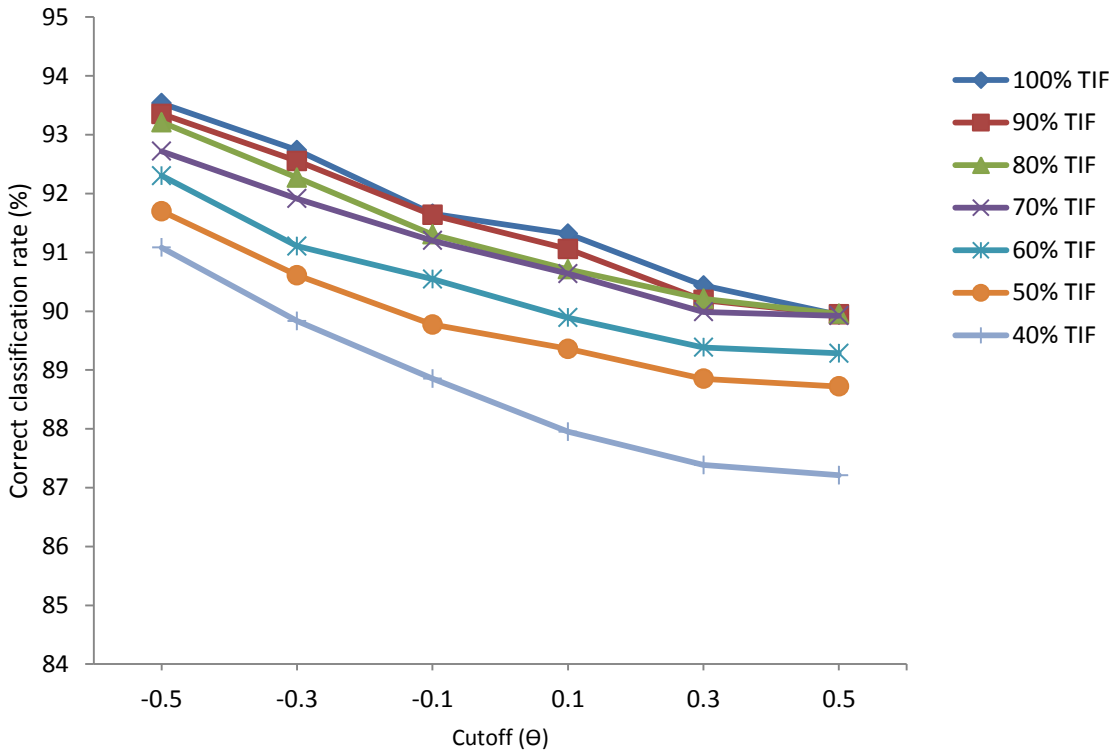

Figure 2.Correct classification rates for various levels of TIFs across cutoffs for the easy difficulty test averaged across 100 replications.

Note. Each of the 100 replications contained 1,000 observations. TIF= test information function.

