Effects of Selected Multi-Stage Test Design Alternatives on Credentialing Examination Outcomes [1,2]

April L. Zenisky and Ronald K. Hambleton

University of Massachusetts Amherst

March 29, 2004

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

[2] The authors would like to thank the American Institute of Certified Public Accountants for both the financial and technical support that enabled us to complete the study

Abstract

One of the main concerns in designing credentialing examinations is the psychometric quality of the pass-fail decisions. The computerized multi-stage test design has been shown in previous research to provide accurate results relative to other computerized test designs in addition to possessing certain operational advantages favored by candidates (e.g., item review). The purpose of the current research was to investigate how selected design variables in multi-stage testing interact with one another to impact on the psychometric quality of pass-fail decisions and ability estimation. The four design variables studied in a computer simulation were (1) module arrangement (4 designs), (2) amount of overall test information (4 levels), (3) distribution of information over stages (2 variations), and (4) strategies for between-stage routing (4 levels), for a total of 128 conditions. Very large sample sizes were used to essentially eliminate the influence of sampling error on the findings. Many of the results were as expected—with small to negligible effects, but two of the findings seemed especially important because of their consequences for practice: (1) with limited amounts of overall test information, it appears best to distribute available information unevenly with more of it going to stage one to improve the accuracy of classifications, and (2) there appeared to be little advantage of exceeding test information much above 10 since the gains in psychometric quality were very small—this finding has implications for the selection of test lengths and/or more effective uses of an item bank.

Background and Purpose

The interest expressed by many testing programs in recent years for an adaptive test that still permits test developers to maintain some level of control over operational 'forms' has led psychometricians to investigate in some depth the properties of the test design known as multi-stage testing (MST). At once both adaptive (between stages) and fixed (within stages), the stage-based approach to measurement has been shown to provide results for both ability estimation and decision outcomes that are comparable to CAT and improved over linear computer-based tests (Rotou, Patsula, Steffen, & Rizavi, 2003; Jodoin, Zenisky, & Hambleton, 2002; Patsula, 1999; Reese & Schnipke, 1999; Reese, Schnipke, & Luebke, 1999; Schnipke & Reese, 1999; Kim & Plake, 1993).

As a topic for psychometric research, MST has been the increasing focus of research efforts in a variety of contexts, especially credentialing exams (e.g., Luecht & Nungester, 1998; Luecht, Brumfield, & Breithaupt, 2002; Luecht, 2000). This is the case because many credentialing exams are now being administered at a computer, and this creates possibilities for new test designs. MST is a particularly agile test design to investigate in that there are numerous design variables that must be considered in the development and operational use of such a test. Lord's (1970, 1971, 1980) work with adaptive tests (and specifically the two-stage variety) yielded a number of such considerations generalizable to a test of *n* stages, including the total number of items on the test, the number of items in initial and each *n*-stage module, the difficulty of the initial module, the number (and difficulty) of alternative modules in each *n*-stage, the cut-points for routing examinees to modules, and the choice of method for scoring stages and each *n*-stage test. To this list can be added several additional variables that have emerged including the number of stages, the ability distribution of the candidate population, the extent of target information overlap for modules within stages, whether random module selection (at appropriate difficulty level) or panel-based administration is used, whether content-balancing is done at the module or total test level, choice of

method for automated test assembly, the size and quality of the item bank, how test information is distributed across stages, placement of cut-scores for pass-fail decisions, the issue of item review by candidates, and control of item exposure levels. The number of possible studies is huge because of the number of variables and combinations of variables that can be investigated.

While Lord (1971) suggested that it was not possible to identify truly statistical optimal designs for each and every operational testing context, it seems entirely reasonable to explore how combinations of these variables provide high-quality results as needed in the context of a particular testing purpose (e.g., credentialing or achievement testing) and the kinds of interpretations to be made based on test scores. The purpose of this study was to consider MST in light of specific design variables and comparatively evaluate the results obtained relative to the measurement outcomes valued by credentialing agencies – decision accuracy, decision consistency, and precision of ability estimates.

The simulation study was designed as a study of selected variables for implementing MST, and was intended to advance previous research into the psychometric properties of multi-stage testing (MST) with credentialing exams. The primary design variables of interest were 1) test structure, 2) the amount of test information, 3) the distribution of test information across stages, and 4) between-stage routing strategies. These four variables were considered in the context of a specific cut-off score and pass rate (30%), chosen to model a highly selective test setting. In total, 128 conditions were evaluated (4 levels of total test information by 2 levels of the distribution of test information across stages by 4 levels of test designs by 4 levels of routing strategies).

Method

In order to focus on the interactions of the four design variables described in the previous section and in more detail below, several aspects of the multi-stage test were held fixed. All

modules in all stages of the MSTs were 20 items long, and all simulated administrations were composed of three stages (one module per stage per examinee) for a total test length of 60 items for all examinees. Each examinee received a module of moderate difficulty in stage one, and the average difficulty difference between modules at stages two and three was a half a standard deviation. The test information function (TIF) that was the baseline for this study was obtained from operational paper-and-pencil forms of a large-scale credentialing examination. Each of the 128 conditions in this study involved 9,000 examinees and was replicated twice. The sample size was sufficient to produce highly stable statistics (this was proven during pilot study runs), and the replication was necessary to compile decision consistency estimates. The examinee distribution was simulated as N(0,1)

Design variables. Generally speaking, most multi-stage tests begin with a module of moderate difficulty. In the second and subsequent stages, however, the possibilities are more varied, as the number of sets of items varying by average difficulty, known in MST-speak as modules (and sometimes called "testlets"), could range from two or three to many more (depending on the depth and breath of the item bank to support finer differentiation). Practically speaking, the typical approach to MST cited in the literature involves the use of two or three modules per stage in a two or three stage MST. Considered in this study were the following four arrangements of modules in a three-stage MST: the 1-2-2, the 1-2-3, the 1-3-2, and the 1-3-3 (Figure 1).

