ED 411 297 TM 027 312

AUTHOR	Patsula, Liane N.; Steffen, Mandred
TITLE	Maintaining Item and Test Security in a CAT Environment: A Simulation Study. Laboratory of Psychometric and Evaluative Research Report No. 309.
PUB DATE	1997-03-00
NOTE	32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; *Item Banks; *Multiple Choice Tests; Probability; Simulation; Test Construction; Test Items
IDENTIFIERS	*Test Security

ABSTRACT

One challenge associated with computerized adaptive testing (CAT) is the maintenance of test and item security while allowing for daily testing. An alternative to continually creating new pools containing an independent set of items would be to consider each CAT pool as a sample of items from a larger collection (referred to as a VAT) rather than as a permanent collection of items. This research investigated the viability of creating overlapping CAT pools from a static VAT by sampling items with replacement. The VAT used in this simulation consisted of approximately 5,000 precalibrated verbal multiple-choice items. Item pools were created, and simulations were performed to configure each of these item collections into CAT pools and to estimate the proportion of examinees at each of 13 ability levels that were expected to see each item. Probabilities of administration were computed for each item and different numbers of examinees and item pools were simulated. As the number of pools increased, their quality decreased, and the proportion of items reaching the retirement threshold for test items decreased. As more pools were created, the potential for control increased. Results support the use of a VAT system as a way of drawing a compromise between test quality and test security. (Contains five figures, six tables, and five references.) (SLD)

Maintaining Item and Test Security in a CAT Environment: A Simulation Study[1,2]

Liane N. Patsula[3]
University of Massachusetts at Amherst
&
Manfred Steffen[4]
Educational Testing Service

[1] Paper presented at the National Council on Measurement in Education, Chicago, IL, March 1997.

[2] Laboratory of Psychometric and Evaluative Research Report No. 309. University of Massachusetts at Amherst, School of Education.

[3] This research was initiated while the first author was a summer intern with the Graduate Research Examination Program at Educational Testing Service.

[4] The authors would like to thank Ronald Hambleton and Charlie Lewis for their comments on an earlier draft of this paper.

Maintaining Item and Test Security in a CAT Environment: A Simulation Study

Liane N. Patsula

University of Massachusetts, Amherst

&

Manfred Steffen

Educational Testing Service

As is evident in the measurement literature of the past ten years, the use of computerized adaptive testing (CAT) by credentialing agencies to measure knowledge and skills in a particular domain has become increasingly prominent. This can be attributed to the many advantages of CAT. From an assessment standpoint, computerized adaptive tests are more efficient than conventional paper-and-pencil (P&P) tests, typically requiring about half as many items to attain an equivalent level of precision. This is due to the fact that questions in a computerized adaptive test are tailored to an individual examinee's ability level. In addition, CAT offers advantages to test developers of improved test reliability, improved test security and data collection, better opportunity to control cheating, and cost savings with regard to printing and shipping. Advantages to test takers include the convenience and flexibility of scheduling an appointment to test, year-round testing, immediate knowledge of scores, faster score reporting service, and potentially shorter tests (Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990).

However, despite the many stated advantages of CAT, there are also some challenges to be confronted in implementing CAT. These include such practical issues as selecting the first item, choosing the stopping rule, scoring adaptive tests, administering items belonging to sets, controlling item exposure and overlap, providing item review to candidates, dealing with item omissions, allowing for incomplete tests, developing CAT pools, maintaining CAT pools, and complying with disclosure requirements (Mills & Stocking, 1995). One such challenge that is the focus of this study is the maintenance of item and test security while allowing for daily testing. The major challenge facing large-scale CAT programs that test daily is to redefine item and test security with regard to the access of items to examinees.

It is important to emphasize that CAT does not necessarily present the new security challenge, but rather it is the daily access to testing. One could potentially conceive of thousands of students gathering in a gymnasium, one Saturday morning, each provided with his or her own computer to take a CAT. There would only need to be one modest sized item pool, perhaps not very much larger than the length of the P&P version of the exam. This would avoid many of the CAT security issues. The real challenge to CAT test security is the daily access to testing or what some call continuous testing.

Test security issues are not unique to CAT. To lead one to believe that the concern over test security is new with the advent of CAT is misleading. P&P testing programs have and have always had security concerns. Indeed, the driving force for parallel forms is to address security concerns. To insure fairness to examinees by insuring that test forms are equivalent, P&P testing programs need to equate test forms as new forms are introduced at each administration. In addition, there are security concerns over pretest items. Including pretest items in a test causes these items to be exposed or known to candidates, and possibly to coaching schools and future candidates. An item need only be exposed to one person for its security to be sacrificed. The concern over pretest items also carries over to CAT. (As an aside, Sheehan and Mislevy (1994) and Billeaud and Steffen (1996) have been examining ways to avoid the need for pretesting items by calibrating item with little or not data using collateral information.) Security issues are not unique to CAT.

There are three main concerns relating to test security whether within a CAT or P&P environment – repeaters, cheaters, and coaching schools. The first concern is that of *repeaters*, those students who choose to take the test more than once. If the same items appear on multiple tests, repeaters have an advantage of foreknowledge of items. A second concern involves *cheaters*. In a P&P environment, this refers to those who copy answers from neighboring examinees. In a CAT environment, while it is harder to copy from a neighbor since each examinee would most likely be seeing a different test and examinees are seated in separate cubicles, candidates can share the items they saw with a candidate scheduled to take the test in the future. In either the P&P or CAT environment, examinees from one time zone could communicate via the internet, fax, or phone sharing the items that appeared on their test. However, this would be more prevalent with P&P programs that use only a single form. Again, this would give the second candidate the advantage of foreknowledge of items. The last concern and perhaps the largest concern involves *coaching schools*. In cases where coaching schools have examinees report to them the items they were administered, these items can then be shared with future candidates and the validity of the test has been compromised. Any sharing of items is cheating, as each candidate signs an agreement stating they will not disclose any items. Still, the practice is known to go on.

Historically, large scale P&P testing programs have maintained test security by closely guarding access to test items by regularly introducing new forms and judiciously reusing old forms over relatively large intervals of time at small numbers of annual administrations. The reuse of these forms was done in such a way that even if a potential examinee had foreknowledge of a specific form; it is unlikely that the examinee would be administered the exact form at the test session. The size of the problem was assumed to be quite small.

Now, with the daily access to item pools necessary to support large volume CAT testing, this type of controlled access is no longer possible. Because only a finite number of precalibrated test items are available to be administered daily; a program must reuse items and must do so in a judicious manner to maintain test security.

