Running head: CBT DESIGNS FOR CREDENTIALING

Comparison of the Psychometric Properties of Several Computer-Based Test Designs for Credentialing Exams

Michael Jodoin, April Zenisky, and Ronald Hambleton

University of Massachusetts at Amherst

April 8, 2002

Abstract

Many credentialing agencies today are either administering their examinations by computer or are likely to be doing so in the coming years. Unfortunately, although several promising computer-based test designs are available, little is known about how well they function in typical examination settings. The goal of this study was to compare fixed length examinations (both operational forms and newly constructed forms) with several variations of multi-stage test designs for making pass-fail decisions. Results were produced for three passing scores. Four operational 60-item examinations were compared to (1) three new 60-item forms, (2) 60-item three-stage tests, and (3) 40-item two-stage tests, all constructed using automated test assembly software. The study was carried out using computer simulation techniques that were set, in all respects, to mimic common examination practices. All 60-item tests, regardless of design or passing score, produced accurate ability estimates, and acceptable and similar levels of decision consistency and decision accuracy. One interesting finding was that the 40 item test results were poorer than the 60 item test results, as expected, but were very much in the range of acceptability. This raises the practical policy question of whether content valid 40 item tests with lower item exposure levels and/or savings in item development costs are an acceptable trade-off for a small loss in decision accuracy and consistency.

Comparison of the Psychometric Properties of Several Computer-Based Test Designs for Credentialing Exams

Many credentialing examination boards today are either administering their examinations via computer, or expect to be doing so within the next few years. There are many well-known reasons for credentialing examinations to switch their examinations from paper and pencil administrations to computer administrations: flexibility for candidates to schedule their exams at convenient times, the option for immediate score reporting, exam score validity can be increased by introducing new item formats to measure higher level cognitive skills, and more. There are a few negative consequences of computer-administered examinations as well. For example, often the cost to candidates is increased, and there is the increased threat of item exposure since examinations are being administered every day (e.g., van der Linden & Glas, 2000; Wainer, et al., 2000). Regardless of these and other shortcomings, the transition of credentialing exams from pencil and paper administrations to computer-based administrations seems inevitable. This leads to many questions, including "Which computer-based test design might be best?"

Computer-based exams can be implemented with a number of test designs. At one end of the continuum, a continuum reflecting the extent to which a computer-based exam adapts item selection to the performance of candidates during an examination administration, are *fixed-forms examinations* that are closely matched to each other in content and item statistics. These examinations can be assigned on a random basis to candidates and the exposure level of individual test items can be controlled by the number of fixed test forms that are available. A popular variation on fixed forms examinations are “linear-on-the-fly tests” (LOFTs) where each candidate receives a unique set of test items that is strictly matched to content and statistical

specifications. Here, again, item exposure levels can be calculated easily and controlled at desired levels through the content and statistical specifications as well as the available item bank size. The securities industry, which administers computer-based exams to 100,000 candidates a year, has used the LOFT design for more than 15 years. Neither fixed-form nor LOFTs vary the exam length or match item selection to the performance levels of candidates during the exam administration.

At the other end of the continuum are *computer-adaptive tests* (CATs). Here, content specifications remain essential and items are selected to optimize the measurement properties of the test administered to each candidate. A variety of stopping rules for the examination are available including achieving a desired level of measurement error or achieving a desired level of decision accuracy (e.g., 90% probability of a correct decision). Often, the same level of measurement precision can be achieved with CATs with 40 to 50% fewer items than fixed-forms or LOFTs (Wainer, et al., 2000; Weiss, 1983). Item exposure can be controlled overall, or conditional on ability (e.g., Stocking & Lewis, 2000).

A middle position on the degree of test adaptation continuum is multi-stage testing (MST), where test items are administered to candidates in fixed sets of items that are called modules (also called item blocks or testlets). Modules may differ in their level of difficulty but are matched on content to the extent possible. Branching of candidates from one module to another may be done to match examinee ability estimates with the difficulty of items in the prepackaged modules. Often, item exposure levels can be determined in advance, or at least predicted if a reasonably accurate estimate of the candidate score distribution is available. This might be available from a previous administration of the exam, or a worst-case scenario approach can be taken so that item exposure levels will not exceed desired levels.

With MST designs, candidates may change answers or skip test items and return to them, prior to actually finishing a module and moving on to another. This limited control by the candidate (i.e., being able to omit test items, and/or skip test items and return to them later) is a test administration feature that is responsive to one of the main criticisms of CAT administrations (Vispoel, 1998). Measurement precision may be gained over fixed- form examinations or LOFTs without an increase in test length by adapting the exam administration to the performance levels of the candidates (see, for example, Lord, 1980; Patsula, 1999; Patsula & Hambleton, 1999). Of course, in practice, MST designs have many degrees of freedom: they can vary in the numbers of modules that are administered, the lengths of modules, branching rules employed, amount of item overlap in the modules, item exposure levels, etc. In addition, the belief exists that it may be easier to meet content specifications with MST designs than with CATs, and at least to date, some candidates have expressed a preference for MSTs over CATs. Clearly then, a number of computer-based test designs are available, and MSTs appear to be a promising option to fixed-forms, LOFTs, or CATs. But more research is needed, and perhaps surprisingly given its potential for improving assessment practices, MST designs have received much less research attention than CAT designs. Also, only a few studies have considered the use of the MST design when classification of candidates is the main purpose of the testing. For prominent exceptions, see Luecht and Nungester (1998) and multiple chapters in the edited book by van der Linden and Glas (2000).

A number of prior studies have compared computer-based test designs in terms of the reliability and validity of ability estimates. Such studies are helpful, but they do not directly address the fundamental measurement problem with credentialing exams and that is to make reliable and valid pass-fail decisions. In principle, it is quite possible that a computer-based test

design might lead to relatively poor ability estimates, though the test design may be quite acceptable for making sufficiently reliable and valid pass- fail decisions. This might be the case, for example, with fixed-form examinations that typically do less than optimal estimation of low and high ability scores, but the level of precision achieved with this examination design may be quite sufficient for making reliable and valid pass-fail decisions (especially if the pass rate is somewhere in the middle of the candidate score distribution). Xing (2000) and Xing and Hambleton (2002) have compared computer based test designs to investigate the reliability and validity of pass-fail decisions, but their primary variables of interest were item quality and item bank size.

Computer-based test designs for credentialing exams definitely have their advantages, but ultimately, they are judged by their content validity and by the levels of decision consistency and decision accuracy that are achieved. These levels depend on many factors including item quality, the location of the passing score in relation to the candidate score distribution, exam length, and the test information function. The choice of computer-based test design is also very important, and this choice was the focus of the present study. Specifically, the goal of the present study was to compare fixed length tests (both operational forms and newly constructed forms) with two MST designs for making pass-fail decisions. Results were produced for several passing scores.

Method

Data from four operational administrations of a large-scale, high-stakes, national certification examination, each consisting of approximately 60 dichotomously scored multiple-choice items classified into three primary content areas, were used in the study. The mean number of items in each content area (across the four operational forms) was used to determine the content specifications for the development of all subsequent forms in the study. Exam items

were calibrated using the three-parameter IRT model to develop an item bank that consisted of 238 items. These item parameters were used for subsequent automated test forms assembly and for the simulation of candidate item response data.

