# An Investigation of Two Combination Procedures of SPRT for Three-category Classification Decisions in Computerized Classification Test

Hong Jiao
Shudong Wang
C. Allen Lau

Harcourt Assessment, Inc.

Paper presented at the annual meeting of American Educational Research Association, San Diego, CA. April 2004. Please make correspondence to: Hong Jiao, Harcourt Assessment, Inc. 19500 Bulverde Road, San Antonio, Texas 78259-3701. E-mail: hong_jiao@harcourt.com

**Acknowledgement**

We would like to express our sincere thanks to Dr. Judy Spray for her generosity in sharing her FORTRAN computer program. Thanks also go to Wendy Lam and Insu Paek for their help.

## Objectives

Two combination procedures of sequential probability ratio test (SPRT) (Wald, 1947) can be applied in computerized classification test (CCT) to make three-category classification decisions. One was proposed by Spray (1993), while the other was proposed by Eggen & Straetmans (2000). Little information is available regarding the performance of these two approaches. This study intends to investigate the properties of these two methods of combining the SPRT in CCT to make three-category classification decisions. The investigation of the performance of the two combination procedures of SPRT, Spray (1993) and Eggen & Straetmans (2000), under the same test conditions in CCT environment will provide information regarding the classification accuracy and efficiency of these two SPRT combination methods. This study will help test practitioners to make a decision regarding which method is better when CCT is utilized to make three-category decisions.

## Theoretical Framework

SPRT was first proposed by Wald (1947) for quality control. Reckase (1983) applied it to cognitive tests to make two-category classification decisions. Spray (1993) extended this method to make three-category classification decisions in CCT using SPRT. Spray's method was based on the SPRT combination procedure developed by Armitage (1950) by simultaneously testing three SPRTs. Eggen & Straetmans (2000) method was based on testing two SPRTs for three-category classification decisions, which was developed by Sobel & Wald (1949).

### Wald's Dichotomous Classification Decision

Originally, Wald's SPRT procedure makes a dichotomous classification decision by testing two statistical hypotheses:

$H_0$: $\theta = \theta_0$ vs. $H_1$: $\theta = \theta_1$.

where $\theta$ is an unknown parameter, and $\theta_0$ and $\theta_1$ are the lower and upper points around the cut score ($\theta_c$). The region between $\theta_0$ and $\theta_1$ is called the indifference region. The likelihood function is a function expressing the likelihood of an observed response pattern.

$$\pi = \prod P^u Q^{1-u}, \qquad (1)$$

where $P$ is the probability of a correct response given a certain $\theta$ value and the item parameter(s) defined by an item response theory (IRT) model. $Q$=1-$P$, which is the probability of an incorrect response given the ability and the item parameter(s). $u$ is the response of an examinee with the ability $\theta$ to an item with the parameter(s) designated by an IRT model. For the null and the alternative hypothesis, the likelihood function is $\pi(\theta_0)$ and $\pi(\theta_1)$ respectively. The ratio of these two functions is called the likelihood ratio, $\pi(\theta_1)/\pi(\theta_0)$.

A classification decision is made by comparing the observed likelihood ratio to the two boundaries, which are determined by the nominal Type I and Type II error rates (Wald, 1947). If the likelihood ratio is larger than or equal to the upper boundary, the alternative hypothesis is accepted and the examinee is classified as a master. If the likelihood is smaller than or equal to the lower boundary, the null hypothesis is accepted and the examinee is classified as a non-

master. If the likelihood ratio is between the two boundaries, no decision can be made and test continues.

Spray's Method for Three-Category Classification Decision

Spray (1993) proposed to use SPRT to make three-category decisions based on Armitage's method (1950). For making a three-category decision, two cut points, $\theta_{c1}$ and $\theta_{c2}$, need to be set up as *a priori*, where $\theta_{c2}$ denotes a higher latent ability level than $\theta_{c1}$. The indifference region is constructed based on the distance between the two cut scores. The half length of the distance is

$$\theta_m = \frac{\theta_{c2} - \theta_{c1}}{2} . \quad (2)$$

Then,

$$\theta_1 = \theta_{c1} - \theta_m , \quad (3)$$

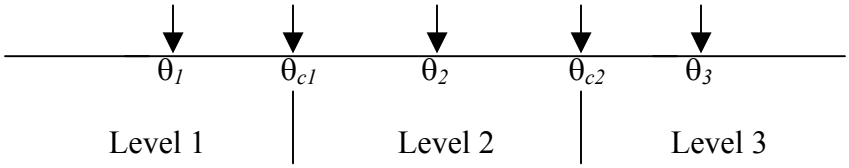
$$\theta_2 = \theta_{c1} + \theta_m , \quad (4)$$

$$\theta_3 = \theta_{c2} + \theta_m . \quad (5)$$

Three sets of SPRT hypotheses can be formulated as follows.

