ED 389 740 TM 024 369

AUTHOR Stocking, Martha L.; Lewis, Charles
TITLE Controlling Item Exposure Conditional on Ability in Computerized Adaptive Testing.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-24
PUB DATE Aug 95
NOTE 41p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; *Adaptive Testing; Algorithms; *Computer Assisted Testing; Simulation; *Test Items
IDENTIFIERS *Item Exposure (Tests); Large Scale Assessment; Paper and Pencil Tests; Randomization; Test Security

ABSTRACT

The interest in the application of large-scale adaptive testing for secure tests has served to focus attention on issues that arise when theoretical advances are made operational. Many such issues in the application of large-scale adaptive testing for secure tests have more to do with changes in testing conditions than with testing paradigms. One such issue is that of insuring item and pool security in the continuous testing environment made possible by the computerized administration of a test, as opposed to the more periodic testing environment typically used for linear paper-and-pencil tests. In the continuous testing environment of adaptive testing, methods have been developed in the past to use the computer to control the rate at which particular items are exposed to test-takers. These methods have typically employed randomization schemes, sometimes in reference to a particular target distribution of test-taker ability. This paper presents a new multinomial method of controlling the exposure rate of items conditional on the ability level of an individual test-taker. The properties of such conditional control on the exposure rates of items, when used in conjunction with a particular adaptive testing algorithm, are explored through five studies with simulated data. (Contains 1 table, 7 figures, and 17 references.) (Author)

RR-95-24

RESEARCH REPORT

CONTROLLING ITEM EXPOSURE CONDITIONAL ON ABILITY IN COMPUTERIZED ADAPTIVE TESTING

Martha L. Stocking
Charles Lewis

Educational Testing Service
Princeton, New Jersey
August 1995

CONTROLLING ITEM EXPOSURE CONDITIONAL ON ABILITY IN COMPUTERIZED ADAPTIVE TESTING

Martha L. Stocking
Charles Lewis

Educational Testing Service
Princeton, New Jersey 08541

August, 1995

CONTROLLING ITEM EXPOSURE CONDITIONAL ON ABILITY IN COMPUTERIZED ADAPTIVE TESTING

Abstract

The interest in the application of large-scale adaptive testing for secure tests has served to focus attention on issues that arise when theoretical advances are made operational. Many such issues have more to do with changes in testing conditions rather than testing paradigms. One such issue is that of insuring item and pool security in the continuous testing environment made possible by the computerized administration of a test, as opposed to the more periodic testing environment typically used for linear paper-and-pencil tests. In the continuous testing environment of adaptive testing, methods have been developed in the past to use the computer to control the rate at which particular items are exposed to test-takers. These methods have typically employed randomization schemes, sometimes in reference to a particular target distribution of test-taker ability. This paper presents a new method of controlling the exposure rate of items conditional on the ability level of an individual test-taker. The properties of such conditional control on the exposure rates of items, when used in conjunction with a particular adaptive testing algorithm, are explored through five studies with simulated data.

Key Words: computerized adaptive testing, conditional exposure control, exposure rates, test security.

CONTROLLING ITEM EXPOSURE CONDITIONAL ON ABILITY IN COMPUTERIZED ADAPTIVE TESTING

Introduction

Recent advances in psychometrics and computing technology have led to the development of a testing paradigm that is very different from linear paper-and-pencil testing -- computerized adaptive testing (CAT; see, for example, Eignor, Way, Stocking & Steffen, 1993; Lord, 1977; Schaeffer, Steffen & Golub-Smith, 1993, Stocking & Swanson, 1993; Wainer, Dorans, Flaugher, Green & Mislevy, 1990). As interest in large-scale implementation of modern adaptive testing has increased, particularly for high-stakes testing programs (Jacobson, 1993), increasing attention has been focussed on issues that arise when theoretical advances are made operational (see, for example, Mills & Stocking, 1995).

Some of these issues stem less from changes in testing paradigms and more from changes in testing conditions. An example of such an issue is insuring the security of items and tests. In linear paper-and-pencil testing, large numbers of candidates take a single form or parallel forms of a test at administration dates scheduled throughout some time period. In this context, the frequency with which a single item is seen by a single test-taker can be tightly controlled in advance of testing through policies that regulate both the reuse of test forms and the frequency with which candidates may retake the test. This system of test administration may be called periodic testing.

Adaptive tests are tests in which items are selected from a large pool of items to be appropriate for a test-taker (the test "adapts" to the test-taker). All but a few proposed designs have assumed that items would be chosen and administered to test-takers on a computer. In the context of adaptive testing, the computer itself, and in particular the item selection algorithm, can be used to implement measures that control the frequency with which an individual test-taker encounters a particular item. In this environment, continuous, as opposed to periodic, testing becomes feasible.

It is, of course, possible to conceive of conventional paper-and-pencil testing in a continuous testing environment, although the security problems of such an administrative mode may be difficult to overcome for reasonable cost. Likewise, it is also possible to conceive of CAT in a periodic testing environment, although this would fail to capitalize on the convenience of computer administration.

This paper addresses the issue of insuring item and item pool security in the environment of continuous testing offered by CAT. By this we do not mean the physical security of items and pools on computers at remote testing sites with data transmission among them. These issues are presumably addressed by encryption and security methodologies from the field of computer science as well as by administrative procedures designed to control the physical security of computers and computer storage devices. Rather, we address issues related to the frequency with which particular items are selected for inclusion in an adaptive test.

All of the published methods designed to control the frequency of item administration in CAT do so without regard to individual test-taker ability. Some of them, in particular the Sympson & Hetter (1985) procedure and the multinomial procedure (Stocking & Lewis, 1995) control the frequency with reference to a particular target population of test-taker ability. These methods may insure, say, that a particular item is not administered to more than 20% of the test-takers in a target population. However, a more detailed examination may show that this item is administered to 100% of the high ability test-takers, even though it is administered to no more than 20% of the test-takers overall. In this paper we present a new method of exposure control that attempts to rectify this problem by limiting the frequency of administration conditional on ability level.

This is in contrast to the method outlined in Davey & Parshall (1995). The Davey & Parshall methodology, which may also be termed 'conditional', limits the frequency with which an item can be administered, conditional on all other items that have already been included in an adaptive test. This conditional approach, while intriguing, may be difficult to implement in a practical context. In addition, it does not directly control the too-frequent administration of some items to test-takers at some ability levels.

In the next section we briefly provide more details about CAT and a particular adaptive testing algorithm that we will use in the later examples. The following sections describe the unconditional multinomial method of exposure control of Stocking & Lewis (1995) and present its conditional version. The final section contains an exploration of some of the properties of conditional exposure control.

Adaptive Testing With the Weighted Deviations Model

As pointed out by Davey & Parshall (1995) high-stakes adaptive testing has at least three goals: 1) to maximize test efficiency by selecting the most appropriate items for a test-taker, 2) to assure that the tests measure the same composite of multiple traits for each test-taker by controlling the nonstatistical nature of items included in the test, and 3) to protect the security of the item pool by controlling the rates at which items can be administered. These goals often compete with one another.

Different approaches to each of these goals yield different algorithms for adaptive testing. The particular algorithm used in this paper is the Weighted Deviations Model (WDM) developed by Swanson & Stocking (1993) and Stocking & Swanson (1993). This paradigm is characterized by flexible approaches to all three goals of adaptive testing.

In general, any CAT algorithm implicitly orders the items in the pool in terms of their desirability for selection as the next item. Differences in ordering typically reflect particular definitions of item optimality and particular methods of estimating test-taker ability. Any attempt to control the exposure of items can then be viewed as modifications imposed on this ordering.

In the WDM the item pool is ordered by employing a methodology from the decision sciences that models the behavior of expert test specialists. The WDM ordering explicitly takes into account nonstatistical item properties or features along with the statistical properties of items. This is to insure that each adaptive test produced from a pool matches a set of test specifications and is therefore as parallel as possible to any other test in terms of content and type of items, while being tailored to an individual test-taker in terms of appropriateness. The desired balance between measurement and construct concerns is reflected by the weights given to them, which are chosen by the test designer. The WDM approach also allows specification of overlapping items that may not be administered in the same adaptive test. In addition, it is possible to restrict item selection to blocks of items, either because they are associated with a common stimulus or common directions or any other feature that test specialists deem important. Thus at each item selection in the WDM, the pool or an appropriate subset of the pool is ordered from most desirable (smallest weighted deviations from desirable test properties) to least desirable (largest weighted deviations from desirable test properties).

