Content-Stratified Random Item Selection in Computerized Classification Testing

Robin Guille †
Rebecca S. Lipner †‡
John J. Norcini ‡
Jane C. Folske †

†American Board of Internal Medicine
‡Institute for Clinical Evaluation

Paper presented at the Annual Meeting of the National Council on Measurement in Education, 2002, New Orleans, LA

Objectives

An optimally efficient item selection rule for computerized adaptive testing (CAT) chooses items with maximum information at each examinee's estimated level of ability (Parshall, Davey, & Nering, 1998). However, this approach ignores practical concerns of operational testing programs such as ensuring proper content balance and limiting the number of examinees permitted to view each item, known as item "exposure." For testing programs that administer examinations on a continuous basis throughout the year, permitting too many examinees to view an item increases the possibility that it can be reproduced from recall, potentially compromising the security of future examinations (Wainer, 2000).

To address these concerns, research in the field of computerized testing has introduced several selection rules that use compound rules for item selection. A set of items with maximum information is first chosen and then items are selected from this set based on exposure and content balance considerations (Stocking, 1993; Parshall, Davey & Nering, 1993). Alternatively, the item pool can be stratified by content area and then sets of items with maximum information can be chosen from the strata.

Although compound rules have been implemented in many testing programs and research in the 1990's was focused on the subtle differences between them (Chang & Twu, 1998; Parshall, Davey & Nering, 1998), Wainer (2000) contends that continuing to use currently popular rules could lead to a shortage of pool items that would be difficult to replenish. In high-stakes testing, the number of items by which one would have to increase an item pool in order to ensure its security is much too large to be practical, given the cost and time necessary to develop new items.

Wainer suggests using alternative methods which more effectively control item exposure, including the possibility of administering tests linearly with all items drawn randomly from the pool prior to administration [similar to "prestructured" item selection (Kingsbury & Zara, 1983)]. These alternative methods are of particular interest for computerized classification tests (CCT), which are computerized adaptive tests with an ultimate pass-fail or multicategorical decision, since they tend to have small item pools.

The present study looks at a less radical proposal—simply making item selection rules much more random. Research studies, to date, often have treated total randomness in item selection as a baseline for comparison with compound rules (e.g., Parshall, Davey & Nering, 1998; Chang & Twu, 1998), but have not examined randomness within content strata as a possibility for implementation. Specifically, we ask the research question "Do computerized classification tests for professional certification featuring content-stratified random item selection make reliable pass/fail decisions, while minimizing item exposure rates and test lengths?"

To address the research question, three selection rules were compared: (1) content-stratified random selection, (2) conditional Sympson & Hetter, and (3) the maximum information for cut score level of ability. Each selection rule identified the content area that was least fulfilled, based on the examination blueprint, and then proceeded to select an item from items within this content area.

To evaluate item exposure control, the conditional Sympson & Hetter rule was included in the study for comparison with random selection, since this rule is commonly used for item exposure control (see Parshall, Davey & Nering, 1998 for a full description of the Sympson & Hetter rule). To evaluate the ability to make reliable pass/fail

decisions, the maximum information for cut score level of ability rule was included in the study, since it is the rule most likely to make reliable pass/fail decisions for classification testing (Spray & Reckase 1996; Parshall, Davey, Spray & Kalohn, 1999).

Data Sources

Data from a recent administration of a 200-item pencil-and-paper professional certification examination were used as the basis for generating simulated data for this investigation. The examination contained five content areas. Item parameter estimates for the simulation were obtained by calibrating the responses of the 511 candidates who took the real examination, using the 2-parameter item response model in BILOG (Mislevy & Bock, 1991), given by,

$$P_i(\theta) = \frac{e^{Da_j(\theta - b_j)}}{1 + e^{Da_j(\theta - b_j)}} \quad \text{(Formula 1)}$$

where $P_i(\theta)$ is the probability of a correct response to item j given θ; a_j corresponds to the item discrimination for item j; b_j is the difficulty of item j; θ is the true examinee proficiency; D is the constant 1.7.

Then, in order to simulate a larger item pool, the 200 item parameter estimates were replicated 5 times and treated as the true item parameters for a simulated 1,000-item pool. Binary (0/1) response data were then generated for 1,000 simulated examinees to the 1,000 simulated items using the 2-parameter item response model.

More specifically, for each simulated examinee for each item, p_i was compared to a random variable drawn from a uniform (0,1) distribution; if p_i exceeded the value of the random variable, the response to item i by examinee j was scored as a correct response.

If p_i was less than or equal to the value of the random variable, the response to item *i* by examinee *j* was scored as an incorrect response.

