**New Efficient and Practicable Adaptive Designs for Calibrating Items Online**

**Appendix A. Online calibration procedures**

The detailed procedures for administering new items online are described as follows.

**Step 0:** Predetermine tuning parameters*.*

a. Set $N\_{T}$, the total number of responses that each item receives when it is retired from the calibration process.

b. Set $n\_{r}$, the total number of examinees participating in the random calibration stage.

c. Set $n\_{b}$, the number of newly received responses for each item to update its item parameters during the adaptive calibration stage.

**Step 1:** Initialization.

a. Set $H=\{k|k=1, 2, \cdots m\}$, where $m$ is the total number of new items and $H$ indicates the ID set of all new items.

b. Set $N\_{k}=0$, where $N\_{k}$ is the cumulative number of responses that the $k$th ($k\in H$) new item has received.

c. Set $m\_{t}=0$, where $m\_{t}$ represents the number of retired items.

d. Set $n=0$, where $n$ denotes the cumulative number of simulated examinees.

e. Set $S\_{k}=0$, where $S\_{k}$ indicates the number of newly received responses on the $k$th ($k\in H$) item after its latest update during adaptive calibration stage.

**Step 2:** Random calibration stage.

a. If $m\_{t}=m$, stop; else, simulate an examine and set $n=n+1$, and then continue.

b. If $n>n\_{r}$, estimate the initial item parameter values based on the collected responses, and go to Step 3a; else, randomly assign one of the new item in $H$ to the test taker, and collect his/her response. Suppose new item $k$ ($k\in H$) is selected, then $N\_{k}=N\_{k}+1$, and continue.

c. If $N\_{k}\geq N\_{T}$, then $m\_{t}=m\_{t}+1$, and retire the new item $k$ by removing the index $k$ from $H$, and then go to Step 2a; else, go to Step 2a.

**Step 3:** Adaptive calibration stage.

a. If $S\_{k}\geq n\_{b}$, ($k\in H$), update the new item $k$’s parameters and reset $S\_{k}=0$, and then continue; else, continue.

b. A new item in $H$ is assigned to the current examinee based on the D-VR criterion or the four proposed ED criteria. Suppose item $k$ ($k\in H$) is selected, then set $N\_{k}=N\_{k}+1$ and $S\_{k}=S\_{k}+1$, and continue.

c. If $N\_{k}\geq N\_{T}$, then $m\_{t}=m\_{t}+1$, and retire the new item $k$ by removing the index $k$ from $H$, and continue; else, continue.

d. If $m\_{t}=m$, stop; else, simulate an examine and update $n$ as $n+1$, and go to Step 3a.

As seen from the above-described procedures, each new item is retired after it has been assigned to $N\_{T}$ examinees. Obviously, more examinees participating in the random calibration stage implies that fewer examinees will participate in the coming adaptive calibration stage. In other words, large random sample size $n\_{r}$ will improve the calibration precision of initial item parameter values while weakening the role of different adaptive design criteria in the whole calibration process. Furthermore, it will automatically become the random design if all new items are retired in random calibration stage before going further to the following adaptive calibration stage.

**Appendix B. Average computation costs for online calibration**

Table B1 records the average computation times of the whole online calibration process, including selecting operational items, updating ability estimates, assigning new items based on different criteria, simulating responses, updating item parameters, etc.

Table B1. Average computation time (*minute*) for the entire calibration process (20 new items)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| $$N\_{T}$$ | Sed. | D-VR | ED-o | ED-min (1 se) | ED-min (3 se) | ED-mean (1 se) | ED-mean (3 se) | ED-lw |
| 200 | Front | 2.01 | 2.40 | 3.60 | 3.51 | 2.46 | 2.82 | 2.48 |
| Middle | 2.04 | 2.60 | 3.57 | 3.12 | 2.91 | 2.63 | 2.45 |
| Rear | 2.06 | 2.56 | 3.33 | 2.82 | 2.18 | 2.29 | 2.40 |
| 400 | Front | 4.69 | 4.97 | 8.52 | 7.65 | 5.61 | 6.17 | 5.52 |
| Middle | 4.91 | 5.66 | 8.61 | 5.67 | 6.06 | 5.62 | 5.59 |
| Rear | 4.84 | 5.56 | 8.60 | 5.94 | 5.75 | 5.09 | 4.97 |
| 600 | Front | 7.21 | 8.15 | 13.48 | 10.89 | 9.99 | 9.47 | 9.60 |
| Middle | 8.02 | 8.22 | 13.40 | 10.81 | 9.93 | 9.58 | 9.20 |
| Rear | 7.74 | 7.97 | 13.35 | 9.12 | 7.29 | 7.75 | 7.72 |

Note: Sed. = Seeding location; (1 se) and (3 se) represent that $δ\_{p}$ is obtained as one and three standard error(s) of ability estimate, respectively.

**Appendix C. Bias against true abilities**

Plotted in Figures C1 to C3 are the average bias values against true abilities with three calibration sample sizes. Note that the ability estimation process was replicated 100 times to reduce random errors. To provide a baseline for comparison, abilities recovered by using true item parameters (this can be done in simulation) are also provided in these figures (see the results represented by the "True" legends).

Insert Figure C1 about here

Insert Figure C2 about here

Insert Figure C3 about here

**Appendix D. Results for calibrating the 3PL model**

A simple simulation study was conducted to apply the ED-o design to calibrate the 3PL model, in which parameters $a$ and $c$ were respectively generated from the uniform distribution (0.4, 1.4) and (0.05, 0.35), and parameter $b$ was randomly sampled from the standard normal distribution as in the study of Kim (2006). As seen from the results presented in Table D1, the ED-o design was also superior to the D-VR design when calibrating the 3PL model online.

Table D1. Relative efficiencies of D-VR and ED-o designs for calibrating 3PL model

|  |  |  |  |
| --- | --- | --- | --- |
| $$N\_{T}$$ | $n\_{r}$ = 100 | $n\_{r}$ = 300 | $n\_{r}$ = 500 |
| D-VR | ED-o | D-VR | ED-o | D-VR | ED-o |
| 200 | 1.0770 | 1.1766 | 1.0663 | 1.1534 | 1.0464 | 1.1051 |
| 400 | 1.0401 | 1.1481 | 1.0388 | 1.1386 | 1.0438 | 1.1283 |
| 600 | 1.0320 | 1.1490 | 1.0344 | 1.1484 | 1.0387 | 1.1352 |