In summary, in the WDM, the next item selected for administration is the item that simultaneously

1) is the most appropriate possible at a test-taker's estimated ability level, while

2) contributing as much as possible to the satisfaction of all other constraints.

At the same time, it is required that the item

3) does not appear in an overlap group containing an item already administered, and

4) is in the current block (if the previous item was in a block), starts a new block, or is in no block.

In the particular version of the WDM used in this paper, the measure of the appropriateness of the item is the Fisher item information function (Lord, 1980, equation 5-9) and the estimate of ability is maximum likelihood (Lord, 1980, equation 4-31), although other measures of the statistical properties of items (see for example, Chang, 1995) and other estimates of ability (see for example, Davey & Parshall, 1995) are possible.

Controlling Item Exposure With the Unconditional Multinomial Method

The multinomial method of Stocking & Lewis (1995) can be viewed as having two distinct, although simultaneous, phases. In the 'adjustment' phase, exposure control parameters are developed for each item, using the methodology of Sympson & Hetter (1985). We call this the adjustment phase in that we are adjusting the exposure control parameters after each iterative simulation. During the adjustment phase, as well as after it, the exposure control parameters for each item are used in a 'selection' phase to over-ride the optimal selection of the next item employing the multinomial method. These two phases are detailed in this section.

The Adjustment Phase

In the adjustment phase, the Sympson & Hetter procedure considers a test-taker randomly sampled from a typical group of test-takers and distinguishes between the probability $P(S)$ that an item is selected as the best next item by some item selection algorithm, and $P(A|S)$, the probability that an item is administered, given that it has been selected. The procedure seeks to control the overall probability that an item is administered, $P(A)$, where $P(A) = P(A|S) * P(S)$, and to insure that the maximum value over all $P(A)$s is less than some specified maximum value r.

The 'exposure control parameters', $P(A|S)$, one for each item, are determined through a series of adjustment simulations using an already established adaptive test design and simulated examinees (simulees) drawn from a typical distribution of ability. Following each simulation, the proportion of times each item is selected as the best item, $P(S)$, and the proportion of times each item is administered, $P(A)$, are separately tallied. If $P(S)$ is less than or equal to the desired maximum, then $P(A|S)$ is set to one for the next iteration, insuring that $P(A) = P(A|S)*P(S) \leq r$. If $P(S)$ is greater than r, then $P(A|S)$ is set to $r/P(s)$ for the next iteration, again insuring that $P(A) \leq r$. The iterative adjustment simulations continue until the $P(A|S)$ have stabilized and the maximum observed $P(A)$ for all items is approximately equal to the desired value of r.

The Selection Phase

When selecting the next item in each iterative adjustment simulation (and with the final exposure control parameters once they have been established), the simplest version of multinomial item selection phase proceeds as follows (see Stocking & Lewis, 1995, for more complex extensions):

a) Form a list of elements ordered by their desirability. We use 'elements' here to indicate that the list may be a mixture of discrete items and stimuli for sets of items. Any method of ordering by desirability is, of course, possible, although in this paper we use the WDM described previously.

b) For each element i in the list, form the operant probabilities k_i, where

$$k_i=\{\prod_{j=1}^{i-1} (1-P_j)\}*P_i \quad .$$

P_i is $P_i(A|S)$, the exposure control parameter for item i. k_i is the joint probability that all items before item i are rejected given selection and item i is administered given it is selected.

c) If necessary, adjust the operant probabilities so that they sum to one by dividing them by their unadjusted sum.

d) Form the cumulative distribution.

e) Generate a random number uniformly distributed between zero and one. Find the corresponding element in the cumulative distribution.

f) Remove all elements proceeding the one selected from further consideration in this adaptive test.

g) If the element selected is a stimulus for a set of items, repeat steps a) through e) for items belonging to this set.

The Conditional Multinomial Method

The unconditional multinomial method described above results in an exposure control parameter, $P_i(A|S)$, for each element i in a pool. The adjustment phase develops these P_i in reference to a particular distribution of ability θ in a relevant population of test-takers.

An approach to developing a conditional version of the multinomial method is to consider the range of θ covered by the distribution of ability, and divide this range into, say, M different discrete values of θ that cover the range of interest. Consider one particular discrete value in this range, θ_m. We can perform iterative adjustment simulations to develop exposure control parameters for each element i in the pool using the multinomial procedure and in reference to only those simulees drawn that have true ability equal to θ_m. (There can be, of course, as many simulees as desired whose true ability is θ_m.)

These iterative adjustment simulations result in a vector of exposure control parameters appropriate for all the elements in the pool for individuals with ability θ_m. If the adjustment simulations are performed simultaneously for all values of θ_m, m = 1, . . . , M, we produce a matrix of conditional exposure control parameters with elements $P_i(A|S,\theta_m)$ for the ith row (one for each element in the pool) and the mth column (one for each of the θ_m that span the range of true abilities in the target population of interest).

During the adjustment phase, when the conditional exposure control parameters are being developed, the tallies of item selection and administration are kept separately by θ level. For the item selection phase, the simulee's estimated ability level, $\hat{\theta}$, is used to select the appropriate column in the matrix of conditional exposure control parameters to use in the selection of the next item.

The advantage of conditional multinomial exposure control is that it allows direct control of item exposure for different levels of ability. Moreover, desirable maximum exposure rates, r_m, can be chosen to be different values for different ability levels, perhaps reflecting the availability of items in the pool. In addition, the conditional parameters are now independent of any target population of test-takers.

However, the conditional multinomial exposure control method retains the disadvantages of unconditional multinomial exposure control, as well as the Sympson & Hetter adjustment procedure. That is, the conditional exposure control parameters are dependent upon the specific item pool and test structure used in the iterative adjustment simulations. Moreover, the conditional control of exposure makes the adjustment process of developing the exposure control parameters even more time-consuming and tedious than when exposure control is unconditional, although there is some practical evidence that this need not be so (see question 4 in the next section).

Explorations of Conditional Multinomial Exposure Control

In this section we explore a series of questions concerning the properties of conditional exposure control using the multinomial method. All of the explorations depend upon simulation studies, not data from real test-takers. Furthermore, all of the explorations used a single pool and adaptive test design.

The Item Pool

Available to the authors was a large pool of items and sets of items measuring various aspects of quantitative reasoning. There were a total of 516 elements in the pool -- 494 items and 22 stimuli. Of the 494 items, 153 were associated with the 22 stimuli and the remaining 341 items were discrete items. The items were calibrated on large samples (2000+) of test-takers from the paper-and-pencil testing program using the 3-parameter logistic item response model and the computer program LOGIST (Wingersky, 1983). They were placed on a common IRT metric using the transformation methods of Stocking and Lord (1983). The mean item discrimination was .82 with a standard deviation of .34; the mean item difficulty was .03 with a standard deviation of 1.23; and the mean pseudo-guessing parameter was .14 with a standard deviation of .11.

The Adaptive Tests

Items were drawn from this pool using the WDM to form (fixed length) adaptive tests of 28 items, subject to 27 constraints on their content. The constraints had relative weights that varied from 11, indicating that it was very important for an adaptive test to have items and/or stimuli with these features, to 1, indicating that it was less important for an adaptive test to have these features. The importance of measurement appropriate for a test-taker was reflected in the weighting of the item information function at 10.

In addition to this test structure, item selection was further restricted by the specification of 83 overlap groups. Items and stimuli belonging to an overlap group may not appear in the same adaptive test with other items and stimuli appearing in the same overlap group. When a stimulus appears in an overlap group, all items associated with that stimulus are included by implication. There were a total of 326 entries in the 83 overlap groups.

