# Effects of changes in the examinees' ability distribution on the exposure control methods in CAT

Shun-Wen Chang
National Taiwan Normal University
Bor-Yaun Twu
National Tainan Teachers College

Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, April 2001.

**Abstract**

Item security is one of the practical issues that substantially concern the makers of high stakes tests for the continuous testing context of CATs. To satisfy the security requirements of CATs, efforts have been made to directly control the exposure rates of optimal items by incorporating statistical methods into the item selection procedure. Since differences likely occur between the exposure control parameter derivation stage and the operational CAT administrations, the exposure control methods may not fully accomplish the goal of controlling item exposure. This study explored the effects of the distribution changes on the performance of the Sympson and Hetter (SH) and the Davey and Parshall (DP) procedures and provided an examination of the assertion that the Stocking and Lewis conditional multinomial (SLC) procedure would function independently of any real examinee population.

Simulations were carried out for this study using various combinations of the intended and real ability distributions. The results showed that the changes in the examinees' ability distribution affected both SH and DP procedures in their control of the observed maximum exposure rates and the effects were more profound for the SH method. The performance of the SLC method was demonstrated to be independent of the examinees' ability distributions in the operational CAT administrations.

Keywords: computerized adaptive testing, item exposure rate, exposure control method, ability distribution

Computerized adaptive testing (CAT) is a relatively recent trend in the testing society, which has called enormous attention to its growing number of large-scale applications. Many practical issues have emerged while theory is translated into practice with the continuous testing context of computerized adaptive tests (CATs). Controlling item exposure is a critical issue that must now be addressed (Chang, 1998; Chang, Ansley, & Lin, 2000; Davey & Fan, 2000; Davey & Parshall, 1995; Hetter & Sympson, 1997; Mills & Stocking, 1996; Parshall, Davey, & Nering, 1998; Pastor, Dodd, & Chang, 2000; Stocking, 1993; Stocking & Lewis, 1995, 1998; Way, 1997). Under CAT item selection rules, some popular items might be too frequently administered to test-takers within a short period of time. When test-takers have access to the questions before test administrations, frequently appearing test items may soon be compromised. High rates of item exposure lead to a great item security risk.

To date, methods have been developed that attempt to reduce the amount of item exposure of the operational CATs by embedding statistical mechanisms into the item selection procedure. The goal is to control the exposure rates of items to a desired maximum value, *r*, that is specified in advance of testing. The Sympson and Hetter (SH) procedure (Hetter & Sympson, 1997; Sympson & Hetter, 1985) employs an exposure control parameter for each item in the pool, which is determined after a series of iterative simulations. Given that an item has been selected, whether to administer this item to the examinee depends upon the exposure control parameter of this item. For very popular items, the exposure control parameters could be as low as the pre-specified desired exposure rate, indicating that these items cannot be freely administered when they are selected. For items rarely appearing, the associated exposure control parameters could be as high as 1.0, meaning that once these items are selected, they are almost always presented. Items that have been selected but not administered are excluded from the pool of remaining items for the examinee. The probabilistic model of Sympson and Hetter seeks to achieve the goal that there is no item administered to more than a pre-specified fraction of examinees.

The Davey and Parshall (DP) methodology (Davey & Parshall, 1995; Parshall et al., 1998) provides an exposure parameter for each item that is conditioned on all other items previously administered to the examinee. To utilize this method, an exposure table needs to be prepared through a series of simulations. Diagonal elements of the table indicate the probability limits with which individual items can be administered given selection. These values will be small if the corresponding items tend to be selected frequently. The off-diagonal elements represent the probability limits with which a pair or set of items can appear together given selection. Similarly, the off-diagonal values will be small if the pairs of items tend to occur together very often. It was concluded in Chang (1998) and Davey and Parshall (1995) that this procedure satisfactorily reduces the extent to which the items overlap across tests administered for examinees with similar ability and for examinees of differing ability.

The values of the exposure parameters of these two methods are determined through a series of iterative simulations using an already established adaptive test design and a large group of simulees representative of the real examinee population. However, unless the real examinee distribution in operational testing is identical to that in the iteration stage, the maximum observed exposure rates often exceed the desired values. While the DP strategy was shown to control the overall maximum observed exposure rates to the desired level under the situation of the same ability distributions (Chang et al., 2000), the results in Chang (1998) demonstrated the effects of population differences on the performance of the SH and DP approaches, respectively. Based on the situation that a $N(0,2)$ was the ability distribution for the derivation of the exposure control parameters and a $N(0,1)$ was the distribution for the operational testing, the maximum exposure rates for these methods were observed to be higher than the pre-specified rates (see Chang, 1998). As indicated in Stocking and Lewis (1995), since the exposure control parameters are developed with respect to a particular intended ability distribution, the parameters might need to be redeveloped if the intended distribution is very different from the real distribution. Davey and Parshall (1995) also cautioned that the exposure rates of items may actually be controlled only to the extent that the real ability distribution of examinees is similar to that intended. Since a perfect match between the two distributions of examinees' ability is highly unlikely, questions regarding the effect of the distribution change demand systematic investigations. How does the change in the examinees' ability distribution affect the performance of the exposure control algorithms and which change would bring a more profound effect? What ability distribution would be more conservative to be adopted for developing the exposure control parameters in order to mitigate the effect of the change? What exposure control method would be more robust to the changes? The current study is designed to examine how the distribution changes affect the performance of the SH and DP procedures in their control of the maximum item exposure rates using various combinations of the ability distributions at the derivation stage and in the operational administrations. Depending on the extent to which the method is affected by the differences in the distributions, ways may be suggested to remedy the situation such that, for example, the maximum exposure rate specified prior to the derivation of the parameters could be set at a value slightly lower than actually desired.

Unlike the SH and DP methods where a global exposure control parameter is developed for an item in order to limit the item's overall appearance in reference to the examinee sample drawn from a target population, the Stocking and Lewis conditional multinomial (SLC) procedure (Stocking & Lewis, 1998) derives for each item an exposure control parameter with respect to a particular level of examinee ability. Accordingly, the SLC procedure results in different exposure control parameters for each item to be applied to various ability levels. While the performance of the SH and DP methods might be affected by a non-optimal specification of examinees' ability distribution, the performance of the SLC procedure

may not be affected since their exposure control parameters are established independently of any real examinee population. This was claimed in Stocking and Lewis (1998) to be an appealing advantage of employing the SLC procedure. However, evidence has not been provided to make such a conclusion. Whether the exposure control parameters of the SLC strategy perform similarly well for various real examinee populations in operational CAT demands more convincing results. This study is intended to provide a check on this claimed advantage by applying the SLC exposure control parameters to various ability distributions in operational testing.

## The Purpose

This study was designed to explore how the item exposure control algorithms in the CAT context are affected by non-optimal specifications of examinees' ability distributions; that is, the exposure control parameters are developed using an examinees' ability distribution that is different from the real or operational examinee population. Specifically, this study attempted to achieve the following objectives:

1. to examine how the performance of the exposure control methods is affected by the differences between the examinees' ability distributions for the derivation of exposure control parameters and for the operational testing situation;

2. to verify whether the performance of the conditional exposure control procedure is independent of the ability distribution of the real examinee population.

## Method and Data

Simulations were employed to carry out this research. The plan of the study design and methods for data analyses are described below.

### I. Design of the Study

Specification of Decisions for the CAT Components

The 3-PL model of IRT forms the basis for the investigations of this study. A real pool of 600 discrete items was used, in which the item parameters were calibrated from multiple forms of a large scale standardized paper-and-pencil test using BILOG (Mislevy & Bock, 1990) on a single ability scale. The mean and SD of the *a* parameters were 1.02 and .34; the mean and SD of the *b*s were .16 and 1.05; and the mean and SD of the *c*s were .17 and .08. In this item pool, there existed more items at the middle of the ability continuum than at the two ends. Also, this pool contained items that discriminate moderately well with more discriminating power at the middle difficulty levels than at the extreme levels. The CAT administration process was initiated by assigning each examinee a common ability estimate of zero to simulate a situation where no a priori information was available about the individuals.

Items were selected based on the maximum item information criterion with the various exposure

control algorithms incorporated into the selection process. The target maximum exposure rate was specified to be .20. The content presented to the examinees was balanced according to the Kingsbury and Zara's mechanism (1989). The first item was administered following the algorithms of item exposure control, regardless of the item's content attribute. The percentage of items that had been administered in each content category was calculated and compared to the corresponding pre-specified percentage. The content area with the largest discrepancy between the empirical and the desired percentage was then identified, from which the next item was selected and administered based on the algorithms of the various exposure control methods. To estimate the examinees' abilities, Owen's Bayesian strategy (Owen, 1975) was utilized for the provisional ability estimation and the maximum likelihood estimation method (Birnbaum, 1968) was employed for the final ability estimation. Each examinee was administered 30 items.

Specification of Factors Included in the Study

1. Item exposure control methods

The SH, DP and SLC procedures were the exposure control methods to be investigated.