[Insert Figure 1 about here]

Next, based on the previous studies of MST (Jodoin, Zenisky, & Hambleton, 2002; Jodoin, 2002; Xing & Hambleton, 2004; Zenisky, 2002), a critical need in the MST literature concerned further investigation of the impact of varying amounts of test information. In previous research, the level of test information specified was in most cases set to be the test information function used with the linear version of the test, and this resulted in MSTs producing high rates of decision

accuracy and consistency. A reasonable direction for further research concerning this aspect of MST design would be to vary the amount of test information in relation to the target or baseline test information function. The view was that these variations would provide valuable results to guide future MST designs. How much practical benefit does increasing the target or baseline TIF by 50% bring, and likewise, what does proportional reductions of the target TIF mean for ability estimation and decision classification? This is a variable of interest to test developers working with the MST design because with lower levels of target test information, the test assembly mechanism has the potential for greater flexibility in putting together test forms, and item exposure levels can be reduced because more items now from an item bank can be used. Since the algorithm can draw on items with more varied discrimination and difficulty values in order to meet its target TIF, it will translate into greater use of the item bank in terms of both breadth and depth, thereby reducing item exposure. In this way, if the ability estimation and decision classification results are of somewhat comparable quality to the results obtained with current levels of information, test security can be enhanced while still providing high-quality assessment results in terms of making pass-fail decisions. The four level of test information implemented in this study were 1) the full (operational) information level, 2) a 50% increase over full information, 3) a 25% decrease from full information, and 4) a 50% decrease from full information.

The third variable for this study involved research into how test information should be distributed across stages. In previous studies, such target information has generally been split equally among the three stages with satisfactory results. But previous research (i.e., Zenisky, 2002; Jodoin, Zenisky, & Hambleton, 2002) has also indicated that alternative distributions such as 1/2-1/4-1/4 may provide improved decision accuracy and consistency. By obtaining better ability estimates after the first stage, it may be possible to make routing to second- and third stage tests more efficient, thereby improving score precision at the conclusion of testing for many candidates.

It seems clear that further research into this aspect of MST can substantially clarify how this variable impacts the quality of measurement obtained. In summary, two different ways of dividing test information across stages (1/3-1/3-1/3 and 1/2-1/4-1/4) were studied.

The last goal of this study was to evaluate routing strategies in the context of several MST designs. Several different methods for routing can be found in the psychometric literature, although there is little in the way of empirical comparisons of the methods. This is a variable of importance because the choice of method used for routing examinees between stages is fundamental to the process of adapting a multi-stage test to candidate ability, and empirical determination of whether certain strategies provide better results than others can help to ensure that a test design is implemented efficiently and effectively. Four routing methods were evaluated in this study: defined population intervals, the proximity method, a number-correct strategy, and a random assignment method. The methodology associated with each of these four routing strategies is described and the use of the MSTSIM5 (Jodoin, 2003) software is also explained below.

Item bank. The item bank used in this study was created to reflect the conditions of the items used in an operational form of a nationally administered credentialing exam. Based on previous research (Jodoin, Zenisky, & Hambleton, 2002; Zenisky, 2002), a bank size of 2,500-3,000 items was identified as providing sufficient breadth and depth for supporting automated test assembly (ATA) to insure that the variables of interest could actually be studied without confound from not meeting target information functions. Though this bank size is considerable, the target information functions being specified were typical of those observed in practice for credentialing exams of approximately the same test length.

The IRT item parameter estimates for six archival forms of the exam we worked with (a total of 358 items, all calibrated with the three-parameter logistic model) were used for realistically modeling the item bank for this simulation study, and the means, SDs, and the correlational

structure among these estimates needed to be maintained in any generated bank. To build this bank, a statistical technique to 'clone' the 358 current items was employed. This was easily accomplished by randomly sampling item difficulties, discriminations, and pseudo-guessing parameters from the expected sampling distributions of item parameter estimates obtained from the known item parameter estimates and their standard errors. For each of the 358 items in the original bank, seven clones were generated creating a bank totaling 3,222 items (358 original items and the 2,864 'cloned' items). Descriptive statistics for the original set of items and the generated items were identical across the original and generated item bank, with the means (and standard deviations) for the *a*-, *b*-, and *c*-parameters constant at 0.62 (0.25), -0.12 (1.11), and 0.00 (0.30). In the original set of 358 items, the correlation between the *a*- and *b*-parameter estimates was 0.36, between the *a*- and *c*-parameters it was 0.35, and between the *b*- and *c*-parameters the correlation was 0.31; among the items in the generated bank, the correlations were 0.36, 0.34, and 0.30, respectively. Clearly the generated bank was a reasonable statistical reflection of the current, operational bank, only larger.

Automated test assembly. The computer program CASTISEL (Luecht, 1998) was used for ATA. CASTISEL is an automated test assembly program that takes statistical and other content constraints into account and automates the process of formulating modules and panels for MST using the available item bank. With CASTISEL, MST modules were simulated by specifying target information functions, the number of stages to be included in a form or module, the number of modules per stage, the number of items per stage, the total test length, and the primary content specifications for content balancing being implemented. CASTISEL created such forms or modules by using the normalized weighted absolute deviations heuristic (NWADH) described by Luecht (2000) to optimize item selection for forms or modules given the target TIFs and other form or module-level considerations.

Computer simulation method. For each condition simulated, the base target TIF was re-centered at .521 to maximize information for examinees at the cut-score (to reflect the operational conditions of high-stakes credentialing agencies where 30% of examinees would pass, given the N(0,1) distribution simulated here. If the condition called for 50% increase in total test information, 50% information was added to this base target information function; likewise, if a 25% or 50% reduction was needed, the base target TIF was reduced by the appropriate amount. To specify the information function for each stage of a three-stage MST, this overall test information was either divided into 1) three equal module information functions (the equal information condition) or 2) three module information functions where one contained of the available information and the other two retained one-quarter of the information each (the 1/2-1/4-1/4 condition). Differences between module difficulties at stages two and three were fixed at one-half of a standard deviation. (A schematic example of this as implemented in the 1-3-3 design is provided in Figure 2.) With a medium-difficulty module centered at .521, the module information function for the easier module would be centered at .021 and the function for the harder module would be centered at 1.021.