The adaptive test was designed to be as parallel as possible to an existing 60-item linear test of quantitative reasoning, both in terms of the construct measured and measurement properties. To facilitate comparisons, the adaptive test score was the estimated number-correct true score on the metric of the linear test, rather than $\hat{\theta}$, which would be a more natural test score when working in an Item Response Theory (IRT) framework.

The Simulations

The adaptive test simulations were conducted with uniform distributions of simulated examinees (simulees) across (nearly) equally spaced values on the score reporting metric, starting from about the chance score level and ending close to the top of the range. This results in values on the θ metric that are unequally spaced. In addition, to facilitate unconditional comparisons, a particular target population was established. The target population was estimated for these same (nearly) equally spaced values on the score reporting metric, using the method of Mislevy (1984) and a sample of over 6000 real test-takers who took an exemplar form of the linear test.

Question 1: Do the adjustment iterations converge for conditional exposure control parameters?

Answer: Yes, but not to the targeted conditional values.

The iterative adjustment simulations were first performed with the unconditional multinomial method of exposure control using a target maximum probability of administration of .2. Ten adjustment simulations were conducted, each of which used 100 simulees at each of 11 values spaced along the score reporting metric, for a total of N = 1100 simulees per iteration. (Sympson & Hetter recommend the use of at least 1000 simulees.)

The results of the iterative adjustment simulations are shown in Figure 1a. Four curves are plotted in this figure: the largest observed probability of administration for each adjustment simulation for discrete items, stimuli, items in sets, and the target maximum probability of .2. The maximum observed probabilities for the unconditional procedure converge smoothly to values slightly above the target value of .2. It is likely that fewer iterations would have been satisfactory, since there is little difference between the results of, say, the fifth iteration and the tenth iteration. At iteration 10, the adaptive test reliability for the target population, computed from the conditional standard errors using the estimate suggested by Green, Bock, Humphreys, Linn & Reckase (1984, equation 6), was .93. In this final adjustment iteration, 14 out of 22 stimuli were used from the pool as well as 263 out of the 494 items.

This series of adjustment simulations was then repeated using the conditional multinomial method of exposure control. The same number of simulees at the same ability levels was used and the target conditional maximum observed probabilities of administration were specified as .2. All other aspects of the unconditional adjustment simulations were left unchanged. The results of the conditional exposure control adjustment simulations are shown in an unconditional fashion in Figure 1b. Even though the unconditional probability of administration was not controlled in these iterations, the same smooth convergence to values around .2 is seen in Figure 1b as in Figure 1a.

At iteration 10, the adaptive test reliability for the target population is .92 in contrast to .93 for unconditional exposure control. In addition to decreasing test reliability slightly, conditional exposure control also increases pool utilization -- 19 stimuli and 403 items were used in contrast to 14 stimuli and 263 items when exposure control was unconditional.

The satisfactory measurement of the intended construct for both series of adjustment simulations is indicated by the average weighted deviation shown in Figure 1c for each adjustment simulation. To compute the average weighted deviation for an adjustment simulation, information concerning the extent of violations in each adaptive test for each nonstatistical constraint on item selection was weighted by the relative weight assigned to that constraint and then averaged over the 1100 simulees to give an estimate of the average weighted deviation for the target population. The price for conditionally controlling exposure rates is seen in the increase in the average weighted deviation, which reaches about .08 per simulee in contrast to .06 per simulee when using unconditional exposure control. This decrease in desirable test properties is due to the fact that for some ability levels the pool is less satisfactory than for other ability levels.

These sets of iteration adjustment simulations indicate that conditional exposure control may slightly increase the deviations from desirable test properties, slightly decrease test reliability, and increase pool usage. The increase in deviations and decrease in reliability were not unexpected and their magnitude for this pool and test structure were judged tolerably small. The increase in pool usage is a positive consequence of conditional exposure control in that it is implied by the obtained reduction in item exposure.

Overall, conditional exposure control appears to give satisfactory results. However, these results must also be analyzed in a conditional fashion. Figure 2a displays the conditional results, for the discrete items in the pool, in a manner comparable to Figures 1a and 1b. In Figure 2a, the behavior of the observed maximum for each adjustment simulation is plotted separately for each of the 11 true score levels. Also plotted is the target of .2 used for each true score level. Similar to looking at the results unconditionally in Figure 1b, the conditional procedure appears to be converging relatively smoothly. However, it seems to be converging to a value around .3 rather than the target of .2.

Figure 2b shows a different method of examining the conditional results. In this figure, the horizontal axis represents the 11 true score levels at which the procedure attempted to control the maximum probability of administration. The vertical axis is the maximum observed probability of administration for each iterative adjustment simulation at each true score level. The thinner lines plotted are the results of each conditional adjustment simulation. The thick line is the result of controlling item exposure unconditionally, as in Figure 1a.

After the fourth iteration, conditional exposure control produces conditional observed maximum probabilities of administration that are smaller than those obtained when the exposure control is unconditional, particularly at extreme (high or low) true score levels. The conditional maximum observed probability of administration for these extreme levels does not approach that for middle true score levels until iteration eight. The last three iterations (8, 9, and 10) produce similar values of conditional maximum observed probabilities of administration, indicating again that the conditional procedure is converging. However, these last three iterations again indicate that the conditional procedure is converging to a value greater than the target value.

Question 2: Why does the conditional procedure converge to unexpected values? Answer: Distributions of maxima of sets of numbers can produce unexpected results.

One possible explanation for these unexpected results is the use of small sample sizes at each ability level in the above experiment -- 100 simulees. It is plausible to speculate that these relatively small conditional sample sizes lead to large sampling variability. Indeed, Figures 2a and 2b show that the conditional procedure converges less smoothly than the unconditional procedure (with 1100 simulees) shown in Figure 1a. However, it is unlikely that sampling variability as we usually think of it could simultaneously fail to cause severe interference with convergence while also causing the procedure to approach incorrect values.

It is possible that perhaps there is something unique about the structure of this item pool, these test specifications, and the target observed maximum probabilities of administration of .2 that interact to make it possible to meet such a target of .2 when controlling exposure unconditionally, but not conditionally. To test this hypothesis, the conditional iterative adjustment simulations were repeated with conditional targets of .1 and .3 for all true score levels. These adjustment simulations converged conditionally to slightly above .2 and .4, respectively. No explanation could be generated concerning the structure of the problem that would explain consistent convergence to values about .1 higher than target values.

Neither sampling variability as we usually think of it nor problem structure yielded satisfactory explanations. The likely answer, then, lies in the properties of what is being estimated, namely the maximum of a set of numbers. To better understand the distribution of a maximum of a set of numbers, repeated simulations with different random number seeds using already established exposure control parameters were considered. (For simplicity, we assumed that these are unconditional, although this is not central to the argument.) These repeated simulations are called 'evaluation trials' to distinguish them from adjustment iterations.

For an evaluation trial, each element i has a true probability of administration π_i, and we obtain $P_i(A)$, the proportion of times element i is administered. We wish to explore how the observed maximum compares with the true maximum, that is, how the maximum $P_i(A)$ compares with the maximum π_i. To aid in this exploration, assume that the $P_i(A)$ are mutually independent random variables with normal distributions having means π_i and variances of $\{\pi_i*(1-\pi_i)/N\}$, where N is the total number of simulees (replications) in one evaluation trial.

In reality, the $P_i(A)$s are not independent of each other, although the amount of dependence in a set of exposure control parameters should be small if the adaptive test length is very much smaller than the pool size, as it is here. Moreover, the $P_i(A)$s are not normally distributed, but rather binomially distributed. However, these simplifying assumptions are made to provide a context in which results can actually be computed.

With these assumptions, the distribution of the maximum, that is, the probability that the maximum observed $P_i(A)$ is less than some arbitrary value, Z $(0 \le Z \le 1)$, is

$$Prob(\max(P_i(A)) \le Z) = \prod_{i=1}^{n} Prob(P_i(A) \le Z) = \prod_{i=1}^{n} \Phi\left(\frac{Z-\pi_i}{\sqrt{\pi_i \times (1-\pi_i)/N}}\right). \qquad (1)$$

In this equation, n is the number of elements in the pool, N is the number of simulees in an evaluation trial, and Φ is the cumulative normal distribution function.