2. Ability distributions at the derivation stage

Four normal distributions of $N(0,1)$, $N(0,2)$, $N(1,1)$ and $N(-1,1)$ on the theta metric were utilized to develop the exposure control parameters of the SH and DP methods. A normal distribution is, in many cases, a representative ability distribution of real examinee populations. In addition to the employment of a standard normal distribution, the normal distribution with a mean of 0 and a variance of 2 was used to allow more examinees at the two ends of a typical ability range of interest in the development of the parameters. The idea was that by using an ability distribution of greater variability, the exposure control parameters of items especially informative at the extreme ability levels were developed based on more simulees and the effect of sampling error might be lessened. Also, a $N(1,1)$ was used as being representative of a normal distribution where the examinees' abilities were on average one theta point higher than those of a $N(0,1)$, and a $N(-1,1)$ was tried to project the real examinee population that was one theta point lower on average. This study examined which one of these four iteration distributions would be more conservative for developing the exposure control parameters in order to mitigate the effect of a change in ability distributions.

3. Ability distributions at the operational testing stage

Five normal distributions of $N(0,1)$, $N(1,1)$, $N(-1,1)$, $N(0,0.5)$ and $N(0,2)$ on the theta scale were simulated to be the real examinee populations in the operational CAT administrations. Besides using a standard normal distribution, a $N(1,1)$ was employed to simulate a population of more able examinees and a $N(-1,1)$ was used to represent a less able population, with one theta point difference on the ability continuum. Also, a $N(0,0.5)$ was adopted to represent a less variable population and a $N(0,2)$ was used for a population of greater variability in ability, where both groups were still distributed normally with a

mean of 0.

**II. Procedures for Data Simulation**

Two stages were involved in the data simulations. First, the exposure control parameter for each item in the pool was developed according to the specific exposure control algorithm and the CAT design proposed for this study. Then, simulations were carried out for the operational CAT administrations.

The Stage of Developing Exposure Control Parameters

For both SH and DP approaches, the adaptive tests were administered to a sample of 50,000 examinees drawn from the four normal distributions of $N(0,1)$, $N(0,2)$, $N(1,1)$ and $N(-1,1)$, respectively. For the SLC procedure, the development of exposure control parameters was in reference to a particular level of proficiency. The adaptive tests were administered to the conditional sample of 5,000 examinees at each of the theta levels equally spaced over the interval of –3.2 and 3.2 with an increment of .40 (i.e., -3.2, -2.8,..., 3.2), a total of 17 ability points. A conditional sample size of 5,000 should be sufficient for producing stable exposure control parameters for this method (see Chang et al., 2000). The desirability of items for the SLC method was ordered only based on item information values, not on the weights as specified in the original algorithm for which the Stocking/Swanson weighted deviations model (WDM) (Stocking & Swanson, 1993; Swanson & Stocking, 1993) was employed to select items. For each of these three methods, the process repeated until the observed maximum exposure rates were approximately equal to the desired level and the exposure control parameters were stabilized in the subsequent iterations. The stabilized parameters at the final round of iterations were the exposure control parameters to be used in operational adaptive testing.

The Stage of Simulating Operational CAT Administrations

For all the procedures of SH, DP and SLC, the adaptive test was delivered to each examinee of a sample of 50,000 examinees respectively drawn from $N(0,1)$, $N(1,1)$, $N(-1,1)$, $N(0,0.5)$ and $N(0,2)$, following the CAT design established for this study. Items were selected and administered according to the specific algorithm of an exposure control strategy. The exposure control parameters developed from the previous stage for the three methods were utilized here to manage the administration frequencies of the selected items. The adaptive tests were also administered to a conditional sample of 3,000 examinees at each of the ability levels equally spaced over the theta metric to obtain the conditional maximum observed exposure rates and the test overlap rates, as well as to evaluate the measurement properties conditional on each ability point.

**III. Methods for Data Analyses**

The criteria for evaluating the effects of the distribution changes in examinees' ability for the various conditions were observed maximum exposure rates, pool utilization, test overlap and the conditional standard errors of measurement (CSEMs). The maximum exposure rates observed were

investigated in reference to the entire examinee group and also to the examinees of a particular ability point. The utilization of the item pool was assessed using the numbers and/or percentages of items in the pool administered at least once. The values of the test overlap rates were classified into the peer-to-peer overlap rates and the test-retest overlap rates to show the extent to which pairs of items appeared together across tests taken by examinees of differing abilities and similar ability, respectively. The peer-to-peer mean overlap rate was obtained by first calculating the overlap percentage of tests taken by two examinees generated from the respective ability distributions in the operational testing situation, then averaging the overlap percentages over all paired examinees. The test-retest mean overlap rate was obtained by first computing the percent of items that overlapped between the adaptive tests given to an examinee of ability theta twice, then averaging the overlap percentages over all examinees at this theta level. The CSEMs were the errors of the ability estimation as a result of introducing the various exposure control algorithms into the item selection process.

To obtain the overall maximum exposure rates observed, the pool utilization and the peer-to-peer test overlap rates, items with the four sets of exposure control parameters were respectively administered to the five real examinee populations of $N(0,1)$, $N(1,1)$, $N(-1,1)$, $N(0,0.5)$ and $N(0,2)$. In order to compute the conditional observed maximum exposure rates, the test-retest overlap rates and the CSEMs, items with the four sets of the exposure control parameters were administered respectively to 3,000 examinees at each ability level (i.e., to examinees of the same ability) rather than to the entire examinee populations. Figures were plotted to display the results at the respective ability levels.

## Results and Discussion

### I. The Results of Developing Exposure Control Parameters Using Different Iteration Distributions

The exposure control parameters were developed through a series of adjustment simulations for both SH and DP approaches employing the four iteration distributions of $N(0,1)$, $N(0,2)$, $N(1,1)$ and $N(-1,1)$. The maximum exposure rates observed are displayed in Figure 1 for each iteration stage to illustrate the converging status of these two procedures with the various ability distributions. For the SLC procedure, the exposure control parameters were derived in reference to a particular ability level. The results of the maximum exposure rates observed at each ability point are also presented in Figure 1. A horizontal line is presented in each figure to indicate the pre-specified exposure rate of .20.

As shown in Figure 1, the SH and DP methods demanded different numbers of iterations in developing the exposure control parameters, but the observed maximum exposure rates of these two procedures all converged to a value very close to the target rate of .20. For the SH iterations, the observed maximum exposure rates were slightly different among the four distributions before the convergence took place, and the $N(-1,1)$ iteration took one or two more iterative steps to converge than

the other distributions. For the DP methodology, the iteration results were almost indistinguishable for the four ability distributions.

Figure 1 shows that for the SLC strategy, the iteration curves did not approach the desired rate of .20. Also, the conditional observed maximum exposure rates for the various ability points seem to converge to different values. The closer the iterations were at the extreme ability levels, the higher the values the conditional observed maximum exposure rates converged to. These results reflected the nature of this item pool, in that there were fewer items appropriate for administration for the extreme ability levels than for the middle levels.

**II. The Results of Using Various Normal Distributions in Operational CAT Administrations**

The four sets of the SH and DP exposure control parameters resulting from the final rounds of the respective iterations were associated with the corresponding items to be used in the operational testing situations, in which the examinees' abilities were variously distributed as $N(0,1)$, $N(1,1)$, $N(-1,1)$, $N(0,0.5)$ and $N(0,2)$. The simulation of the operational CAT administrations proceeded similarly for the SLC method, but with only one set of parameters that were derived using no intended examinees' ability distributions, but with respect to each level of the ability continuum.

Described below are the results of both SH and DP methods under the various combinations of ability distributions at the derivation stage and the operational CAT administrations with respect to each of these criteria: the observed maximum exposure rates, the utilization of the item pool, the peer-to-peer and the test-retest overlap rates, and the CSEMs. For the SLC method, only the overall maximum exposure rates observed, the pool utilization and the peer-to-peer test overlap rates were employed as the criteria to detect whether this conditional exposure control strategy functioned independently of any real examinee population. The five real ability distributions of $N(0,1)$, $N(1,1)$, $N(-1,1)$, $N(0,0.5)$ and $N(0,2)$ were used.

**The SH and DP Methods**

Observed Maximum Exposure Rate

Table 1 reports the summary statistics of the observed exposure rates as a result of applying the different sets of the SH exposure control parameters to the various examinee populations in the operational testing situations. The N column lists the numbers of items that were administered to examinees at least once. Based on these items, the summary statistics were obtained.

It can be seen in Table 1 that for the SH procedure, there were effects from deriving the exposure control parameters using distributions that were different from the real examinee distributions. Only when both distributions matched perfectly were the observed maximum exposure rates as low as the pre-specified rate of .20. For the iteration distributions that did not match closely to the real examinees'

ability distribution, the observed maximum exposure rates could reach a high value, although the average observed exposure rates were still similar. The degree of the maximum exposure rate observed exceeding the desired value appeared to depend on how the intended and real distributions differed. The results in Table 1 seem to suggest that when the two distributions possessed similar means, the observed maximum exposure rates were lower than those obtained when the distributions had different mean values. That is, if the population mean values could be well projected, the maximum observed exposure rates may not far exceed the desired value or in other words, the optimal item security could be better ensured, at least for the design proposed in this study.