Forms Assembly

In order to develop realistic target information functions for new fixed form and MST designs with the existing item bank, test information functions for each of the four operational forms were determined, and are displayed in Figure 1. Subsequently, the mean of the four operational test information functions was used a basis for the development of target information functions for automated test assembly using CASTISEL (Luecht, 1998).

[Insert Figure 1 About Here]

Three parallel fixed length test forms consisting of 60 items were created using the content specifications and the target information function (which was the mean of the test information functions from the four operational forms). This resulted in three content-balanced 60-item test forms that will subsequently be referred to as *LOFT forms* 1, 2, *and* 3, respectively, to distinguish them from the four operational (REAL) test forms. Three forms were generated since it would allow an item exposure rate of 33% and utilize three-quarters of the available item pool. The authors felt that given the size of the item pool this reflected a reasonable compromise between item exposure and pool usage. In operational situations, a larger item pool and lower exposure controls would be needed. For example, doubling the item bank would enable the exposure level of items to be cut to 17.5%. Doubling the bank again would reduce exposures below 10%.

Figure 2 displays the test information functions for each LOFT form and the target test information function. It is apparent from Figure 2 that the test forms assembled with CASTISEL

provided less than the specified information for higher abilities and for two of the three forms more information at lower abilities.[1] These are not ideal forms, but they represent the best that could be constructed with the available item bank, the content specifications, and the test assembly software.

[Insert Figure 2 About Here]

Next, a 3-stage MST design following a 1-3-3 stage structure was constructed as illustrated in Figure 3. Again, the mean test information function from the operational administrations was used to form target information functions for each module. Target information functions for the medium difficulty modules for stages one, two, and three were set to one-third the value of the mean test information function from the operational forms. The easy modules in stages two and three were set to one third of value of the mean test information function with a negative horizontal transformation of one half standard deviation. The hard modules in stages two and three were set to one third of the value of the mean test information function with a positive horizontal transformation of one half standard deviation. That is, the easy and hard modules were identical to the medium difficulty module in terms of information but shifted left and right by one half standard deviation, respectively.

[Insert Figure 3 About Here]

In order to maintain an item exposure level of 33%, three parallel-forms were also constructed for the medium difficulty stage one module. This made it possible to create three MST panels: Each MST panel consisted of one of the three unique stage one medium difficult modules and the same six modules in the second and third stages and are referred to as panels 1, 2, and 3. Thus, nine 20-item modules were simultaneously created resulting in item pool usage and item exposure rates that would be comparable to the LOFT condition. This 3-stage MST

design with module target information functions based on one-third of the mean operational test information function is subsequently referred to as *3-Stage MST Design 1*. Figure 4 shows the module information functions for the first panel of the 3-Stage MST Design 1. Again, the automated test assembly process had difficulty meeting the target information for higher abilities but in general provided additional information at lower abilities.

[Insert Figure 4 About Here]

Subsequently, a second 3-Stage MST with a 1-3-3 design as outlined in Figure 3 was developed. In this MST design, target information functions for stage 1 modules were reduced to one quarter and stage 2 and 3 modules were increased to three eighths of the mean operational test information function. This design had the desirable effect of placing the more discriminating items in stage 2 and 3 modules when better matching of candidate ability estimates and item difficulties is possible. The idea is simple: To gain an advantage from the most discriminating items, it is best to have good ability estimates for optimal assignment. More accurate ability estimates are available after the first and second stage than before the routing test administration, and so moving the more discriminating items to the second and third stages of an MST should improve its measurement precision.

Once again, easy and hard module target information functions were horizontally translated by one half standard deviation to the left and right, respectively. Similarly, three 20-item stage 1 modules were created simultaneously with six 20-item stage 2 and 3 modules to create three MST panels. This MST design is subsequently referred to as *3-Stage MST Design 2*. Figure 5 shows the module information functions and similar to the other designs, module information functions are below the target information functions for higher abilities.

[Insert Figure 5 About Here]

In preliminary analyses of both 3-Stage MST designs, ability estimates computed at the end of each stage in the three-stage design were correlated with the true ability for each MST panel for two replications and are summarized in Table 1. This analysis had a major implication for the design of the study. While there was a substantial improvement in the ability estimates between the end of the first and second stages, the improvement between the second and third stages was modest. This result suggested that two additional MST test designs would be informative. A *2-Stage MST Design* 1 was created by simply eliminating the stage 3 modules from the 3-Stage MST Design 1. This resulted in a 2-Stage MST Design with three panels. Each panel consisted of one of three unique 20-item medium difficulty modules in the first stage and the same easy, medium, and hard modules in stage 2 of the 3-Stage equivalent. Similarly, the 3-Stage MST Design 2 was converted to a *2-Stage MST Design* 2 *by* simply dropping the stage 3 modules. With these two additional designs, it was possible to compare 40 item and 60 item MSTs to each other and to the operational forms.

[Insert Table 1 About Here]

Data Generation

Using the item parameters from the large-sample three-parameter item response theory calibration of the operational forms, examinee response data for each test form by test design combination (4 operational test forms; 3 LOFT forms; 3 panels each of 3- Stage MST Design 1 and 2, and 2-Stage MST Design 1 and 2) were simulated using MSTSIM (Jodoin, 2002). For each test form, the identical random sample of 5000 examinees from a normal distribution with mean 0 and standard deviation 1 were used to closely reflect the actual distribution of the examinee population. By using a very large examinee sample size, sampling errors in the statistics of interest could be kept very small and reduce problems in interpreting the main

results. In addition, different response seeds were used with the random generator when simulating examinee response data to the 3-Stage and 2-Stage MST designs to create independent response vectors for subsequent analyses. Subsequently, a replication of every condition was carried out to compute test-retest reliabilities.

For all analyses, maximum likelihood estimates were used for computing ability estimates (see Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). Based on ability estimates at the end of a previous stage, examinees in the lower, middle, and upper third of ability estimates were assigned to the easy, medium, and hard modules respectively for MST designs. This allocation plan ensured a common item exposure rate of 33%. Also, pathways were restricted so that examinees were not able to move from the easy to hard or from hard to easy modules in successive stages.

To evaluate the accuracy of ability estimates from each test design, three correlational analyses were conducted. First, the correlation between the true ability and final ability estimates were computed for each replication of each test design by form. Second, test-retest reliability was computed by correlating the final ability estimates between replications of the first form of each test design. Third, "parallel-forms" [2] reliability was computed by correlating the final ability estimate between each form within test design.

Although accurate ability estimates are generally important in testing, the quality of certification examinations is judged by the level of decision accuracy and decision consistency that is achieved. To investigate the capabilities of each test design to properly classify examinees as qualified or unqualified candidates, analyses were conducted with pass rates of approximately 30%, 40%, and 50%. True and estimated abilities above 0.521, 0.223, and 0.000 were classified as true or observed certifiable candidates and true and estimated abilities below these values

were classified as true or observed non-certifiable candidates. The three pass rates span the range of pass rates that are expected with this credentialing agency. Furthermore, it should provide an important basis for generalizing the results of this study since the number of examinees near a passing score affects the decision consistency and decision accuracy.

To assess the capability of each test design to properly classify examinees as certifiable or non-certifiable, decision accuracy, false-positive and false negative rates, and Kappa coefficient were calculated. Finally, the decision consistency and Kappa coefficient were calculated by comparing the classification decisions for examinees on forms one through three of each test design at each passing score. For both the decision accuracy and consistency analyses, results were reported for replication one only. Similar findings were obtained with the additional forms and replications.