To actually compute and plot this distribution requires the π_i. To approximate these reasonably well, 10 evaluation trials were performed using the exposure control parameters from the fifth iteration of the unconditional adjustment iterations shown in Figure 1a and the results averaged over the 10 evaluation trials. Since each evaluation trial involved 1100 simulees, this is equivalent to averaging over 11,000 simulees. The theoretical approximate cumulative distributions of the observed maximum $P_i(A)$, given in equation (1), are plotted in Figure 3 for different values of N.

The maximum of the approximate π_i was .2386. As sample size is increased, the 50th percentile of the distribution of the observed maximum moves closer to the true maximum. Selected percentiles of the four distributions are given in Table 1.

Table 1: Percentiles of the Distributions of Observed Maximum Values

Percentiles	N=110	N=1100	N=11,000	N=110,000
25th	.2923	.2430	.2367	.2377
50th	.3031	.2477	.2390	.2386
75th	.3157	.2532	.2415	.2395
99th	.3564	.2702	.2481	.2416

If the true maximum is around .24, a sample size of 11,000 is necessary to insure that the percentiles between the 25th and the 75th of the distribution of the observed maximum correspond to values that are accurate to two decimal places. The use of a sample size of 110 gives a 50th percentile of around .30, and an interquartile interval of approximately .29 to .32. In the iterative adjustment simulations for the conditional procedure the sample size at each level of ability is only 100. This suggests that the convergence of the conditional procedure to values of around .3 when targets are .2 may be improved by increasing the conditional sample sizes.

Question 3: Does increasing the sample size for each ability level in the iterative adjustment simulations rectify the situation?

Answer: Yes.

The iterative adjustment simulations for both the unconditional (Figure 1a) and conditional (Figures 2a and 2b) methods were repeated with 1000 simulees at each of the 11 ability levels. The results for the conditional procedure are shown in Figures 4a, 4b, and 4c. Figure 4a, when compared to Figure 2a, shows less variability in the conditional maximum observed exposure rate from iteration to iteration. More importantly, the iterations converge to values slightly above .2, as did the unconditional iterations (not shown here). At the tenth iteration, this adaptive test had an estimated reliability of .91, in contrast to .92 for the smaller sample conditional simulations. Using larger conditional sample sizes also increased pool utilization slightly -- 20 stimuli and 418 items were used, in contrast to 19 stimuli and 403 items with the smaller conditional sample sizes.

Figure 4b displays these large conditional sample size results comparable to the smaller conditional sample size results in Figure 2b. In this figure, the thick line is the result of the large sample unconditional iterations. Again the larger conditional sample sizes result in smoother convergence of the procedure to values closer to the target maxima.

Figure 4c displays the average weighted deviation for the small conditional sample size adjustment simulations (solid line, repeated from Figure 1c) and the large conditional sample size iterations (dotted line). The average weighted deviation remains similar across the adjustment simulations for the two conditional experiments.

For both the small and large sample size conditional iterative adjustment simulations, it seems clear that at least eight iterations are needed. This is in contrast to the unconditional results shown in Figure 1a, where five iterations seemed sufficient. Thus the procedure to develop the conditional control of exposure rates in adaptive testing seems to require not only larger (conditional) sample sizes than for unconditional control, but also more iterative adjustment simulations.

Question 4: Given that larger (conditional) sample sizes and more adjustment simulations are required, is there anything that we can do to shorten the process of obtaining conditional exposure control parameters?

Answer: Yes.

In their original description of the adjustment simulations to develop exposure control parameters for a target population, Sympson & Hetter recommended that initial values of all P(A|S) be set to 1.0. This recommendation is made, at least in part, because in the selection procedure used by Sympson & Hetter, there must be at least n items in the pool with exposure control parameters of 1.0 to guarantee the administration of an n-item adaptive test. Since it cannot be known in advance which n exposure control parameters are best set to 1.0, Sympson & Hetter recommended that the iterative adjustment simulations be started with all of them set to 1.0, and adjusted from this starting value as outlined previously.

With the multinomial method of exposure control, this requirement disappears. Using this method, the administration of an n item test is guaranteed by step e) above, in which the sum of the operant probabilities is adjusted to 1.0 if necessary. Therefore it makes sense to explore the use of different starting values for the adjustment simulations.

The conditional exposure control adjustment simulations with conditional sample sizes of 1000 were repeated, with starting values for all $P_i(A|S,\theta)$ set to .2. The results for the conditional maximum probabilities of administration for six iterations are shown in Figures 5a and 5b, which can be compared to Figures 4a and 4b. The differences are striking; starting values close to target values are obviously effective in reducing the number of iterations required.

Careful examination of Figures 5a and 5b shows that the conditional maximum observed probabilities of administration were low for the first iteration and are slightly higher for subsequent iterations. This indicates that a single iteration is not enough to be confident that the procedure has converged. However, this examination indicates that the procedure probably has converged by the third simulation, that is, after two adjustments to the starting values. For the third simulation, the test reliability was .91, all 22 stimuli were used, as were 434 items, compared to test reliability of .91, 20 stimuli, and 418 items used after 10 iterations with starting values of 1.0

for the conditional exposure control parameters.

Figure 5c displays the average weighted deviation for the (large sample) conditional exposure adjustment iterations when the starting value for the exposure control parameters was 1.0 (solid line, partially repeated from Figure 4c) and when it was .2 (dotted line). The average weighted deviation for the first iteration when starting values are .2 is so large that it cannot be plotted on the same scale as the values for the other iterations (over .50). This is because values of .2 for all conditional exposure control parameters at all abilities are not appropriate for this item pool or test structure. However, after the second iteration, the values for the average weighted deviation become similar to those obtained with starting values of 1.0.

Figures 6a, 6b, and 6c are scatterplots of the conditional exposure control parameters used in the third iteration (after two adjustments) when starting values are all .2 (vertical axis) and those used in the eighth iteration (after seven adjustments) when starting values are all 1.0 (horizontal axis), for three different true score levels. Not shown on these plots are the points that have conditional exposure control parameters equal to 1.0 in both conditions. Out of 516 such possibilities, there are 368 for the lowest true score level, 345 for the middle true score level, and 373 for the highest true score level. In the Sympson & Hetter design, conditional exposure control parameters of 1.0 occur when P(S) is less than or equal to the target r. More elements for the extreme (low and high) true scores have conditional exposure control parameters of 1.0 for both conditions than for the middle true score level, indicating that more elements are selected infrequently for these extreme true score levels. This accords with the nature of the item pool in that there are more items appropriate for administration to middle ability simulees than for simulees with extreme true score levels.

The lowesc value for a conditional exposure control parameter is .2. This value is assigned to elements that are so desirable that they are always selected (P(S) = 1.0), in order to insure that their overall probability of administration, P(A), does not exceed the target maximum of .2. The line on each figure is the 45-degree line. It is not surprising to find that starting from .2 produces lower exposure control parameters than starting from 1.0. The procedures are approaching optimum values from below and from above.

A challenge, then, is to discover some comparison between the two sets of results that leads to a clear choice between the two different starting values. The eighth iteration with starting values of 1.0 had an average weighted deviation of .076 and a test reliability estimated to be .91. The third iteration with starting values of .2 had an average weighted deviation of .075, and test reliability also estimated to be .91, with comparable use of the item pool, and the adjustment iterations took less than 1/3 the time it took for starting values of 1.0. Since there does not appear to be any basis to reject starting close to the desired maximum conditional exposure rates, on the basis of time alone it seems profitable to do so.

Question 5: How does conditional exposure control compare at different ability levels with unconditional Sympson & Hetter exposure control in terms of statistical properties and conformance to the test plan?

Answer: Extreme (high and low) ability levels show more effects than middle ability levels.

We have noted that conditional exposure control decreases overall test reliability slightly and increases the overall lack of conformance with desirable test properties slightly. Thus far, we have made conditional comparisons only in terms of the exposure rates of items. In this section, we examine both measurement and nonstatistical properties conditional on ability and compare the results to the more familiar unconditional Sympson & Hetter exposure control procedure.