Methods

The three selection rules were compared on the following criteria: (1) average item exposure rate, (2) average test length, (3) percentage of forced decisions, (4) percentage of correct decisions (PCD), and (5) standard error.

CAT Simulations Using SPRT Stopping Rule

To make the first three comparisons—average item exposure rate, average test length and percentage of forced decisions—we simulated administrations of CATs employing the three item selection rules and analyzed the results.

An exposure table for later use by the conditional Sympson & Hetter selection rule was prepared in advance, using another Monte Carlo simulation, which was run for 30 iterations of 1000 simulees each. The table was conditioned on examinee ability quartiles. The Sympson & Hetter rule selected a maximum of 400 items from the 1,000-item pool using an arbitrarily set 30% maximum exposure rate, which is the arguably the upper limit exposure rate for actual practice (Chang & Twu, 1998), but still low enough to draw from a 1,000 item pool.

The Sequential Probability Ratio Test (SPRT) (Reckase, 1983) was used to make classification decisions since it has been shown to possess certain advantages over other approaches in CCT (Parshall, Davey, Spray & Kalohn, 1999). Forced pass/fail decisions were made after 400 items. Examinees closest to the SPRT upper boundary after 400

items were classified as passing, while those closest to the lower boundary were classified as failing. The rather high 400-item limit was chosen to minimize forced pass-fail decisions in the study.

To identify the indifference region, which is the ability region between the decision boundaries, the following SPRT (Reckase, 1983) definitions were used:

Lower bound $= \beta / (1 - \alpha)$
Upper bound $= (1 - \beta) / \alpha$

(Formula 2)

where α is acceptable decision error of passing a non-master
where β is acceptable decision error of failing a master

In this study, these were set to levels of $\alpha = 0.05$ and $\beta = 0.05$. The indifference region corresponded to one standard error of measurement on either side of the cut score, as determined from 200-item pencil-and-paper results.

It's important to stress that, in SPRT, the upper and lower bounds, along with the current likelihood ratio, completely determined the decision to either stop early with a pass, stop early with a fail, or continue. The domain scale (θ) scores representing 1 SEM above the cut score and 1 SEM below the cut score were *not* used to set the boundaries. They were used instead in the computation of the current likelihood ratio—since we were testing which of two hypotheses hold:

H_1: the examinee's ability is greater than or equal to 1 SEM above the cut score
H_0: the examinee's ability is less than or equal to 1 SEM below the cut score

After each response, a revised likelihood ratio was computed as the following product:

$$\left(\frac{p_1}{p_0}\right)^{\Sigma x_i} \left(\frac{1-p_1}{1-p_0}\right)^{n-\Sigma x_i}$$

(Formula 3)

where n is the number of items administered thus far,
where p_1 is $P(\theta)_i$ from Formula 1 when θ = the point 1 SEM

above the cut score (H_1),
where p_0 is $P(\theta)_i$ from Formula 1 when θ = the point 1 SEM
below the cut score (H_0),
where x_i is the (1,0) score on ith item presented thus far,
and the values of a and b from Formula 1 are the parameters
for item x_i.

This is simply a binomial calculation, where the current likelihood ratio for the examinee is maintained as a logarithm. Each examinee's ratio starts out at the 0.0 estimate. After each response, using Formula 3, the logarithm of the leftmost element (p_1/ p_0) is added to the logarithm of the current likelihood ratio if item i is answered correctly or the logarithm of the rightmost element [(1-p_1)/(1- p_0)] is added if item i is answered incorrectly. When the revised likelihood ratio moves out of the region of indifference (i.e., the ratio is greater than or equal to the upper bound or is less than or equal to the lower bound), the CAT stops early, otherwise it keeps on repeating this revision of the likelihood ratio until all 400 items have been administered.

Fixed-Length CAT Simulations

To adequately make the fourth and fifth comparisons—percentage correct decisions (PCD) and standard error— the simulations employing the three selection rules were re-run, but required that a fixed set of items be presented to each examinee instead of permitting early stopping. A fixed test length of 200 was chosen. Standard error was computed for each simulee using the following formula (Baker, 1985), where i is the item number, N is the number of items seen, a is the item discrimination, $P(q)$ is the probability of a correct response and $Q(q)$ is the probability of an incorrect response:

$$SE(\theta) := \frac{1}{\sqrt{\sum_{i=1}^{N} (a_i)^2 \cdot P(\theta) \cdot Q(\theta)}} \qquad \text{(Formula 4)}$$

The percentage of correct pass/fail classification decisions made by each of the item selection rules was computed to compare pass/fail decision accuracy across the three methods under the fixed-length CAT condition. First, the true ("correct") pass/fail classification was determined for each simulee by comparing the simulee's true ability (used for simulating the response data) to the cut score (set at the point on the ability scale corresponding to the standard of the actual pencil-and-paper administration). Simulees with a true ability value at or above the cut score were classified as passing, while the others were classified as failing. Pass/fail classification decisions subsequently were made based upon the estimated abilities via the three CAT simulations and then compared to this "correct" decision to calculate the PCD.