[Insert Figure 2 about here]

Content balancing of modules was specified in the CASTISEL input files to ensure proportional representation of the three primary content dimensions in each of the modules created, regardless of module difficulty, CASTISEL was used to automatically assemble the modules for the MST. After completing use of CASTISEL, the sets of modules produced represented 32 unique test conditions (created by implementing three of the four design variables under consideration: 4 levels of total test information by 2 levels of the distribution of test information across stages by 4 levels of test designs).

The next step was the specification of input files for the MSTSIM5 program in order to implement the fourth and final design variable under consideration: routing strategies during

administration of the simulated MST. In the MSTSIM5 input files, denoted were the examinee and response seeds, the number of examinees in the sample (9000), the distribution of the examinee population (~N(0,1)), the number of panels (3), the number of stages (3), the number of modules per stage (which varied depending on the design condition), the number of items per module (20), and the specifications for routing examinees through modules.

As explained previously, four methods for routing examinees between stages were identified and implemented in this study:

- In the defined population interval (DPI) approach, the idea is to divide the examinees equally among the modules by ordering all examinees' IRT ability estimates after completion of one stage and before moving on to the next. The values of -0.43 and 0.43 on the ability scale represent the two cut-points for which a normally-distributed population of examinees would be divided into thirds for a stage consisting of three modules. If the examinee population needed to be halved (with two modules varying by difficulty within a stage), a cut-score of 0.0 was specified.
- For the proximity approach, the means of modules in successive stages were computed. After the first and second stages, the examinee's estimated ability (as computed by MSTSIM5 using maximum likelihood estimation) was compared to the average module difficulties at the next stage, and then the module that would provide the most information at the next stage would be chosen for administration to the candidate – this amounted to choosing the module with the average difficulty level that was closest to the candidate's ability estimate.
- In the number-correct strategy, for routing from stages 1 to 2, the test characteristic curve of the routing module was used to find the expected number-correct score or scores that would lead to assigning roughly equal number of candidates to modules at the second

stage. So, with two modules at stage 2, the expected number correct score of a candidate with ability = 0.0 was determined, and used to divide candidates into the two modules at stage two. The more difficult module was assigned to candidates who met or exceeded the number-correct cut score (which would be about 50% of the candidates). With three modules at stage two, the expected number correct scores of candidates with ability levels at -.43 and +.43 were determined (these ability levels divide a normal ability distribution into thirds), and these scores were used to divide the candidates into modules at stage 2 based on their number-correct scores at stage 1. Those candidates meeting or exceeding the highest cut-score were assigned to the most difficult of the modules at stage 2. Those candidates with number correct scores below the lowest cut score were assigned the easiest module at stage 2. Everyone else was assigned to the middle difficult module. The result was approximately 33% of the candidates being administered each module at stage 2. Routing candidates after stage 2 with this method to insure roughly equal numbers of candidates at each stage 3 module was a bit more complicated. If two modules were available at stage 3, for each possible route involving stages 1 and 2, the expected score of the middle ability candidate (ability = 0.0) was determined. This number-correct score was then applied to candidates who took the particular route and those candidates scoring equal to or higher than the cut-score were assigned to the more difficult module at stage 3. Candidates scoring lower than the cut-score were assigned to the easier module at stage 3. A similar process was used if three modules were available at stage 3. For each possible route through stages 1 and 2, the expected number correct score of candidates with abilities at -.43 and +.43 was determined. These cut-scores were then used to assign candidates to the three modules at stage 3. Those candidates achieving or exceeding the highest cut-score were assigned to the most difficult module at stage 3. Thos falling below the lowest cut-score were assigned to

the easiest module at stage 3. Remaining candidates were assigned to the middle difficult module at stage 3. The goals of the number-correct strategy were (1) to avoid calculating ability estimates after each stage (obtaining number correct scores is more for candidates is more convenient in practice) and (2) to assign roughly equal numbers of candidates to each module in the MST. This was accomplished with the strategy just described.

- For the random method, a random process was used for assigning candidates to modules such that equal numbers of candidates were assigned to each module—no consideration was given to candidate performance in the assignments of modules. These results served as a baseline for judging the "value-added" of MST designs that capitalized on ability estimates after each stage for optimal assignment of modules to candidates.

MSTSIM5 was run twice for each condition to provide two replications, which allowed for decision consistency analyses to be carried out. The only difference in the two replications was in the response seed specified in the MSTSIM5 input file for initialing candidate responses to the test items.

The results from the MST simulations were then analyzed with respect to several criteria of interest. The first and second outcomes of interest in this study were the levels of decision accuracy (DA) and decision consistency (DC) observed with these conditions at different pass rates. Decision accuracy is the extent to which the decisions made using candidates' estimated abilities are consistent with decisions made based on true abilities (which are known in a simulation study), and decision consistency provides information about the reliability of classification decisions for candidates. To compute DA, the true and estimated classification of examinees within conditions were compared to produced the percent of examinees who were appropriately classified as well as the proportion of Type I and Type II errors. To compute DC across examinees in each condition, two replications of every condition were completed and the classification results across the two

administrations were compared to identify the proportions of consistent and inconsistent classifications. The passing score for this study was set at 0.521 on the ability scale and this was associated with a 30% pass rate. [Other pass rates were investigated and results are reported in Zenisky (2004) but since the findings were similar, results for only one passing score will be reported here.]

Thirdly, the quality of ability estimation after Stage 3 for each combination of conditions was evaluated with correlations between true and estimated abilities and an analysis of root mean squared errors (RMSE).

The last set of results presented is an analysis of routing paths for the 1-2-2 design. Reported is the frequency of examinees being routed to each possible path for each condition in this study. The purpose in including these results is to provide insight into the nature of the routing decision being made for candidates and also the levels of exposure that might be seen with each of the routing strategies implemented here. The choice in focusing on the 1-2-2 design here is to discuss results related to one particular MST application (results for the other designs can be found in Zenisky, 2004).

Results

The results are organized below with respect to each of the four outcomes of interest (decision accuracy, decision consistency, ability estimation, and routing path frequency).

Decision accuracy and decision consistency. Tables 1 to 4 provide DA and DC results broken out by the four MST designs. For DA the results in Tables 1 to 4 are reported in terms of the percent of examinees misclassified in each condition, and for DC the results given are the percent of examinees inconsistently classified.