One CAT analog to the P&P procedure of reusing items might be to prepare a large number of CAT pools and rotate them in a random fashion. However, two factors need to be considered. First, CAT pools tend to be much larger than traditional test forms, typically the equivalent of 8-12 test forms. To attempt to parallel traditional security measures results in a demand for new items so great that it far exceeds the financial resources of even large programs. Thus, with a limited item supply, the number of unique pools is restricted. Second, an artifact of CAT pools is that examinees of comparable ability tend to be administered many of the same items. Since CATs are shorter than their P&P counterparts, this similarity can become a serious threat to test security. With daily as opposed to quarterly testing, opportunities to recognize the similarity of tests increases and correspondingly the probability of foreknowledge increases of particular items.

An alternative to continually creating new pools containing an independent set of items, would be to consider each CAT pool as a sample of items from a larger collection (subsequently referred to as a VAT) rather than as a permanent collection of items. By sampling items with replacement, a theoretically infinite number of overlapping CAT pools might be created. In addition, a set of item reuse rules intended to minimize the benefit that might be expected from foreknowledge of a portion of the VAT could be instituted. One example of an item reuse rule may be not to include any item in a pool that has been seen by more than 200 people in the last two months.

The purpose of this research was to investigate the viability of creating overlapping CAT pools from a static VAT by sampling items with replacement. Several item reuse conditions were explored; mainly item overlap and item exposure rates. Item overlap rates refer to the number of pools in which an item appears. Item exposure rate refers to the number of examinees who have seen a particular item. An item is usually retired (*i.e.*, removed from future use) after it has been seen by a certain number of examinees over some specified period of time.

To increase the generalizability of the results of this study to different size testing programs, different annual test taking volumes were studied. In addition, it was hypothesized that testing programs with different volumes would not require as many pools per year. Therefore, the number of pools created per year was also manipulated. The primary research question was: For a fixed-size VAT, if items in a CAT pool are retired under 'X' conditions and pools are seen by 'N' examinees, how many times can the pool creation process be repeated before the selected pool will fail to support the delivery of CATs meeting certain delivery time constraints (*e.g.*, one pool per week)?

Method

This section is divided into four parts: description of data, test conditions, simulation procedure, and data analysis.

Description of Data

This section describes the simulations that were used to prepare the data for use in this study.

The VAT of items used in this study consisted of approximately 5,000 precalibrated Verbal multiple-choice items. These items included both discrete items and items belonging to sets, as well as stimuli. To use these items to study item exposure rates, it was necessary to estimate predicted usage rates for each item. This was accomplished by creating ten pools, consisting of approximately 450 items each, from this collection of 5,000 items. Simulations were performed in order to configure each of these ten item collections into CAT pools and to estimate the proportion of examinees at each of 13 ability levels that were expected to see each item. These proportions are referred to as conditional (on ability) probabilities of administration.

As not every item in the VAT appeared in at least one of the ten pools, some items did not have probabilities of administration (PRB's) associated with them. Thus, it was necessary to assign conditional probabilities of administration to these items in some way. The method used to assign PRB's to stimuli was different than that used to assign PRBs to discrete items. Each method is presented separately.

Discrete Items. To assign a PRB vector to every discrete item in the VAT, one simplifying assumption was made. It was assumed that the interaction of the major content area and the difficulty parameter of the item adequately captured the PRB vector. This assumption allowed us to associate the PRB vector of a discrete item with a certain content area and difficulty level to another discrete item with the same content area and difficulty level.

To associate a PRB vector to every discrete item in the VAT, all of the items from the ten pools were sorted by content area and difficulty level. The items were placed into levels of difficulty ranging from -3.0 to 3.0 in increments of 0.5. This resulted in five content areas and 13 difficulty levels and therefore 65 (5x13) content by difficulty combinations. Where there were less than ten items of one combination, two or more difficulty levels were aggregated. The result was 44 different content by difficulty combinations. An analysis of these PRB vectors indicated our assumption was plausible, that there was significant similarity in the vectors for the items with similar combinations of content classifications and difficulty parameters. Thus, it was possible to assign a PRB to each discrete item in the VAT that did not appear in one of the ten pools.

Stimuli. For the CAT algorithm under study, stimuli are not directly selected. With few exceptions, it turns out that within every set, there are items that mirror the overall selection characteristics of the stimulus. This fact was used to predict PRB's for stimuli. To assign PRB's to stimuli in the VAT, several variables – item in set with maximum information, item in set with largest *a* estimate, and item in set with smallest *c* estimate – were analyzed to determine which item's PRB vector best predicted the PRB vector of the stimulus. This varied for the different ability levels and content areas. In all cases, depending on what variable accounted for the most variance, either the PRB vector associated with the item in the set that had the maximum information or the item in the set with the largest *a* estimate was assigned to each stimulus in the VAT. In the remainder of the study, only discrete items and stimuli were included in the VAT.

In summary, the VAT consisted of 2,978 elements with each discrete item and stimuli have a PRB vector associated with it. The probabilities associated with discrete items depended upon the content area and level of difficulty. The probabilities associated with stimuli depended upon the content area and the item in the set that had the maximum information or the item in the set with the largest *a* estimate was assigned to each stimulus in the VAT.

Test Conditions

The remainder of the study involved simulating 15 different test conditions for each item reuse rule set (see Table 1). The item reuse rules are described below.

Table 1

Summary of 15 Test Conditions

Test Condition	Annual Test-Taker Volume	Number of Pools Created per Year
1	20,000	6
2	50,000	6
3	100,000	6
4	20,000	12
5	50,000	12
6	100,000	12
7	20,000	18
8	50,000	18
9	100,000	18
10	20,000	24
11	50,000	24
12	100,000	24
13	20,000	30
14	50,000	30
15	100,000	30

Each test condition was defined by some combination of two factors: (1) number of pools to be created per year and (2) annual test taking volume. The number of pools to be created per year was chosen to be 6, 12, 18, 24, or 30. The lower number of pools to be created of 6 was chosen, as it seemed to be the bare minimum number of pools to be used to keep item exposure rates to a minimum. The upper number of 30 pools to be created per year was chosen to reflect what is thought might be reasonable in practice. Second, three total test taker volumes were used: N=20,000, 50,000, and 100,000. The lowest sample size of 20,000 was chosen to reflect the test taker volume of smaller testing programs and the upper sample size of 100,000 was chosen to reflect the test taker volume of larger testing programs.

Simulation Procedure

The simulation phase consisted of creating item pools from the VAT. This involved three steps.

Step 1. Before being passed to the pool creation process, items were evaluated with respect to each of four reuse rules. Failure to meet any one of these rules precluded the item from being reused (*i.e.,* included in the present pool). There was one item overlap rule and three item exposure rules. First, concerning item overlap, an item could not be included in the present pool if it appeared in the last pool created. For example, during the development of Pool 9, no item from Pool 8 was considered. The purpose of this rule was to minimize the effect of candidates who have taken the test sharing with candidates planning on taking the test. Concerning item exposure rates, an item could not be used in the current pool if:

i) the total number of examinees who have seen the item since its introduction exceeded 10,000; or

ii) the number of examinees who have seen the item in the past four months exceeded 1,000; or

iii) the number of examinees who have seen the item in the past two months exceeded 100, 200, 300, 400, or 500.