Figure 7a makes these comparisons for the conditional standard errors of measurement (CSEM). In this figure, the estimated distribution of true ability is displayed as a bar graph, with values to be read from the right-hand vertical axis. Two comparison curves are plotted as solid lines -- the CSEM curve for the 60-item conventional test to which the adaptive test is designed to be parallel, scored as number correct (the upper solid curve) and as estimated number correct true score (the lower solid curve). Changing the scoring of the conventional test decreases the conditional standard error of measurement.

The dotted line in Figure 7a is the CSEM curve at the end of eight unconditional Sympson & Hetter adjustment simulations with the target maximum exposure rate of .2, and 100 simulees at each ability level. The dashed line is the CSEM curve at the end of six conditional multinomial exposure control iterations with starting and target maximum exposure rates of .2 at all levels of ability and 1000 simulees at each ability level. This latter curve is smoother than the Sympson & Hetter curve because of the larger conditional sample size.

For the lowest and highest three levels of ability, which represent about 13% and 15%, respectively, of a typical population, the conditional standard error of measurement for conditional exposure control is typically about 30% higher than that for the Sympson & Hetter unconditional exposure control; at no point is it greater than about 43%. For the middle five levels of ability, which represent about 72% of a typical population, the CSEM for conditional exposure control is typically about 4% higher than that for the Sympson & Hetter unconditional exposure control; at no point is it greater than about 13%. Thus the greatest measurement penalty due to conditional exposure control is at the extremes of the ability distribution, where the CSEM is typically small to begin with. For the middle 72% of the ability distribution, the penalty is substantially smaller, but generally positive. This differential effect on different ability levels is due to the fact that the item pool contains more items appropriate for middle ability levels. The CSEM for the conditional exposure control condition could be reduced if it were economically feasible to obtain a larger pool with more items statistically appropriate at each ability level, particularly the extreme (high and low) ability levels.

Figure 7b displays the lack of conformance with desirable test properties for the Sympson & Hetter unconditional and multinomial conditional exposure control conditions in terms of the conditional weighted deviations summed across all 28 constraints and averaged across the simulees at a particular ability level. As before, the estimated distribution of true ability is plotted as a bar graph with proportional frequencies to be read from the right-hand vertical axis.

There were no constrain violations in either condition for constraints that received high relative weights. All constraint violations occurred for 13 constraints that had the lowest relative weight of 1. At all ability levels except the top 15% of a typical population, there is some penalty to be paid for conditional exposure control in terms of the average weighted deviation. This penalty is largest for the lowest ability levels. Again, if it were economically feasible to obtain a pool containing more items that reflect desirable test properties at all ability levels, the penalty seen in the conditional exposure control condition could be made to disappear.

Discussion

The interest in the application of large-scale adaptive testing for secure tests has served to focus attention on issues that arise when theoretical advances are made operational. One such issue is that of insuring item and pool security. This paper has concentrated on addressing such concerns through the control of the frequency with which items in a pool are administered.

Previous research concentrated on addressing exposure control either in a random fashion, or with respect to items already administered, or with respect to an overall target distribution of ability. None of these approaches directly controls the exposure of items to test-takers with the same or similar abilities, which can be quite high even though the overall exposure rate is low. In this paper, we demonstrated how a particular method of controlling item exposure, the multinomial method of Stocking & Lewis (1995), can be extended to control item exposure conditional on ability.

The examples demonstrate a number of salient features of this method of exposure control conditional on ability:

1) The conditional sample sizes used in the development of conditional exposure control parameters must be large, preferably 1000 or larger.

2) For the adaptive test studied here, the control of item exposure conditional on ability (holding pool size and test length fixed) decreases estimated test reliability slightly, increases the lack of conformance with desirable test properties slightly, and increases pool utilization substantially. The ≀ctual size of observed effects for any other pool should depend upon the structure of the pool and test specifications at all ability levels at which conditional control is sought.

3) The control of item exposure conditional on ability increases both the CSEM and the lack of conformance to desirable test properties differently at different levels of ability. Generally the deterioration is largest at extreme (high and low) levels of ability and smaller for middle levels of ability. Again, the actual size of observed effects for any other pool should depend upon the conformance of pool and test structure.

4) Substantial time savings can accrue in the iterative adjustment simulations required to establish conditional exposure control parameters by using starting values close to or equal to the desired maximum conditional values of the probability of exposure for each item at each ability level.

Much remains to be done before this method of exposure control conditional on ability can be universally recommended. First, this method needs to be extensively applied to other pools with different structures and test specifications. In particular, studies should be done with other pools that confirm the utility of different starting values. Second, experience with actual item exposure rates obtained from real, as opposed to simulated, adaptive testing on many different pools is strongly desirable.

In addition, it should be noted that although exposure rate is controlled conditional on ability level, it is not controlled with respect to candidate volume. An item with an exposure rate of .1 at the highest ability level will only be seen by approximately 10% of the most able test-takers, but if there are a million highly able test-takers, the absolute exposure could be quite high. Any exposure control methodology that seeks to control exposure rates as opposed to absolute exposure suffers from this criticism, including that of Davey & Parshall. This suggests that future research might profitably begin to focus on the effects of absolute exposure for expected candidate volumes. The development of pool rotation schedules and partial or complete pool replacement methods are crucial to this effort. In addition, it seems important that the consequences for test scores of administering items about which test-takers may have some pre-knowledge must be thoroughly understood before continuous adaptive testing can be seen as a secure alternative to paper-and-pencil testing.

References

Eignor, D. R., Way, W. D., Stocking, M. L., & Steffen, M. (1993). Case studies in computer adaptive test design through simulations. (Research Report 93-56). Princeton, NJ: Educational Testing Service.

Chang, H. (1995). A global information approach to computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education. April, 1995, San Francisco, CA.

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

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347-360.

Jacobson, R. L. (1993, september 13). New computer technique seen producing a revolution in testing. *The Chronicle of Higher Education*, p A22.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, *1*, 95-100.

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

Mills, C. N., & Stocking, M. L. (1995). Practical issues in large-scale computerized adaptive testing. (Research Report 95-xx). Princeton, NJ: Edcuational Testing Service.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.

Schaeffer, G., Steffen, M., & Golub-Smith, M. (1993) Introduction of a computer adaptive GRE general test (Research Report (93-XX). Princeton, NJ: Educational Testing Service.

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7(2), 201-120.