$H_{10}$: $\theta_i = \theta_1$, $H_{20}$: $\theta_i = \theta_2$, $H_{30}$: $\theta_i = \theta_1$,
$H_{11}$: $\theta_i = \theta_2$; $H_{21}$: $\theta_i = \theta_3$; $H_{31}$: $\theta_i = \theta_3$.

The classification decisions can be schematically represented as follows.




Nominal error rates are specified as follows to test the three sets of hypotheses.

$$p_{h|j} = (1 - p_{h|h}) \frac{d^{-1}{}_{hj}}{|D_h|}, \; h \neq j \quad (6)$$

where $p_{h|j}$ is the probability that $\theta = \theta_h$ is accepted but that $\theta = \theta_j$ is correct, $h$ =1, 2, 3; $j$ = 1, 2, 3. The power of any singe set of SPRT is $p_{h|h}$ and $p_{j|j}$. $d_{hj}$ is the absolute distance between $\theta_h$ and $\theta_j$, $h \neq j$. $|D_h| = \sum \frac{1}{d_{hj}}$ is summed over $j$, which is the norm of these distances.

To test $H_{0:}\,\theta = \theta_h$ vs. $H_{1:}\,\theta = \theta_j$, the upper boundary of the likelihood ratio is $\frac{p_{j|j}}{p_{j|h}}$ and the lower boundary is $\frac{p_{h|j}}{p_{h|h}}$, where $h = 1, 2, 3; j = 1, 2, 3;\ h \neq j$. Thus,

- If $L(x_{1,}x_2,...,x_n|\theta_{1,}\theta_2) \leq \frac{p_{1|2}}{p_{1|1}}$ and $L(x_{1,}x_2,...,x_n|\theta_{1,}\theta_3) \leq \frac{p_{1|3}}{p_{1|1}}$, then $\theta_i < \theta_{c1}$, the examinee is classified as Level 1;
- If $L(x_{1,}x_2,...,x_n|\theta_{2,}\theta_3) \leq \frac{p_{2|3}}{p_{2|2}}$ and $L(x_{1,}x_2,...,x_n|\theta_{1,}\theta_2) \geq \frac{p_{2|2}}{p_{2|1}}$, then $\theta_{c1} \leq \theta_i < \theta_{c2}$, the examinee is classified as Level 2;
- If $L(x_{1,}x_2,...,x_n|\theta_{2,}\theta_3) \geq \frac{p_{3|3}}{p_{3|2}}$ and $L(x_{1,}x_2,...,x_n|\theta_{1,}\theta_3) \geq \frac{p_{3|3}}{p_{3|1}}$, then $\theta_i \geq \theta_{c2}$, the examinee is classified as Level 3; otherwise, test continues.

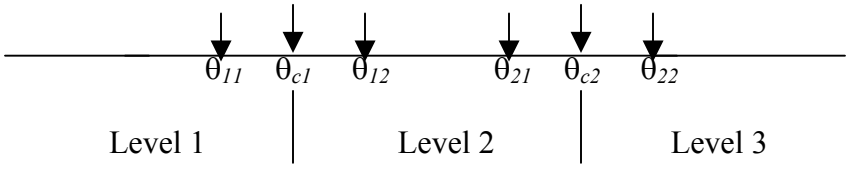
Eggen and Straetmans' Method for Three-Category Decision

Eggen & Straetmans (2000) proposed a generalized procedure based on Sobel & Wald (1949) work. Two sets of hypotheses are tested.

$H_{10}: \theta_i \leq \theta_{11}$, $\quad H_{20}: \theta_i \leq \theta_{21}$,
$H_{11}: \theta_i \geq \theta_{12}$; $\quad H_{21}: \theta_i \geq \theta_{22}$;

where $\theta_{11}$ and $\theta_{12}$ are the lower and the upper boundary of the indifference region around the lower cut score $\theta_{c1}$; and $\theta_{21}$ and $\theta_{22}$ are the lower and the upper boundary of the indifference region around the higher cut score $\theta_{c2}$.

