Measurement and Research Department Reports 2006-3

Preventing the Bankruptcy of an Item Bank

Angela J. Verschoor

Measurement and Research Department Reports 2006-3

Preventing the Bankruptcy of an Item Bank

Angela J. Verschoor

Cito
Arnhem, 2006

Abstract

This paper discusses the long-term effect of using Automated Test Assembly methods on item banks. Depletion of items with high discriminations can be observed, causing to shorten the life span of the item banks. Item bank management strategies, focused on modification of the test assembly methods, are investigated using simulation studies performed in the context of two large-scale projects.
Keywords: Item Response Theory, Test assembly, Item bank management

1 Introduction

The popularity of computer based testing (CBT) and automated test assembly (ATA) has created various opportunities for testing agencies to improve the cost-effectiveness of their "silverware" by intensifying the use of test items over a prolonged period of time. For this improved use, testing agencies have been investing considerable amounts of money into the development of item banks. These investments are justified by the foreseen extended use of the items in these banks. Therefore, it is not surprising that attention has recently been paid to the problem of item bank management to safeguard long-term item bank quality, and thus to maximize the return on the investments that have been made.

Recent research by Way and Steffen (1998), Belov and Armstrong (2004) and Ariel (2005) focused either on item bank design or on training and selection of item writers to construct items with favorable characteristics. In some testing programs, however, the possibilities to implement these measures might be limited, for example, by budget restrictions. Moreover, optimization techniques, whether performed by hand or by using ATA models and accompanying software, will select the most favorable items that are available. Therefore, these methods put a pressure on the item construction process to construct more items of even better quality than before. As test assembly methods continue to select the best available items, improvement of the quality of newly constructed items will not prevent this problem. The quality of the tests might improve but depletion of high-quality items in the item bank continues to occur. In fact, as ATA models are likely to be more efficient in identifying favorable items than manual optimization techniques, this pressure on item construction tends only to increase when ATA is introduced.

Therefore, in order to maintain a uniform quality of an item bank, test assembly must be harnessed in such a way that the quality of the items used in the tests does not exceed the expected quality of newly constructed items or previously used items that are returned to the bank. Several strategies to harness the assembly of tests are proposed in this paper and a simulation study is conducted to evaluate these strategies in the context of two testing programs for which item bank management is a vital issue: the Unified State Examinations (USE) for secondary education in the Russian Federation, and the State Examinations for Dutch as a Second Language (DSL) in the Netherlands.

1.1 The Russian Unified State Examinations

Started in 2001, the USE program is gradually being introduced in the Russian Federation. Nationwide, schools for secondary education are in the process of adapting to the school leaving and university entrance examinations in the program, providing their students with test results comparable all over the country (Russian Federation Ministry of Education and Science, 2005). The Federal Institute for Pedagogical Measurement (FIPI) develops the examinations. Approximately 700,000 students have been tested in 2006, and this number is likely

to grow to over 1,500,000 annually as of 2012.

Since the Russian Federation extends over nine time zones, two issues are important for the success of the USE program: test equivalence and item security. It has been decided that 100 parallel test forms will be developed yearly, while each form will be administered to roughly equal numbers of students. A limited overlap between the test forms is acceptable in the USE program. Currently the item banks contain approximately 3000 - 4000 items that have been calibrated under the OPLM model, that is, a 2PL model with integer discrimination indices (Verhelst, 1995). The item banks are planned to be expanded annually with 1000 calibrated items until the target size of 10000 items has been reached.

The parameter distributions of the items in the current USE item banks for Russian and Physics can both be approximated by $\log(\alpha) \sim N(1, 0.36)$ and $\beta \sim N(0, 0.28)$. Analysis of current data has shown that the ability of the population is approximately normally distributed with $\mu = 0.01$ and $\sigma = 0.23$.

1.2 Dutch as a Second Language Examinations

While the purpose of the USE testing program is to give an advise on possible further education for a relatively broad population, the tests of the DSL program can be characterized best as certification tests. Each year, approximately 5000 DSL-learners take an examination on either level 1 or 2 of four different language skills: Reading, Writing, Listening, and Speaking. Citc has been commissioned to produce the Listening and Speaking examinations three times a year. The test lengths for these examinations are 40 and 15 items, respectively. Starting in 2007, a new system of examinations will be implemented, giving candidates higher flexibility in taking the examinations. For each skill and level combination, an item bank is developed that should contain 1000 - 2000 items, from which six parallel examinations will be assembled annually. In this case, the test forms should be parallel according to Samejima (1977). Contrary to the USE testing program, no overlap is allowed.