Results

The mean item exposure rate for the content-stratified random rule was 15.3% (SD=1.0%), for the conditional Sympson & Hetter rule it was 14.1% (SD=8.3%), and for the maximum information at cut score rule it was 18.4% (SD=23.3%). The distribution of item exposure in Figure 1a shows that the content-stratified random selection rule spread the items out consistently across all levels of difficulty, while the other two rules shown in Figures 1b and 1c did not. Also, the content-stratified random rule showed much less spread than the other two rules, particularly than maximum information rule shown in Figure 1c, which was uncontrolled for item exposure.

The mean test length was 153 items for the content-stratified random rule, 116 for the conditional Sympson & Hetter rule, and 73 for the maximum information at cut score rule.

Forced pass/fail decisions were made for 10.1% of examinees using the content-stratified random rule, 8.9% under the conditional Sympson & Hetter rule and 3.6% under the maximum information at cut score rule.

The conditional standard error was plotted for the variable-length simulations in Figure 2. These results, however, tend to be similar at the cut score inflection point because of the use of SPRT. When the special 200-item simulation was evaluated instead, the conditional standard error of measurement at the cut score was 0.28 for the content-stratified random rule, 0.25 for the Sympson & Hetter rule, and 0.20 for the maximum information rule.

Table 1 shows the results of the percent correct decision analyses. Again, the special fixed-length test was run to provide a less biased result. The content-stratified random rule made correct decisions 82% of the time; Sympson & Hetter 86% of the time; and the maximum information rule made correct decisions 91% of the time.

Discussion

These results suggests that if item exposure concerns are paramount in the classification testing environment, using a simple random-selection-within-content-strata rule to control item exposure could produce reasonable, albeit less than optimal results. If the purpose of the test were to screen examinees for stage two testing, then the content-

stratified random rule might work acceptably because forced decisions wouldn't have to be made.

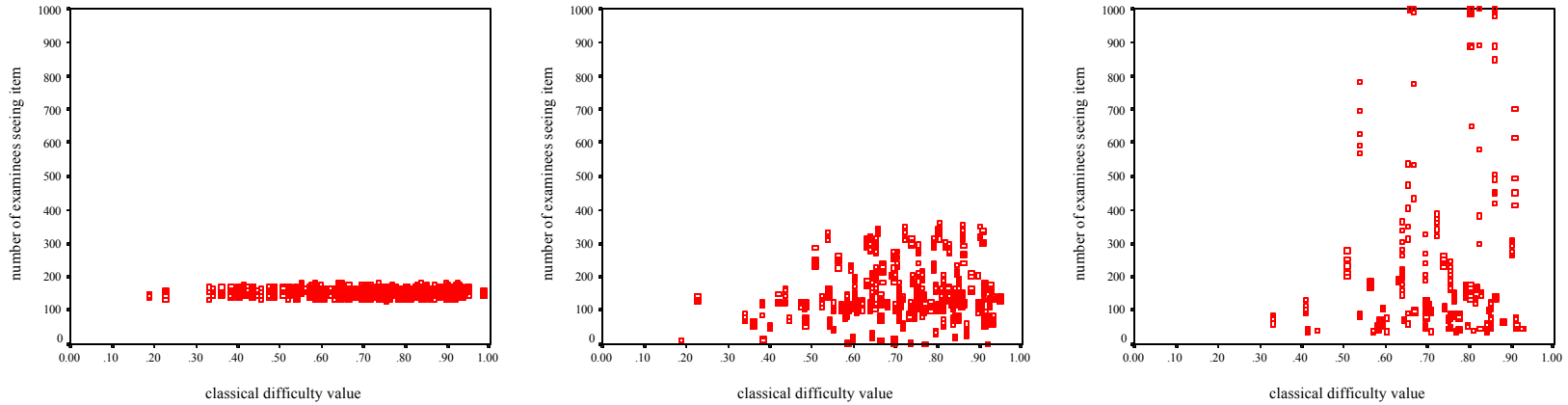
The particular environment that the simulations were based upon—professional certification testing—tends to have small item pools and a tradition of lengthy fixed-length tests. Although in these simulations the content-stratified random rule produced lengthier tests than the other selection rules, it did excel at item exposure control. The average test length was still much less than that of the pencil and paper examination used as the basis of calibration for the simulations.

A methodological limitation of the study is that the data were not simulated multi-dimensionally, which is recommended to make it a bit more realistic (Davey, Nering & Thompson, 1997).