Reasons for these phenomena may be as follows. During the iteration process, items discriminating well near the center of the intended ability distribution were likely to be selected frequently, so these items had lower exposure control parameters. Items discriminating well only at the two ends of the intended ability distribution were seldom used, so these items had higher exposure control parameters. As the ability distribution in the operational testing situation was different from that used for the parameter derivation, the exposure rates of items administered to the real population were not appropriately controlled by their exposure control parameters. It is possible that items associated with high parameters were administered to the large number of examinees near the center of the real examinee distribution (where, instead, they should have been more strictly controlled) and accordingly, were overexposed. When the mean of the real examinee population was close to that of the iteration group (i.e., most examinees were centered at similar locations), a large proportion of the examinees were still administered items with low exposure control parameters. These items were under a tough control so the risk of exceeding the desired rate would be diminished. Since the discriminating power of items is related to the location along the ability continuum, the item pool characteristics such as the quality and the number of items at each ability level must be a compelling factor in explaining these effects also. It is likely that values of the observed maximum exposure rates were affected by factors such as over what range of the ability scale most items discriminate well and how well these items discriminate, and how many of them are available for selection as well.

The results in Table 1 seem to suggest that a $N(0,2)$ distribution might be slightly more conservative to use for the exposure control parameter derivation than a $N(0,1)$ to ensure item security in operational CAT administrations. Because a $N(0,2)$ has relatively more examinees at both ends of the ability scale than a $N(0,1)$, the exposure control parameters resulting from a $N(0,2)$ for the items discriminating well only at the two ends would not probably be as high as those from a $N(0,1)$. When these items became popular for a large number of examinees in real testing, their exposure rates were under relatively better control than those from a $N(0,1)$. However, it is inevitable that the results were confounded by the properties and structures of the particular item pool also.

For the DP method, while a perfect match between the two distributions for the derivation and the real examinee population led the observed maximum exposure rates close to the pre-specified rate of .20, differences in the distributions also caused the maximum observed rates to be higher than the desired value (see Table 2). However, this effect of a non-optimal specification of the examinee ability distribution was smaller than that for the SH procedure. The DP method seemed more robust to the effect of changes in the ability distributions. It can also be seen from Table 2 that when the mean values of both distributions were located at similar points, the risk of observing high maximum exposure rates may not be as great as that when the mean values were further apart; the explanations discussed with the SH method are applicable here.

The conditional maximum observed exposure rates at each ability level are displayed in Figure 2. For the SH method, the maximum exposure rates conditionally observed at each theta point varied greatly among the four iteration distributions. When the exposure control parameters were derived using the $N(0,1)$ distribution, the conditional maximum exposure rates were low at the ability points around 0.0 but were increased dramatically towards both ends. Similar patterns were found with both $N(1,1)$ and $N(-1,1)$ iteration distributions. The conditional maximum exposure rates were low at the points near the mean of the distribution used for the exposure control parameter derivation, but the values increased at points farther away from that mean. For the $N(0,2)$ iterations, the pattern appeared to be less dramatic.

These findings demonstrated that, for the SH algorithm, the conditional maximum observed exposure rates were best being controlled at the ability points near the target of the distribution used for the iteration. Near the target of the intended distributions in the iteration, items tended to be selected more often, so their exposure control parameters would be lower in order to limit the operational administration frequencies of these items. This led to the lowest conditional maximum exposure rates being observed at ability points near the respective mean values of the iteration distributions. At points farther away from the mean, items were less often selected and the issue of overexposing these items became a less concern, so the corresponding exposure control parameters were derived to be at relatively higher values. But when these items were the only items discriminating well for examinees at a particular ability level in operational testing, the fact that the exposure rates of these items were not tightly controlled allowed the maximum observed exposure rates to reach a high value for this particular ability level. When a distribution of greater variability was employed in the iterations, the exposure control parameters were derived to be at more moderate values. Accordingly, the curve of the conditional maximum exposure rates for the $N(0,2)$ distribution varied less dramatically. However, it is important to recognize that all of these results might be complicated with the characteristics of the item pool, since the numbers and quality of the items in the pool were not the same at each ability level.

For the DP method, the effects of employing the four iteration distributions to develop the exposure control parameters on the conditional maximum observed exposure rates were much smaller than for the SH method (see Figure 2). Somewhat similar patterns were still detected with the DP procedure where the conditional maximum exposure rates were lower at the ability levels near the center of the iteration distributions, but the values only increased slightly at the points away from that mean. The distributions used for the derivation of the DP exposure control parameters did not appear to affect too much the performance of this procedure in controlling the conditional maximum exposure rates. The feature of the DP algorithm seems to perform similarly well in controlling the exposure rates of items at each ability level over the entire continuum.

Utilization of the Item Pool

The numbers and/or the percentages of items that were administered at least once are reported in Table 1, which provides information about the extent to which the pool was utilized. The greater the number of used items relative to the whole pool (i.e., the used percentage), the greater the utilization of the item pool. It can be seen that a non-optimal specification of the distribution does not seem to have noticeable effects on the utilization of the item pool. A perfect match of the two distributions did not make the best use of the item pool for the SH procedure. Indeed, Table 1 reveals that no matter what distribution was employed for the exposure control parameter derivation, the real examinee distribution of $N(0,2)$ always led to the greatest pool utilization while the $N(0,0.5)$ resulted in the least degree of the pool utilization by a small amount.

To achieve the goal of greater pool usage, it is necessary to increase the administration frequencies of those rarely used items to at least one time, since the pool usage was defined in this study as the numbers and/or the percentages of items administered at least once. The extent to which the pool utilization was increased was a result of utilizing more items with high exposure control parameters (especially those with a value of 1.0) rather than utilizing items with low parameters for which they might have already been used often because of their popularity. When the real examinee population was distributed with greater variability, items appropriate for examinees at the two ends of the ability scale would be in greater demand. If these items were associated with relatively high exposure control parameters, they were more likely to be administered once selected. Accordingly, the chance of utilizing items of the higher exposure control parameters was increased, leading to a greater extent of pool utilization. While a perfect match between the intended and real distributions resulted in the popular items being well controlled to the pre-specified level, a perfect match did not yield the greatest pool utilization. However, it is important to recognize the effects of the item pool characteristics on the outcomes of the pool utilization also.

The DP procedure resulted in a greater extent of pool utilization than the SH method (see Table

2), which has been shown in Chang (1998) and Chang et al. (2000) to be an advantage of employing this algorithm. The results also showed that an optimal match did not lead to the greatest pool utilization for this procedure. The real examinee distribution of $N(0,2)$ always produced the greatest pool utilization while $N(0,0.5)$ yielded the least utilization. How the real examinee population was distributed seems to have a noticeable effect on the utilization of the pool for the DP strategy. The same rationale provided earlier for the SH method might account for these varying degrees of the pool usage.

The Peer-to-Peer Test Overlap Rates

The values of the peer-to-peer test overlap rates represent the overlap percentages of items administered to examinees of randomly dissimilar abilities. Table 3 reports the summary statistics of the peer-to-peer test overlap rates for the SH algorithm under the various combinations of the ability distributions. It can be seen in Table 3 that the real examinee distribution of $N(0,2)$ led to the smallest average peer-to-peer overlap rates for all iteration distributions, except for the $N(1,1)$ iteration. An optimal specification of the ability distributions did not produce the smallest values of the average peer-to-peer overlap rates. This finding is similar to that of the pool utilization reported in Table 1. Again, when the distribution of the real examinee population was more variable, items appropriate at the two ends of the ability scale might be used relatively more often, so the chance of having paired examinees being administered the same items could be reduced. In addition, for the two distributions having similar means, the average peer-to-peer test overlap rates tend to be lower than for those having different means. This finding indicated that for most examinees in the operational testing situation, the items administered were more appropriately controlled for their appearance frequencies by the exposure control parameters so the overlap rates remained at lower percentages. These phenomena might imply that the peer-to-peer test overlap rates were determined by the ability distribution of the real examinee population as well as the extent to which both intended and real distributions were similar.

As shown in Table 4, the mean values of the peer-to-peer test overlap rates for the DP procedure were smaller than those for the SH method, which was an advantage of employing the DP algorithm (Chang, 1998; Davey & Parshall, 1995). Also, the differences of these mean values for the DP procedure were small. Similarly to the SH method, the real examinee distribution of $N(0,2)$ for the DP method also led to the smallest average peer-to-peer test overlap rates for all four iteration distributions. The real distribution of $N(0,0.5)$ resulted in the largest values of the peer-to-peer overlap rates for all iteration distributions, except for the $N(-1,1)$ iteration. As has been addressed earlier with the SH procedure, how the examinee populations are distributed in the operational CAT administrations and how similar the intended and real distributions are might interact to affect the results of the peer-to-peer test overlap rates for the DP procedure. However, the DP method has been shown to be more robust to the changes in the distribution in controlling item exposure. The differences among the mean values of

these peer-to-peer test overlap rates might be more of a result of using the various real examinees' ability distributions in the operational administrations.