In order to assemble the examinations, ATA methods will be used. The goal is to assemble shorter examinations, preferably with no more than 30 - 35 items for Listening, while the TIFs will be required to be equal to those of the current examinations. Every year approximately 100 new items will be constructed, pretested, and added to the banks. In addition, items will not be allowed to be used again within two years after administration. In this way, the addition and reuse of items will compensate for the loss of the items that will be used in the examinations or will become outdated.

The current item bank for Listening at level 1 contains 767 items, while the parameter distribution can be approximated by $\log(\alpha) \sim N(1.5, 0.28)$ and $\beta \sim N(-0.17, 0.24)$. As it was decided that the cut-off point on the raw score scale was at 28 out of 40, while simultaneously it was defined to be $\vartheta = 0.089$ on the ability scale of a reference test, the majority of the items are relatively easy in order to accommodate the combination of cut-off points. At the same time, the distribution of the target population was observed to be approximately

normally distributed with $\mu = 0.12$ and $\sigma = 0.28$.

2 The Dynamics of Item Bank Management

It is assumed that the item bank can be split into two separate item pools: an active item pool from which items may be selected, and a passive item pool containing items that were previously used and for security reasons should not be used again for a certain period of time. Then, the dynamics involved in item bank management can be modelled as two processes that interact between the two pools. The first process comprises of the construction and pretesting of new items, and, if applicable, of the return of items to the active pool. The items from this process form the input of the active pool. The second process is an output process. After items have been used in a test, they are stored in the passive pool for possible future reuse. In the remainder of this paper, the term item pool refers to the active item pool, unless stated otherwise.

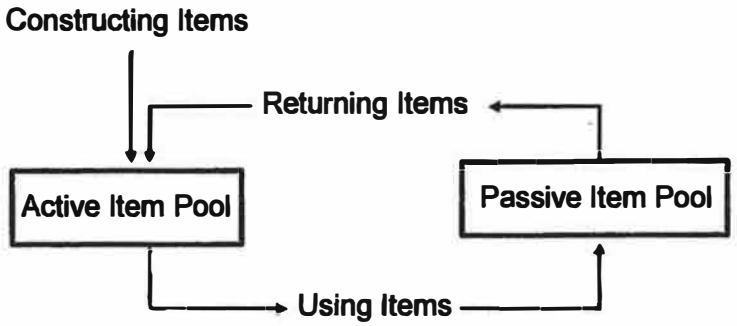

Figure 1: The Processes of Item Bank Management

Three distributions can be distinguished: the parameter distribution of items from the input process, of those from the output process, and of those in the item pool. Note that the last distribution depends on the other two.

It is clear that the parameter distribution in the input process must at least match that of the output process to avoid depletion and to maintain an item pool that can supply items according to the test specifications over an extended period of time. Strategies for item bank management could be said to be successful if they control the input and output processes such that a stable pool is maintained.

As strategies to improve the input are expected to influence the quality of the tests only, and not to prevent depletion of high-quality items, the input distribution is assumed to be stable throughout this study. Therefore, it suffices to regard either the quality of the tests or the distribution of the items in the item pool to conclude whether an item bank management strategy is successful or not.

3 Item Bank Depletion

Before we discuss strategies to prevent item bank depletion, it is necessary to identify the precise nature of the depletion. Which types of items are selected first, and which types of items tend to remain behind when no precautions are taken? For commonly used ATA models, it is easy to determine which items are most favorable. A look at the objective function of the maximin model (van der Linden and Boekkooi-Timminga, 1989) reveals that highly discriminating items are more likely to be selected than lowly discriminating items.