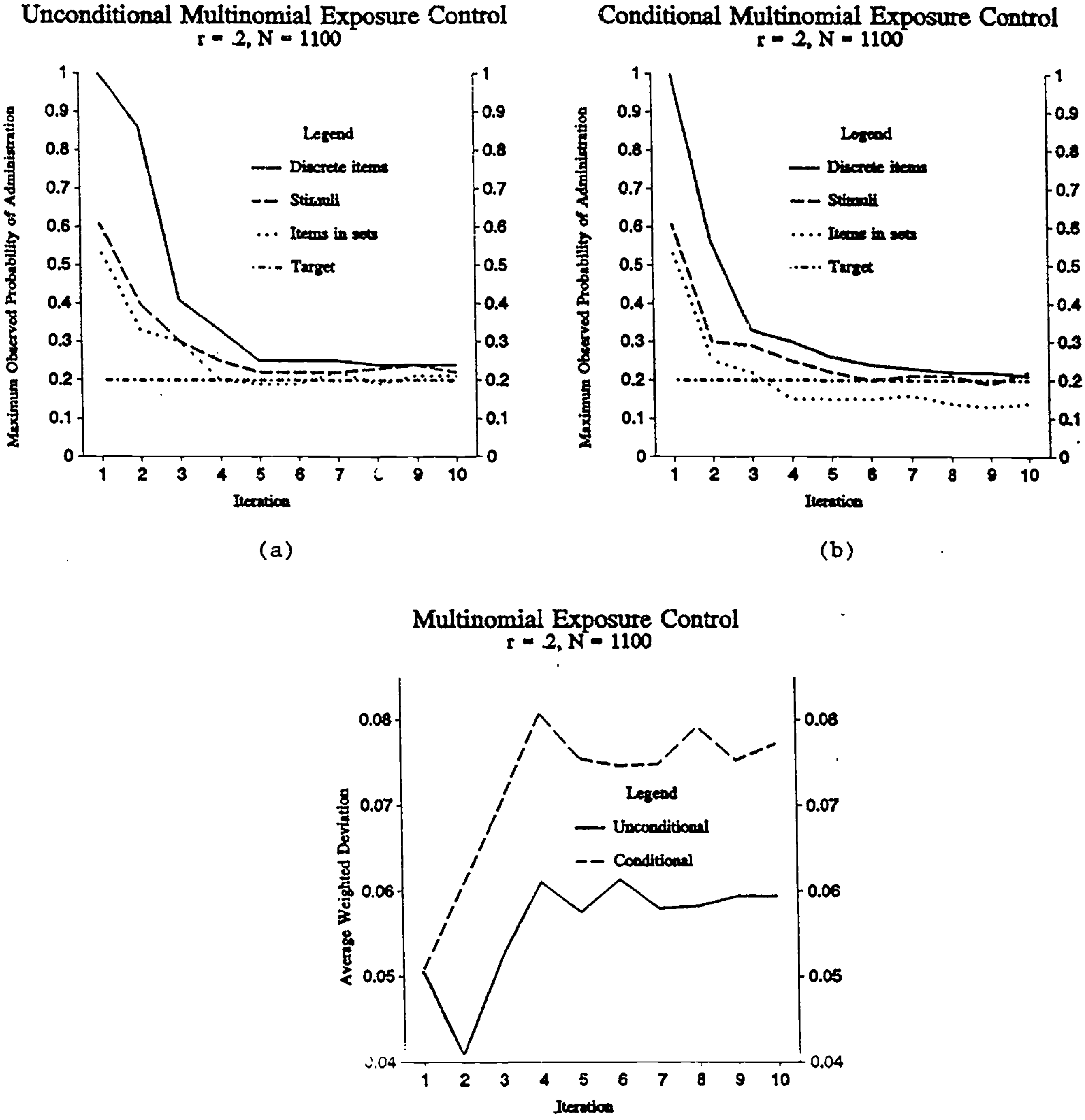
Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17(3), 277-292.

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

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing, as described in Wainer, et al., (1990).

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J.,Steinberg, L., & Thissen, D. (1990). Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wingersky, M. S. (1983) LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

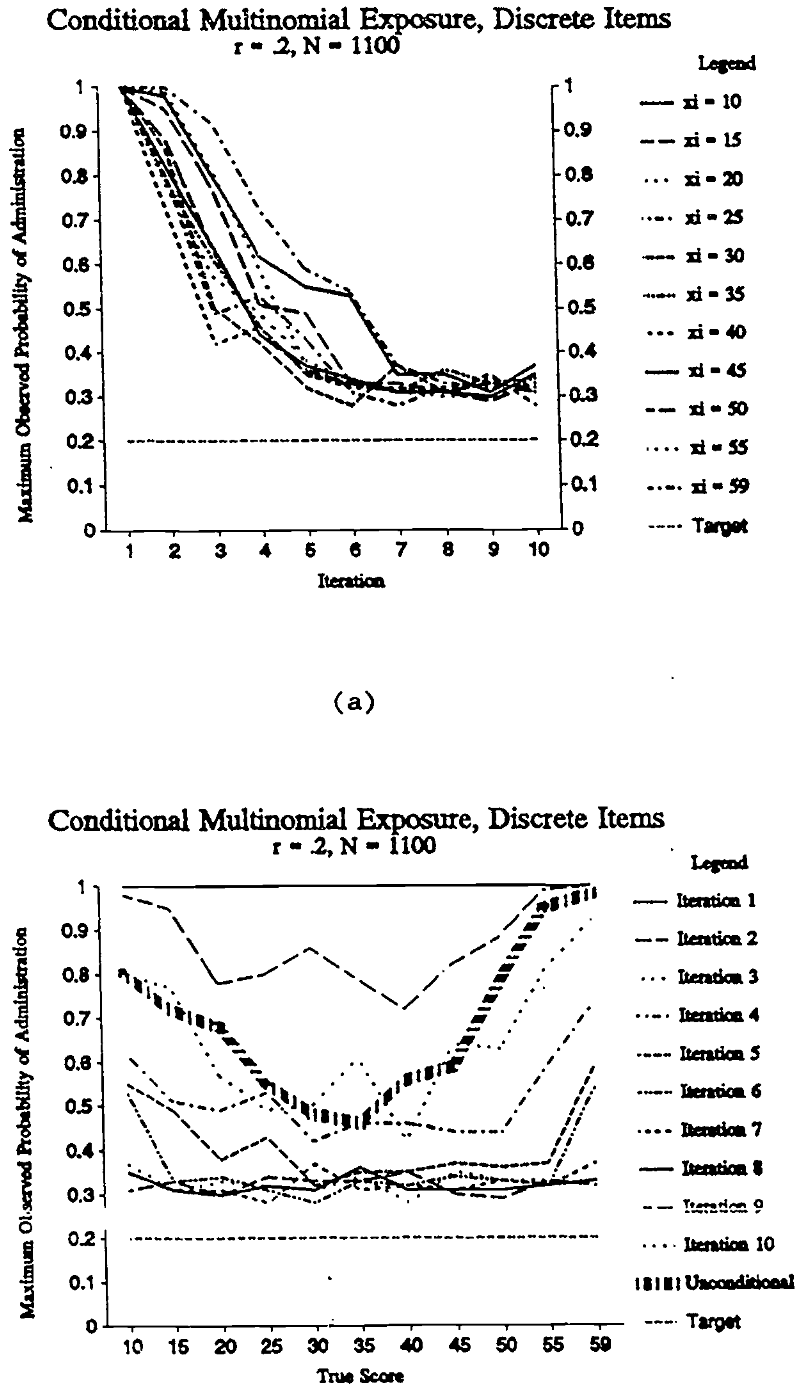


(a)

(b)

(c)

Figure 1: Unconditional and conditional results of adjustment iterations for r = .2 and N = 1100. (Conditional results are shown unconditionally, see text.)



(a)

(b)

Figure 2: Conditional results of adjustment iterations for r = .2, N = 1100.