A practical limitation of using a random rule is that it implies that all items in the pool are quality items, since any one of them has an equal probability of being presented. This means that test developers must take more care in weeding out questionable items from the pool when using a random algorithm than when using a maximum information rule.

Future research might evaluate the content-stratified random rule with some measuring of test overlap (Chen, Ankenmann & Spray, 1999). In addition, future research may focus on comparing other rules, such as the "5-4-3-2-1 randomization" rule (McBride & Martin, 1983), with the content-stratified random rule. It is less elaborate than the conditional Sympson & Hetter rule, but it too selects from a set of optimally informative items and is the kind of rule more commonly found in CCT (Lin & Spray,

2000). Also, future research might examine the effects of random item selection from pools of different quality.



*-

Figure 1. Item exposure across levels of item difficulty. In the maximum case, 1000 simulated examinees see the item. (a) Random rule, (b) conditional Sympson & Hetter rule, and (c) maximum information at cut score level ability rule (no exposure control).

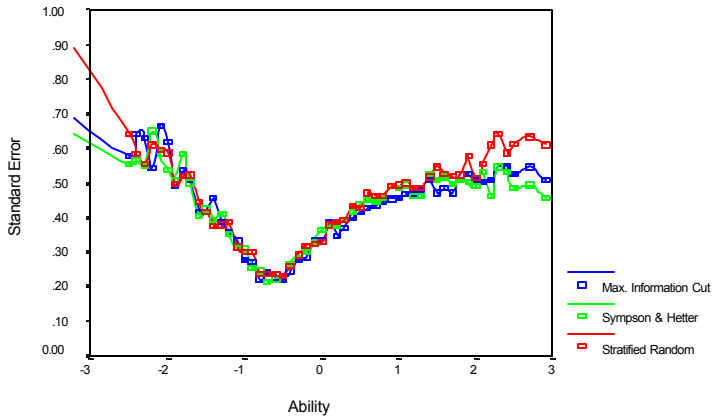


Figure 2. Standard errors at each level of ability for variable-length CAT simulations

	Variable-Length CAT Simulation	Special Fixed 200-Item Simulation
Content-Stratified Random	95%	82%
Sympson & Hetter	94%	86%
Maximum Information	97%	91%

Table 1. Percent Correct Decision (PCD) for the three selection rules under the variable-length and fixed-length CAT simulations.

References

Ackerman, T. (2000). "Practical applications of item response theory" Short course manual (pp. 32). College Park: Department of Measurement, Statistics & Evaluation, University of Maryland.

Baker, F. (1985). *The basics of item response theory.* Portsmouth, NH: Heinemann.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model.* Hilldsdale, NJ: Lawrence Erlbaum Associates.

Chang, S. & Twu, B. (1998). A comparative study of item exposure control methods in computerized adaptive testing. (Research Report 98-3). Iowa City, IA: American College Testing.

Chen, S., Ankenmann, R. D. & Spray, J. A. (1999). Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing. (Research Report 99-5). Iowa City, IA: American College Testing.

Davey, T., Nering, M. & Thompson, T. (1997). Realistic simulation of item response data (Research Report 97-4). Iowa City, IA: American College Testing.

Hambleton, R. K., & Swaminathan, H. (Eds.). (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

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

Kingsbury, G. G. & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 257-283). New York: Academic Press.

Lin, C. & Spray, J. A. (2000). Effects of item-selection criteria on classification testing with the Sequential Probability Ratio Test. (Research Report 2000-8). Iowa City, IA: American College Testing.

Linacre, J. M. & Wright, B. D. (1991). *BIGSTEPS Rasch-model computer program*. MESA Press: Chicago.

McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223-236). New York: Academic Press.

Mislevy, R. & Bock, R. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. (2nd Ed)., Mooresville, IN: Scientific Software.

Parshall, C., Davey, T., Spray, J., & Kalohn, J. (1999). "Computerized testing - Issues and applications" Mini-course manual, *Annual Meeting of the National Council on Measurement in Education*. Montreal.

Parshall, C. G., Davey, T. & Nering, M. L. (1998). Test development control for adaptive testing. *Annual Meeting of the National Council on Measurement in Education*. San Diego.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 237-255). New York: Academic Press.

Spray, J. A. & Reckase, M.D. (1996). Comparison of SPRT and Sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, 21(4), 405-414.

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

Wainer, H. (Ed.). (1990). *Computerized adaptive testing: A primer*. Hilldsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (2000). Rescuing computerized testing by breaking Zipf's Law. *Journal of Educational and Behavioral Statistics, 25*(2), 203-224.

Author Note

This research was supported by the ABIM but does not necessarily reflect its opinions.