A simple simulation of a large-scale test assembly problem, resembling the case of the Physics examinations in the USE program, confirms the favorableness of highly discriminating items. From an item pool consisting of 10000 items with parameter distribution of $\log(\alpha) \sim N(1, 0.36)$ and $\beta \sim N(0, 0.28)$, a number of tests was assembled. Annual groups of 100 tests were assembled sequentially for a target population with ability distribution $\vartheta \sim N(0.01, 0.23)$. Each test contained 30 items and required a maximum overlap of three items with any other test within its group, and no overlap with other tests. Each item should appear in no more than two tests. The PARIMAX model was used to formulate the test assembly problem involved. The test information function (TIF) target was defined to be equal at three ability points $\vartheta_1 = -0.34$, $\vartheta_2 = 0.0$, and $\vartheta_3 = 0.34$. Equations (1) through (6) specify the test assembly problem used. $I_i(\vartheta_k)$ is the information function value of item i at point ϑ_k, while x_{ij} is the decision variable indicating whether item i is selected in test j:

$$\begin{aligned}
\text{maximize} \quad & y && && (1)\\
\text{subject to:} \quad & y \leq \sum_i I_i(\vartheta_k) x_{ij} && \forall k, j && (2)\\
& \sum_i x_{ij} \leq 30 && \forall j && (3)\\
& \sum_i x_{ij} x_{il} \leq 3 && \forall j, l && (4)\\
& \sum_j x_{ij} \leq 2 && \forall i && (5)\\
& x_{ij} = \begin{cases} 1, & \text{item } i \text{ in test } j \\ 0, & \text{else} \end{cases} && \forall i, j. && (6)
\end{aligned}$$

The tests were assembled through the DOT 2005 software (Verschoor, 2005), while the depletion of the item bank was investigated through the parameter distribution of the items remaining in the pool. These specifications were observed to require between 1500 and 1600 items annually, thus an item pool of 10000 items could support the test demand for six years before new items had to be added or previously used items had to be returned. The assembled tests were evaluated using two criteria: the height of the TIF values on the three ability

points and the average standard error of measurement for the target population:

$$\omega = \int_{-\infty}^{\infty} \frac{g(\vartheta)}{\sqrt{I(\vartheta)}} d\vartheta. \tag{7}$$

Table 1: Average TIF-values and ω-values (USE Physics)

Year	$I(\vartheta_1)$		$I(\vartheta_2)$		$I(\vartheta_3)$		ω	
	M	SD	M	SD	M	SD	M	SD
1	91.7	1.3	122.8	6.9	91.5	1.4	0.102	0.001
2	66.5	0.6	95.7	1.0	66.6	0.6	0.118	0.001
3	51.6	0.3	64.5	0.0	51.6	0.3	0.136	0.000
4	45.8	0.5	49.1	0.7	45.1	0.3	0.148	0.001
5	26.6	0.0	29.8	0.0	26.6	0.0	0.192	0.000
6	25.5	0.1	28.1	0.0	25.6	0.1	0.196	0.000

Table 1 contains the TIF values at the specified ϑ's, averaged over the 100 tests per year, as well as the average ω-values. Next to the average TIF values, the standard deviations are reported, giving an indication of to what extent the tests can be assumed to be parallel. It can be clearly seen that over six years, the average height of the TIF decreases to approximately 25% of the average TIF in the first year. At the same time, the average ω almost doubled. Thus, it could not be maintained that the tests in the last year were parallel to the tests in the first year, and signs of depletion were obvious.

Figure 2 shows the distributions of the a- and b-parameters of the available items at the start and directly after assembly of the tests for the second, fourth and sixth year. It can be observed that items with high discrimination parameters are favoured, thus depleting the pool with respect to these items. This effect is in line with similar findings for computerised adaptive tests, as were found, for example, by Lord (1980). As a result, the lower discriminating items tended to remain, from which only low-informative tests could be assembled. A second observation is that no significant depletion with respect to item difficulty occurred. For every difficulty level a reasonably large choice of items remained available during the entire test assembly process.