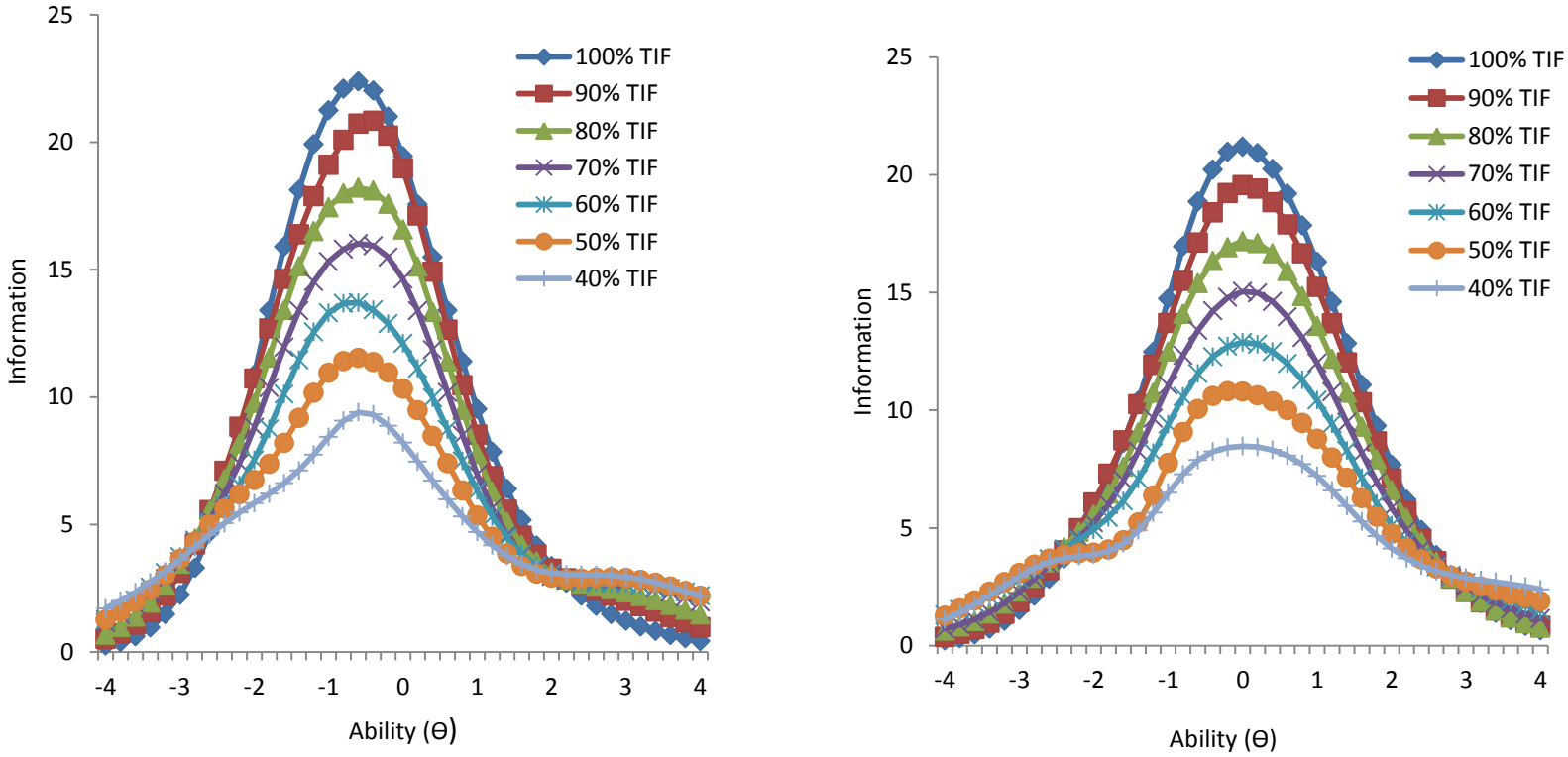

Figure 3. Various levels of test information functions for medium–medium–medium (left) pathway and medium–hard–hard pathway (right) with *for the medium difficulty test*.
Note. TIF= test information function.

Table 2: *Comparisons for the Classification Error and Accuracy Rates of MST simulation with Varied Levels of TIFs, Cutoffs for the medium difficulty test (Averaged Across 100 Replications)*

	CO = -0.5			CO = -0.3			CO = -0.1			CO = 0.1			CO = 0.3			CO = 0.5		
L-TIF	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER
100%	93.82	3.99	2.19	93.58	3.11	3.31	92.59	2.28	5.13	92.57	4.84	2.59	92.84	3.78	3.38	93.26	2.90	3.83
90%	93.77	3.68	2.55	93.03	3.27	3.70	92.33	2.69	4.98	92.26	4.85	2.88	92.50	4.03	3.47	92.73	3.51	3.77
80%	93.19	3.70	3.10	92.60	3.48	3.92	91.80	3.25	4.95	91.74	4.97	3.29	91.93	4.39	3.68	92.28	3.96	3.76
70%	92.65	3.65	3.70	91.97	3.75	4.28	91.20	3.80	5.00	91.02	5.19	3.79	91.35	4.77	3.89	92.01	4.43	3.56
60%	92.22	3.56	4.22	91.47	4.06	4.47	90.61	4.34	5.04	90.48	5.27	4.25	90.78	5.16	4.06	91.51	4.88	3.61
50%	91.48	3.79	4.73	90.41	4.40	5.19	89.77	4.86	5.37	89.73	5.52	4.75	90.04	5.51	4.45	90.67	5.32	4.01
40%	90.73	3.97	5.30	89.63	4.68	5.68	88.72	5.45	5.83	88.55	6.01	5.44	88.55	6.31	5.14	89.37	6.27	4.35