Results

The correlations between the true and estimated ability scores are provided in Table 2. Clearly, all test designs performed comparably to the original four test forms. However, correlations were highest and most consistent across the Real, LOFT, and 3- Stage MST designs, ranging between 0.93 and 0.94. Although still large, the correlations for the 2-stage MST designs were somewhat lower, on the magnitude of 0.91. Additionally, no differences of note were obtained between the MST designs 1 and 2 in either the 2- or 3-stage MSTs.

[Insert Table 2 About Here]

Tables 3, 4, and 5 provide the test-retest and parallel-forms reliabilities for the LOFT, 3- and 2-stage MST designs, respectively. The parallel-form and test-retest reliabilities were similar for the LOFT and 3-Stage MST designs. Parallel-form and test-retest reliabilities for the 2-stage MST designs were somewhat smaller than the other designs. However, this is expected since the

2-stage MST design has 20 fewer items than the other designs and thus the examinee ability estimates computed contain more error. Finally, within the 2- and 3-stage designs, no differences were observed between MST Designs 1 and 2.

[Insert Tables 3, 4, and 5 About Here]

The correlation evidence presented in Tables 2 to 5 suggests that for the most part LOFT and the 3-stage MST designs provided highly consistent results in terms of ability estimation. Though the 2-stage results were not as high as observed with the other designs, the results for all test designs were decidedly similar to those observed with the original test forms. All results were stable across replications within designs, and a high level of consistency was exhibited between the two MST design approaches.

However, for credentialing and licensure testing programs, precision in ability estimation is less important than the accurate classification of examinees into certifiable and non-certifiable groups. Table 6 contains the decision accuracy, false positive and false negative rates, and Kappa for the first form and replication of the Real and LOFT designs for 30%, 40%, and 50% pass rates. The decision accuracy was high and exceeded 90% for all pass rates for both test designs. As expected the false-positive and false negative rates increased slightly from 30% to 50% since classification errors increase with the proportion of examinees near the passing score. Finally, Kappa, a measurement of agreement in classification corrected for chance level agreement, was approximately 0.80 across designs and pass rates indicating a high level of correct classification. Although not reported, rates for other forms and replications were comparable.

[Insert Table 6 About Here]

Decision accuracy, false-positive and false-negative rates, and Kappa are reported for the 3- and 2-stage MST designs in Tables 7 and 8. Again, decision accuracy decreased, and false-

positive and false-negative rates increased as the pass rate increased for all designs. Surprisingly, the 3-stage MST designs had only slightly higher decision accuracy than the LOFT design and lower decision accuracy than the operational form. However, the differences were slight with misclassifications below 10% and correct classification in excess of 90% for the 3-stage MST, LOFT, and operational forms at all three pass rates. The 2-stage MST results fared only slightly worse with misclassifications between 10% and 12% and correct classifications exceeding 88% across pass rates. These results are striking since the 2-stage designs are only 40 items in length. These decision accuracy results are summarized by design and pass rate in Table 9.

[Insert Table 7, 8, and 9 About Here]

Decision consistency results are reported in Tables 10, 11, and 12. The percentage of inconsistently classified examinees was somewhat higher when the 2-stage MST was used regardless of the passing score compared to the LOFT and 3-stage MST designs. Similarly, Kappa was also substantially lower for the 2-stage MST designs than the LOFT and 3-stage MST designs. Finally, as expected, inconsistent classifications increased as the pass rate increased since more examinees have true abilities near the center of the distribution in this study.

[Insert Tables 10, 11, and 12 About Here]

Conclusions

This study was designed to investigate the merits of several MST designs over fixed-form designs for a large-volume credentialing exam. Every effort was made to simulate realistic data for this single examination. The item bank was created from four operational administrations. The operational IRT model was used to calibrate the items for the item bank and to generate the examinee responses. Content specifications were specified as the mean of current operational

exams. A realistic distribution of ability scores was chosen and the passing rates were chosen to span those in recent exam administrations. The MST designs too were realistic in that common 2- and 3-stage designs were implemented. Each module was of fixed length, which is a practical consideration and common policy, and the total number of items for the fixed-form and 3-stage MST designs was 60 items; the same length as the operational forms. A common finding across test designs was that ability estimates and pass-fail classifications were highly accurate. Clearly, item quality and 60 item exams are sufficient to produce exams with excellent psychometric properties. Despite the realism of the simulations, some of the other findings were surprising and potentially very important. At the very least, the findings suggest some promising next steps for future research.

First, the 3-stage MST designs and the LOFTs produced results that were comparable to the current operational forms but certainly were not better. Higher levels of decision accuracy and consistency had been expected. Several possible explanations for this surprising finding can be offered. The MST designs and LOFTs were held to a somewhat higher standard of content matching than the operational exams and this constraint made it difficult for the test assembly software to closely match the intended targets. Figures 4 and 5 show that the test assembly software consistently had difficulty in meeting the target test information functions. Since only three forms were constructed from a pool derived from four forms, it follows that this difficulty was the result of the variation in the content specification variability of the operational forms that was not permitted in the automated test assembly procedure. This had the net effect of creating an item bank that simply was not deep enough in items (recall that 75% of the available items were used in item selection) to allow the test assembly software to do a good job of constructing modules to meet the intended targets. Finally, the bank was not deep enough in quality or

numbers to try, with the same number of test items as the operational forms, to exceed the average operational test information function. Thus, the MST designs implemented here had little opportunity to exceed the fixed-forms or to provide as much adaptation as desired. Figure 6 illustrates this point since the test information functions from form 1 of the Real, LOFT, and each pathway in the 3-stage MST Design 1 do not vary as much was expected. Had higher levels of adaptation, or more aggressive target information functions that could be successfully met been possible, higher decision accuracy and consistency results would have been observed.

[Insert Figure 6 About Here]

Consequently, it seems apparent that a common challenge in computer adaptive tests is applicable in MST designs as well. Testing programs are either going to need to develop more test items to meet content and statistical specifications (than was available in this study), or to simply write many more items in hopes that some of them will be useful in meeting the targets of MST and LOFT designs. MSTs and LOFTs will not be optimal unless there is an item bank rich enough in items to support these test designs.

Second, the two variations of the MST designs produced highly comparable results. This was a surprising finding since it was expected that design 2 would enable construction of modules that better aligned to the target information functions in an automated test assembly environment with a restricted item bank and forced the use of highly discriminating items when ability estimates were most accurate. Nevertheless, the idea of moving the more discriminating items to the later stages remains a good one (Chang, Qian, & Ying, 2001). It is possible that an advantage was not seen in this study because a large enough difference was not created between MST Design 1 and Design 2. The difference in allocation of information across three stages to be 33%, 33%, 33% versus 25.0%, 37.5%, 37.5%, was simply not large enough for the effect to

materialize in the psychometric quality of the exam. In a future study, another MST design that might be more like 15.0%, 42.5%, and 42.5% would be worth evaluating to provide additional results.