The Test-Retest Overlap Rate

The test-retest mean overlap rates (i.e., the average values of the test-retest overlap rates) are displayed for each ability point in Figure 3. For the SH strategy, the patterns for the test-retest mean overlap rates across the ability continuum were similar to those for the conditional maximum observed exposure rates (see Figure 2). The test-retest mean overlap rates were low around the theta points near the mean of the respective distributions used in the iteration, but the values were increased outwards. For items appropriate for the examinees near the mean of the iteration distribution, the corresponding exposure control parameters would be derived to be smaller values since these are the items likely to be overused in real testing. These exposure control parameters performed to limit the exposure rates of items and at the same time, reduce the test-retest overlap rates for examinees at these levels.

As for the DP procedure, the effects of employing the various iteration distributions on the test-retest mean overlap rates were not as substantial as those for the SH procedure (see Figure 3). A careful inspection shows that the pattern for the DP method was not similar to those in Figure 2 or to the SH method displayed in Figure 3. Instead, the values of the test-retest mean overlap rates were somewhat higher at the points near the center of the respective iteration distributions. One possible explanation for this phenomenon might be that the DP method managed the test-retest overlap rates similarly well for each ability level, but due to the more frequent administrations of items appropriate for the middle part of the respective iteration distributions, the chance of having such items overlapped was relatively greater. In addition, the performance of the DP method must have been affected by the properties and structures of the item pool.

Conditional Standard Errors of Measurement

The CSEM curves produced by using the various iteration distributions are displayed in Figure 4. The large conditional sample sizes of 3,000 examinees led to the smooth curves in the configuration. Because there were fewer items appropriate for administration for both extreme ability levels, the SEMs at the two ends were higher than those at the middle part of the ability scale. The effect of guessing caused higher SEMs at the lower end than at the upper end of the scale.

Figure 4 reveals that the CSEM curves yielded by using the various iteration distributions were almost the same for the SH algorithm, except for the $N(-1,1)$ where the CSEMs were somewhat higher at the lower end. The application of the various sets of exposure control parameters resulted in almost the same SEMs at each ability level of the scale. The reason might be that the CSEMs, the square root of the average squared difference between the estimated and the expected value of the estimated ability

across CAT administrations at a particular theta point, were mainly determined by the features of the exposure control algorithm itself. As long as the exposure control parameters were stabilized, as was the case in the current study, the errors would be similar for this procedure.

Figure 4 displays that the DP strategy resulted in higher CSEMs than the SH procedure. These results were stronger at the two extremes than at the middle range. The price for better control of the item exposure rates with the DP algorithm was seen in the loss of measurement precision. This is, of course, not unexpected. The CSEM curves for the DP strategy were fairly similar also, but with more variation at the two ends of the ability continuum among the four iteration situations.

**The SLC Method**

The results of applying the SLC exposure control parameters to the various ability distributions in operational testing are presented in Tables 5 and 6. Table 5 shows that the mean, minimum and maximum values of the observed exposure rates were almost the same among the various real examinee populations, with only slight differences in the SD values. The numbers and/or percentages of used items were fairly similar for the first four distributions presented, except that the pool utilization for the $N(0,2)$ distribution was somewhat greater. Compared to the degree of similarity of the results in Table 5, the variation of the average peer-to-peer test overlap rates among the five distributions appears to be slightly greater, as can be seen in Table 6. The $N(-1, 1)$ distribution resulted in the greatest mean overlap rate, followed by the $N(0,0.5)$. The $N(0,2)$ distribution led to the smallest mean overlap rate. The real population distributed as a $N(0,2)$ had relatively more examinees at the two ends of the ability scale, items appropriate for these two ends were therefore forced to be utilized, which increased the pool utilization and at the same time, reduced the overlap percentages. On one hand, these findings seem contradictory to the results in Table 5 that the performance of the SLC procedure was independent of the real examinee populations in controlling the maximum observed exposure rates. On the other hand, however, these phenomena may reinforce the previous argument that the distribution of the real examinee population played a major role in determining the results of the pool usage and the peer-to-peer overlap rates.

Nevertheless, these observations should have provided sufficient evidence to support the claim that the performance of the SLC procedure in its control of the maximum observed exposure rates is not affected by the examinee populations in the operational CAT situations. By developing the exposure control parameters with respect to each ability point, the SLC procedure possessed this appealing advantage of being independent of any real examinee population in achieving item security.

## Summary and Conclusions

### Summary of Results

The current study examined how the item exposure control algorithms in the CAT context were affected by developing the exposure control parameters using an examinees' ability distribution that was different from the real examinee population. This study also attempted to verify whether the performance of the conditional exposure control method of SLC was independent of the ability distribution of any real examinee population. The results of this study are summarized below.

#### The Effects of Non-Optimal Specifications of Examinees' Ability Distribution for the SH and DP Methods

For the SH and DP procedures, only when the distributions used for the parameter derivation and the real examinee population matched perfectly were the observed maximum exposure rates controlled to the desired rate of .20. Especially with the SH algorithm, the observed maximum exposure rate could otherwise reach a very high value, causing a great item security concern. The DP methodology was less vulnerable to the changes in the ability distributions. For these two methods, when the mean values of both intended and real distributions were closer, the degree to which the observed maximum exposure rates exceeded the desired level seemed to be at a smaller amount than that when the means were farther apart. These outcomes indicated that a well-projected mean value of the real examinees' ability distribution might reduce the chance of running into a high item security risk.

The curves of the conditional maximum observed exposure rates at each ability level for the SH algorithm varied significantly among the four iteration distributions of $N(0,1)$, $N(0,2)$, $N(-1,1)$ and $N(1,1)$. The curves for the DP strategy were less different among these distributions. It was seen that the conditional maximum exposure rates were lower at points near the center of the respective iteration distributions and the values increased at points away from that mean. That the results were much stronger with the SH procedure than the DP procedure was an indication that the performance of the SH algorithm was more sensitive to the iteration distributions.

In terms of the item pool utilization, the findings revealed that a perfect match between the distributions used for the parameter derivation and the real examinee population did not result in the best use of the item pool for either procedure. Instead, it was the real examinee distribution of $N(0,2)$ that led to the best pool usage. Similar phenomena were discovered in the results of the average peer-to-peer overlap rates for both SH and DP algorithms. The smallest values of the average peer-to-peer overlap rates were not seen under the optimal specification of the distributions, but mostly under the specific real distribution of $N(0,2)$. How the ability of the examinees was distributed in the operational CAT administrations seemed to play a major role in determining the results of the pool utilization or the

peer-to-peer overlap rates. Of course, the characteristics of the item pool is an important factor that cannot be ignored.

As to the results regarding the test-retest overlap rates for the SH method, the patterns across the entire ability scale were similar to those for the conditional maximum observed exposure rates. The test-retest mean overlap rates were low at theta points near the mean of the respective iteration distributions and the values increased outwards, suggesting that the location of the mean value of the iteration distribution seems to be a determining factor in the results of the test-retest overlap rates for the SH algorithm. However, the patterns for the DP method were in a somewhat opposite direction for this type of overlap rate, which might be an implication that the power of the DP procedure in controlling the test-retest overlap rates offset the distribution implemented for the parameter derivation.

For either of the SH and DP methods, imposing different iteration distributions for the parameter derivation did not result in substantially different measurement precision. The employment of the different sets of exposure control parameters derived through the various intended ability distributions produced almost the same CSEMs at all ability levels.

The Performance of the SLC Method

The SLC method yielded almost the same results of the observed exposure rates when the five distributions of the real examinee populations were applied. These findings provided sufficient evidence to confirm that the performance of the SLC algorithm in controlling the observed maximum exposure rates was independent of any examinee population in the operational CAT administrations. Similar results among the five real examinee distributions were found for the pool utilization, but for the $N(0,2)$, the utilization of the pool was somewhat greater. The differences of the average peer-to-peer test overlap rates were relatively larger where the $N(0,2)$ distribution resulted in the smallest overlap rate among the five ability distributions. The outcomes seem to imply that the pool utilization and the peer-to-peer overlap rates were more determined by the nature of the real examinee distributions. Also, the effects of the item pool properties and structures might have been substantial, although not being investigated in the present study.

## Conclusions

The effects of the changes in the examinees' ability distributions on the performance of the SH and DP methods were explored in the current study. The findings suggested that the differences between the examinees' ability distributions for the derivation of the exposure control parameters and in the operational testing situations caused the observed maximum exposure rates to be higher than the pre-specified rate. Especially with the SH strategy, the maximum exposure rates could far exceed the desired level, posing a great item security concern. One way to remedy this situation is to set the

maximum exposure rate specified prior to the derivation of the parameters at a value slightly lower than actually desired. By doing so, the maximum observed exposure rates resulting from the operational CAT administrations would be closer to the desired level, although still being subject to the effects of the distribution changes. As to the question of how much lower would be deemed appropriate, further studies are needed.