Note. Each of the 100 replications contained 1,000 observations. CO= cutoff; D = ability distribution; L-TIF = levels of test information function; CCR = correct classification rate; FNER = false negative error rate; FPER = false positive error rate; N= normal distribution.

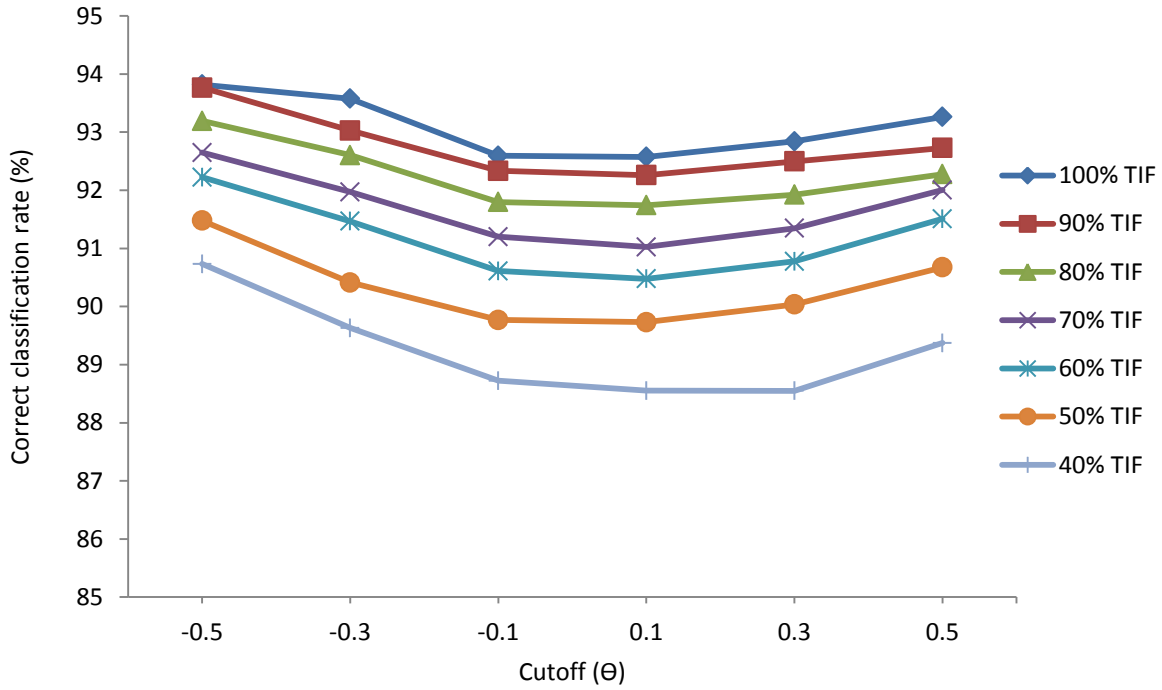

*Figure 4.*Correct classification rates for various levels of TIFs across cutoffs for the medium difficulty test averaged across 100 replications

*Note.*Each of the 100 replications contained 1,000 observations. TIF= test information function.