Third, the psychometric results from the 2-stage test designs were only a bit lower (on the average about 2%) than the results from the 3-stage MST designs and LOFTs. Somewhat bigger differences had been expected. Thus, assuming content and validation constraints could be met, a test that is 40 items in length with any of the designs considered in the study might be more than adequate to meet the needs of the credentialing agency. An exam that is 50% shorter (40 items compared to 60 items) could reduce exam costs for candidates and the credentialing agency, reduce testing time, lower item exposure levels, possibly require smaller item banks, and/or might make the current banks better able to meet the needs of the shorter examinations. Of course, this issue of balancing content coverage and such operational concerns involves careful weighting of the specific needs of the testing agency, characteristics of the examinee population, and the inferences to be drawn from the examination.

One follow-up study that is planned is to repeat the study using a simulated item bank that better meets the content and statistical specifications of the exam. With this ideal item bank, the full advantages of MSTs and LOFTs can be determined. If the results are similar, then under the current circumstances, these new designs may have little to offer the credentialing agency. If the psychometric results from the MSTs and LOFTs are substantially better than the operational forms, then the obvious direction for the credentialing agency is to write many more test items, and preferably test items that meet the needs of the ideal item bank for the exam (van der Linden, Veldkamp, & Reese, 2000; Veldkamp & van der Linden, 2000).

References

Chang, H. H., Qian, J., & Ying, Z. (2001). A-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement, 25,* 333-341.

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

Jodoin, M. J. (2002). *MSTSIM* [Computer software]. Amherst, MA: University of Massachusetts, School of Education.

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

Luecht, R. M. (1998). *CASTISEL* [Computer software]. Philadelphia, PA: National Board of Medical Examiners.

Luecht, R. M., & Nungester, R. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35,* 239-249.

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

Patsula, L. N., & Hambleton, R. K. (1999, April). *A comparative study of ability estimates obtained from computer-adaptive and multi-stage testing.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Canada.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Boston, MA: Kluwer Academic Publishers.

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

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

Veldkamp, B. P., & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 149-162). Boston, MA: Kluwer Academic Publishers.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement, 35,* 328-347.

Wainer, H., et al. (Eds.). (2000). *Computerized adaptive testing: A primer* (2nd. Ed.). Mahwah, NJ: Lawrence Erlbaum Publishers.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Publishers.

Xing, D. (2000). *Impact of several computer-based testing variables on the psychometric properties of credentialing examinations.* Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Xing, D., & Hambleton, R. K. (2002, April). *Impact of item quality and item bank size on the psychometric properties of computer-based credentialing exams.* Paper presented at the meeting of NCME, New Orleans.

Author Note

This paper was presented at the 2002 meeting of the National Council on Measurement in Education, New Orleans, LA.

Please address all correspondence to Michael G. Jodoin, Room 152 Hills South, University of Massachusetts, Amherst, MA 01003. Email: mjodoin@psych.umass.edu

Footnotes

[1] The inability of CASTISEL to meet the test information target generated from the mean of the operational test forms while using only three quarters of the items is likely due to variability of the content specifications between operational forms. This had the effective result of placing tight content specification for the given item bank on this simulation study.

[2] Parallel forms is used loosely in this context for the MST designs. The parallel forms could potentially consist of a large number of the same items depending on the pathway taken by an examinee through each MST panel.

Table 1

Correlations Between True and Estimated Ability by Stage and by Replication for 3- Stage MSTs

Design	Form	Replication 1			Replication 2		
		Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3
1	1	.840	.914	.935	.840	.915	.935
	2	.849	.916	.936	.850	.915	.936
	3	.844	.915	.936	.849	.915	.936
2	1	.814	.913	.938	.814	.916	.938
	2	.809	.910	.937	.810	.911	.937
	3	.813	.911	.935	.817	.912	.937

Table 2

Correlations Between True and Final Ability Estimates

Type	Design	Form	Replication	
			1	2
Real		1	.937	.938
		2	.928	.932
		3	.931	.930
		4	.934	.933
LOFT		1	.941	.938
		2	.937	.936
		3	.927	.922
3-Stage MST	1	1	.935	.935
		2	.936	.936
		3	.936	.936
	2	1	.938	.938
		2	.937	.937
		3	.935	.937
2-Stage MST	1	1	.915	.914
		2	.916	.918
		3	.916	.913
	2	1	.917	.912
		2	.916	.910
		3	.913	.911

Table 3

Test-Retest and Parallel Form[1] Reliability Estimates for LOFT Forms

Form	LOFT Form		
	1	2	3
1	.878	.881	.869
2		.878	.867
3			.856

[1] Parallel forms reliability for replication 1

Table 4

Test-Retest and Parallel Form[1] Reliability Estimates for 3-Stage MSTs

Form	Design 1			Form	Design 2		
	1	2	3		1	2	3
1	.875	.872	.876	1	.881	.879	.877
2		.873	.882	2		.876	.875
3			.877	3			.878

[1] Parallel forms reliability for replication 1

Table 5

Test-Retest and Parallel Form Reliability[1] Estimates for 2-Stage MSTs

Form	Design 1			Form	Design 2		
	1	2	3		1	2	3
1	.838	.835	.837	1	.838	.834	.831
2		.842	.842	2		.837	.829
3			.830	3			.830

[1] Parallel forms reliability for replication 1

Table 6

Decision Accuracy for Real and LOFT Test Designs by Pass Rate[1]

Type		30% Pass Rate			40% Pass Rate			50% Pass Rate		
		$Fail_E$	$Pass_E$		$Fail_E$	$Pass_E$		$Fail_E$	$Pass_E$	
Real	$Fail_T$	65.8%	4.6%	70.4%	55.8%	5.2%	61.0%	45.7%	5.3%	51.0%
	$Pass_T$	3.2%	26.4%	29.6%	3.4%	35.6%	39.0%	4.6%	44.4%	49.0%
		69.0%	31.0%	K=.813	59.2%	40.8%	K=.822	50.3%	49.7%	K=.801
LOFT	$Fail_T$	65.4%	5.0%	70.4%	55.8%	5.2%	61.0%	46.1%	4.9%	51.0%
	$Pass_T$	4.4%	25.2%	29.6%	4.4%	34.6%	39.0%	4.7%	44.3%	49.0%
		69.8%	30.2%	K=.775	60.2%	39.8%	K=.800	50.8%	49.2%	K=.807

[1] Statistics are based on the first form and first replication.
[E] Estimate
[T] True

Table 7

Decision Accuracy for 3-Stage MST Designs by Pass Rate[1]

Design	Decision	30% Pass Rate			40% Pass Rate			50% Pass Rate		
		$Fail_E$	$Pass_E$		$Fail_E$	$Pass_E$		$Fail_E$	$Pass_E$	
1	$Fail_T$	65.8%	4.6%	70.4%	56.0%	5.0%	61.0%	45.8%	5.2%	51.0%
	$Pass_T$	3.6%	26.0%	29.6%	4.2%	34.8%	39.0%	4.6%	44.4%	49.0%
		69.4%	30.6%	K=.803	60.2%	39.8%	K=.809	50.4%	49.6%	K=.806
2	$Fail_T$	66.1%	4.3%	70.4%	56.0%	5.0%	61.0%	45.9%	5.1%	51.0%
	$Pass_T$	3.8%	25.8%	29.6%	4.0%	35.0%	39.0%	4.5%	44.5%	49.0%
		69.9%	30.1%	K=.807	60.0%	40.0%	K=.811	50.4%	49.6%	K=.809

[1] Statistics are based on the first form and first replication
[E] Estimate
[T] True