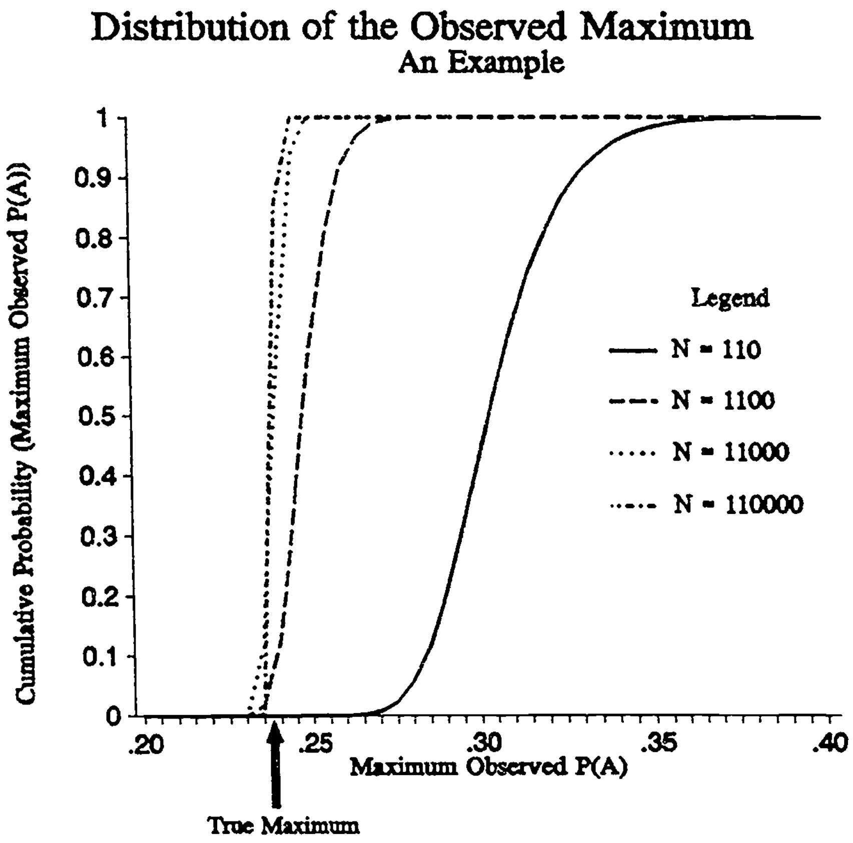
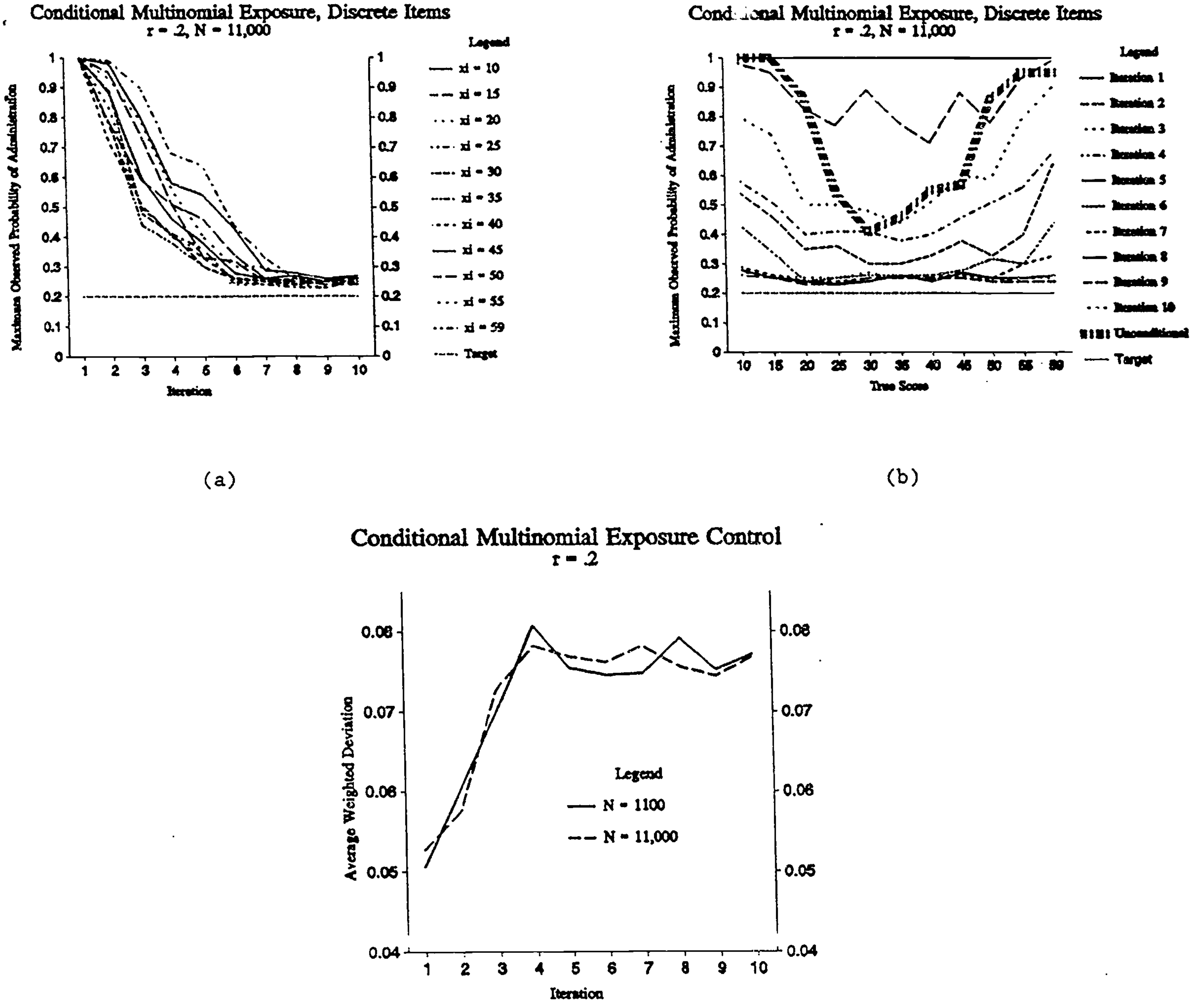


Figure 3: The cumulative distribution function for the observed maximum when the true maximum is .2386, for different sample sizes.

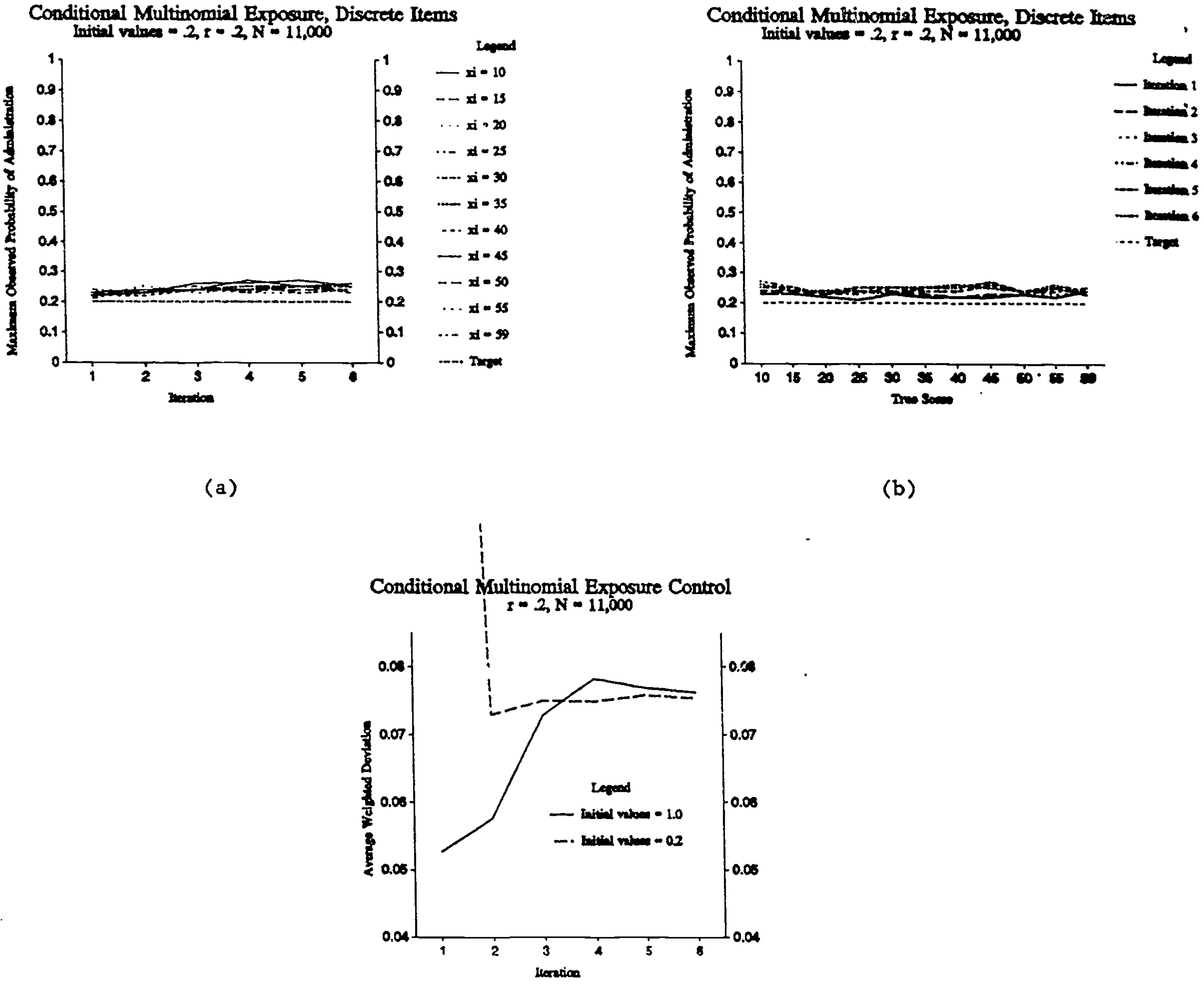


(a)

(b)

(c)

Figure 4: Conditional results of adjustment iterations for r = .2, N = 11,000.