The classification decisions can be schematically represented as follows.




The classification error rates are $\alpha_1$, $\beta_1$, $\alpha_{2,}$ $\beta_2$, are

$P(\text{accept } H_{10} | \ H_{10} \text{ is true}) \geq 1\text{-}\alpha_1$ $\quad P(\text{accept } H_{10} | \ H_{11} \text{ is true}) \leq \beta_1$
$P(\text{accept } H_{20} | \ H_{20} \text{ is true}) \geq 1\text{-}\alpha_2$ $\quad P(\text{accept } H_{20} | \ H_{21} \text{ is true}) \leq \beta_2$

For a three-category classification decision, when the nominal error rates for the two hypotheses $\alpha_1$, $\beta_1$, $\alpha_2$, $\beta_2$ are equal to $\alpha$, and half of the indifference region is set to be $\delta$, and $A=\ln(1-\alpha)/\alpha$, decisions will be made as follows.

- If $\sum_{i=1}^{k} a_i x_i \le \frac{-A-\sum_{i=1}^{k}\ln\frac{Q_i(\theta_{c1}+\delta)}{Q_i(\theta_{c1}-\delta)}}{2\delta}$, then $\theta_i < \theta_{c1}$, the examinee is classified as Level 1;
- If $\frac{A-\sum_{i=1}^{k}\ln\frac{Q_i(\theta_{c1}+\delta)}{Q_i(\theta_{c1}-\delta)}}{2\delta} \le \sum_{i=1}^{k} a_i x_i \le \frac{-A-\sum_{i=1}^{k}\ln\frac{Q_i(\theta_{c2}+\delta)}{Q_i(\theta_{c2}-\delta)}}{2\delta}$, then $\theta_{c1} \le \theta_i < \theta_{c2}$, the examinee is classified as Level 2;
- If $\sum_{i=1}^{k} a_i x_i \ge \frac{A-\sum_{i=1}^{k}\ln\frac{Q_i(\theta_{c2}+\delta)}{Q_i(\theta_{c2}-\delta)}}{2\delta}$, then$\theta_i \ge \theta_{c2}$, the examinee is classified as Level 3; otherwise, test continues.

<u>Classification Decision Making For Test Length Constraint</u>

When the maximum test length is reached, a forced decision has to be made. The examinee's weighted score, $\sum_i a_i x_i$, will be compared with two boundaries. Classification will be made as follows.

- If $\sum_{i=1}^{k} a_i x_i \le \frac{-\sum_{i=1}^{k}\ln\frac{Q_i(\theta_{c1}+\delta)}{Q_i(\theta_{c1}-\delta)}}{2\delta}$, then the examinee is classified as Level 1;
- If $\frac{-\sum_{i=1}^{k}\ln\frac{Q_i(\theta_{c1}+\delta)}{Q_i(\theta_{c1}-\delta)}}{2\delta} \le \sum_{i=1}^{k} a_i x_i \le \frac{-\sum_{i=1}^{k}\ln\frac{Q_i(\theta_{c2}+\delta)}{Q_i(\theta_{c2}-\delta)}}{2\delta}$, then the examinee is classified as Level 2;
- If $\sum_{i=1}^{k} a_i x_i \ge \frac{-\sum_{i=1}^{k}\ln\frac{Q_i(\theta_{c2}+\delta)}{Q_i(\theta_{c2}-\delta)}}{2\delta}$, then the examinee is classified as Level 3.

This forced decision procedure was applied to both Spray's and Eggen & Straetmans' methods.

## Methods

To compare the performance of these two combination methods of SPRT in making three-category classification decisions, a series of CCT test conditions were simulated using S-

PLUS. Then, the two procedures were compared in terms of classification accuracy and the average test length or the average number of items used for classification decisions.

## Research Design

*Fixed Variable*

In all simulated CCT test conditions, the two true cut scores were set at –0.5 and 0.5 on the latent ability scale. Half width of the indifference region for the SPRT was fixed at 0.1 for Eggen & Straetmans method. The nominal type I and type II error rates for the two hypotheses testing were all set at the 0.1 level. The power for a single SPRT was 0.9.

*Independent Variable*

Simulation was implemented under 4 study conditions that were the combination of three manipulated factors. The three manipulated factors were the item selection method, the constraints of the test length, and the stratum depth indicating the item exposure rate control in the randomization scheme.