Table 8

Decision Accuracy for 2-Stage MST Designs by Pass Rate[1]

Design	Decision	30% Pass Rate			40% Pass Rate			50% Pass Rate		
		$Fail_E$	$Pass_E$		$Fail_E$	$Pass_E$		$Fail_E$	$Pass_E$	
1	$Fail_T$	65.0%	5.4%	70.4%	55.4%	5.6%	61.0%	45.5%	5.5%	51.0%
	$Pass_T$	4.2%	25.4%	29.6%	5.1%	33.9%	39.0%	6.2%	42.8%	49.0%
		69.2%	30.8%	K=.773	60.5%	39.5%	K=.775	51.7%	48.3%	K=.766
2	$Fail_T$	65.0%	5.4%	70.4%	55.1%	5.9%	61.0%	45.2%	5.8%	51.0%
	$Pass_T$	4.8%	24.8%	29.6%	5.2%	33.8%	39.0%	6.0%	43.0%	49.0%
		69.8%	30.2%	K=.757	60.3%	39.7%	K=.768	51.2%	48.8%	K=.761

[1] Statistics are based on the first form and first replication
[E] Estimate
[T] True

Table 9

Decision Accuracy By Pass Rate by Design

Type	Pass Rate			Number of
	30%	40%	50%	Items
Real	92.2	91.4	90.1	60
LOFT	90.6	90.4	90.4	60
3 Stage MST Design 1	91.8	90.8	90.2	60
3 Stage MST Design 2	91.9	91.0	90.4	60
2 Stage MST Design 1	90.4	89.3	88.3	40
2 Stage MST Design 2	89.8	88.9	88.2	40

Table 10

Decision Consistency for Real Data and LOFT Design by Pass Rate[1]

Type	Decision	30% Pass Rate			40% Pass Rate			50% Pass Rate		
		$Fail_{F2}$	$Pass_{F2}$		$Fail_{F2}$	$Pass_{F2}$		$Fail_{F2}$	$Pass_{F2}$	
Real	$Fail_{F1}$	63.6%	5.4%	69.0%	52.9%	6.3%	59.2%	43.6%	56.7%	50.3%
	$Pass_{F1}$	6.1%	24.9%	31.0%	6.9%	33.9%	40.8%	7.1%	42.6%	49.7%
		69.7%	30.3%	K=.729	59.8%	40.2%	K=.726	50.7%	49.3%	K=.725
LOFT	$Fail_{F1}$	64.4%	5.5%	69.9%	53.7%	6.3%	60.0%	43.7%	6.7%	50.4%
	$Pass_{F1}$	5.8%	24.3%	30.1%	6.8%	33.2%	40.0%	6.9%	42.7%	49.6%
		70.2%	29.8%	K=.731	60.5%	39.5%	K=.728	50.6%	49.4%	K=.728

[1] Statistics are based on the first and second form, first replication
[F1] Form 1
[F2] Form 2

Table 11

Decision Consistency for 3-Stage MST Design by Pass Rate[1]

Design	Decision	30% Pass Rate			40% Pass Rate			50% Pass Rate		
		$Fail_{F2}$	$Pass_{F2}$		$Fail_{F2}$	$Pass_{F2}$		$Fail_{F2}$	$Pass_{F2}$	
1	$Fail_{F1}$	63.6%	5.8%	69.4%	53.4%	6.8%	60.2%	42.7%	9.0%	51.7%
	$Pass_{F1}$	6.0%	24.1%	30.6%	6.8%	33.0%	39.8%	7.5%	40.8%	49.3%
		70.1%	29.9%	K=.706	60.2%	39.8%	K=.715	50.2%	49.8%	K=.698
2	$Fail_{F1}$	64.4%	5.5%	69.9%	53.7%	6.3%	60.0%	43.7%	6.7%	50.4%
	$Pass_{F1}$	5.8%	24.3%	30.1%	6.8%	33.2%	40.0%	6.9%	42.7%	49.6%
		70.2%	29.8%	K=.731	60.5%	39.5%	K=.728	50.6%	49.4%	K=.728

[1] Statistics are based on the first and second form, first replication

[F1] Form 1

[F2] Form 2

Table 12

Decision Consistency for 2-Stage MST Design by Pass Rate[1]

Design	Decision	30% Pass Rate			40% Pass Rate			50% Pass Rate		
		$Fail_{F2}$	$Pass_{F2}$		$Fail_{F2}$	$Pass_{F2}$		$Fail_{F2}$	$Pass_{F2}$	
1	$Fail_{F1}$	62.3%	6.9%	69.2%	52.4%	8.1%	60.5%	42.7%	9.0%	51.7%
	$Pass_{F1}$	7.1%	23.7%	30.8%	7.3%	32.2%	39.5%	7.6%	40.7%	49.3%
		69.4%	30.6%	K=.670	59.7%	40.3%	K=.680	50.3%	49.7%	K=.698
2	$Fail_{F1}$	61.8%	8.0%	69.8%	51.5%	8.9%	60.4%	42.0%	9.2%	51.2%
	$Pass_{F1}$	7.2%	23.0%	30.2%	7.8%	31.8%	39.6%	8.3%	40.5%	48.8%
		69.0%	31.0%	K=.643	59.3%	40.7%	K=.653	50.3%	49.7%	K=.650