These numbers were chosen arbitrarily. Initially, the first two rules were to be manipulated by increasing the numbers from 10,000 to 25,000 and 50,000; and from 1,000 to 2,000 and 3000. However, based on some preliminary analyses, very few items were seen by more than 5,000 people and so increasing this number was not warranted. After 10,000 or more people saw an item, it was retired. This is to say that the item could no longer be included in another pool and so it was removed from the VAT.

In addition, some preliminary analyses revealed that for small volumes of test-takers, nothing was caught by the four-month rule. However, it did come into play somewhat for larger volumes. Nevertheless, this rule was not manipulated.

Only the two-month rule was manipulated. To minimize the effect of sharing among candidates, an item was not included in the subsequent two month pools if it had been seen by more than 100 people. This rule was relaxed to allow up to 200, 300, 400, or 500 people to see an item in the past two months to examine the impact of this rule on the quality of subsequent pools.

Step 2. Pools were created with respect to content specifications, the overlap rule, item exposure rules, and the total pool information function. Each of these constraints had a weight associated with it to reflect its importance. In this way, test specialists can quantitatively prioritize the constraints. Weights varied from 5 being relatively unimportant to 60 being very important. A summary report of rule violations and overall index of the degree to which the pool met specifications was produced. This overall index was a total weighted deviation and was used as an overall index of pool quality in subsequent analyses.

Step 3. The last step involved updating the cumulative pool history record. It is at this stage that annual test taking volume was manipulated. For each item included in the current pool, the number of examinees expected to see the item was computed. Computation of the expected volume for each item required three pieces of information:

i) the PRB vector,

ii) a vector of weights (WTS) associated with each PRB. This vector represents the proportion of a specified examinee population that is expected to appear at each of the 13 ability levels. A normal distribution of ability was modeled.

iii) the number of examinees to whom pool *P* will be administered (VOL_P). Manipulation of this number allowed us to model annual test taking volume (20,000, 50,000, and 100,000).

In summary, the volume for discrete item or stimulus *j* was computed as:

$$VOLUME_j = VOL_P * \sum_I (PRB_i * WTS_i).$$

This simulation cycle was repeated 6, 12, 18, 24, and 30 times to model the number of pools to be created per year.

Data Analysis

The data were analyzed according to the effects of the number of pools created per year and the annual test-taking volume on the quality of pools and retirement of items. First, the quality of the resultant pools in terms of their total weighted deviations was examined. As described earlier, the total weighted deviation is an index of the amount of deviation there is in the pool from the set of constraints placed upon the pool. Since it is not always possible to satisfy all constraints simultaneously, each constraint is weighed to reflect the importance of the constraint. Any deviation from the constraint is reflected in the total weighted deviation. As an example, one constraint may be that the pool should consist of at least 50 but no more than 55 antonym items. This constraint might have a weight of 60 associated with it. In some instances, perhaps due to many antonym items included in the previous pool, there may only be 48 antonym items available for inclusion in the present pool. In this case, the deviation would be two and the weighted deviation would be 120 (2x60). The concern was not the absolute magnitude of the total weighted deviation but the relative value across pools created within a year (6, 12, 18, 24, or 30 pools per year).

This total weighted deviation for each pool was plotted. An ideal situation would be to have minimal total weighted deviations across all pools. The first irregularity of the plotted values indicated that the number of "optimal" pools had been selected. The purpose of the simulation cycle was to find the threshold where the violations produced an unacceptable pool. In this context, unacceptable means that it was unlikely that a CAT design could be created that would allow this collection of items to produce acceptable CATs. That is, at some point, the selected pools may not adequately meet all of the constraints.

Second, the rate at which items are predicted to be retired from use (due to more than 10,000 people being administered an item) was examined. As items are reused over a number of pools.

Results

The results are summarized according to the effects of the number of pools created per year and the annual test-taking volume on the quality of pools and retirement of items.

Quality of Pools. Recall that only the two-month rule was manipulated. Figures 1 to 5 correspond to the total weighted deviations of pools created in a year using the two-month rule of not including an item in a pool if more than 100, 200, 300, 400, and 500 people saw the item within the last two months, respectively.

Upon inspection of Figure 1, it is apparent that in general, with annual test-taking volume held constant (looking down columns), as the number of pools created per year increased, on average, the quality of the pools decreased as is evident by larger total weighted deviations. In general, with number of pools created per year held constant (looking across rows), pool quality also decreased as test-taker volume increased from 50,000 to 100,000 with the exception of creating six pools per year. However, the same does not hold true for all conditions when going from 20,000 to 50,000 test-takers per year. Except for the "spike" that occurs around three months, it appears as though there are better quality pools with 20,000 than 50,000 test-takers per year when creating 24 or 30 pools per year. Also evident in Figure 1 is that when more than six pools were created per year, pools #6, 7, or 8 had a large total weighted deviation.

In changing the two-month rule from not including an item in a pool if more than 100 people saw the item in the last two months to 200 people, the same results hold. One marked difference between Figures 1 and 2 is Condition 13 (30 pools created per year and 20,000 test-takers). The diminishing pattern of the quality of pools is more evident in Figure 2 Condition 3. In addition, the "spikes" are not as high for all conditions in Figure 2 as compared to those in Figure 1.

Figures 3-5 show the same patterns as Figures 1 and 2 with the exception of the conditions with a test-taker volume of 20,000 people per year. In these cases, test quality appears to be as good or better with 100 versus 300, 400, or 500 people two-month rule. In addition, as the two-month rule shifts from 200 to 500 people, the spikes begin to flatten.

Overall, as the two-month rule is changed from 100 to 500 people, the quality of the pools seemed to increase with the exception of the lowest test-taking volume of 20,000 where quality diminished.

Retirement of Items. Recall that an item was retired if more than 100,000 people saw the item since its inception. Tables 2 to 6 correspond to Figures 1 to 5. These tables provide the percent of the total number of items seen by different numbers of people and the total number of items seen at the end of the year for different test-taking volumes.

Regardless of the two-month rule, as the number of pools created per year increased, the total number of items used increased for all annual test-taking volumes (see Tables 2-6). As well, as the annual test-taking volume increased, the total number of items seen by people within a year also increased. This was true regardless of the number of pools created per year or the two-month rule (see Tables 2-6). In addition, as the annual test-taking volume increased, there was an increase in the percentage of total items seen by a larger number of people. For example, in Table 6, with six pools created per year, only 4.9% as opposed to 71.4% of the items were seen by between 2,500-4,999 people with annual test-taking volumes of 20,000 and 100,000, respectively.