For item selection, two methods, random and midpoint methods were employed. When the random method was used, an item to be administered was the one randomly selected from the item pool excluding the ones used before. When the midpoint method was applied, the item selected was the one with the maximum information at the midpoint of the two cut scores.

Regarding test length constraint, one condition did not impose any constraint. The other condition set the minimum and the maximum test length as 30 and 60 items respectively to represent a realistic testing situation. The levels of the three manipulated factors are as follows:

1. Item selection algorithm:
   i). Random method,
   ii). Midpoint method,
2. Test length constraint:
   i). Minimum=30; maximum=60
   ii). Minimum=1; maximum=300
3. Stratum depth (item exposure control):
   i). 1
   ii). 5

The resulted test conditions were summarized in Table 1.

CCT3KS stands for Spray's method. CCT3KE stands for Eggen & Straetmans' method. A1L1S1 stands for the test condition where random item selection method was applied and no constraint was imposed. A1L2S2 stands for the test condition where midpoint item selection method was applied, test length was constrained, and item exposure was controlled. A2L1S1 stands for the test condition where midpoint item selection method was applied, and no test constraints were imposed. A2L2S2 stands for the test condition where midpoint item selection method was applied, test length was constrained, and item exposure was controlled.

## Data Generation

The latent ability parameters for 10000 examinees were generated from a normal distribution with a mean of 0 and a standard deviation of 1. Item parameters were generated for 2PL IRT model. The item discrimination parameters (*a*-parameter) were generated from a log normal distribution with mean log of 0 and standard deviation log of 0.4 and within the range from .4 to 1.6. The true item difficulty parameters (*b*-parameter) were generated from a normal distribution with mean of 0 and standard deviation of 1 and within the range of –1.5 to 1.5. To guarantee the created *a*- and *b*-parameters within the desired range, 600 values were generated for both *a*- and *b*- parameters. Then, the values outside the bounds were filtered off. Lastly, a random sample of 300 was selected. The item and true person parameters remained the same for all simulated CCTs.

During the implementation of a CCT, an item response was generated as follows. First, the probability of a person correctly answering a selected item was obtained by incorporating the item and the true ability parameters into the IRT model. This probability was then compared with a random number generated from a uniform distribution from 0 to 1. If the probability was larger than or equal to the random number, a correct response of 1 was obtained, otherwise, an incorrect response of 0.

The item pool was calibrated using a sample of 2000 examinees' responses to 300 items in the item pool. Given the item responses of these 2000 examinees to the 300 items, the items were calibrated using 2PL IRT model via the BILOG-MG program. The obtained item parameters from BILOG-MG were the item parameters used in CCT simulations.

Simulation of CCT

When the random item selection method was applied in CCT implementation, the first item was randomly selected for administration. If the midpoint item selection method was applied, the items in an item pool were ranked in a descending order based on the amount of item information at 0, the midpoint of the two cut scores. If no item exposure constraint was imposed, the first item selected was the item in the pool with the largest information. Each item selected was always one with the highest item information at 0 in the item pool. When the stratum depth of 5 was used to control item exposure rate, 300 items in the item pool were divided into 60 strata, with each stratum consisting of 5 items based on the amount of item information at 0. In this case, the first item administered was an item randomly selected from the 5 items in Stratum1. The second item was an item randomly selected from the 5 items in Stratum 2. Once an item was selected from an item pool, the response to the item was simulated based on the true person's ability and item parameters. Then, different algorithms were used to make a classification decision for Spray's method and Eggen & Straetmans' method. If test length constraint was imposed, the likelihood ratio was calculated after the minimum number of items, 30 items were administered. If the maximum test length of 60 or 300 items had been reached, and no classification decision could be made, then a classification decision was forced. In each simulation, the classification decision and the number of items used to come up with a classification decision were recorded for each examinee.

## Results and Discussion

The descriptive statistics of the generated data for the person and the item parameters are summarized in Table 2. All these parameters are close to their pre-specified values.

The classification accuracy and the average test length for making three-category classification decisions for Spray (1993) and Eggen & Straetmans (2000) methods were compared under each simulated CCT condition. Among the 10000 examinees, 3048 (30.5%) examinees should be in level 1, 3879 (38.8%) examinees should be in level 2, and 3073 (30.7%) examinees should be in level 3.