[1] Statistics are based on the first and second form, first replication

[F1] Form 1

[F2] Form 2

Figure Caption

Figure 1. Operational form information functions

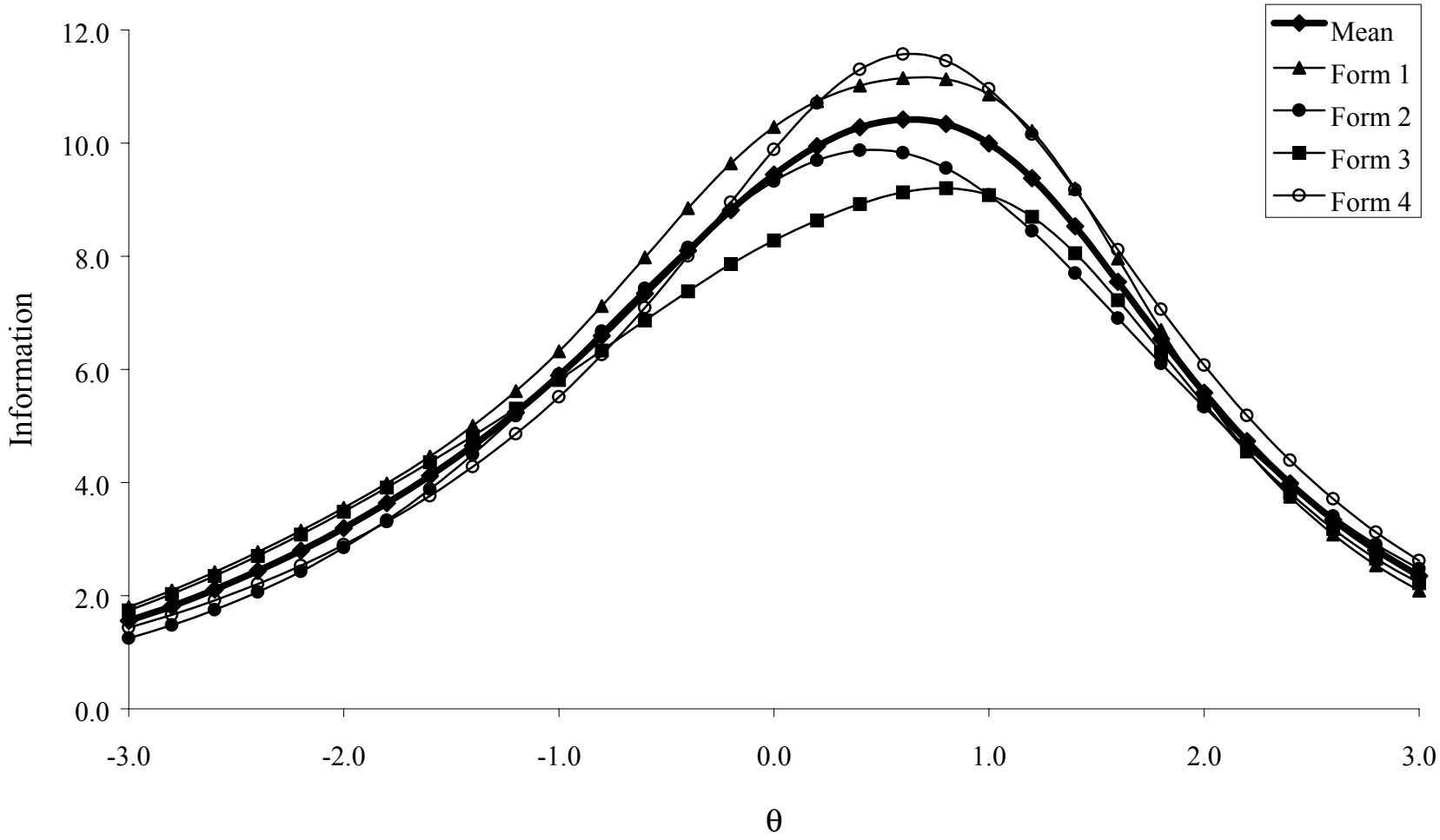
Figure 2. LOFT information functions

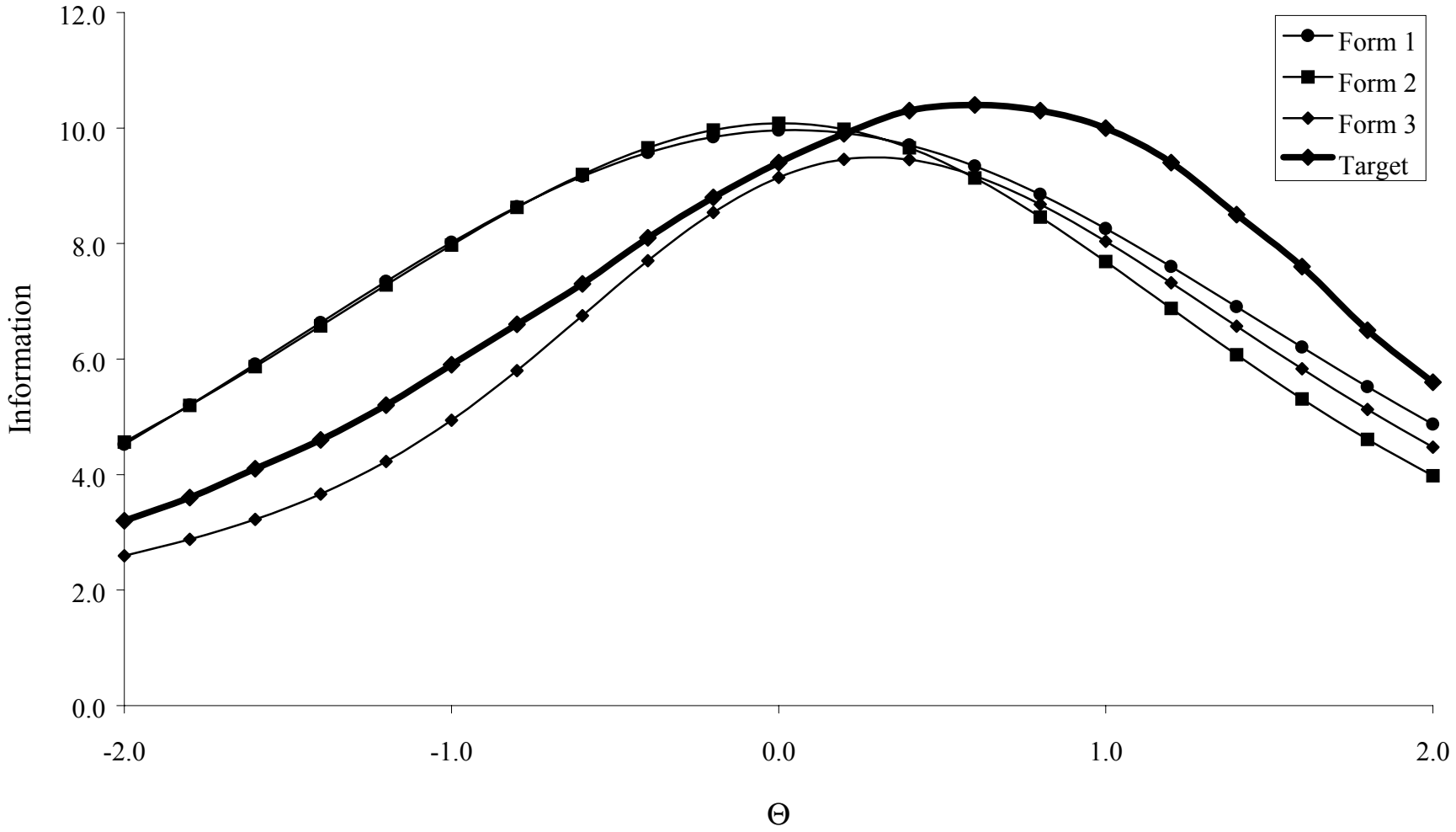