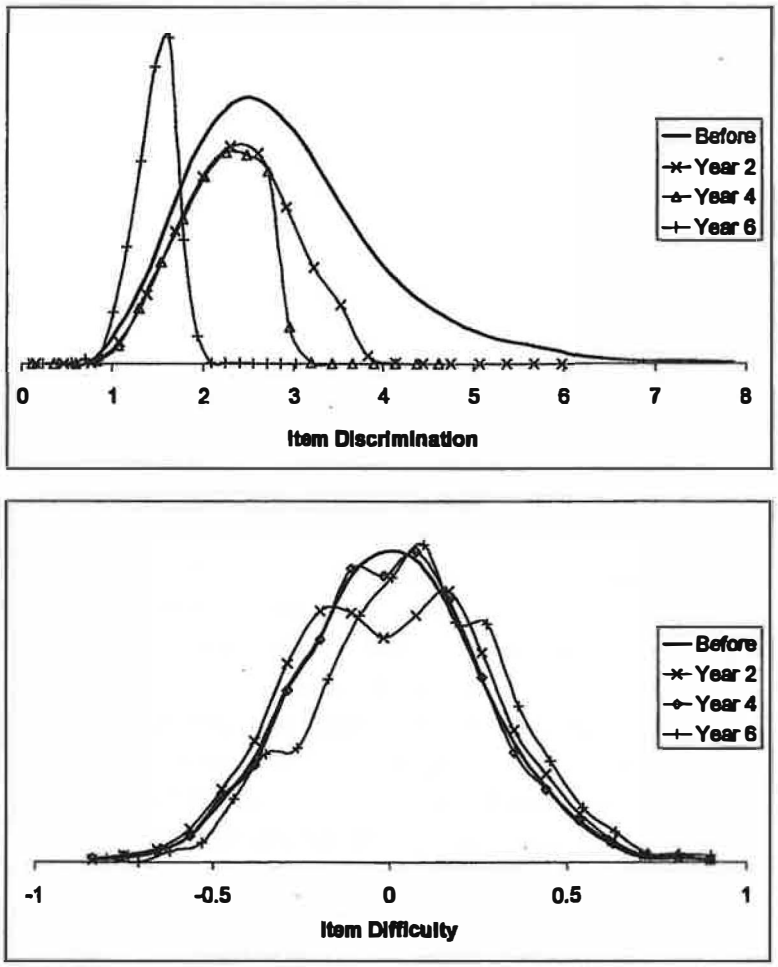

FIGURE 2. Parameter Distributions (USE Physics)

A second simulation, resembling the Listening examinations of the DSL program, was conducted. The current examinations have, on average, a TIF value of 122.6 at the cut-off point. The test specifications require a TIF target defined at two ability points. The first point is the cut-off point at which a TIF target of 122.6 should be reached, and the second point was chosen at -0.15, with an equal target. This results in a TIF that will be high in the interval $(-0.15 \leqslant \vartheta \leqslant 0.089)$, and a cut-off score at approximately 70% of the maximum raw score.

Thus, for the simulations the target was defined to be 122.6 at $\vartheta_1 = -0.15$ and $\vartheta_2 = 0.089$, while the test length was minimized. Test overlap was not allowed. Equations (8) through (12) specify the test assembly problem:

$$\text{minimize} \quad y \tag{8}$$

$$\text{subject to:} \quad y \geq \sum_i x_{ij} \qquad \forall j \tag{9}$$

$$\sum_i I_i(\vartheta_k) x_{ij} \geq 122.6 \qquad \forall k, j \tag{10}$$

$$\sum_j x_{ij} \leq 1 \qquad \forall i \tag{11}$$

$$x_{ij} = \begin{cases} 1, & \text{item } i \text{ in test } j \\ 0, & \text{else} \end{cases} \qquad \forall i, j. \tag{12}$$

Every year 100 items were added to the item bank drawn from the target parameter distribution: $\log(\alpha) \sim N(1.5, 0.28)$ and $\beta \sim N(0, 0.24)$. At the same time, every year 50 items were removed because they were no longer considered to be appropriate. Although in the DSL program it is allowed to reuse items after two years, reuse was not allowed in the simulations at all.

Table 2: Average TIF-values and ω-values (DSL)

Year	$I(\vartheta_1)$		$I(\vartheta_2)$		ω		Length
	M	SD	M	SD	M	SD	
1	130.2	3.5	129.2	1.0	0.143	0.004	12
2	125.1	1.2	125.1	1.4	0.124	0.002	17
3	131.9	3.4	125.6	0.4	0.108	0.002	22
4	124.6	1.9	123.6	1.1	0.107	0.001	27
5	136.4	4.4	124.7	0.7	0.102	0.002	30

In Table 2, it can be seen the test length rose rapidly over the years, while the average TIF values remained rather stable. At the same time, the average standard error of measurement, ω, dropped significantly over the years, despite the fact that the TIF at the specified abilities was almost constant. These signs indicate that large differences in information might be found at extreme abilities, and hence that the tests might not be parallel at these ability levels. If the differences between the TIFs are deemed unacceptable, it might be wise to specify the target at more ability points in order to decrease these differences. But even then, it might be expected that the test length will have to rise over

time in order to maintain stable values for both the TIF and ω.