[Insert Tables 1, 2, 3, and 4 about here]

Generally speaking, the DA and DC results followed expectations. First, the DA results, on average, were high across conditions. This finding was not unexpected but the results probably overestimate results that might be expected in practice. In practice, of course, the three-parameter logistic model is not going to completely fit the data.

As the amount of test information decreased, the levels of misclassification and inconsistent classification respectively increased, and this was also expected. Since decisions are based on estimated abilities, less test information produces less precise ability estimates, while higher levels of test information mean that more highly informative items will be selected during module and test assembly to ensure that the higher target information functions are met, which translates into better estimation for individuals. From a 50% increase in test information to full information, misclassification rates increased by about 1% across conditions, and the drop from full information to a 25% decrease in information resulted in an additional 1.5% to 2% of candidates being misclassified. When moving from the 25% to 50% decrease in information, a similarly large percentage increase in misclassifications was observed, on the magnitude of 1.5 to over 2 percentage points. As to inconsistent classification, the first incremental decrease in information yielded an additional 1 or 1.5% of candidates with classification errors, while for the second decrease in information the increase in inconsistent classification was often over 2% and for the drop from 25% decrease to 50% decrease in information about 3% more candidates were inconsistently classified as passing or failing. To give context to such percentages, recall that the sample size was 9,000 candidates per condition. A 1% difference affects 90 candidates and a 3% difference affects 270 candidates.

In specifying test information functions, the lesser amounts of information resulted in lowered DA levels, and as TIF levels decreased the decline in DA grew, but even at the 50%

decrease level DA rates of 87% to 88% were observed. These results with respect to DA thus clearly provide test developers with important information about how much loss in DA could be expected with proportional reductions in test information functions at the overall test level.

Among the routing strategies, the DA and DC results observed were revealing. By a slight margin, the Proximity and NC methods were associated with the lowest levels of misclassification and inconsistent classification, followed by the DPI method. Overall, DA and DC results were poorest with the Random routing methods, as they should have been. In contrast, since the Proximity and NC methods base routing decisions on candidate performance, they more economically use the statistical information in the adaptive test to advance the examinee through the stages of the test in the most difficulty-appropriate way. Thus, these results are significant in that it is expected that higher levels of DA would be observed when either of those approaches are implemented as compared to random or strictly population-based methods.

However, the magnitude of the DA differences between the Random method and the other three routing strategies was generally small or about one-half of one percentage point. To give that meaning, in an operational testing setting with an annual testing population of 10,000 examinees, such a difference translates into perhaps 45 to 90 more misclassifications than would be seen with the other routing strategies. While this number is not trivial, and certainly would not be to the candidates affected, somewhat larger differences had been expected. Perhaps the small differences point to the robustness of estimates of ability and candidate assignments to pass-fail states based on 60 test items, even with reduced test information.

No differences of note were detected with respect to misclassification and inconsistent classification among the test design structures of 1-2-2, 1-3-3, 1-2-3, and 1-3-2. However, with respect to the amount of test information (either an equal split across stages or a 1/2-1/4-1/4 division), clear trends to the misclassification and inconsistent classification results were present.

At high levels of test information (either a 50% increase or full information), the equal split of information outperformed the approach where relatively more information was provided at stage 1. However, with less information (either a 25% or 50% decline in the size of the TIF), the 1/2-1/4-1/4 strategy was more in line with the results from the equal information method. The key point here seems to be that with less information available in a test design, getting candidates correctly assigned to modules at stages 2 and 3 takes on increased importance. This can be accomplished by allocating more of the available information to the stage 1 routing test. With better assignments of candidates based on the stage 1 routing test, the lesser available information at stages 2 and 3 can be utilized effectively.

Accuracy of ability estimation. The accuracy of candidate ability estimates from a test is always a major concern, even in the context of certification and licensure assessment where the decision outcome for each individual is the paramount outcome of interest. Accurate ability estimates provide failing candidates with information about how close they came to passing, and the quality of diagnostic feedback can be improved. In this study, accuracy results for each of the 384 conditions were reported with respect to correlations between true and estimated candidate abilities. Correlation results are reported in Table 5.

[Insert Table 5 about here]

The correlations are simple indicators of the strength of the relationship between the true and final estimates of ability in each simulation condition. Overall, the correlations observed were quite high, from about 0.96 to a low of 0.91 or 0.90, suggesting that even in cases where less information or less optimal routing was used, the final ability estimates were very accurate approximations of candidate abilities.

Across conditions, several informative patterns relating to the ability estimation process with these design variables were evident. First, with respect to the implementing equivalent information

across stages or a strategy with 1/2 information in stage 1 and 1/4 information in the two subsequent stages, the results indicated that in most conditions small differences on the magnitude of approximately 0.01 to 0.03 were present depending on which division of test information was used, regardless of the other variables. This trend indicates that slightly higher correlations between true and estimated abilities were generally associated with the practice of dividing the test information function equally among the stages in the test.

A second trend of note concerns the differences relating to choice of routing strategy. The random method of assigning candidates to modules in the second and third stages generally provided the lowest correlations, although the differences in the magnitude of the correlations for this routing strategy and the others were for the most part equal to about 0.05. A likely explanation for this is that candidate abilities were well estimated with 60 items, and optimal assignment of items at stages 2 and 3 provided little additional information about ability estimation. Interestingly, the gains due to non-random assignment were more noticeable with ability estimation than in making pass-fail decisions. This finding is not surprising, and is important because even with a primary focus on pass-fail decisions, often ability estimation is important too, at least for failing candidates.

These results also suggested that in many cases the DPI method of routing candidates from stage to stage performed as well or marginally better than either the proximity or the NC methods (all methods were in all conditions superior to random routing, though as described above the magnitude of those differences were generally 0.02 to 0.03). This is an interesting result in that the DPI method simply orders theta estimates for candidates and assigns modules on that basis (a very norm-referenced approach).

Routing path analysis. The results presented in Tables 7 – 10 are averages of the percentages of candidates being routed in each path across the two replications in each condition.