For test condition A1L1S1 where items were randomly selected from the item pool and no constraint was imposed, the comparison results are summarized in Table 3. When Spray's method was applied, 90.2% of level 1 examinees, 72.6% of level 2 examinees, and 90.6% of level 3 examinees was correctly classified. The classification error rates were distributed as 3% in level 1, 10.62% in level 2, and 2.9% in level 3. When Eggen & Straetmans' method was used, 94.5% of level 1 examinees, 89.5% of level 2 examinees, and 94.1% of level 3 examinees was correctly classified. The classification error rates were distributed as 1.69% in level 1, 4.07% in level 2, and 1.81% in level 3.

For test condition A1L2S2 where items were randomly selected from the item pool, test length was constrained and item exposure was controlled, the comparison results are summarized in Table 4. When Spray's method was applied, 88.3% of level 1 examinees, 73.8% of level 2 examinees, and 87.6% of level 3 examinees was correctly classified. The classification error rates were distributed as 3.58% in level 1, 10.16% in level 2, and 3.81% in level 3. When Eggen & Straetmans' method was used, 87.0% of level 1 examinees, 75.5% of level 2 examinees, and 87.2% of level 3 examinees was correctly classified. The classification error rates were distributed as 3.96% in level 1, 9.5% in level 2, and 3.94% in level 3.

For test condition A2L1S1 where items were selected based on the largest item information at the midpoint of the two cut scores and no constraint was imposed, the comparison results are summarized in Table 5. When Spray's method was applied, 90.8% of level 1 examinees, 72.9% of level 2 examinees, and 91.9% of level 3 examinees was correctly classified. The classification error rates were distributed as 2.8% in level 1, 10.53% in level 2, and 2.49% in level 3. When Eggen & Straetmans' method was used, 93.8% of level 1 examinees, 89.6% of level 2 examinees, and 94.2% of level 3 examinees was correctly classified. The classification error rates were distributed as 1.89% in level 1, 4.02% in level 2, and 1.79% in level 3.

For test condition A2L2S2 where items were selected based on the largest item information at the midpoint of the two cut scores, test length was constrained, and item exposure was controlled, the comparison results are summarized in Table 6. When Spray's method was applied, 88% of level 1 examinees, 75% of level 2 examinees, and 87% of level 3 examinees was correctly classified. The classification error rates were distributed as 3.65% in level 1, 9.68% in level 2, and 3.98% in level 3. When Eggen & Straetmans' method was used, 87.1% of level 1 examinees, 74.6% of level 2 examinees, and 87.3% of level 3 examinees was correctly classified. The classification error rates were distributed as 3.94% in level 1, 9.86% in level 2, and 3.9% in level 3.

Table 7 summarizes the overall correct classification rates, classification error rates, and the average test length for each simulated CCT condition using each of the SPRT combination method. When there was no constraint, no matter random or midpoint item selection method was employed, Eggen and Straetmans' method yielded higher correct classification rates but much longer average test length. When test length was constrained and item exposure was controlled through the randomization scheme, these two methods produced about the same error rates. However, Spray's method used shorter average test length.

To further explore the location that classification error occurred, the classification results for the ability interval from –1 to +1 were analyzed. The frequency for the classified levels is summarized in Table 8 to Table 11. The classification error rates for the ability interval from –1 to +1 was shown in Table 12. These error rates were relative to the total number of examinees in one CCT simulation. It was obvious that almost all the classification errors were committed in the interval from –1 to +1. That is, the classification errors most frequently occurred around the two cut scores, -0.5, and +0.5. This is consistent with what was observed in Spray (1993).

## Summary

Spray's method and Eggen & Straetmans' method of combining SPRT both can be applied to make a three-category classification decision in CCT environment. In general, Spray's method required fewer average number of items to come up with a classification decision. When test length was constrained and item exposure was controlled, these two methods did not make differences in terms of correct classification rates. When these two constraints were not imposed, Eggen & Straetmans's method yielded higher correct classification rates. No matter which SPRT combination method was applied, classification errors were most often committed around the two cut scores.

The generalizability of the study results is limited to the simulation conditions of the CCT tests. In order to obtain a more comprehensive picture of these two SPRT combination methods for three-category classification decision in CCT, more realistic test conditions probably need to be simulated to better understand the properties or characteristics of Spray's method and Eggen & Straetmans' method

## References

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society*, *B(12)*, 137-144.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, *60*, 713-734.