The appeal of developing the exposure control parameters with respect to each ability level was supported by the results of this study. The performance of the SLC algorithm was independent of the examinee populations in the operational CAT administrations in maintaining the observed maximum exposure rate at the target level or in other words, in achieving the item security anticipated. Therefore, considering the effects of a non-optimal specification of the real examinees' ability distribution, the conditional exposure control method of SLC has much to recommend it, not to mention the advantages that have been reported in the CAT literature (e.g., Chang, 1998; Chang et al., 2000; Stocking & Lewis, 1998).

The findings for both SH and DP methods revealed that when the mean values of the real examinee populations were better projected, the effects of a non-optimal specification of the distribution might be mitigated. A better knowledge of the examinees' ability distribution in operational testing, particularly about their average ability levels, might help to ease the concern of overexposing some popular items. However, since only combinations of the normal distributions for both intended and real distributions were simulated, the generalizations may be limited to the types of ability distributions that were being varied in the present study. Systematic investigations using more combinations of the distributions are needed to expand the current results. Asymmetric distributions such as being negatively and positively skewed are, in fact, often seen in the real examinee populations; explorations might be continued with these types of distributions. Also, it would be beneficial if the ability distributions of the examinees in operational testing could be simulated based on realistic data, so the effects of the ability changes on the item exposure rates would be better predicted.

The results indicated that the distributions of the real examinee populations played a decisive role in determining the degree to which the pool was utilized and the items overlapped. An optimal match of the intended and real distributions did not produce the most desirable outcomes. The distribution, $N(0,2)$, in the operational testing situations mostly led to the largest extent of the pool utilization as well to the smallest extent of the peer-to-peer overlap percentages. Different distributions of the real examinee populations might be simulated to ensure that these outcomes were mostly affected by the ability distributions of examinees in the operational CAT administrations, rather than the differences between the intended and real ability distributions.

Since the discriminating power of items is associated with the difficulty location along the ability

distribution, the quality and/or the number of items appropriate for each ability point in the pool must be a factor in the outcomes of this study. Further studies might also be focused on the effects of changes in the examinees' ability under item pools of different characteristics or complex adaptive testing situations. Also, studies are suggested to provide more information about the effects of the ability changes in relation to the item pool size and the desired exposure rate. Could the impacts of the changes be lessened by enlarging the item pool size or relaxing the desired exposure rate? Questions such as these remain to be answered.

In conclusion, the changes in the examinees' ability distribution affected the performance of both unconditional procedures of SH and DP where the exposure rates of items were actually controlled to the extent that the real ability distribution of examinees was similar to that intended. The effects were more profound for the SH procedure than for the DP method. By developing the exposure control parameters in light of each ability level, the SLC procedure functioned independently of the real examinee populations. The SLC procedure best served the purpose of controlling the observed exposure rates to the desired values.

Efforts have been made in CAT to directly control the exposure rates of optimal items by incorporating statistical methods into the item selection procedure. Since differences likely occur between the derivation stage and the operational testing situation, the exposure control methods may not fully accomplish the goal of controlling item exposure. The results of this study have provided valuable insights on how the exposure control methods are affected by the distribution changes and have provided evidence to warrant the advantage of the conditional exposure control method. Suggestions have been offered to mitigate such an effect to better control the exposure rates of items in practice.

## References


Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Chang, S. W. (1998). *A comparative study of item exposure control methods in computerized adaptive testing.* Unpublished doctoral dissertation, The University of Iowa. Iowa City, IA.

Chang, S. W., Ansley, T. N., Lin, S. H. (2000, April). *Performance of item exposure control methods in computerized adaptive testing: Further explorations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Davey, T., & Fan, M. (2000, April). *Specific information item selection for adatpvie testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359-375.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*(4), 287-304.

Mislevy, R. J., & Bock, R. D. (1990). Item analysis and test scoring with binary logistic models. *BILOG 3*. Chicago, IL: Scientific Software, Inc.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*(350), 351-356.

Parshall, C. G., Davey, T., & Nering, M. L. (1998, April). *Test development exposure control for adaptive testing*. Paper presented at the annual meeting the National Council on Measurement in Education, San Diego.

Pastor, D. A., Dodd, B. G., Chang, H. H. (2000, April). *A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*(1), 57-75.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*(3), 277-292.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*(2), 151-166.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

Way, W. D. (1997, March). *Protecting the integrity of computerized testing item pools*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Table 1. Results of Observed Exposure Rates for the SH Method
by Real Examinee Distribution and Iteration Distribution

| Iteration Distribution | Real Examinee Distribution | N(%) | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|---|---|
| N(0,1) | N(0,1) | 279(46.5%) | 0.10753 | 0.08527 | 0.00002 | 0.20634 |
| | N(1,1) | 277(46.2%) | 0.10830 | 0.11574 | 0.00002 | 0.53684 |
| | N(-1,1) | 275(45.8%) | 0.10909 | 0.12936 | 0.00002 | 0.52178 |
| | N(0,0.5) | 274(45.7%) | 0.10949 | 0.09862 | 0.00002 | 0.26600 |
| | N(0,2) | 282(47.0%) | 0.10638 | 0.07841 | 0.00002 | 0.26698 |
| N(0,2) | N(0,1) | 271(45.2%) | 0.11070 | 0.09629 | 0.00002 | 0.27634 |
| | N(1,1) | 269(44.8%) | 0.11152 | 0.11819 | 0.00002 | 0.42002 |
| | N(-1,1) | 271(45.2%) | 0.11070 | 0.12726 | 0.00002 | 0.41124 |
| | N(0,0.5) | 265(44.2%) | 0.11321 | 0.11583 | 0.00002 | 0.36648 |
| | N(0,2) | 279(46.5%) | 0.10753 | 0.08311 | 0.00002 | 0.20714 |
| N(1,1) | N(0,1) | 267(44.5%) | 0.11236 | 0.11224 | 0.00002 | 0.48792 |
| | N(1,1) | 269(44.8%) | 0.11152 | 0.08443 | 0.00002 | 0.20380 |
| | N(-1,1) | 262(43.7%) | 0.11450 | 0.17318 | 0.00002 | 0.81794 |
| | N(0,0.5) | 259(43.2%) | 0.11583 | 0.13214 | 0.00002 | 0.55312 |
| | N(0,2) | 272(45.3%) | 0.11029 | 0.09659 | 0.00002 | 0.49064 |
| N(-1,1) | N(0,1) | 278(46.3%) | 0.10791 | 0.09927 | 0.00002 | 0.49116 |
| | N(1,1) | 277(46.2%) | 0.10830 | 0.14342 | 0.00002 | 0.72592 |
| | N(-1,1) | 281(46.8%) | 0.10676 | 0.08746 | 0.00002 | 0.20490 |
| | N(0,0.5) | 272(45.3%) | 0.11029 | 0.12039 | 0.00002 | 0.50778 |
| | N(0,2) | 284(47.3%) | 0.10563 | 0.08276 | 0.00002 | 0.45942 |

Note. The descriptive statistics were based on items that were used at least once.

Table 2. Results of Observed Exposure Rates for the DP Method
by Real Examinee Distribution and Iteration Distribution

| Iteration Distribution | Real Examinee Distribution | N(%) | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|---|---|
| N(0,1) | N(0,1) | 556(92.7%) | 0.05396 | 0.05993 | 0.00002 | 0.20876 |
| | N(1,1) | 554(92.3%) | 0.05415 | 0.05737 | 0.00002 | 0.30710 |
| | N(-1,1) | 548(91.3%) | 0.05474 | 0.06702 | 0.00002 | 0.32870 |
| | N(0,0.5) | 546(91.0%) | 0.05495 | 0.07028 | 0.00002 | 0.25602 |
| | N(0,2) | 559(93.2%) | 0.05367 | 0.05026 | 0.00002 | 0.19764 |
| N(0,2) | N(0,1) | 573(95.5%) | 0.05236 | 0.05933 | 0.00002 | 0.25392 |
| | N(1,1) | 570(95.0%) | 0.05263 | 0.05680 | 0.00002 | 0.30498 |
| | N(-1,1) | 570(95.0%) | 0.05263 | 0.06518 | 0.00002 | 0.33724 |
| | N(0,0.5) | 563(93.8%) | 0.05329 | 0.06885 | 0.00002 | 0.30336 |
| | N(0,2) | 578(96.3%) | 0.05190 | 0.04996 | 0.00002 | 0.21064 |
| N(1,1) | N(0,1) | 559(93.2%) | 0.05367 | 0.05361 | 0.00002 | 0.25786 |
| | N(1,1) | 554(92.3%) | 0.05415 | 0.06136 | 0.00002 | 0.21028 |
| | N(-1,1) | 554(92.3%) | 0.05415 | 0.05617 | 0.00002 | 0.30066 |
| | N(0,0.5) | 551(91.8%) | 0.05445 | 0.06188 | 0.00002 | 0.29882 |
| | N(0,2) | 562(93.7%) | 0.05338 | 0.04554 | 0.00002 | 0.22638 |
| N(-1,1) | N(0,1) | 559(93.2%) | 0.05367 | 0.05777 | 0.00002 | 0.31348 |
| | N(1,1) | 558(93.0%) | 0.05376 | 0.05101 | 0.00002 | 0.37102 |
| | N(-1,1) | 557(92.8%) | 0.05386 | 0.07074 | 0.00002 | 0.21254 |
| | N(0,0.5) | 551(91.8%) | 0.05445 | 0.06710 | 0.00002 | 0.34058 |
| | N(0,2) | 567(94.5%) | 0.05291 | 0.04901 | 0.00002 | 0.29246 |

Note. The descriptive statistics were based on items that were used at least once.