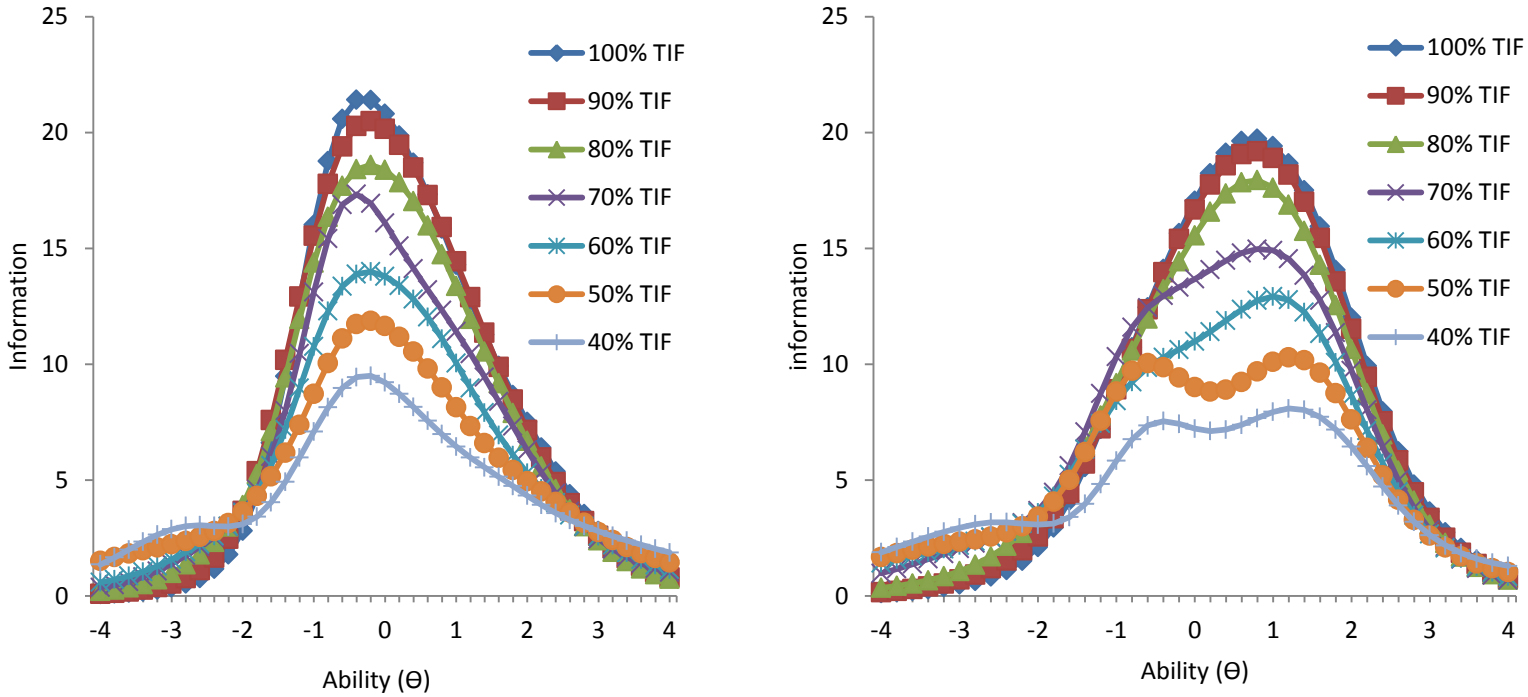

Figure 5. Various levels of test information functions for medium–medium–medium (left) pathway and medium–hard–hard pathway (right) with for the hard difficulty test.
Note. TIF= test information function.

Table 3: *Comparisons for the Classification Error and Accuracy Rates of MST simulation with Varied Levels of TIFs, Cutoffs for the hard difficulty test (Averaged Across 100 Replications)*

	CO = -0.5			CO = -0.3			CO = -0.1			CO = 0.1			CO = 0.3			CO = 0.5		
L-TIF	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER	CCR	FNER	FPER
100%	93.15	4.41	2.27	92.85	3.31	3.84	92.43	2.76	4.81	92.49	4.94	2.61	93.28	3.42	3.30	93.51	2.84	3.65
90%	93.17	4.43	2.25	92.63	3.40	3.97	92.47	2.73	4.80	92.49	5.07	2.44	93.01	3.56	3.43	93.3	2.75	3.95
80%	92.72	4.42	3.30	92.40	3.66	3.94	92.24	3.07	4.69	92.2	5.10	2.70	92.84	3.52	3.64	93.29	2.89	3.82
70%	92.35	4.46	3.30	92.04	4.07	3.89	91.92	3.72	4.36	91.84	5.15	3.01	92.29	3.93	3.78	92.94	3.49	3.57
60%	91.78	4.43	3.79	91.28	4.56	4.16	90.93	4.63	4.44	91.06	5.20	3.73	91.59	4.54	3.88	92.26	4.11	3.63
50%	91.51	4.23	4.26	90.84	4.50	4.65	90.25	4.66	5.09	90.58	4.92	4.50	90.82	4.79	4.39	91.32	4.55	4.13
40%	90.34	4.99	4.67	89.48	5.33	5.19	89.32	5.47	5.21	89.03	5.61	5.36	89.50	5.43	5.07	90.32	5.07	4.62