Generally speaking, in these results, no patterns relating to the conditions in the simulation could be detected. However, with respect to exposure of modules in each stage of the MST, regardless of routing strategy used, exposure levels for individual modules were largely consistent, which is good news for practice.

The Random method functioned as expected, in that candidates were assigned to paths in equal proportions. The DPI method resulted in relatively low proportions of candidates being assigned to modules of different difficulty levels for the second and third stages. As information decreased, for all routing strategies except Random the number of candidates whose module difficulty levels changed between stages increased. In practical terms, as module information is lessened, more error is present in the ability estimation process, and for candidates in the vicinity of the cut-scores for routing, the likelihood of their being routed to one module or another increased because their estimated ability is less precise (i.e., more inconsistent with the true ability).

In terms of the division of the test information function, the results here too varied in an interesting way, although the implications were quite different across routing strategies. With the DPI method, using equal information across stages resulted in slightly more examinees changing module difficulty between the second and third stages regardless of overall amount of test information, as compared to the approach where half of the test information is specified in the first stage and a quarter in each of the later stages.

Discussion and Directions for Future Research

Many previous studies have documented the quality of measurement associated with multi-stage tests relative to other test designs. The current simulation study was carried out to help practitioners understand better some of the psychometric properties of MST given that there are many design variables to consider in constructing and using such test designs.

Across analyses reported here, the results were largely consistent in their implications for operational multi-stage testing. The most unexpected result concerned the choice of how to divide test information among stages. The results indicated that with high overall amounts of information available the preferable approach is to split information equally across the stages. When lesser levels of overall information available, better results both with respect to ability estimation and making pass-fail decisions are likely to be obtained through a strategy in which more of the available information is used at the first stage. Such a strategy results in more candidates being classified correctly, and thus they benefit from the targeted assessment available to them at subsequent stages of the test.

This finding has significance for operational testing in several respects. First, a strategy that allows for comparable measurement results to be obtained with less test information may be quite desirable to testing agencies. The second meaning of this result is that it suggests that employing unbalanced levels of test information across stages may well be beneficial for testing in some contexts, and sometimes gathering higher levels of test information earlier in the test to make better routing decisions may be helpful. However slight the benefit, any improvement in the accuracy of decision outcomes due to increased efficiency of the routing at earlier points in the test is a highly defensible goal.

In addition, while clear differences in the results were observed between the levels of information, the most sizeable differences concerning decisions and ability estimation were noted in moving from full information or a 25% decrease in information to a 50% decrease in information. Such declines in DA, DC, and the accuracy of ability estimation between levels of test information have clear implications for candidates with respect to the quality of the measurement results in a high-stakes context. To the extent that test developers are able to specify and meet high levels of test information, the measurement outcomes of interest are likely to be psychometrically sound.

However, when item development or other operational considerations constrain or negatively impact test assembly, then understanding the trade-off in measurement precision that can be expected becomes necessary. In this case, two percentage points' worth of DA might be lost when information is decreased by half: if a program tests 100,000 candidates for certification per year, that translates into 2,000 more misclassifications. If the decrease in DC is 5 percentage points, that is 5,000 examinees whose decision classifications from one test occasion to another would vary.

The results relating to the routing rules implemented were likewise interesting and have considerable implications for those implementing MST. One strategy, the Random approach, did not take examinee ability into account whatsoever in determining the difficulty of the second and third stage modules to administer, and measurement and decision results across individuals for this method were lower – though not substantially so – than the other methods that did make such decisions based on estimated ability or number correct scores. Were the modules to be positioned further apart within stages, however, measurement results from the Random method would likely be poorer than evidenced here, thus generalizability of these results about the suitability of the Random approach are not warranted.

Among the strategies that did incorporate estimates of ability into the routing decisions made, the DPI method did give results that were slightly poorer than the Proximity and Number-Correct methods. The DPI method, as an approach that appears more norm-referenced in nature, may well be fine for candidates who are clearly very less able, very highly able, or very middle ability, in stages where three different modules differing by difficulty are present. However, for candidates whose estimated ability places them nearer to the cuts for easy, middle, and high ability, the rank-ordering of candidates may be an especially artificial mechanism that forces some candidates into difficulty levels that are not appropriate for them.

At the same time, results obtained by the DPI methods were only slightly less accurate and consistent than using the Proximity or Number-Correct strategies, and these latter two approaches were highly consistent with one another. All things considered, the logic of the Proximity method may be considered to be the most appropriate and defensible of the four methods for high-stakes decisions used in this study, as it involves assigning candidates to the module empirically determined to most nearly match their estimated ability. The Number-Correct method suffers a bit because it does not capitalize on all of the information available for estimating ability – ability estimates would function a bit better than number-correct scores when assigning candidates to modules. The DPI method does not attempt to link assignments of candidates to the information provided by modules at the next stage and so in that sense it, in general, this method would be less desirable that the Proximity method.

There was an absence of differences in the results due to design structure. The choice of two or three modules in the second and third stages seemed to have no significant impact on the quality of the pass-fail decisions. Concerning design strategy, in the absence of clear measurement advantages, the bigger operational concern for programs might be using more than two stages so that the candidates do not have the perception of being unable to pass if they do poorly on the routing test.

A number of important research questions remain. First, a future investigation might well focus more closely on the splitting of information between stages. Interesting patterns to the results on this dimension were reported in this study, and another important dimension concerns how well that information is targeted to candidates. The questions to be considered include how the candidate population is distributed over the ability scale and where the passing score is located relative to both the candidates and the relative difficulties of the modules at each stage.

Second, another important extension of this work is to build more error into the simulations, to better reflect the kinds of errors that would be seen in operational testing. Here, data fit the model, and correspondingly, the results were higher than what would be observed in practice. Simulation approaches such as adding a second dimension correlated to the first should be considered, since with less good model fit due to the second dimension, it would be possible to begin to evaluate to effects of multidimensionality in both the test and candidates. To some extent, multidimensionality is always present in practice.

Third, many testing programs are now assessing skills and abilities that are more complex in nature. This raises the possibility of MSTs with polytomously-scored test items. A stage might consist, for example, of two or more polytomously-scored items. Utilizing the adaptive structure of MST to improve measurement precision by improving selection of such items for administration as part of a stage-based test structure may well be a direction of interest for researchers. Indeed, approaches using polytomous items in an MST could explore the efficacy of both dichotomous and polytomous items or polytomous items alone.