In general, as both the number of pools created per year and the annual test-taking volume increased, there was an increase in the total number of items exposed within the year. Interesting to note is that only in Condition 3 did items need to be retired – six pools created per year with an annual test-taking volume of 100,000. Note however, that this was only 0.7% of the total number of items used in a year.

Discussion

The obtained results are relatively consistent with expectations. First, as the number of pools increases, the quality of pools decreases. This is likely due to the fact that limits were imposed on the proportion of elements from each pool that could be used in any given future pool. Second, as the number of pools increases, the proportion of items reaching (or even nearing) the retirement threshold decreases. Third, as volumes increase, pool quality again decreases. This results from the increase in the number of elements that are frozen from use by the two-month, four-month, and total volume thresholds. Additionally, since higher information items tend to be selected more frequently, subsequent pools will tend to be created from an available item collection with lower average information.

The spikes that occur are likely the result of demands on the VAT, with respect to content, that are incongruent with the content composition of the VAT. More specifically, they are due to the complex interaction of item difficulty and item information with item content. Since we did not balance the psychometric characteristics across content strata, the occurrence of these spikes was inevitable. The only question was when they would appear and the periodicity of the occurrence. It should be kept in mind that these spikes resulted from a static VAT. That is a fixed collection of elements was manipulated throughout. Thus, if the VAT were supplemented at the point where the spikes are predicted to occur, their appearance could be averted. Furthermore, the consistency of pool quality between spikes indicates that several moderate infusions of elements may suffice to maintain pool quality. That is, it is not necessary to augment the VAT before each pool is developed.

Tables 2-6 clearly indicate that the ability to control the use of items is dependent on the number of pools developed. The more pools created, the higher the potential for control. Conversely, as the number of pools developed is decreased, the larger the number of examinees that interact with each pool and the less control is available for managing exposure of individual items. This relates directly to the balancing act of item security. If the goal is to minimize the exposure of each item in small windows of time and reuse items over large windows of time, this is best accomplished by creating many pools per year. However, if the goal is to obtain maximal exposure in short windows of time and then retire items, this is best achieved by creating fewer pools per year.

Taken together, these results seem to show some promise for the viability of managing item security. By manipulating the number of pools developed per year, and manipulating the reuse of items across pools and time, the rate exposure of items can be explicitly manipulated.

Summary

Concerns over test security are always present in large-scale testing programs because breaches in test security impact on validity. But, in a CAT environment, there many need to be a balancing of test security and test quality. As seen in this study, increasing test quality often comes with a decrease in test security. There are always trade-offs to be made.

The concept of using a VAT is only one way to address item and test security concerns. It is not necessarily the best way, but it does offer certain advantages such as controlling item exposure rates by imposing item reuse and item overlap rules. Another distinct advantage of maintaining item test security using a VAT is that it is a preventative approach to minimizing the benefit that might be expected from foreknowledge of items by certain examinees. Other methods, such as person-fit indices (Davis & Lewis, 1996), are more *post-hoc* approaches to addressing test security. Person-fit indices allow one to detect "cheaters" after the administration of the test. Cheaters may be detected by uncovering the fact that they copied their answers from the people beside them or they may be detected by unaccountable gains when they repeat the test. Another alternative includes two-stage testing.

The results obtained in this study give evidence to support the use of a VAT system as one way of drawing a compromise between test quality and test security. However, this study of examining a VAT system was by no means exhaustive. For example, this study could be expanded by examining the effects of different ability distributions of candidates and different time windows on the quality of pools. In addition, one could determine what portion of the VAT becomes deficient (*e.g.,* certain content areas or level of difficulty of items).

While the generalizability of the specific results obtained from study may be limited, the possibility of the use of the sampling with replacement model seems to be a promising alternative to traditional vault-like security procedures.

References

Billeaud, K., & Steffen, M. (1996, April). Collateral information in item calibration. Paper presented at the meeting of the National Council on Measurement in Education, New York.

Davis, L. A., & Lewis, C. (1996, April). Person-fit indices and their role in the CAT environment. Paper presented at the meeting of the National Council on Measurement in Education, New York.

Mills, C. N., & Stocking, M. L. (1995, August). Practical issues in large-scale high-stakes computerized adaptive testing (Research Report 95-23). Princeton, NJ: Educational Testing Service.

Sheehan, K., & Mislevy, R. J. (1994, April). A tree-based analysis of items from an assessment of basic mathematics skills (Research Report 94-14). Princeton, NJ: Educational Testing Service.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, F. J., Steinberg, L., and Thissen, D. (1990). Computerized adaptive testing: a primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Table 2

Percent of Total Number of Items Seen by Different Numbers of People and the Total Number of Items Seen at the End of the Year for Different Test-Taking Volumes Using the **100** Two-Month Rule

No. of Pools /Year	Annual Test-Taker Volume	Number of People							Total
		0-99	100-499	500-999	1,000-2,499	2,500-4,999	5,000-9,999	10,000+	
6	20,000								
	50,000	0.8	1.8	15.0	43.1	38.5	0.9	0.0	1082
	100,000	0.8	1.1	1.5	10.1	71.1	14.6	0.7	1105
12	20,000	0.5	6.0	91.1	2.4	0.0	0.0	0.0	1719
	50,000	0.5	1.3	48.1	38.5	11.6	0.0	0.0	1894
	100,000	0.5	1.2	10.5	41.2	44.0	2.8	0.0	1903
18	20,000	0.5	8.5	86.5	4.5	0.0	0.0	0.0	1925
	50,000	0.4	1.6	57.0	31.2	9.9	0.0	0.0	2040
	100,000	0.4	1.3	11.6	45.8	36.5	4.3	0.0	2056
24	20,000	0.5	18.4	77.5	3.6	0.0	0.0	0.0	2105
	50,000	0.4	2.3	61.4	30.4	5.5	0.0	0.0	2310
	100,000	0.4	1.2	22.5	40.2	34.6	1.1	0.0	2347
30	20,000	0.5	24.8	69.4	5.3	0.0	0.0	0.0	2246
	50,000	0.4	2.9	62.4	31.8	2.5	0.0	0.0	2554
	100,000	0.4	1.2	31.3	33.4	33.6	0.2	0.0	2954

Table 3

Percent of Total Number of Items Seen by Different Numbers of People and the Total Number of Items Seen at the End of the Year for Different Test-Taking Volumes Using the **200** Two-Month Rule