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

Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annuals of Mathematical Statistics*, *20*, 502-522.

Spray, J. (1993). Multiple-category classification using a sequential probability ratio test. ACT research report series, 93-7.

Wald, A. (1947). *Sequential Analysis*. New York: Wiley.

Table 1. Test Conditions.

| Test Condition | Item Selection Method | Test Length Constraint Minimum-Maximum | Stratum Depth |
|---|---|---|---|
| 1 (A1L1S1) | Random (A1) | 1-300 (L1) | 1 (S1) |
| 2 (A1L2S2) | Random (A1) | 30-60 (L2) | 5 (S2) |
| 3 (A2L1S1) | Midpoint (A2) | 1-300 (L1) | 1 (S1) |
| 4 (A2L2S2) | Midpoint (A2) | 30-60 (L2) | 5 (S2) |

Note: The symbol for a specific test condition is inside the parenthesis.

Table 2. Generated Item Parameters.

| | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|
| Ability | -3.4719 | 3.8359 | 0.0049 | 0.9978 |
| Item Difficulty | -1.4640 | 1.4947 | -0.0551 | 0.7234 |
| Item Discrimination | 0.4000 | 1.5900 | 0.9132 | 0.2846 |

Table 3. Classification Error Rates for Test Condition A1L1S1.

| CCT3KS.A1L1S1 | | | Classification 1 | 2 | 3 | Total |
|---|---|---|---|---|---|---|
| True Class | 1 | Count | 2749 | 297 | 2 | 3048 |
| | | % within true class | 90.2% | 9.7% | 0.1% | 100% |
| | | % of total | 27.5% | 3.0% | 0.0% | 30.5% |
| | 2 | Count | 538 | 2817 | 524 | 3879 |
| | | % within true class | 13.9% | 72.6% | 13.5% | 100% |
| | | % of total | 5.4% | 28.2% | 5.2% | 38.8% |
| | 3 | Count | 2 | 288 | 2783 | 3073 |
| | | % within true class | 0.1% | 9.4% | 90.6% | 100% |
| | | % of total | 0.0% | 2.9% | 27.8% | 30.7% |
| CCT3KE.A1L1S1 | | | | | | |
| True Class | 1 | Count | 2879 | 169 | 0 | 3048 |
| | | % within true class | 94.5% | 5.5% | 0.0% | 100% |
| | | % of total | 28.8% | 1.7% | 0.0% | 30.5% |
| | 2 | Count | 206 | 3472 | 201 | 3879 |
| | | % within true class | 5.3% | 89.5% | 5.2% | 100% |
| | | % of total | 2.1% | 34.7% | 2.0% | 38.8% |
| | 3 | Count | 0 | 181 | 2892 | 3073 |
| | | % within true class | 0.0% | 5.9% | 94.1% | 100% |
| | | % of total | 0.0% | 1.8% | 28.9% | 30.7% |

Note: CCT3KS stands for Spray's method. CCT3KE stands for Eggen & Straetmans' method. A1L1S1 stands for the test condition where random item selection method was applied and no constraints was imposed.

Table 4. Classification Error Rates for Test Condition A1L2S2.

| CCT3KS.A1L2S2 | | | Classification 1 | 2 | 3 | Total |
|---|---|---|---|---|---|---|
| True Class | 1 | Count | 2690 | 358 | 0 | 3048 |
| | | % within true class | 88.3% | 11.7% | 0.0% | 100.0% |
| | | % of total | 26.9% | 3.6% | 0.0% | 30.5% |
| | 2 | Count | 510 | 2863 | 506 | 3879 |
| | | % within true class | 13.1% | 73.8% | 13.0% | 100.0% |
| | | % of total | 5.1% | 28.6% | 5.1% | 38.8% |
| | 3 | Count | 1 | 380 | 2692 | 3073 |
| | | % within true class | 0.0% | 12.4% | 87.6% | 100.0% |
| | | % of total | 0.0% | 3.8% | 26.9% | 30.7% |
| CCT3KE.A1L2S2 | | | | | | |
| True Class | 1 | Count | 2652 | 395 | 1 | 3048 |
| | | % within true class | 87.0% | 13.0% | 0.0% | 100.0% |
| | | % of total | 26.5% | 4.0% | 0.0% | 30.5% |
| | 2 | Count | 458 | 2929 | 492 | 3879 |
| | | % within true class | 11.8% | 75.5% | 12.7% | 100.0% |
| | | % of total | 4.6% | 29.3% | 4.9% | 38.8% |
| | 3 | Count | 0 | 394 | 2679 | 3073 |
| | | % within true class | 0.0% | 12.8% | 87.2% | 100.0% |
| | | % of total | 0.0% | 3.9% | 26.8% | 30.7% |