Table 3. The Peer-to-Peer Test Overlap Rates for the SH Method
by Real Examinee Distribution and Iteration Distribution

| Iteration Distribution | Real Examinee Distribution | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|---|
| N(0,1) | N(0,1) | 0.17602 | 0.17867 | 0.00000 | 0.90000 |
| | N(1,1) | 0.23054 | 0.22212 | 0.00000 | 0.96667 |
| | N(-1,1) | 0.26100 | 0.22407 | 0.00000 | 0.96667 |
| | N(0,0.5) | 0.19729 | 0.16427 | 0.00000 | 0.83333 |
| | N(0,2) | 0.16376 | 0.19664 | 0.00000 | 0.96667 |
| N(0,2) | N(0,1) | 0.19306 | 0.19307 | 0.00000 | 0.90000 |
| | N(1,1) | 0.23577 | 0.22320 | 0.00000 | 0.93333 |
| | N(-1,1) | 0.25740 | 0.20856 | 0.00000 | 0.90000 |
| | N(0,0.5) | 0.23076 | 0.19049 | 0.00000 | 0.83333 |
| | N(0,2) | 0.17226 | 0.19902 | 0.00000 | 0.93333 |
| N(1,1) | N(0,1) | 0.22425 | 0.22999 | 0.00000 | 1.00000 |
| | N(1,1) | 0.17520 | 0.17246 | 0.00000 | 0.93333 |
| | N(-1,1) | 0.37433 | 0.28384 | 0.00000 | 1.00000 |
| | N(0,0.5) | 0.26454 | 0.22527 | 0.00000 | 0.96667 |
| | N(0,2) | 0.19438 | 0.23348 | 0.00000 | 1.00000 |
| N(-1,1) | N(0,1) | 0.19973 | 0.20463 | 0.00000 | 0.96667 |
| | N(1,1) | 0.29747 | 0.26134 | 0.00000 | 1.00000 |
| | N(-1,1) | 0.17827 | 0.14580 | 0.00000 | 0.90000 |
| | N(0,0.5) | 0.24279 | 0.20623 | 0.00000 | 0.96667 |
| | N(0,2) | 0.16980 | 0.20241 | 0.00000 | 1.00000 |

Table 4. The Peer-to-Peer Test Overlap Rates for the DP Method by Real Examinee Distribution and Iteration Distribution

| Iteration Distribution | Real Examinee Distribution | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|---|
| N(0,1) | N(0,1) | 0.12059 | 0.09116 | 0.00000 | 0.53333 |
| | N(1,1) | 0.11443 | 0.08167 | 0.00000 | 0.50000 |
| | N(-1,1) | 0.13714 | 0.08290 | 0.00000 | 0.50000 |
| | N(0,0.5) | 0.14495 | 0.09178 | 0.00000 | 0.56667 |
| | N(0,2) | 0.10095 | 0.08794 | 0.00000 | 0.53333 |
| N(0,2) | N(0,1) | 0.11950 | 0.08724 | 0.00000 | 0.50000 |
| | N(1,1) | 0.11343 | 0.07905 | 0.00000 | 0.50000 |
| | N(-1,1) | 0.13263 | 0.07916 | 0.00000 | 0.50000 |
| | N(0,0.5) | 0.14273 | 0.08799 | 0.00000 | 0.66667 |
| | N(0,2) | 0.09972 | 0.08419 | 0.00000 | 0.53333 |
| N(1,1) | N(0,1) | 0.10690 | 0.08296 | 0.00000 | 0.50000 |
| | N(1,1) | 0.12351 | 0.09146 | 0.00000 | 0.53333 |
| | N(-1,1) | 0.11269 | 0.07181 | 0.00000 | 0.53333 |
| | N(0,0.5) | 0.12513 | 0.08192 | 0.00000 | 0.50000 |
| | N(0,2) | 0.09220 | 0.08287 | 0.00000 | 0.50000 |
| N(-1,1) | N(0,1) | 0.11532 | 0.08824 | 0.00000 | 0.60000 |
| | N(1,1) | 0.10220 | 0.07344 | 0.00000 | 0.60000 |
| | N(-1,1) | 0.14712 | 0.09402 | 0.00000 | 0.56667 |
| | N(0,0.5) | 0.13677 | 0.08741 | 0.00000 | 0.53333 |
| | N(0,2) | 0.09758 | 0.08751 | 0.00000 | 0.56667 |

Table 5. Results of Observed Exposure Rates for the SLC Method by Real Examinee Distribution

| Real Examinee Distribution | N(%) | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|---|
| N(0,1) | 488(81.3%) | 0.06148 | 0.04684 | 0.00002 | 0.19984 |
| N(1,1) | 488(81.3%) | 0.06148 | 0.04912 | 0.00002 | 0.19902 |
| N(-1,1) | 489(81.5%) | 0.06135 | 0.06083 | 0.00002 | 0.19858 |
| N(0,0.5) | 486(81.0%) | 0.06173 | 0.05653 | 0.00002 | 0.19996 |
| N(0,2) | 495(82.5%) | 0.06061 | 0.03783 | 0.00002 | 0.19762 |

Note. The descriptive statistics were based on items that were used at least once.

Table 6. The Peer-to-Peer Test Overlap Rates for the SLC Method by Real Examinee Distribution

| Real Examinee Distribution | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|
| N(0,1) | 0.09722 | 0.08078 | 0.00000 | 0.43333 |
| N(1,1) | 0.09999 | 0.08022 | 0.00000 | 0.50000 |
| N(-1,1) | 0.12204 | 0.08331 | 0.00000 | 0.50000 |
| N(0,0.5) | 0.11281 | 0.07920 | 0.00000 | 0.46667 |
| N(0,2) | 0.08452 | 0.08117 | 0.00000 | 0.46667 |