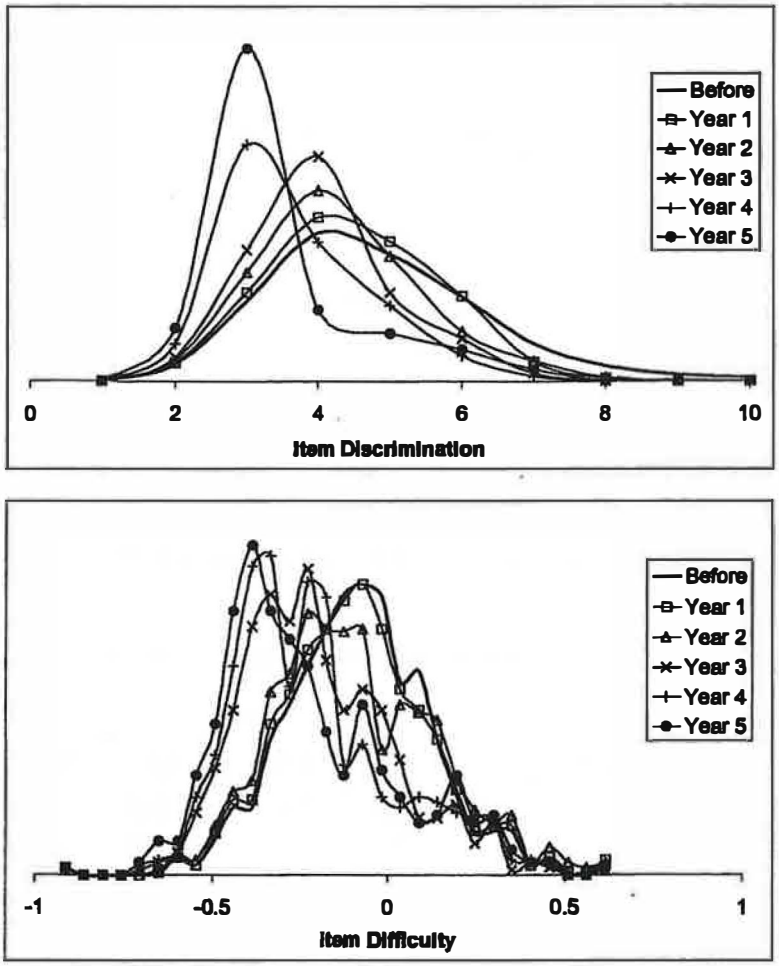

FIGURE 3. Parameter Distributions (DSL)

Figure 3 shows the distribution of the a- and b-parameters of the available items at the start and after test assembly each year. Also for the DSL simulations, items with high discriminations were favoured and tended to be selected early on. Contrary to the absence of depletion with respect to item difficulty in the simulations for the USE program, Figure 5.3 shows that the test assembly process had a tendency to select the somewhat more difficult items, that is, those with a b-parameter near the cut-off ability. Despite this tendency, still sufficient items from all difficulty levels remained available.

4 Item Bank Management Strategies

Item bank management strategies should prevent the depletion of item banks with respect to highly discriminating items. The following four actions could be considered:

Stratification. The first strategy is stratification of the item pool according to the discrimination parameters of the items. By adding restrictions to the test assembly model, the number of selected items from every stratum can be restricted to be proportional to the number of available items in the pool. Thus, an even use of items from various discrimination categories will be enforced. As a result it would be plausible that the parameter distribution in the output process is equal to the parameter distribution in the pool and the latter distribution remains stable.

Subdivision. The second strategy is a random subdivision of the entire item pool into several pools. Each year a pool will be selected to serve as the current pool. The size and number of these pools can be determined by observing how many items are needed for solutions without any item bank management. Each current pool should have this size in order to fulfill the demands for test assembly each year, and the number of years that the testing program could be supported before depletion will occur is thus equal to the number of pools that can be constructed. In every item bank, however, lowly informative items, which do not contribute substantially to the information in the tests, can be found. If the pool sizes are increased, more of these lowly informative items can be left unused. More highly informative items would be selected, resulting in higher-informative tests. Reversing this argument reveals the dilemma that test assemblers are confronted with: In order to assemble more informative tests, the pool sizes should be increased, shortening the life span of the bank.

An alternative subdivision, through parallel item pool assembly, could be considered if differences between the tests appear to be unacceptably high when a random subdivision is used.

Shadow pool. The third item bank management strategy is the big-shadow-test method proposed by van der Linden (2005). In this method, the items are selected either directly in the current tests or in a shadow test. The items in the shadow test are preserved and after test assembly they are returned to the item pool. In the current application the big shadow test can be considered as a shadow item pool.