References

Jodoin, M. G. (2002, June). *Reliability and decision accuracy of linear parallel form and multi stage tests with realistic and ideal item pools*. Paper presented at the International Conference on Computer-Based Testing and the Internet, Winchester, England.

Jodoin, M. G. (2003). *MSTSIM5* [Computer software]. Amherst, MA: University of Massachusetts, School of Education.

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2002, April). *Comparison of the psychometric properties of several computer-based test designs for credentialing exams.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Kim, H., & Plake, B. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the meeting of the National Council on Measurement in Education. Atlanta, GA.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* [pp. 139-183]. New York: Harper and Row.

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

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

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

Luecht, R. M. (2000, April). *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., Brumfield, T., & Breithaupt, K. (2002, April). *A testlet-assembly design for the Uniform CPA Exam*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229-249.

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

Reese, L.M., & Schnipke, D. L. (1999). *An evaluation of a two-stage testlet design for computerized testing* (Law School Admissions Council Computerized Testing Report 96-04). Newtown, PA: Law School Admissions Council.

Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). *Incorporating content constraints into a multi-stage adaptive testlet design* (Law School Admissions Council Computerized Testing Report 97-02). Newtown: PA: Law School Admissions Council.

Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2003, April). A comparison of multi-stage tests with computerized adaptive and paper & pencil tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Schnipke, D. L., & Reese, L. M. (1999). *A comparison of testlet-based test designs for computerized adaptive testing*. (Law School Admissions Council Computerized Testing Report 97-01). Newtown: PA: Law School Admissions Council.

Xing, D., & Hambleton, R. K. (2004). Impact of item quality and bank size on the psychometric quality of computer-based credentialing exams. *Educational and Psychological Measurement, 64(1), 5-24.* .

Zenisky, A. L. (2002). *An empirical investigation of selected multi-stage testing design variables on test assembly and decision accuracy outcomes for credentialing exams* (Center for Educational Assessment Research Report No. 469). Amherst, MA: University of Massachusetts, School of Education.