No. of Pools /Year	Annual Test-Taker Volume	Number of People							Total
		0-99	100-499	500-999	1,000-2,499	2,500-4,999	5,000-9,999	10,000+	
6	20,000	0.9	3.4	75.4	18.8	1.6	0.0	0.0	1033
	50,000	0.8	1.8	14.4	43.4	58.7	0.9	0.0	1075
	100,000	0.8	1.1	1.5	10.1	71.3	14.6	0.7	1104
12	20,000	0.7	9.5	81.7	8.1	0.0	0.0	0.0	1486
	50,000	0.5	1.5	45.7	40.1	12.2	0.0	0.0	1824
	100,000	0.5	1.2	10.4	41.1	44.0	2.8	0.0	1894
18	20,000	0.5	14.2	75.7	8.9	0.8	0.0	0.0	1863
	50,000	0.5	2.3	52.9	34.2	10.2	0.0	0.0	1956
	100,000	0.4	1.2	11.6	45.8	36.6	4.4	0.0	2044
24	20,000	0.4	26.1	67.3	5.5	0.7	0.0	0.0	2203
	50,000	0.5	3.2	54.6	35.8	5.9	0.0	0.0	2183
	100,000	0.4	1.4	20.8	41.3	35.0	1.1	0.0	2310
30	20,000	0.4	32.9	61.8	3.7	1.1	0.0	0.0	2465
	50,000	0.4	3.5	59.0	30.5	6.6	0.0	0.0	2303
	100,000	0.4	1.3	30.2	33.3	34.5	0.2	0.0	2550

Table 4

Percent of Total Number of Items Seen by Different Numbers of People and the Total Number of Items Seen at the End of the Year for Different Test-Taking Volumes Using the **300** Two-Month Rule

No. of Pools /Year	Annual Test-Taker Volume	Number of People							Total
		0-99	100-499	500-999	1,000-2,499	2,500-4,999	5,000-9,999	10,000+	
6	20,000	0.9	4.1	72.0	21.3	1.6	0.0	0.0	998
	50,000	0.8	1.8	13.6	43.9	38.9	0.9	0.0	1070
	100,000	0.8	1.1	1.5	10.2	71.1	14.6	0.7	1102
12	20,000	0.6	12.8	72.0	14.0	0.6	0.0	0.0	1435
	50,000	0.5	1.5	42.4	42.5	13.1	0.0	0.0	1766
	100,000	0.5	1.2	8.6	42.2	44.7	2.8	0.0	1868
18	20,000	0.4	15.1	77.2	5.1	2.2	0.0	0.0	2031
	50,000	0.5	2.5	47.4	38.6	11.0	0.0	0.0	1852
	100,000	0.5	1.4	9.9	46.2	37.6	4.5	0.0	1998
24	20,000	0.4	25.9	69.6	2.4	1.7	0.0	0.0	2331
	50,000	0.5	3.7	57.5	29.1	9.3	0.0	0.0	2092
	100,000	0.4	1.5	20.1	39.8	37.0	1.2	0.0	2268
30	20,000	0.4	31.7	64.3	2.0	1.6	0.0	0.0	2535
	50,000	0.4	3.9	60.0	26.1	9.4	0.2	0.0	2253
	100,000	0.4	1.6	29.9	31.8	36.0	0.2	0.0	2498

Table 5

Percent of Total Number of Items Seen by Different Numbers of People and the Total Number of Items Seen at the End of the Year for Different Test-Taking Volumes Using the **400** Two-Month Rule

No. of Pools /Year	Annual Test-Taker Volume	Number of People							Total
		0-99	100-499	500-999	1,000-2,499	2,500-4,999	5,000-9,999	10,000+	
6	20,000	1.0	4.2	67.7	23.4	3.7	0.0	0.0	968
	50,000	0.9	1.5	11.9	45.1	39.7	0.9	0.0	1060
	100,000	0.8	1.1	1.3	9.7	71.4	14.7	0.7	1097
12	20,000	0.6	13.1	72.5	12.6	1.2	0.0	0.0	1539
	50,000	0.6	1.7	35.3	48.3	14.1	0.0	0.0	1670
	100,000	0.5	1.4	5.6	43.5	46.1	2.9	0.0	1826
18	20,000	0.4	15.3	78.0	4.1	2.2	0.0	0.0	2048
	50,000	0.5	2.6	47.5	38.2	10.9	0.4	0.0	1806
	100,000	0.5	1.5	9.0	44.5	40.0	4.5	0.0	1960
24	20,000	0.4	26.7	68.9	2.4	1.6	0.0	0.0	2363
	50,000	0.4	3.6	55.4	31.6	8.6	0.3	0.0	2062
	100,000	0.5	1.5	19.7	37.1	40.1	1.2	0.0	2195
30	20,000	0.4	34.5	61.9	1.8	1.4	0.0	0.0	2672
	50,000	0.4	3.3	65.7	22.3	7.7	0.6	0.0	2378
	100,000	0.4	2.0	25.4	35.6	34.1	2.5	0.0	2319

Table 6

Percent of Total Number of Items Seen by Different Numbers of People and the Total Number of Items Seen at the End of the Year for Different Test-Taking Volumes Using the **500** Two-Month Rule

No. of Pools /Year	Annual Test-Taker Volume	Number of People							Total
		0-99	100-499	500-999	1,000-2,499	2,500-4,999	5,000-9,999	10,000+	
6	20,000	0.9	4.2	71.0	19.1	4.9	0.0	0.0	1005
	50,000	0.9	1.5	10.3	46.5	39.9	1.0	0.0	1037
	100,000	0.8	1.1	0.9	10.4	71.4	14.7	0.7	1097
12	20,000	0.5	13.6	77.1	6.7	2.1	0.0	0.0	1710
	50,000	0.7	2.0	37.1	40.3	19.4	0.5	0.0	1508
	100,000	0.6	1.5	6.0	41.0	40.2	10.7	0.0	1638
18	20,000	0.4	15.5	78.8	2.9	2.4	0.0	0.0	2121
	50,000	0.5	2.5	49.1	35.3	11.7	0.8	0.0	1811
	100,000	0.5	1.4	7.2	45.8	40.5	4.6	0.0	1929
24	20,000	0.4	26.4	69.6	1.9	1.7	0.0	0.0	2385
	50,000	0.4	3.6	59.7	29.1	6.6	0.7	0.0	2167
	100,000	0.5	1.7	16.9	41.7	35.6	3.6	0.0	2104
30	20,000	0.4	38.6	58.4	1.6	1.1	0.0	0.0	2764
	50,000	0.4	3.6	68.1	22.2	4.6	1.1	0.0	2487
	100,000	0.5	2.0	23.1	37.5	31.2	5.8	0.0	2240

Figure 1. Total Weighted Deviations of Pools with Items not Included in Pools if more than 100 People Saw Item in Last Two Months