Similar to the decision on the pool sizes in case of random subdivision, a decision must be made on the size of the shadow item pool. The fewer items are accepted to be left unused, the larger the shadow pool should be. Once the size of the shadow pool has been determined, the restrictions can be assumed to be proportional to the sizes of the current tests, while no overlap between the shadow pool and the current tests is allowed.

Alternative IRT model. The fourth item bank management strategy is using an alternative IRT model that does not allow for any differences in item discrimination: the Rasch model. When no differences in item discrimination are recognized, then by definition no depletion of highly informative items could occur, thus avoiding the management problems completely.

There is a remark to be made, however. It may be observed in practice that using the Rasch model may result in the rejection of 10 - 15% of the available items during calibration due to poor item fit. Many of those items would fit the 2PL-model because the only deficiency is that their discriminations are too high or too low for the Rasch model.

5 Simulations

A series of simulations was performed to investigate which strategy prevents depletion most efficiently. Three scenarios were used for the two testing programs. Scenario 1 and 2 were used to investigate the proposed item bank management strategies for the USE program and were based on the previously simulated item bank. Scenario 1 represented the situation in the near future, when the item bank is still under development. At the start, 5000 items were available, to which every year 1000 new items were added. Scenario 2 represented a situation in the longer future, without structural item development. All 10000 items were available at the start, while no replacement or extension was assumed. In both scenarios, 100 tests were assembled annually: Each test consisted of 30 items, while the test overlap was restricted to three, and the item exposure restricted to two. Overlap between tests of different years was not allowed. The TIF target was defined to be equal at three ability points $\vartheta_1 = -0.34$, $\vartheta_2 = 0.0$, and $\vartheta_3 = 0.34$.

Scenario 3 simulated the DSL program. At the start of the simulations, 767 calibrated items were available, to which 100 items were added annually to the item bank drawn from the target parameter distribution: $\log(\alpha) \sim N(1.5, 0.28)$ and $\beta \sim N(0, 0.24)$. At the same time, 50 items were removed annually to represent that they were considered to be outdated. Every year, six tests with no overlap were assembled.

The aim of these simulations was to investigate which strategy prevents depletion most efficiently. Therefore, the following conditions were studied:

- No action was undertaken with respect to depletion. This condition served as the base line condition.
- The items were stratified according to their discrimination. In the item bank for the USE program, the range of discrimination parameters was observed to be [1..10]. Five strata were defined:

Stratum		Available Items
1	$\alpha \geqslant 6$	253
2	$\alpha = 5$	564
3	$\alpha = 4$	1573
4	$\alpha = 3$	3624
5	$\alpha \leqslant 2$	3986

Furthermore, four restrictions for every test were added to the test assembly problem in (1) - (6). In each test, a maximum number of 1, 2, 5 and 10 items were allowed from strata 1 through 4, respectively. As items from stratum 5 were expected to be selected only when no alternatives were available, they could be selected freely. The stratification restrictions were formulated as

$$\sum_i c_{im} x_{ij} \leq C_m \qquad \forall m, j$$

whereby coefficient c_{im} indicates the stratification, having value 1 if item i belongs to stratum m, and value 0 otherwise. C_m is the maximum number of items from stratum m in the tests.

In the DSL bank, the discrimination parameters varied from 2 to 10 and the item bank was stratified using 6 categories:

Stratum		Available Items
1	$\alpha \geqslant 8$	65
2	$\alpha = 7$	101
3	$\alpha = 6$	264
4	$\alpha = 5$	468
5	$\alpha = 4$	550
6	$\alpha \leqslant 3$	318

Five restrictions per test were added to (8) - (12), imposing maxima on the number of items to be selected from each stratum. As the test lengths were not fixed, these maxima were expressed as percentages of the realized test length, rounded off upwards. (3%, 6%, 15%, 25% and 30% were allowed in strata 1 through 5, respectively.) Selection of items from stratum 6 was not restricted. These restrictions were formulated as

$$\sum_i c_{im} x_{ij} \leq \left\lfloor C_m \sum_i x_{ij} \right\rfloor \qquad \forall m, j$$

whereby C_m is the maximum fraction of items from stratum m in the tests.