Note: CCT3KS stands for Spray's method. CCT3KE stands for Eggen & Straetmans' method. A1L2S2 stands for the test condition where midpoint item selection method was applied, test length was constrained, and item exposure was controlled.

Table 5. Classification Error Rates for Test Condition A2L1S1.

| CCT3KS.A2L1S1 | | | Classification 1 | 2 | 3 | Total |
|---|---|---|---|---|---|---|
| True Class | 1 | Count | 2768 | 280 | 0 | 3048 |
| | | % within true class | 90.8% | 9.2% | 0.0% | 100.0% |
| | | % of total | 27.7% | 2.8% | 0.0% | 30.5% |
| | 2 | Count | 525 | 2826 | 528 | 3879 |
| | | % within true class | 13.5% | 72.9% | 13.6% | 100.0% |
| | | % of total | 5.3% | 28.3% | 5.3% | 38.8% |
| | 3 | Count | 1 | 248 | 2824 | 3073 |
| | | % within true class | 0.0% | 8.1% | 91.9% | 100.0% |
| | | % of total | 0.0% | 2.5% | 28.2% | 30.7% |
| CCT3KE.A2L1S1 | | | | | | |
| True Class | 1 | Count | 2859 | 189 | 0 | 3048 |
| | | % within true class | 93.8% | 6.2% | 0.0% | 100.0% |
| | | % of total | 28.6% | 1.9% | 0.0% | 30.5% |
| | 2 | Count | 202 | 3477 | 200 | 3879 |
| | | % within true class | 5.2% | 89.6% | 5.2% | 100.0% |
| | | % of total | 2.0% | 34.8% | 2.0% | 38.8% |
| | 3 | Count | 0 | 179 | 2894 | 3073 |
| | | % within true class | 0.0% | 5.8% | 94.2% | 100.0% |
| | | % of total | 0.0% | 1.8% | 28.9% | 30.7% |

Note: CCT3KS stands for Spray's method. CCT3KE stands for Eggen & Straetmans' method. A2L1S1 stands for the test condition where midpoint item selection method was applied, and no test constraints were imposed.

Table 6. Classification Error Rates for Test Condition A2L2S2.

| CCT3KS.A2L2S2 | | | Classification 1 | 2 | 3 | Total |
|---|---|---|---|---|---|---|
| True Class | 1 | Count | 2683 | 365 | 0 | 3048 |
| | | % within true class | 88.0% | 12.0% | 0.0% | 100.0% |
| | | % of total | 26.8% | 3.7% | 0.0% | 30.5% |
| | 2 | Count | 469 | 2911 | 499 | 3879 |
| | | % within true class | 12.1% | 75.0% | 12.9% | 100.0% |
| | | % of total | 4.7% | 29.1% | 5.0% | 38.8% |
| | 3 | Count | 0 | 398 | 2675 | 3073 |
| | | % within true class | 0.0% | 13.0% | 87.0% | 100.0% |
| | | % of total | 0.0% | 4.0% | 26.8% | 30.7% |
| CCT3KE.A2L2S2 | | | | | | |
| True Class | 1 | Count | 2654 | 394 | 0 | 3048 |
| | | % within true class | 87.1% | 12.9% | 0.0% | 100.0% |
| | | % of total | 26.5% | 3.9% | 0.0% | 30.5% |
| | 2 | Count | 485 | 2893 | 501 | 3879 |
| | | % within true class | 12.5% | 74.6% | 12.9% | 100.0% |
| | | % of total | 4.9% | 28.9% | 5.0% | 38.8% |
| | 3 | Count | 0 | 390 | 2683 | 3073 |
| | | % within true class | 0.0% | 12.7% | 87.3% | 100.0% |
| | | % of total | 0.0% | 3.9% | 26.8% | 30.7% |

Note: CCT3KS stands for Spray's method. CCT3KE stands for Eggen & Straetmans' method. A2L2S2 stands for the test condition where midpoint item selection method was applied, test length was constrained, and item exposure was controlled.