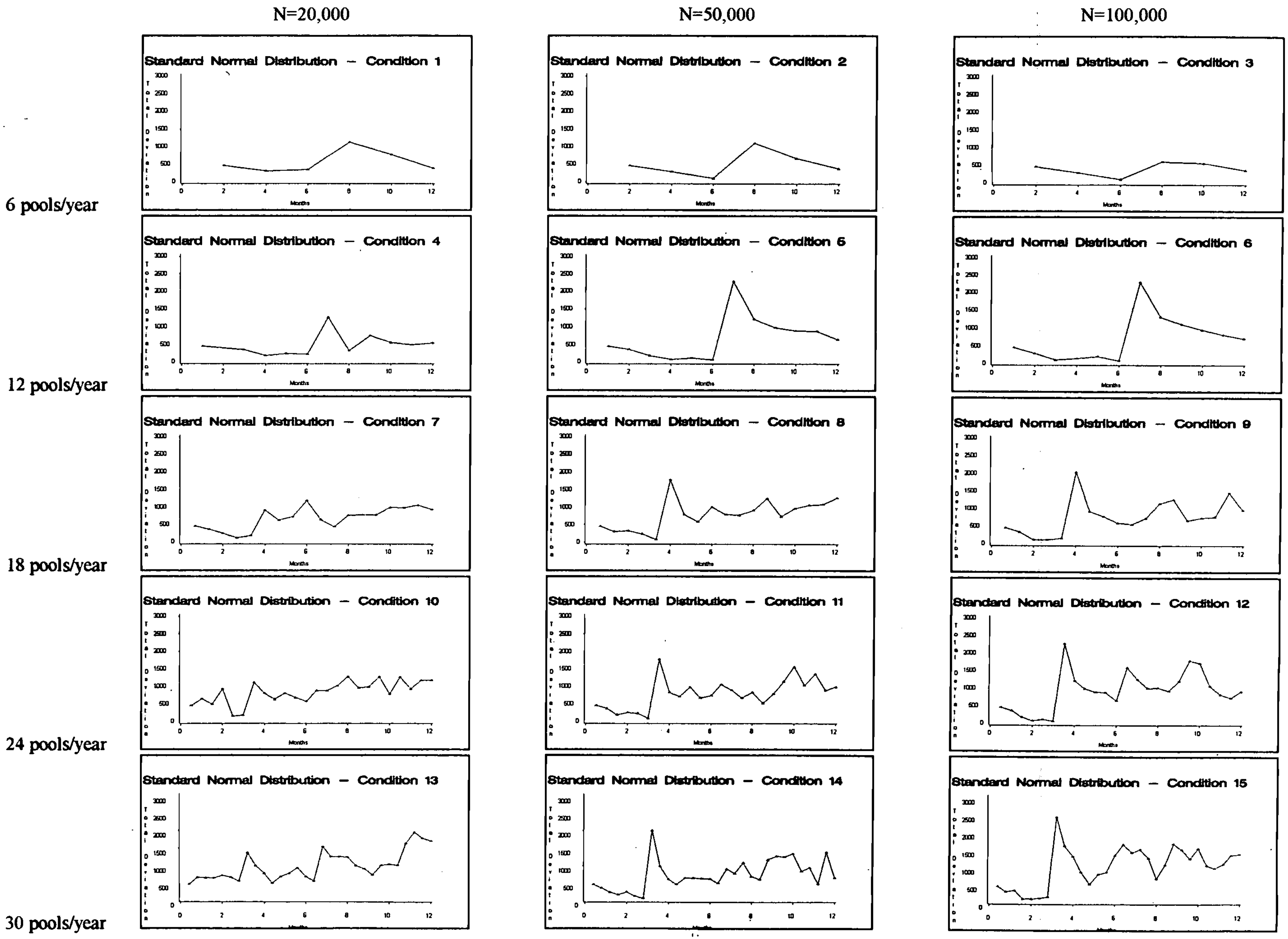


Figure 2. Total Weighted Deviations of Pools with Items not Included in Pools if more than 200 People Saw Item in Last Two Months

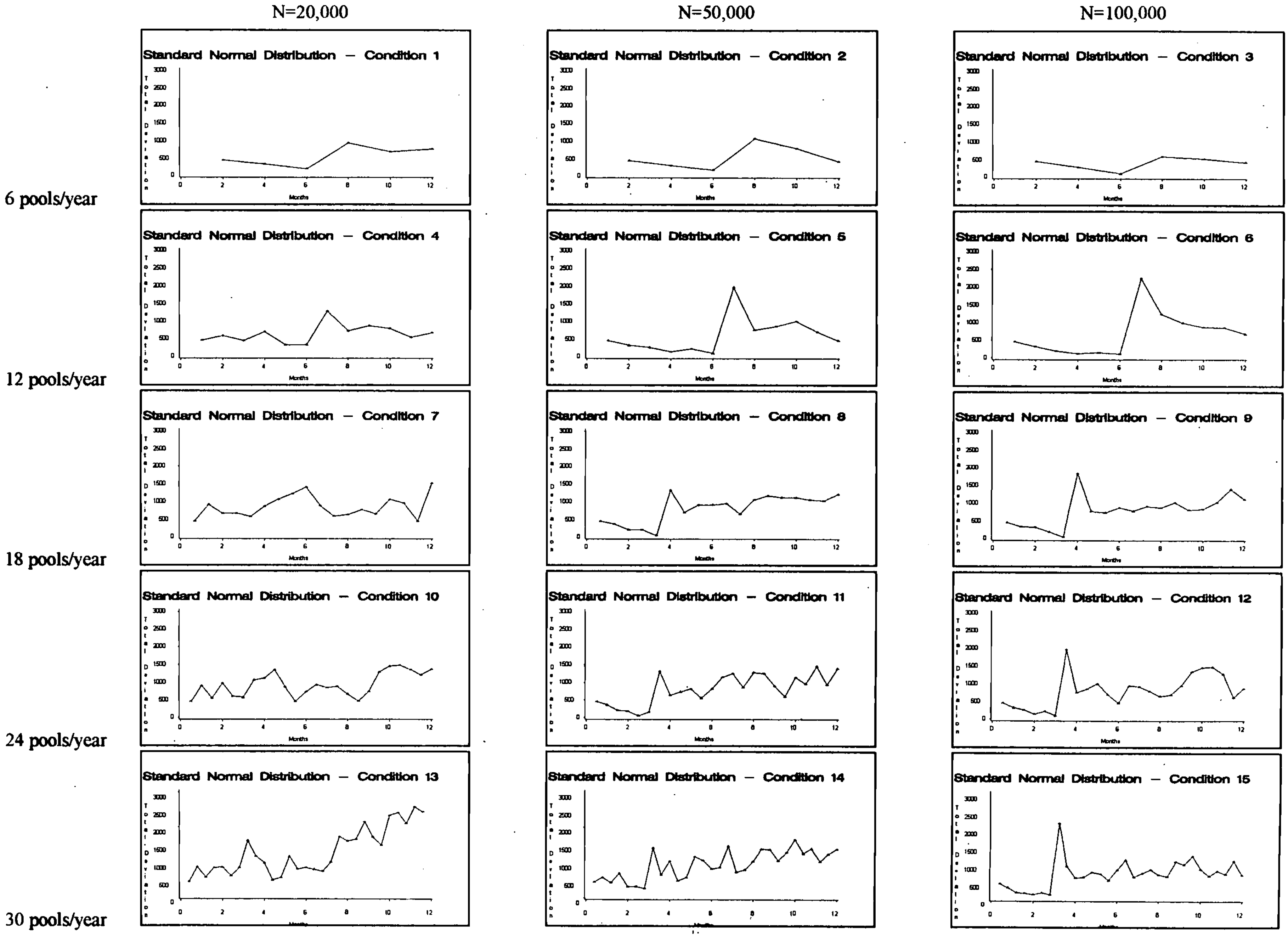


Figure 3. Total Weighted Deviations of Pools with Items not Included in Pools if more than 300 People Saw Item in Last Two Months

