## Appendix A

**Theorem 1**. For a fixed $k$, let $\mathcal{F}_k$ be the sigma algebra generated by the first $k$ items and responses. Then the estimator $\hat{\theta}_n$ defined in Step 2 above is strongly consistent as $n \to \infty$ under the conditional measure $P(.|\mathcal{F}_k)$. Letting

$$A_n = \sum_{i=1}^{n} w_i a_i (1 - c_i) G_i \bar{G}_i \tag{8}$$

and

$$B_n = \sqrt{\sum_{i=k+1}^{n} w_i^2 (1 - c_i)[c_i + (1 - c_i)G_i]\bar{G}_i}, \tag{9}$$

where $w_i = \frac{a_i(1+\sqrt{1+8c_i})}{2c_i+1+\sqrt{1+8c_i}}$, $G_i = 1/[1 + \exp\{-a_i(\theta - b_i)\}]$ and $\bar{G}_i = 1 - G_i$, suppose there is a nonrandom sequence $\nu_n$ such that

$$\frac{A_n}{\nu_n B_n} \xrightarrow{p} 1. \tag{10}$$

Then under the conditional measure $P(.|\mathcal{F}_k)$ for fixed $k$,

$$\nu_n\left(\hat{\theta}_n - \theta - \delta_n\right) \xrightarrow{d} N(0,1). \tag{11}$$

as $n \to \infty$, where

$$\delta_n = \frac{1}{A_n} \sum_{i=1}^{k} w_i [u_i - c_i - (1 - c_i)G_i]. \tag{12}$$

The claim also holds if the $\delta_n$ term in (11) is omitted.

*Proof*:

Consistency follows the proof of Chang and Ying (2009, Theorem 3) with only minor changes. For the limiting normal distribution, the condition (10) looks somewhat different from theirs, although we follow their proof with a few changes. Let $Z_i = w_i[U_i - c_i - (1 - c_i)G_i]$. Note that

$B_n^2$ is the conditional variance of $\sum_{i=k+1}^{n} Z_i$. It follows from the martingale central limit theorem (Pollard, 1984, p. 171) that

$$\frac{1}{B_n}\sum_{i=k+1}^{n} Z_i \xrightarrow{d} N(0,1) \tag{13}$$

as $n \to \infty$. Expanding Equation (4.4) of Chang and Ying (2009) with a Taylor series gives

$$0 = \sum_{i=1}^{n} w_i Z_i - \left(\hat{\theta}_n - \theta\right)\sum_{i=1}^{n} w_i a_i (1-c_i) G_i^* \bar{G}_i^*, \tag{14}$$

where $G_i^*$ and $\bar{G}_i^*$ are the same as the unstarred versions but with $\theta$ replaced by some value $\theta_n^*$ between $\hat{\theta}_n$ and $\theta$. Letting $A_n^*$ denote the starred version of $A_n$ which is asymptotically equivalent since $\theta_n^* \to \theta$, solving (14) for $\hat{\theta}_n - \theta$ and multiplying both sides by $\nu_n$ gives

$$\nu_n\left(\hat{\theta}_n - \theta\right) = \frac{\nu_n}{A_n^*}\sum_{i=1}^{n} Z_i = \frac{\nu_n}{A_n^*}\sum_{i=1}^{k} Z_i + \frac{\nu_n B_n}{A_n^*}\left(B_n^{-1}\sum_{i=k+1}^{n} Z_i\right). \tag{15}$$

The first term on the right-hand-side of (15) is equivalent to $\nu_n \delta_n$, and the second term in (15) approaches a standard normal by (10) and (13), proving (11). The final claim of the theorem holds because $\nu_n \delta_n = \frac{\nu_n}{A_n}\sum_{i=1}^{k} Z_i$ converges to 0 since, by (10), $\nu_n / A_n$ is equivalent to $B_n^{-1} \to 0$, and the sum that follows is finite.

## Appendix B

*Simulation Set 2*

In the second simulation set, two test constraints were imposed, namely content balancing and item exposure control. For the first constraint, each item in the pool was randomly assigned to one of four hypothetical content areas such that all four were equally represented in the pool with 125 items each. The content balancing constraint was such that equal representation of each content area was also preserved in each of the simulated tests. At any stage during the test, the next item to be administered was constrained to be the one from the content area that had been least represented up to the current item.

Item exposure control was based on the Sympson and Hetter (1985) method with the desired maximum exposure rate for all items being 30%. After determining a candidate item for administration, a random number generated from the U(0,1) distribution was drawn and compared to the item's exposure control parameter $P(A|S)$ where $A$ denotes administration and $S$ denotes selection. If the random number was less than the item's exposure control parameter, the item was then administered. Otherwise, the item was discarded from the pool and another item from the same content area became the next candidate for administration subject to its exposure control parameter. This process continued until an item from that content area was administered. The exposure control parameters for all items had been computed through a series of preliminary simulations that mimicked the actual conditions for simulation set 2. In particular, the same number of simulees (= 1,000) was generated at each of the 13 ability values from –0.6 to +0.6. In addition, the same content balancing constraint as described above was also used in the preliminary simulations. Since the CI stopping rule would result in the longest test for all $\theta$ values, it served as the termination criterion for the preliminary simulations.