Figure 3. A 3-Stage MST design

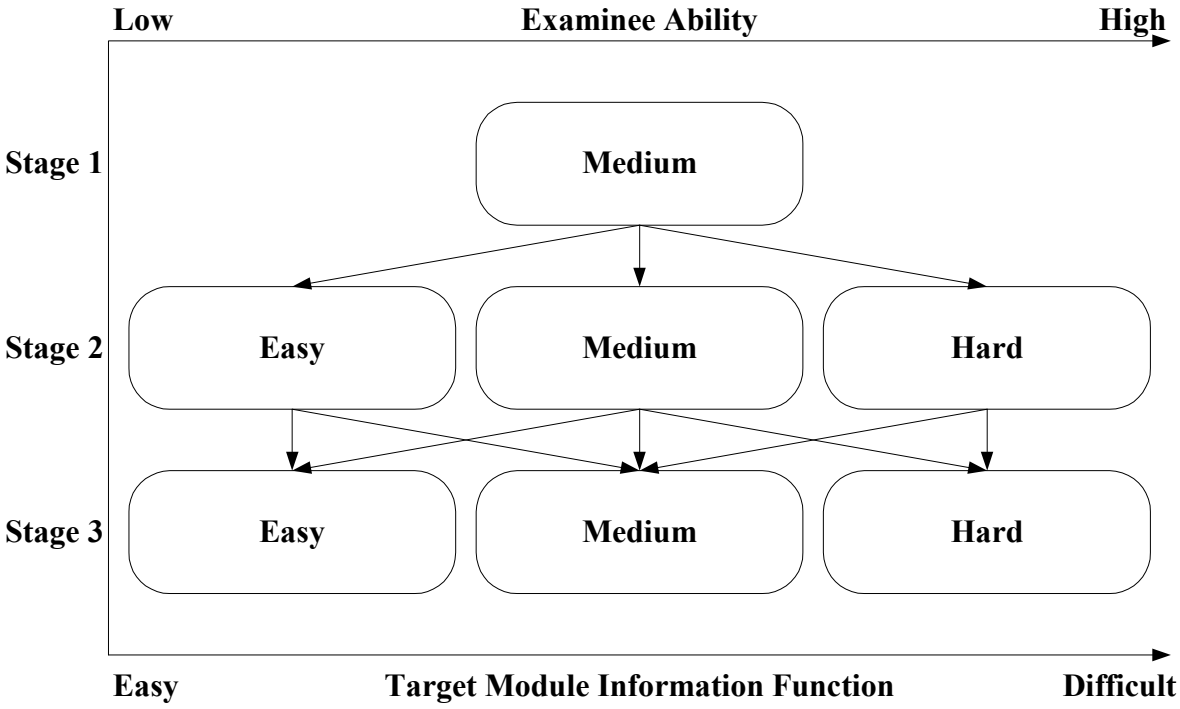
Figure 4. Module information functions, 3-Stage MST Design 1, Form 1

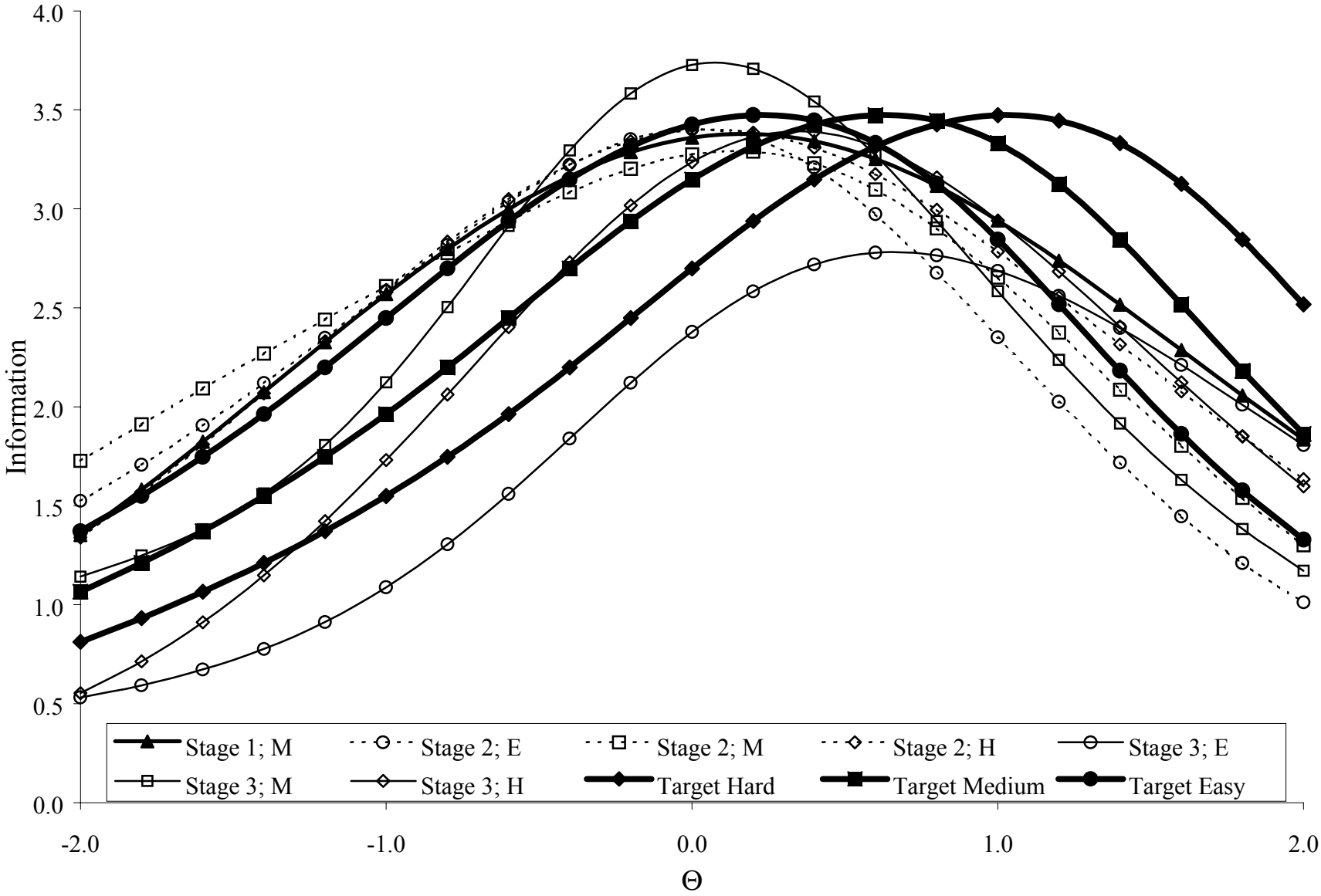
Figure 5. Module information functions, 3-Stage MST Design 2, Form 1