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

Figure 1. Test Structures of Interest

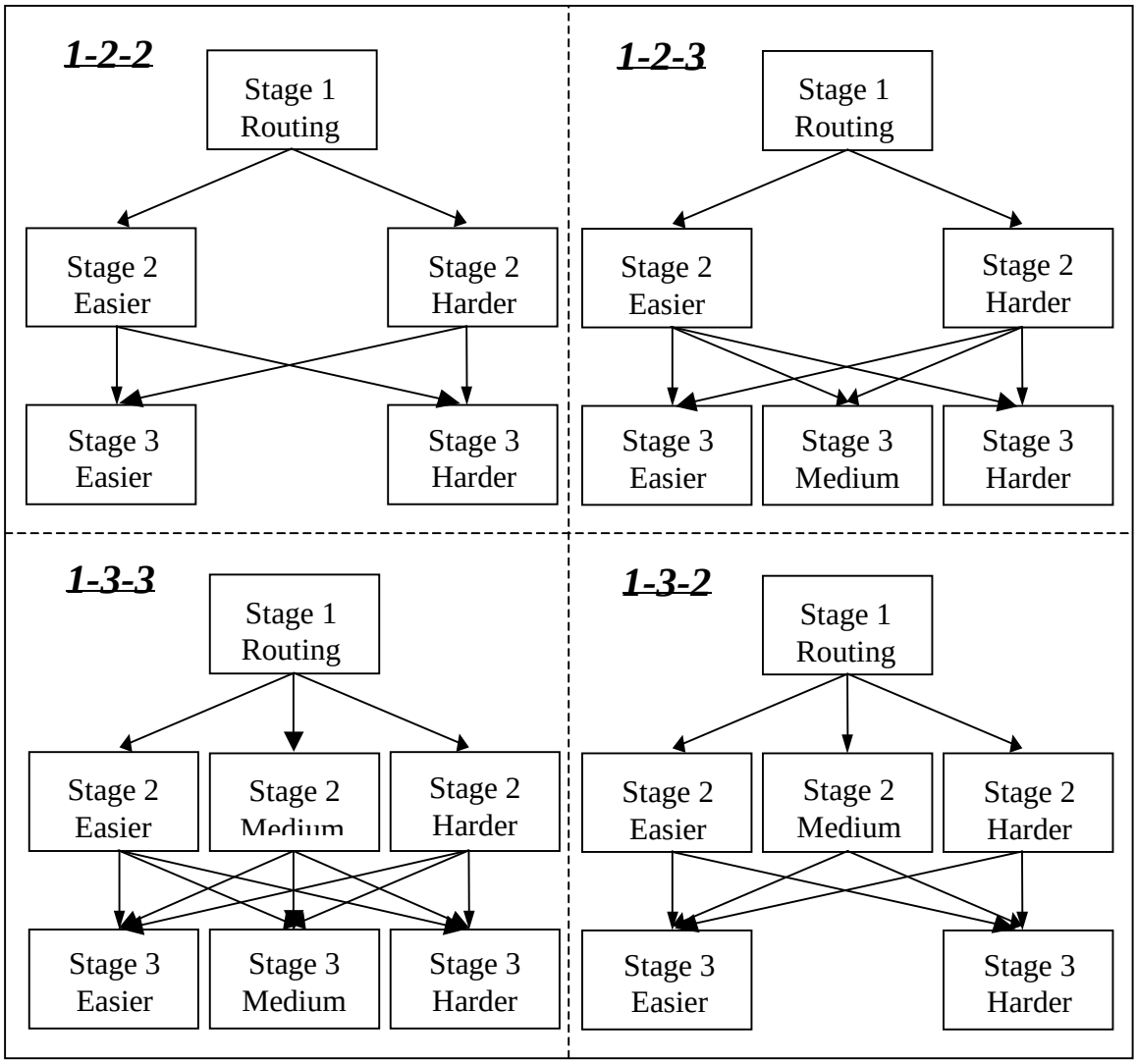

Figure 2. Sample Assignment of Stage-Level Information Functions to Modules

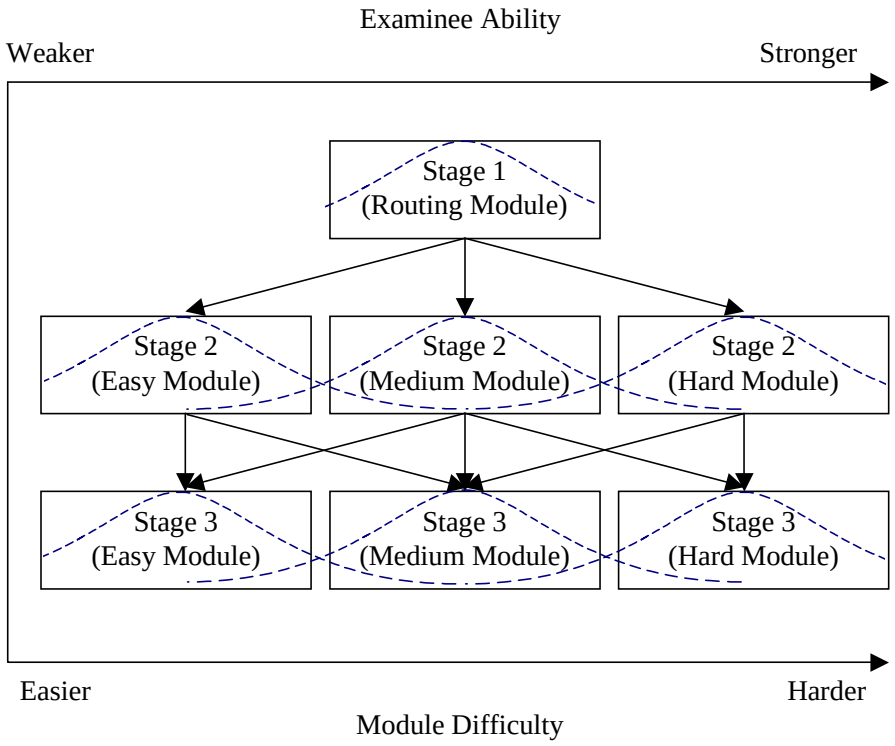

Table 1. Decision Accuracy and Consistency for the 1-2-2 Design

Routing Strategy	Design	TIF Level	Equal Information		1/2-1/4-1/4 Information	
			Misclassified (%)	Inconsistent Classification (%)	Misclassified (%)	Inconsistent Classification (%)
DPI	1-2-2	50% inc.	7.2	10.2	7.3	10.6
	1-2-2	Full	8.0	11.5	9.4	12.2
	1-2-2	25% dec.	10.0	14.4	9.8	14.1
	1-2-2	50% dec.	12.3	16.9	12.0	17.1
Proximity	1-2-2	50% inc.	6.7	9.8	7.1	10.6
	1-2-2	Full	8.1	12.0	10.0	12.0
	1-2-2	25% dec.	9.7	13.8	9.6	14.6
	1-2-2	50% dec.	12.0	17.4	11.9	16.9
Number-Correct	1-2-2	50% inc.	6.7	9.9	7.0	10.6
	1-2-2	Full	8.2	12.2	10.0	12.1
	1-2-2	25% dec.	9.7	13.9	9.7	14.7
	1-2-2	50% dec.	12.1	17.4	11.8	17.0
Random	1-2-2	50% inc.	7.7	10.9	7.1	10.2
	1-2-2	Full	8.6	12.4	9.9	12.7
	1-2-2	25% dec.	10.1	14.5	11.0	14.6
	1-2-2	50% dec.	13.1	18.0	12.3	17.2

Table 2. Decision Accuracy and Consistency for the 1-3-3 Design

Routing Strategy	Design	TIF Level	Equal Information		1/2-1/4-1/4 Information	
			Misclassified (%)	Inconsistent Classification (%)	Misclassified (%)	Inconsistent Classification (%)
DPI	1-3-3	50% inc.	7.1	9.9	7.3	10.4
	1-3-3	Full	8.1	11.7	8.8	12.1
	1-3-3	25% dec.	9.6	13.5	9.9	13.9
	1-3-3	50% dec.	12.4	17.5	11.7	16.5
Proximity	1-3-3	50% inc.	6.8	9.9	7.2	10.6
	1-3-3	Full	8.5	11.9	8.2	11.9
	1-3-3	25% dec.	9.5	13.7	9.9	14.6
	1-3-3	50% dec.	11.8	17.3	11.8	16.9
Number-Correct	1-3-3	50% inc.	6.9	10.1	7.2	10.6
	1-3-3	Full	8.5	12.0	8.2	11.9
	1-3-3	25% dec.	9.5	13.8	9.9	14.6
	1-3-3	50% dec.	11.8	17.2	11.8	17.1
Random	1-3-3	50% inc.	7.5	10.9	7.7	11.0
	1-3-3	Full	9.0	12.2	8.9	12.2
	1-3-3	25% dec.	10.0	14.5	10.1	14.2
	1-3-3	50% dec.	12.3	17.0	12.4	17.4

Table 3. Decision Accuracy and Consistency for the 1-2-3 Design

Routing Strategy	Design	TIF Level	Equal Information		1/2-1/4-1/4 Information	
			Misclassified (%)	Inconsistent Classification (%)	Misclassified (%)	Inconsistent Classification (%)
DPI	1-2-3	50% inc.	7.0	10.0	7.1	10.3
	1-2-3	Full	8.0	11.5	8.4	12.8
	1-2-3	25% dec.	10.1	14.6	10.0	13.7
	1-2-3	50% dec.	11.6	17.2	12.1	17.1
Proximity	1-2-3	50% inc.	7.1	9.9	7.2	10.4
	1-2-3	Full	8.6	12.4	8.7	12.1
	1-2-3	25% dec.	9.8	13.7	10.0	14.5
	1-2-3	50% dec.	11.9	17.3	11.7	16.7
Number-Correct	1-2-3	50% inc.	6.9	9.7	7.3	10.3
	1-2-3	Full	8.4	12.5	8.6	12.2
	1-2-3	25% dec.	9.7	13.6	10.0	14.5
	1-2-3	50% dec.	11.8	17.2	11.5	16.7
Random	1-2-3	50% inc.	7.6	10.5	7.6	10.5
	1-2-3	Full	9.2	12.8	8.7	12.4
	1-2-3	25% dec.	10.5	15.0	10.3	14.6
	1-2-3	50% dec.	12.8	18.3	12.1	17.0