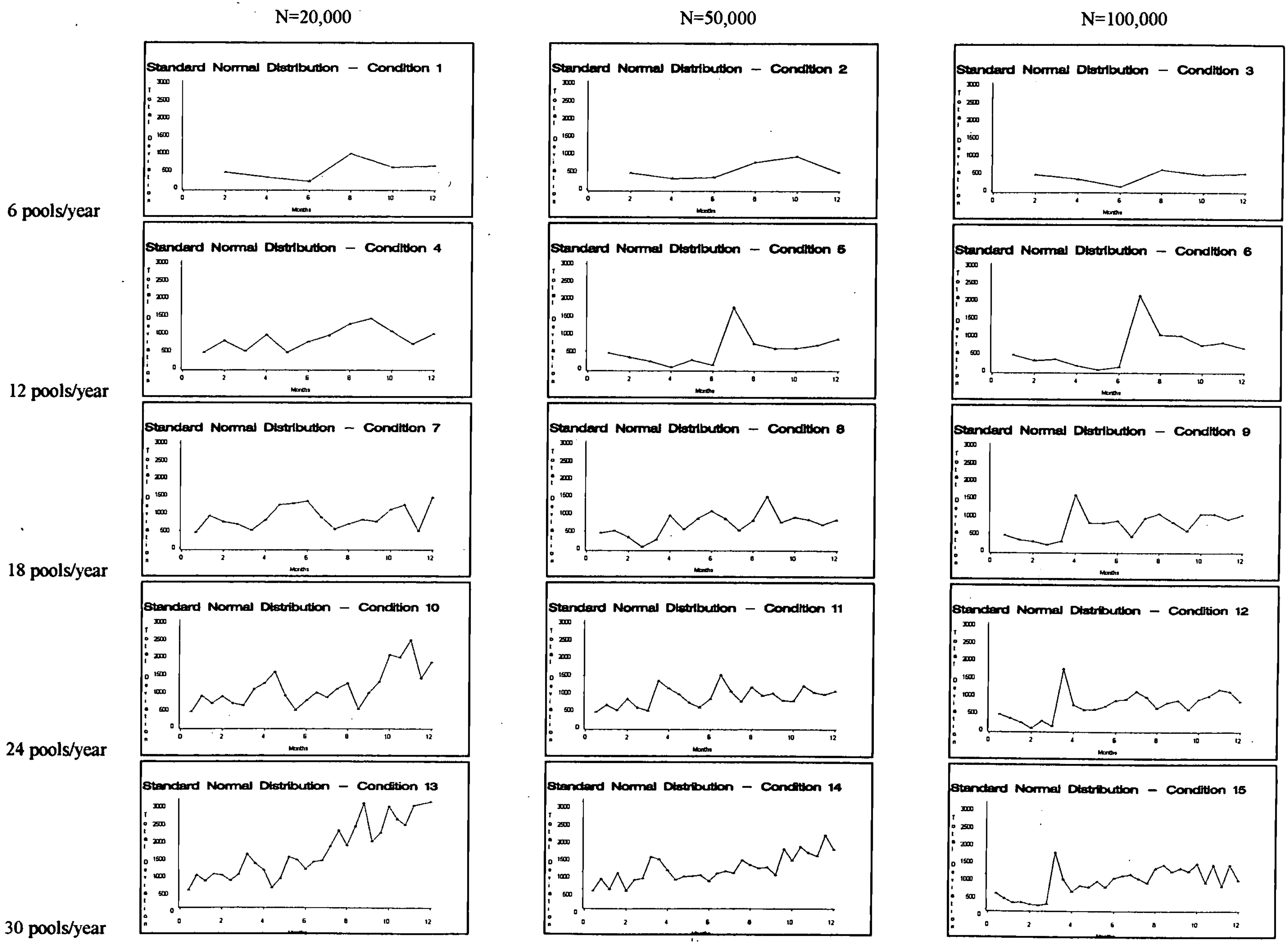


Figure 4. Total Weighted Deviations of Pools with Items not Included in Pools if More than 400 People Saw Item in Last Two Months