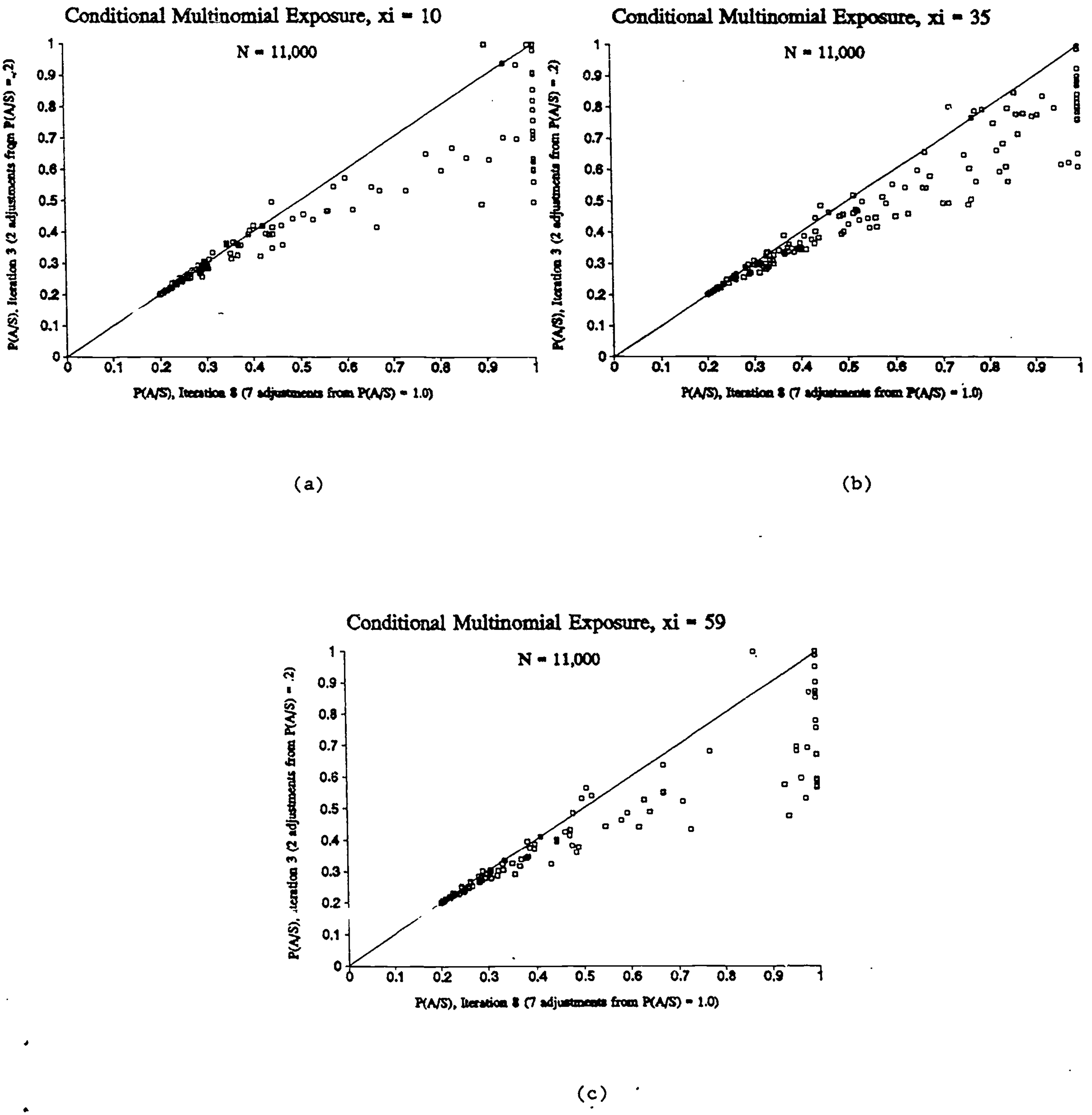


(a)

(b)

(c)

Figure 5: Conditional results of adjustment iterations with starting values of exposure control parameters equal to .2, r = .2, N = 11,000.

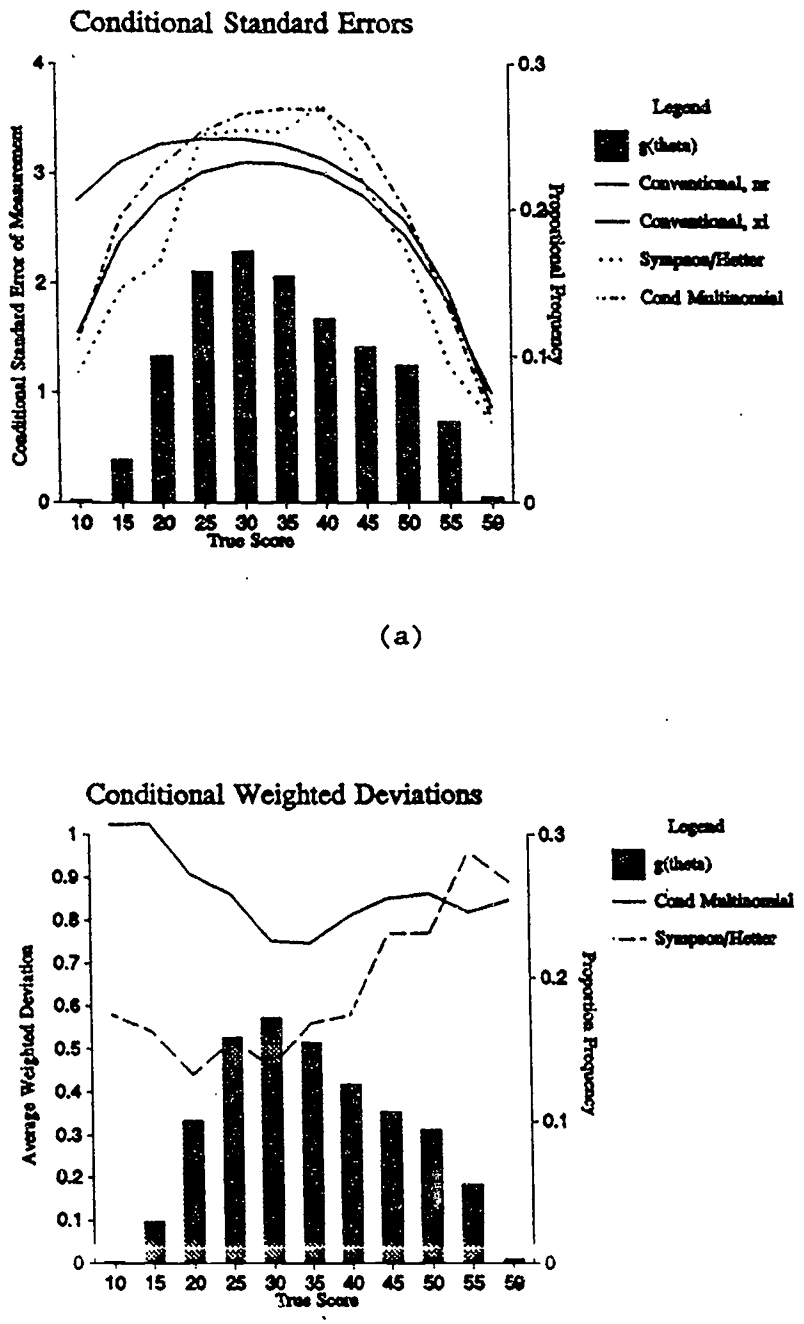


(a)

(b)

(c)

Figure 6: Scatterplots of conditional exposure control parameters after two adjustments from starting values of .2 (vertical axis) vs. seven adjustments from starting values of 1.0, for ξ = 10 (a), 35 (b), and 59 (c).



(a)

(b)

Figure 7: Conditional comparisons of CSEMs (a) and conformance to test properties (b) for unconditional Sympson/Hetter and conditional multinomial exposure control (see text).