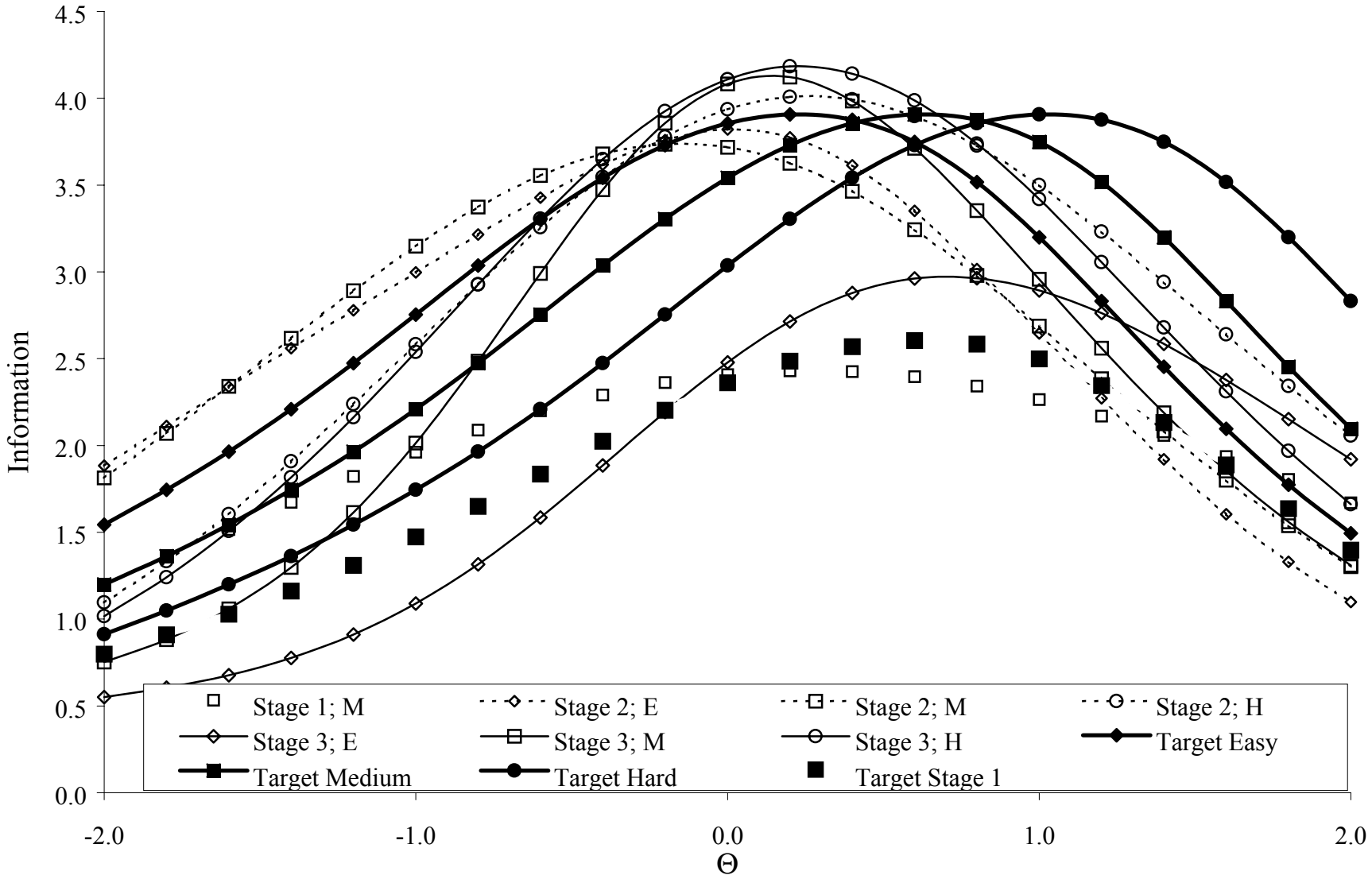
Figure 6. Information functions for Real, LOFT, and 3-Stage MST Design 1 (7 Possible Paths), Form 1

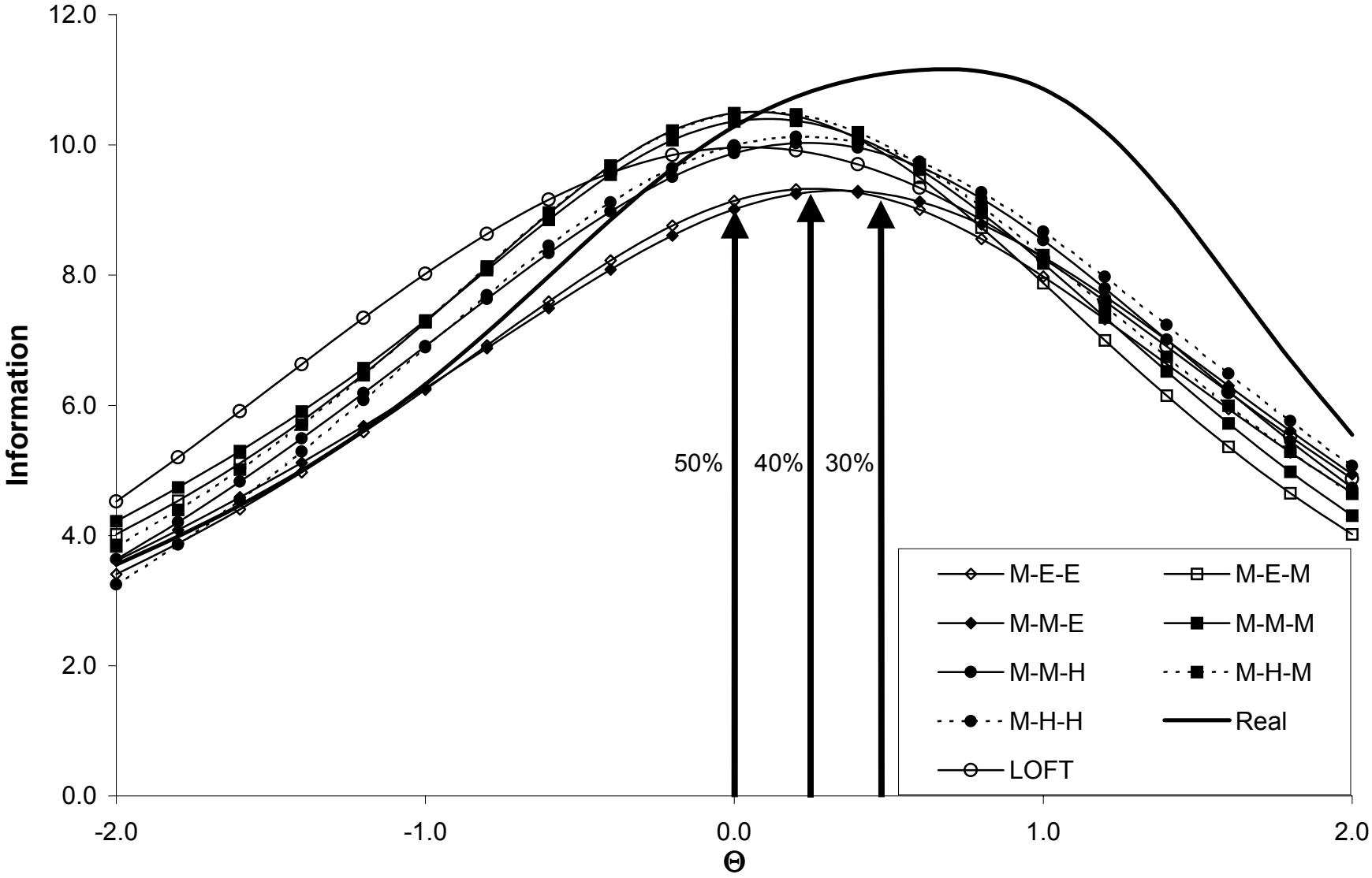

Mean
Form 1
Form 2
Form 3
Form 4
Information
12.0
10.0
8.0
6.0
4.0
2.0
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
θ


12.0
10.0
8.0
6.0
4.0
2.0
0.0
Information
-2.0
-1.0
0.0
1.0
2.0
Θ
Form 1
Form 2
Form 3
Target


Low
Examinee Ability
High
Stage 1
Medium
Stage 2
Easy
Medium
Hard
Stage 3
Easy
Medium
Hard
Easy
Target Module Information Function
Difficult


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Information
-2.0
-1.0
0.0
1.0
2.0
Θ
Stage 1; M
Stage 2; E
Stage 2; M
Stage 2; H
Stage 3; E
Stage 3; M
Stage 3; H
Target Hard
Target Medium
Target Easy


Information
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-2.0
-1.0
0.0
1.0
2.0
Θ
Stage 1; M
Stage 2; E
Stage 2; M
Stage 2; H
Stage 3; E
Stage 3; M
Stage 3; H
Target Easy
Target Medium
Target Hard
Target Stage 1


12.0
10.0
8.0
6.0
4.0
2.0
0.0
Information
-2.0
-1.0
0.0
1.0
2.0
Θ
50%
40%
30%
M-E-E
M-E-M
M-M-E
M-M-M
M-M-H
M-H-M
M-H-H
Real
LOFT