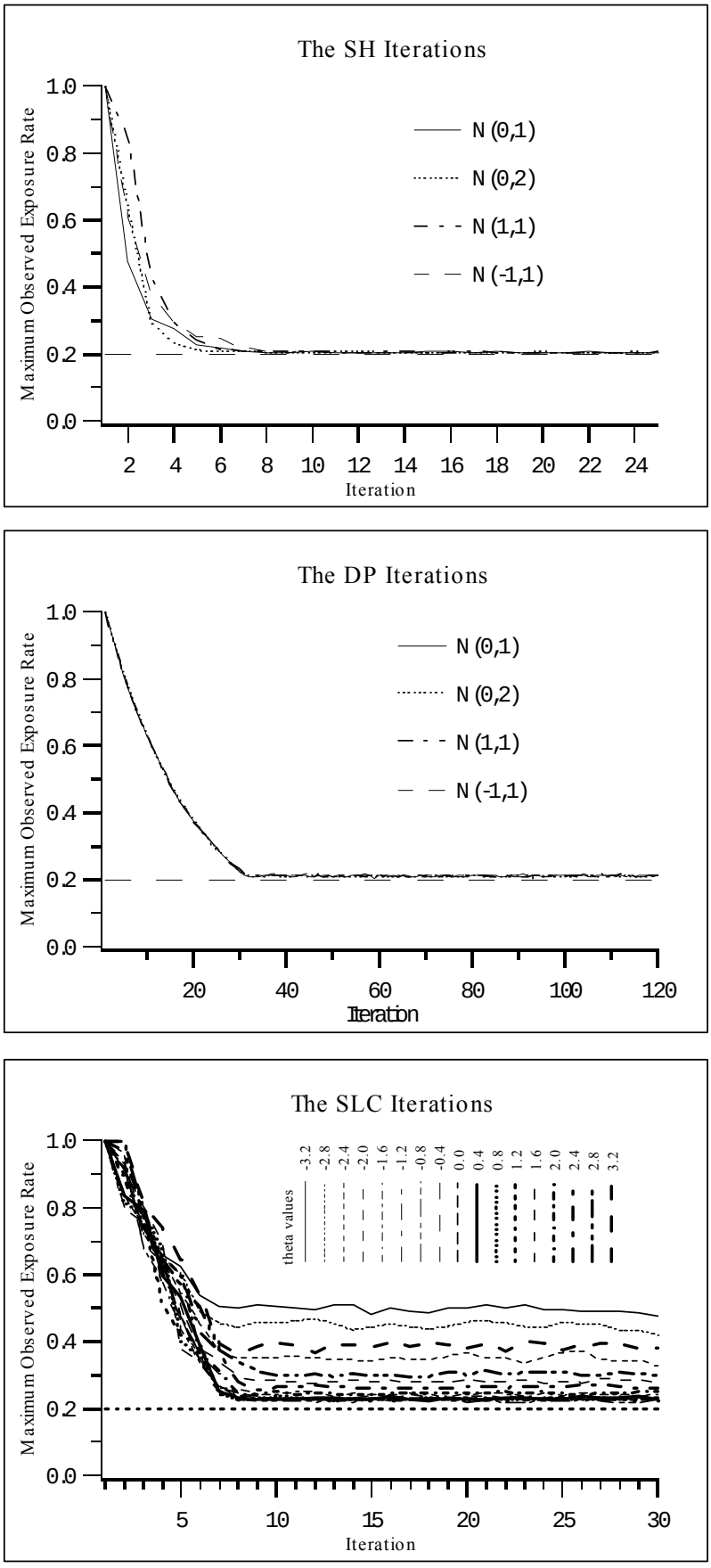
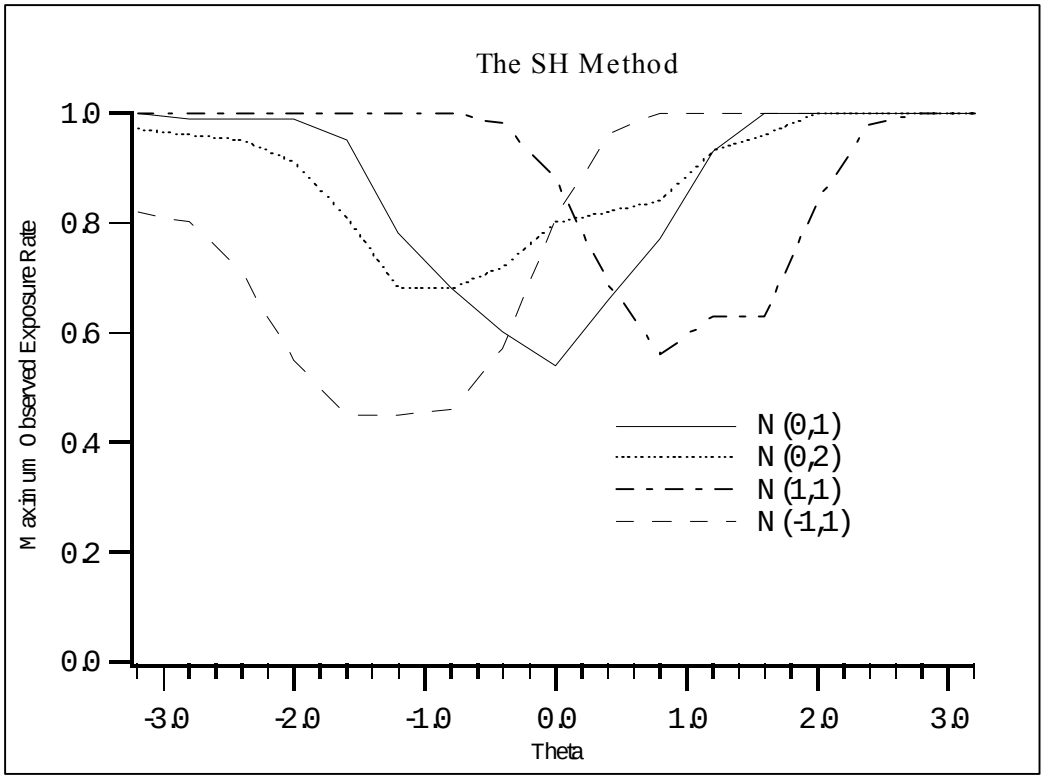



Figure 1. Results of Iterations for the Various Procedures



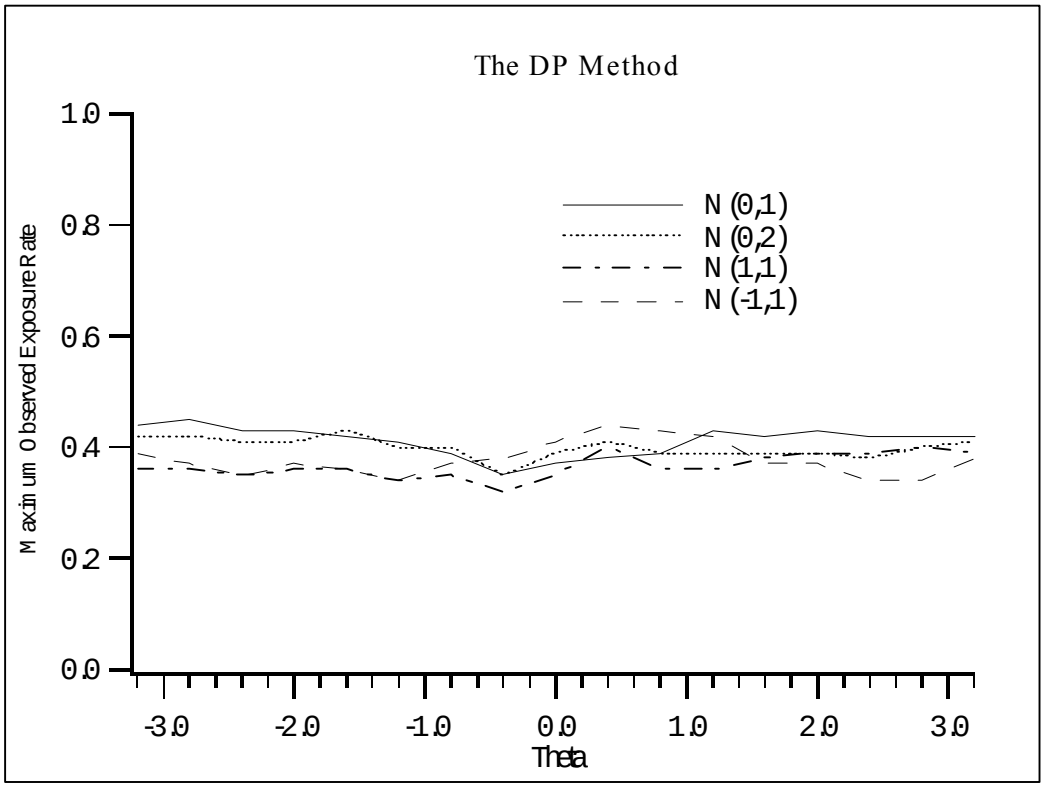
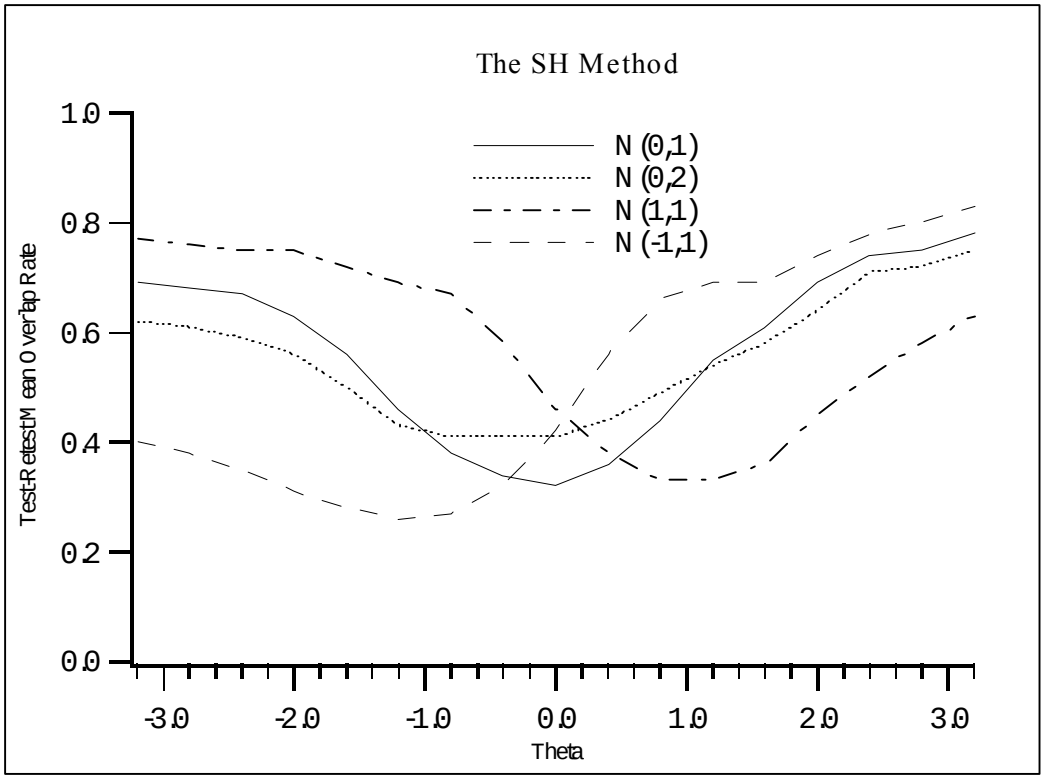