Table 4. Decision Accuracy and Consistency for the 1-3-2 Design

Routing Strategy	Design	TIF Level	Equal Information		1/2-1/4-1/4 Information	
			Misclassified (%)	Inconsistent Classification (%)	Misclassified (%)	Inconsistent Classification (%)
DPI	1-3-2	50% inc.	7.0	9.9	7.5	10.5
	1-3-2	Full	7.8	11.5	8.5	12.0
	1-3-2	25% dec.	9.8	14.1	9.8	14.2
	1-3-2	50% dec.	12.4	17.6	12.2	17.2
Proximity	1-3-2	50% inc.	6.9	9.7	7.5	10.5
	1-3-2	Full	8.4	12.2	8.4	12.0
	1-3-2	25% dec.	9.7	13.8	9.8	14.3
	1-3-2	50% dec.	11.6	16.8	11.8	17.4
Number-Correct	1-3-2	50% inc.	7.0	9.9	7.4	10.6
	1-3-2	Full	8.2	12.1	8.3	12.2
	1-3-2	25% dec.	9.6	13.8	9.8	14.3
	1-3-2	50% dec.	11.6	16.9	12.0	17.4
Random	1-3-2	50% inc.	7.4	10.4	7.3	9.9
	1-3-2	Full	9.2	12.7	9.2	12.5
	1-3-2	25% dec.	10.0	14.3	10.0	14.1
	1-3-2	50% dec.	12.7	17.4	12.3	17.3

Table 5. Correlations Between True and Estimated Abilities

Design	TIF Level	Equal Information				1/2-1/4-1/2 Information			
		DPI $r_{t \cdot e_{final}}$	Prox. $r_{t \cdot e_{final}}$	NC $r_{t \cdot e_{final}}$	Ran. $r_{t \cdot e_{final}}$	DPI $r_{t \cdot e_{final}}$	Prox. $r_{t \cdot e_{final}}$	NC $r_{t \cdot e_{final}}$	Ran. $r_{t \cdot e_{final}}$
1-2-2	50% inc.	0.96	0.96	0.96	0.95	0.96	0.96	0.96	0.96
1-2-2	Full	0.94	0.95	0.95	0.94	0.95	0.94	0.94	0.94
1-2-2	25% dec.	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
1-2-2	50% dec.	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
1-3-3	50% inc.	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95
1-3-3	Full	0.94	0.95	0.95	0.94	0.95	0.95	0.95	0.94
1-3-3	25% dec.	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
1-3-3	50% dec.	0.91	0.90	0.91	0.90	0.91	0.90	0.90	0.90
1-2-3	50% inc.	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95
1-2-3	Full	0.95	0.95	0.95	0.94	0.95	0.94	0.94	0.94
1-2-3	25% dec.	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
1-2-3	50% dec.	0.91	0.91	0.91	0.90	0.91	0.90	0.90	0.90
1-3-2	50% inc.	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95
1-3-2	Full	0.94	0.95	0.95	0.94	0.95	0.94	0.94	0.94
1-3-2	25% dec.	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
1-3-2	50% dec.	0.91	0.91	0.91	0.90	0.91	0.90	0.90	0.90

Table 7. Routing Path Frequencies in 1-2-2 Design with Four Routing Strategies

Division of Information	TIF Level	Module			Routing Strategy			
		s1	s2	s3	DPI	Proximity	NC	Random
Equal Information Across Stages	50% Increase	1	1	1	45.5%	36.3%	38.9%	25.2%
		1	1	2	4.5%	12.2%	7.9%	24.7%
		1	2	1	4.5%	7.8%	4.7%	25.2%
		1	2	2	45.5%	43.7%	47.5%	24.9%
	Full	1	1	1	44.6%	33.8%	37.5%	24.9%
		1	1	2	5.4%	10.9%	8.6%	25.1%
		1	2	1	5.4%	11.8%	4.4%	24.6%
		1	2	2	44.6%	43.5%	49.5%	25.4%
	25% Decrease	1	1	1	43.9%	30.8%	39.4%	25.2%
		1	1	2	6.1%	12.0%	8.7%	25.0%
		1	2	1	6.1%	14.0%	3.3%	24.8%
		1	2	2	43.9%	43.1%	48.6%	25.0%
	50% Decrease	1	1	1	42.7%	29.0%	41.2%	24.7%
		1	1	2	7.3%	13.9%	8.7%	25.2%
		1	2	1	7.3%	14.9%	3.4%	25.0%
		1	2	2	42.7%	42.2%	46.7%	25.1%
1/2-1/4-1/4 Information Across Stages	50% Increase	1	1	1	46.6%	42.8%	34.1%	24.9%
		1	1	2	3.4%	8.6%	12.3%	25.0%
		1	2	1	3.4%	5.7%	9.1%	25.1%
		1	2	2	46.6%	42.9%	44.5%	25.0%
	Full	1	1	1	46.4%	43.0%	34.6%	24.9%
		1	1	2	3.6%	8.9%	13.8%	25.1%
		1	2	1	3.6%	6.3%	8.8%	25.0%
		1	2	2	46.4%	42.8%	42.7%	25.0%
	25% Decrease	1	1	1	45.8%	39.8%	32.1%	24.7%
		1	1	2	4.2%	9.8%	12.4%	25.2%
		1	2	1	4.2%	8.0%	12.6%	25.1%
		1	2	2	45.8%	42.4%	43.0%	24.9%
	50% Decrease	1	1	1	45.0%	38.5%	28.7%	24.9%
		1	1	2	5.0%	9.8%	14.1%	24.8%
		1	2	1	5.0%	9.0%	15.5%	25.2%
		1	2	2	45.0%	43.8%	41.7%	25.1%