Note. Each of the 100 replications contained 1,000 observations. CO= cutoff; D = ability distribution; L-TIF = levels of test information function; CCR = correct classification rate; FNER = false negative error rate; FPER = false positive error rate

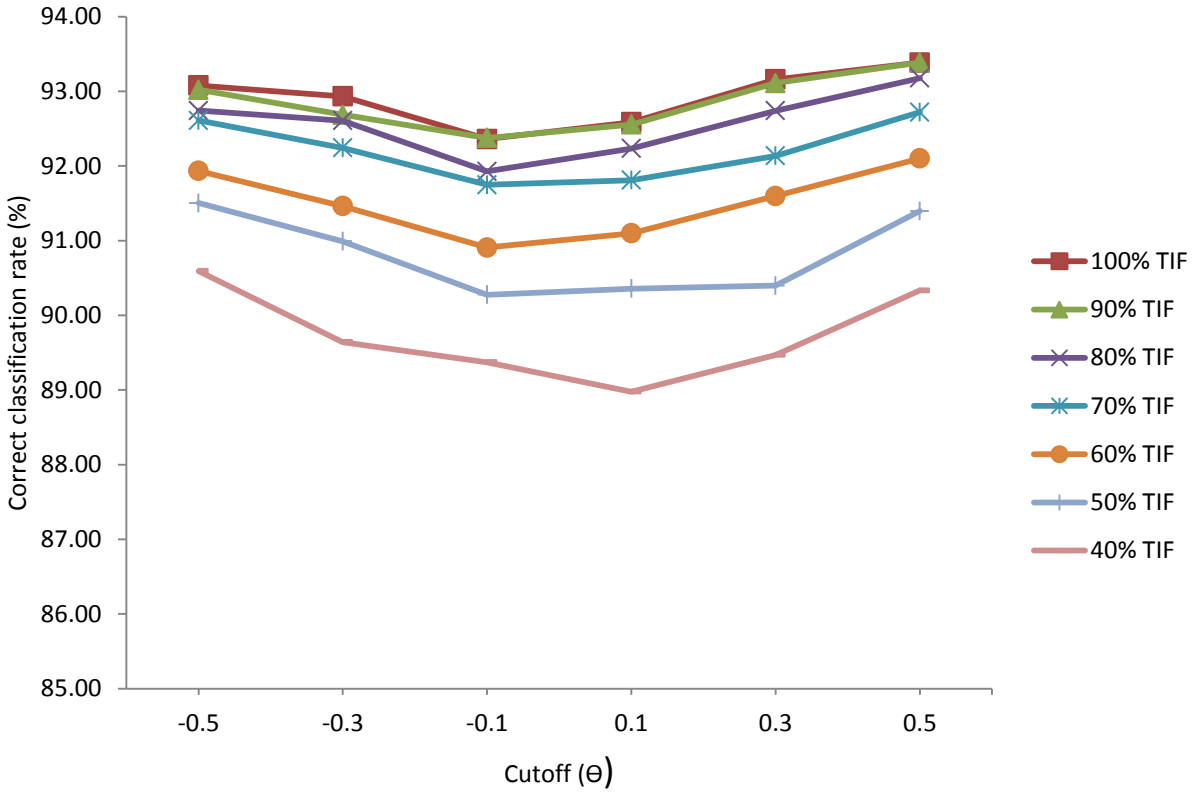

Figure 6. Correct classification rates for various levels of TIFs across cutoffs for the hard difficulty test averaged across 100 replications

Note. Each of the 100 replications contained 1,000 observations. TIF= test information function