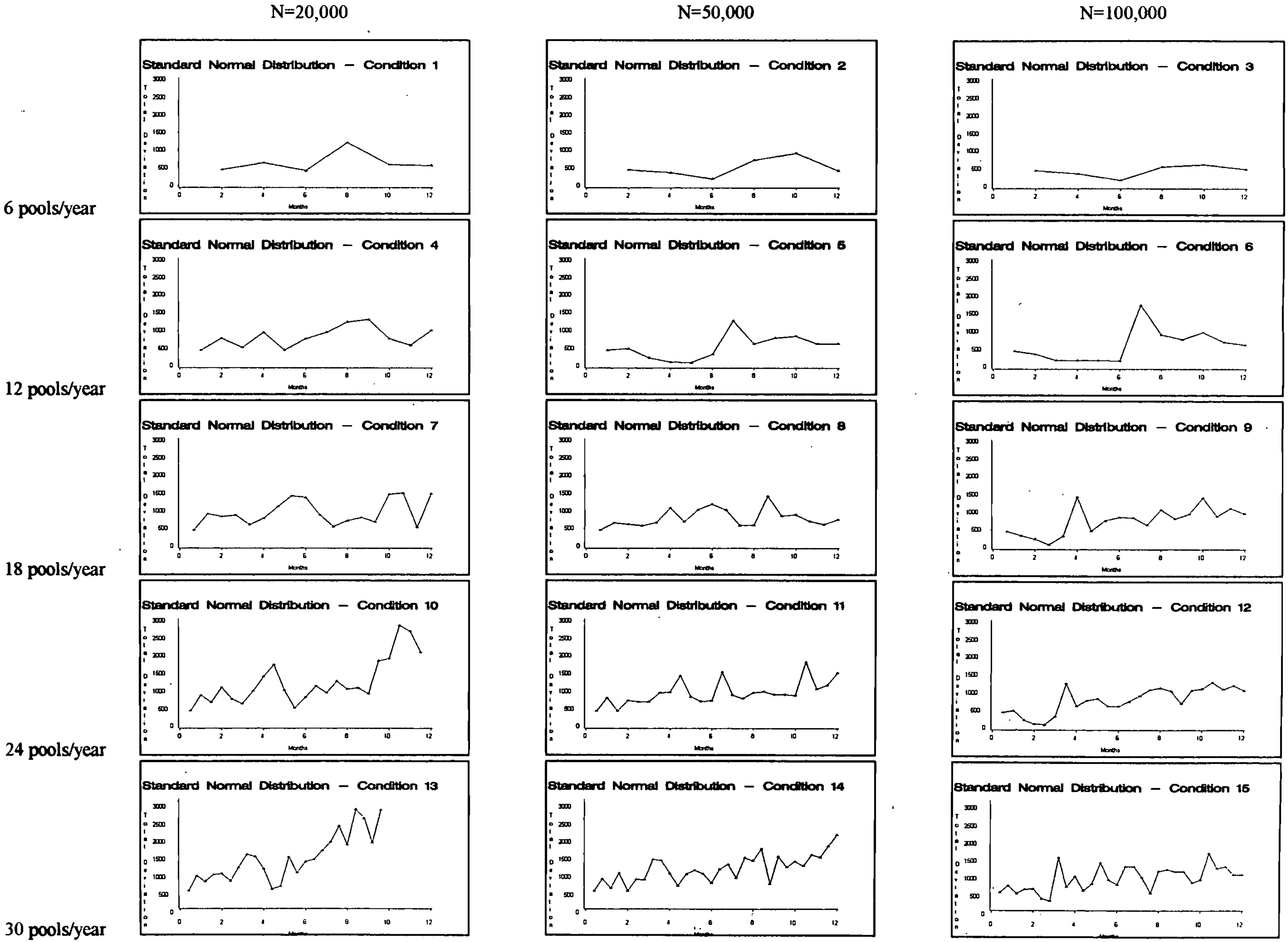
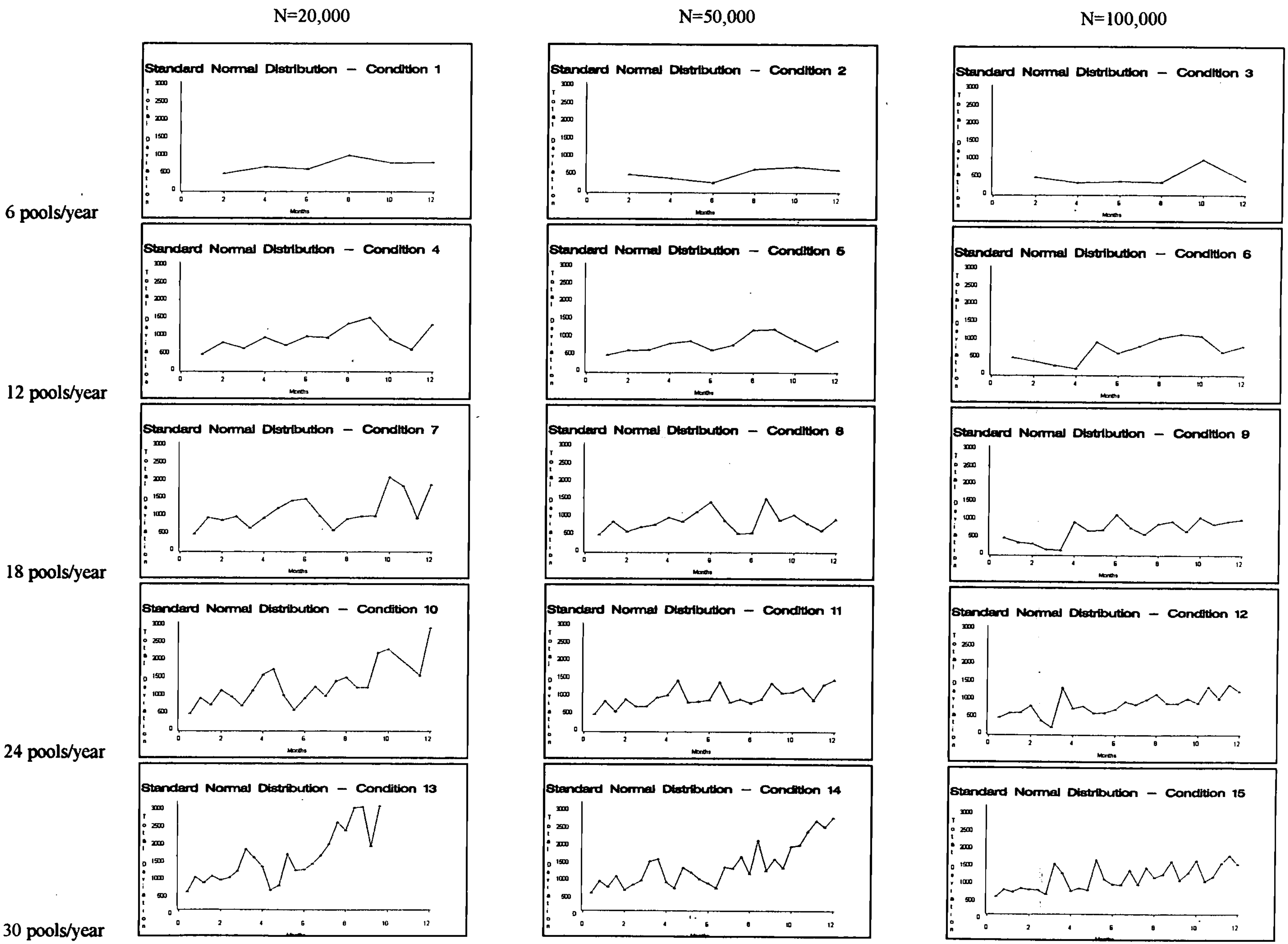


Figure 5. Total Weighted Deviations of Pools with Items not Included in Pools if More than 500 People Saw Item in Last Two Months



NCME 1997

TM027312

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Maintaining Item and Test Security in a CAT Environment: A Simulation Study.	
Author(s): Liane N. Patsula and Manfred Steffen	
Corporate Source:	Publication Date: March 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document □

or here
Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please Liane Patsula

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Liane Patsula	Position: Graduate Student
Printed Name: Liane Patsula	Organization: University of Massachusetts
Address: 152 Hills South UMass Amherst, MA 01003	Telephone Number: (413) 549-8644 Date: March 31, 1997

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 24, 1997

Dear NCME Presenter,

Congratulations on being a presenter at NCME[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

We are gathering all the papers from the NCME Conference. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our process of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

NCME 1997/ERIC Acquisitions
O'Boyle Hall, Room 210
The Catholic University of America
Washington, DC 20064

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an NCME chair or discussant, please save this form for future use.