**Table 2.** Simulation Set 2: Proportion of Correct Decisions, Average Test Lengths, and Average Losses with $c_W = 100$

| θ | Truncated CI | | | Curtailed CI | | | SC CI modified ML | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PCD | ATL | Average Loss | PCD | ATL | Average Loss | PCD | ATL | Average Loss | P |
| -0.60 | 0.994 | 27.5 | 28.10 | 0.994 | 23.23 | 23.83 | 0.987 | 20.82 | 22.12 | 0. |
| -0.50 | 0.979 | 31.19 | 33.29 | 0.978 | 24.85 | 27.05 | 0.973 | 21.62 | 24.32 | 0. |
| -0.40 | 0.958 | 36.24 | 40.44 | 0.959 | 27.40 | 31.50 | 0.945 | 22.96 | 28.46 | 0. |
| -0.30 | 0.904 | 39.93 | 49.53 | 0.907 | 30.60 | 39.90 | 0.884 | 24.49 | 36.09 | 0. |
| -0.20 | 0.805 | 43.67 | 63.17 | 0.814 | 34.25 | 52.85 | 0.794 | 26.74 | 47.34 | 0. |
| -0.10 | 0.660 | 45.63 | 79.63 | 0.673 | 37.27 | 69.97 | 0.669 | 28.07 | 61.17 | 0. |
| 0.00 | 0.496 | 46.12 | 96.52 | 0.486 | 38.67 | 90.07 | 0.488 | 28.82 | 80.02 | 0. |
| 0.10 | 0.668 | 45.64 | 78.84 | 0.658 | 39.27 | 73.47 | 0.640 | 28.38 | 64.38 | 0. |
| 0.20 | 0.801 | 43.47 | 63.37 | 0.793 | 37.88 | 58.58 | 0.785 | 27.25 | 48.75 | 0. |
| 0.30 | 0.915 | 40.20 | 48.70 | 0.911 | 35.55 | 44.45 | 0.904 | 25.77 | 35.37 | 0. |
| 0.40 | 0.956 | 37.21 | 41.61 | 0.953 | 32.99 | 37.69 | 0.948 | 24.01 | 29.21 | 0. |
| 0.50 | 0.985 | 32.26 | 33.76 | 0.983 | 29.22 | 30.92 | 0.978 | 22.15 | 24.35 | 0. |
| 0.60 | 0.994 | 27.07 | 28.30 | 0.993 | 25.95 | 26.65 | 0.991 | 21.10 | 22.00 | 0. |

Note: CI = Confidence Interval, SC = Stochastic Curtailment, ML = Maximum Likelihood, PCD = Proportion of Correct Decisions, ATL = Average Test Length

The results of the second simulation set are presented in Table 2. Under all four stopping rules, the content balancing constraint was satisfied for all simulees, with the most and least represented content areas differing by no more than one item in cases where the test length was not a multiple of four. In addition, the item exposure rate was also controlled below 30% under all four stopping rules. In particular, the item exposure rate was lowest under the more aggressive stochastically curtailed CI stopping rule, followed by the more conservative CI stopping rule, the curtailed CI stopping rule, and the original method, reflecting the ordering of ATL.

Comparing the results in Table 1 and those in Table 2, the PCDs were smaller and the ATLs were higher in the latter, which was reasonable due to the presence of test constraints. Within Table 2, the PCD of the curtailed CI stopping rule was always within 0.01 = 1% of that of the original CI stopping rule. The slightly lower PCD of the former was compensated by an average ATL reduction of 6.13 items across all $\theta$ values. In particular, the curtailed CI stopping rule was able to save at least 4 items at 11 of the 13 $\theta$ values, at least 5 items at 8 values, and at least 8 items at 4 values.

The PCD of the more aggressive stochastically curtailed CI stopping rule was always within 0.028 = 2.8% of the PCD of the original CI stopping rule, with a median difference of only 0.8% in favor of the latter. On the other hand, the former was able to save an average of 13.43 items across all $\theta$ values. In particular, the ATL reduction was at least 10 items at 10 of the 13 $\theta$ values, and at least 15 items at 6 values. Turning to the more conservative stochastically curtailed CI stopping rule, its PCD was always within 0.013 = 1.3% of that of the original CI stopping rule. On average, the ATL reduction under the former was 8.62 items across all $\theta$

values. More specifically, the ATL reduction was at least 5 items at 12 of the 13 $\theta$ values, and at least 10 items at 5 values.

The index (7) was also computed for each of the four stopping rules with $C_W = 100$. The same pattern was observed as in Table 1, namely that the more aggressive stochastically curtailed CI stopping rule proved to be the best termination criterion, followed by the more conservative approach, the curtailed version, and finally the original CI stopping rule.

## Appendix C

**Figure 1**. Comparison between ML and modified ML estimates for various ability levels