- The items were selected from a random subdivision that was slightly larger than the total number of items needed. For USE, every year 1666 items were randomly drawn from the available items, in order to form the current pool. For DSL, the number of items needed was not fixed as the corresponding test specifications were based on a minimization model. The goal was, however, to assemble tests with TIFs comparable to the existing tests, using a maximum of approximately 30 items each. The aim, therefore, was to use no more than approximately 180 items per year. Based on this situation, the size of the item pools was determined to be 200 items.

- The idea behind the shadow pool method is the distribution of characteristics of the items in the tests is equal to that of the items reserved for future use. Therefore, the shadow pool should contain the vast majority of the items remaining available. The size of the shadow pool was determined at 90% of the remaining items. Thus, the size for year i was $0.9(5000 - 500i)$ for Scenario 1 and $0.9(10000 - 1500i)$ for Scenario 2. For Scenario 3, not the size of the shadow pool had to be restricted, but the target for the TIF had to be determined. The target was $122.6 * 6 * 0.9 * N_i/180$, where N_i denotes the size of the item pool in year i.

- Items were calibrated under the Rasch model. Use of this model resulted in the rejection of a total of 1318 items in the USE bank, and of 103 items in the DSL bank.

For Scenario 1, the results of the simulations are presented in Tables 3 and 4. The average ω-values in Table 3 showed a moderate increase for the Rasch model and the stratification approach, while they remained almost constant for the subdivision and shadow pool methods. In Table 4, the minimum TIF values at the three specified ϑ's are presented, averaged over the 100 tests assembled for each year, together with the standard deviation observed across these minima. In accordance with the increase in ω for the Rasch model and the stratification approach, the minimal TIF values showed a decline. Because of the fact that 1318 items had to be rejected for the Rasch model, in year 6 only 77 parallel tests could be assembled instead of the required 100 tests.

Table 3: Average ω-values for Scenario 1

Year	None	Rasch	Stratification	Subdivision	Shadow
1	0.101	0.146	0.134	0.136	0.136
2	0.129	0.147	0.135	0.136	0.136
3	0.142	0.148	0.142	0.136	0.137
4	0.152	0.151	0.136	0.138	0.137
5	0.155	0.154	0.144	0.137	0.137
6	0.150	0.155$\star$	0.147	0.135	0.135

Note: $\star$: Only 77 tests could be assembled

The results for Scenario 2 are shown in Tables 5 and 6. The differences were somewhat larger than for Scenario 1. In Scenario 1, every year a number of items was added according to the original parameter distribution, thus replenishing the highly discriminating items somewhat while Scenario 2 lacked this replenishment. Also in Scenario 2, the subdivision and shadow pool methods showed an output of tests with almost constant properties over the years. Note that for both scenarios, the Rasch method did not only fail to assemble the required number of tests in the last year but the height of the TIFs and the ω-values were also unfavorable compared to the other methods. All poorly fitting

Table 4: Average Minimum $I(\vartheta)$ for Scenario 1

Year	None		Rasch		Strat.		Subdiv.		Shadow	
	M	SD	M	SD	M	SD	M	SD	M	SD
1	78.6	0.8	45.1	0.1	51.2	0.1	50.7	0.4	51.8	0.4
2	56.4	2.7	44.5	0.1	51.2	0.3	51.5	0.5	51.4	0.4
3	47.5	0.6	44.0	0.2	49.5	0.4	51.6	0.6	50.8	0.4
4	41.1	0.5	42.6	0.2	50.9	1.2	49.7	0.6	51.2	0.5
5	40.0	0.5	41.2	0.2	47.1	1.5	50.3	0.5	50.5	0.5
6	42.2	0.6	40.3⋆	0.2	43.3	1.0	52.0	0.6	52.0	0.5

Note: ⋆ : Only 77 tests could be assembled

Table 5: Average ω-values for Scenario 2

Year	None	Rasch	Stratification	Subdivision	Shadow
1	0.102	0.146	0.133	0.136	0.136
2	0.118	0.146	0.134	0.136	0.136
3	0.136	0.147	0.135	0.136	0.136
4	0.148	0.149	0.137	0.138	0.136
5	0.192	0.153	0.141	0.137	0.137
6	0.196	0.165⋆	0.163	0.135	0.138

Note: ⋆ : Only 77 tests could be assembled

items were removed, this involved not only items that had a low a-parameter in the 2PL-model and thus would normally be ignored but also items with a high a-parameter for the 2PL-model, which would have contributed substantially to the information in the tests if they had been available.

The simulation results for Scenario 3 are presented in Tables 7 and 8.