Figure 2. Conditional Maximum Observed Exposure Rates
by Iteration Distribution and Exposure Control Method



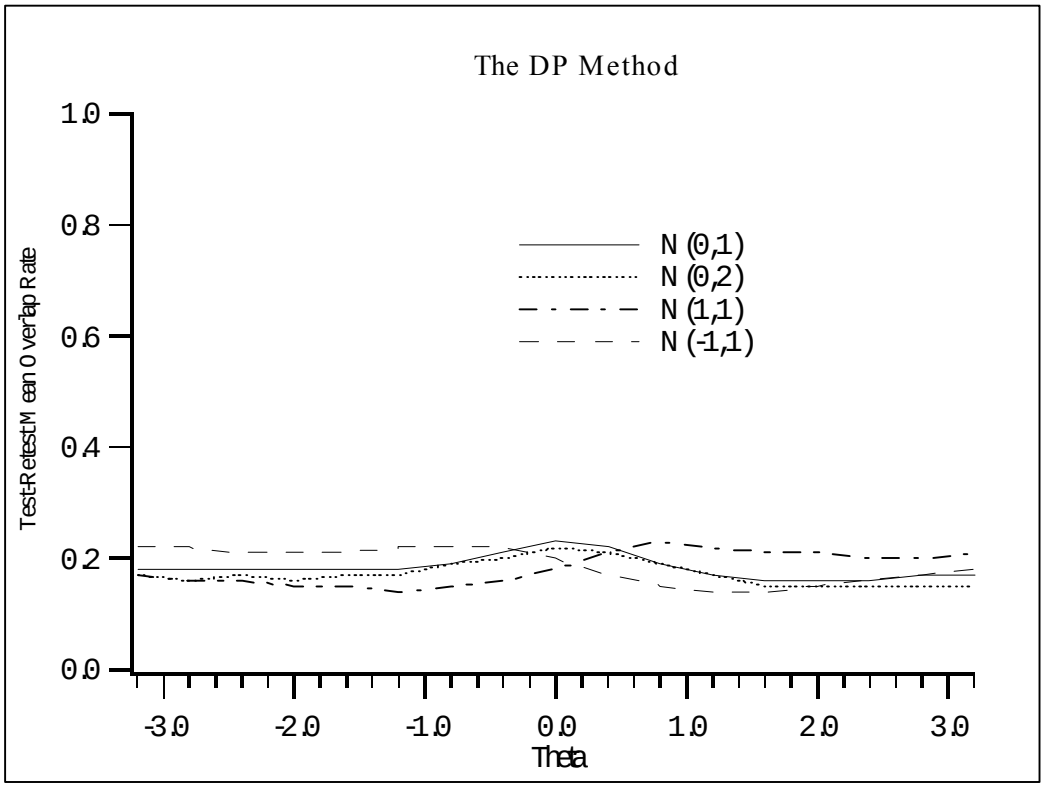
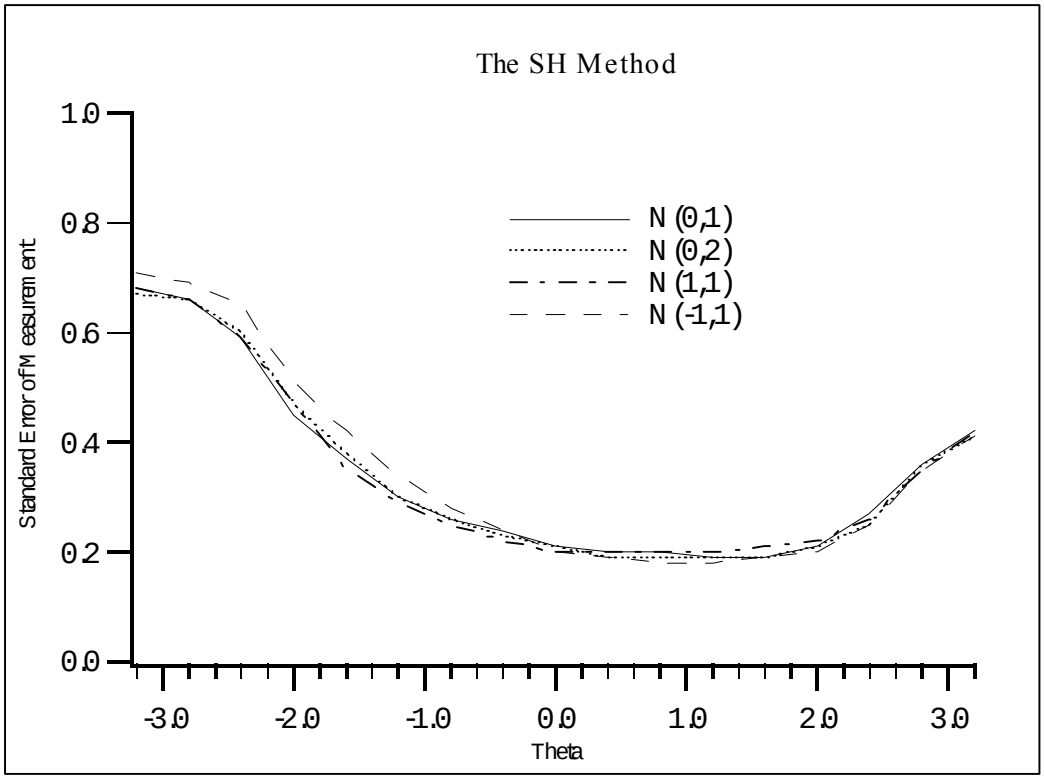




Figure 3. Test-Retest Mean Overlap Rates by Iteration Distribution and Exposure Control Method



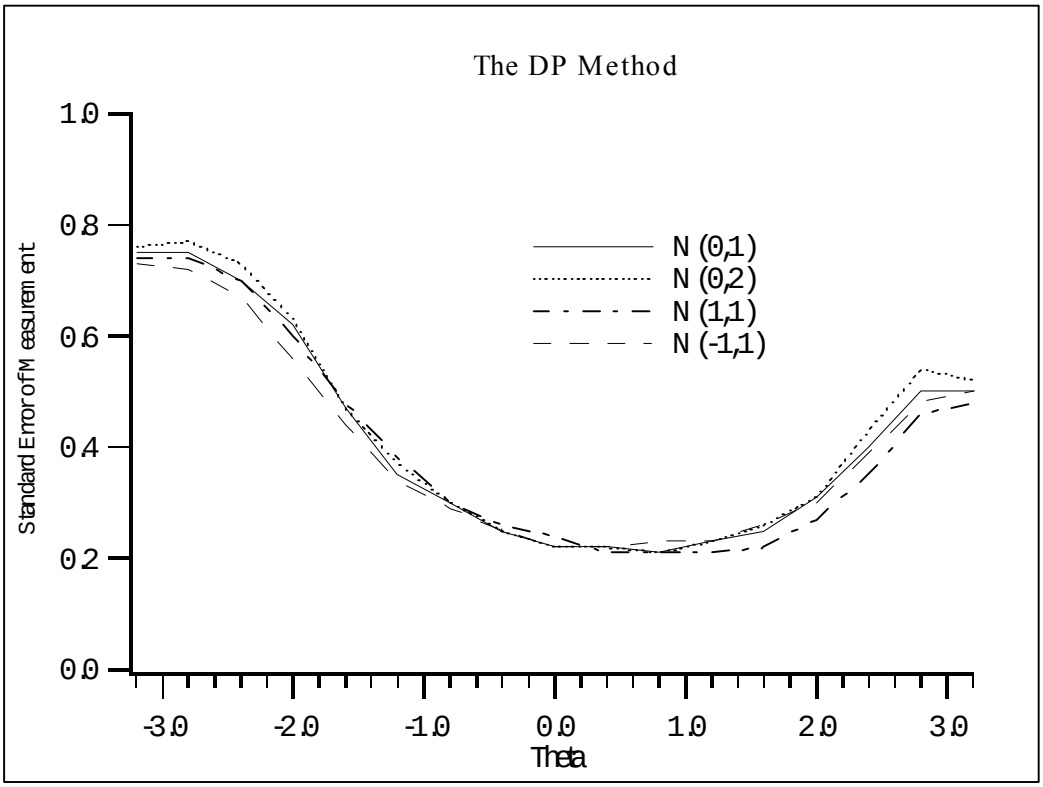




Figure 4. Standard Errors by Iteration Distribution and Exposure Control Method