The average ω-values are given in Table 7. For all strategies, except subdivision, the ω-values tended to decrease, a sign that the tests might not be parallel over the years. Contrary to Scenarios 1 and 2, the shadow pool method for Scenario 3 showed a slight deterioration. This phenomenon can be explained by the use of the test assembly models: In minimization models, the test length is not fixed but must be estimated instead. Yet, this number of items was used as the basis to determine the size of the shadow pool, or in case of minimization models, the target for the TIF of the shadow pool. In hindsight, the numbers of items used annually appeared to be smaller than the estimated 200 and an improved estimation might have improved the performance of the shadow pool method. On the other hand, the subdivision method appeared to be more robust against estimation errors for item usage.

Table 6: Average Minimum $I(\vartheta)$ for Scenario 2

Year	None		Rasch		Strat.		Subdiv.		Shadow	
	M	SD	M	SD	M	SD	M	SD	M	SD
1	91.0	1.2	45.3	0.0	51.2	0.0	50.7	0.4	52.4	0.4
2	65.5	0.6	45.1	0.0	51.2	0.1	51.5	0.5	52.0	0.5
3	51.2	0.1	44.6	0.0	51.4	0.1	51.6	0.6	51.8	0.4
4	45.2	0.3	43.8	0.0	51.4	0.3	49.7	0.6	51.4	0.4
5	26.6	0.0	41.9	0.1	50.2	1.2	50.3	0.5	50.8	0.5
6	25.5	0.0	36.7⋆	0.1	38.2	0.4	52.0	0.6	49.5	0.5

Note: $\star$: Only 77 tests could be assembled

Table 7: Average ω-values for Scenario 3

Year	None	Rasch	Stratification	Subdivision	Shadow
1	0.143	0.110	0.121	0.113	0.117
2	0.124	0.110	0.118	0.113	0.114
3	0.108	0.108	0.110	0.114	0.109
4	0.107	0.102	0.109	0.110	0.109
5	0.102	0.099	0.102	0.115	0.109

6 Discussion

Long-term use of items is one of the main reasons for which testing agencies make use of item banks. At the same time, the use of automated test assembly methods poses a threat to the life span of these banks. In this paper it is shown that item banks can be depleted rapidly if no counter measures are taken. Therefore, it is important for the design of an item bank to use methods for prevention of depletion from the start. In the two testing programs discussed in this paper, the reason of depletion was the tendency to select highly discriminating items. Various item bank management methods were considered, two of which appeared to effectively prevent depletion of highly discriminating items: random subdivision of the bank and the shadow pool method. Of these two, random subdivision of the item bank into several item pools, and subsequently selecting one pool as the active pool to assemble the tests from, is the simplest method. If the test specifications do not contain many restrictions, using this method ensures an output of tests of constant quality over the years. Although test specifications with larger numbers of restrictions were not needed in the programs under consideration, and hence were not considered in this paper, it remains to be seen whether this method still performs well.

The shadow pool method performed at least equally well for maximin models. On the other hand, for minimization models the shadow pool method

Table 8: Average Test Lengths for Scenario 3

Year	None	Rasch	Stratification	Subdivision	Shadow
1	12.0	27.0	20.5	20.3	20.8
2	17.5	27.0	21.7	20.8	22.5
3	22.0	28.7	23.8	20.3	25.3
4	26.7	30.7	25.7	22.2	25.0
5	30.2	35.0	29.3	19.7	24.7

seemed to be less robust than the random subdivision of the item bank into pools.

References

Ariel, A. (2005). *Contributions to Test-Item Bank Design and Management.* PhD thesis, University of Twente.

Belov, D. and Armstrong, R. (2004). A monte carlo approach for item pool analysis and design. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Russian Federation Ministry of Education and Science (2005). *Analytical Report on National Examinations in the System of Educational Quality Assessment.* Moscow: Author.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, 42:193–198.

van der Linden, W. (2005). *Linear Models for Optimal Test Design.* New York: Springer.

van der Linden, W. and Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika*, 54:237–247.

Verhelst, N., Glas, C., and Verstralen, H. (1995). *One-Parameter Logistic Model (OPLM).* Arnhem: Cito. Software and User's Manual.

Verschoor, A. (2005). *DOT 2005.* Arnhem: Cito. Software for Automated Test Assembly.

Way, W. and Steffen, M. (1998). Strategies for managing item pools to maximize item security. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.