Table 7. Correct Classification Rates, Classification Error Rates, and Average Test Length.

| | Percentage of Correct Classification | Average Test Length |
|---|---|---|
| CCT3KS.A1L1S1 | 83.49% (16.51%) | 40.20 |
| CCT3KE.A1L1S1 | 92.43% (7.57%) | 171.08 |
| CCT3KS.A1L2S2 | 82.45% (17.55%) | 40.82 |
| CCT3KE.A1L2S2 | 82.60% (17.40%) | 58.69 |
| CCT3KS.A2L1S1 | 84.18% (15.82%) | 19.13 |
| CCT3KE.A2L1S1 | 92.30% (7.70%) | 137.86 |
| CCT3KS.A2L2S2 | 82.69% (17.31%) | 39.35 |
| CCT3KE.A2L2S2 | 82.30% (17.70%) | 58.14 |

Table 8. Classification Error for the Test Condition A1L1S1 for the Ability Interval from –1 to +1.

| | | Classification | | | |
|---|---|---|---|---|---|
| CCT3KS.A1L1S1 | | 1 | 2 | 3 | Total |
| True Class | 1 | 1189 | 289 | 2 | 1480 |
| | 2 | 538 | 2817 | 524 | 3879 |
| | 3 | 2 | 280 | 1198 | 1480 |
| CCT3KE.A1L1S1 | | | | | |
| True Class | 1 | 1311 | 169 | 0 | 1480 |
| | 2 | 206 | 3472 | 201 | 3879 |
| | 3 | 0 | 181 | 1299 | 1480 |

Table 9. Classification Error for the Test Condition A1L2S2 for the Ability Interval from –1 to +1.

| | | Classification | | | |
|---|---|---|---|---|---|
| CCT3KS.A1L2S2 | | 1 | 2 | 3 | Total |
| True Class | 1 | 1135 | 345 | 0 | 1480 |
| | 2 | 510 | 2863 | 506 | 3879 |
| | 3 | 1 | 362 | 1117 | 1480 |
| CCT3KE.A1L2S2 | | | | | |
| True Class | 1 | 1102 | 377 | 1 | 1480 |
| | 2 | 458 | 2929 | 492 | 3879 |
| | 3 | 0 | 375 | 1105 | 1480 |

Table 10. Classification Error for the Test Condition A2L1S1 for the Ability Interval from –1 to +1.

| | | Classification | | | |
|---|---|---|---|---|---|
| CCT3KS.A2L1S1 | | 1 | 2 | 3 | Total |
| True Class | 1 | 1206 | 274 | 0 | 1480 |
| | 2 | 525 | 2826 | 528 | 3879 |
| | 3 | 1 | 242 | 1237 | 1480 |
| CCT3KE.A2L1S1 | | | | | |
| True Class | 1 | 1291 | 189 | 0 | 1480 |
| | 2 | 202 | 3477 | 200 | 3879 |
| | 3 | 0 | 179 | 1301 | 1480 |

Table 11. Classification Error for the Test Condition A2L2S2 for the Ability Interval from –1 to +1.

| | | Classification | | | |
|---|---|---|---|---|---|
| CCT3KS.A2L2S2 | | 1 | 2 | 3 | Total |
| True Class | 1 | 1137 | 343 | 0 | 1480 |
| | 2 | 469 | 2911 | 499 | 3879 |
| | 3 | 0 | 379 | 1101 | 1480 |
| CCT3KE.A2L2S2 | | | | | |
| True Class | 1 | 1101 | 379 | 0 | 3048 |
| | 2 | 485 | 2893 | 501 | 3879 |
| | 3 | 0 | 374 | 1106 | 1480 |

Table 12. Classification Error Rates for All Ability and for the Ability Interval from –1 to +1.

| | All Ability | Ability Interval from -1 to +1 |
|---|---|---|
| CCT3KS.A1L1S1 | 16.51% | 16.35% |
| CCT3KE.A1L1S1 | 7.57% | 7.57% |
| CCT3KS.A1L2S2 | 17.55% | 17.24% |
| CCT3KE.A1L2S2 | 17.40% | 17.03% |
| CCT3KS.A2L1S1 | 15.82% | 15.70% |
| CCT3KE.A2L1S1 | 7.70% | 7.70% |
| CCT3KS.A2L2S2 | 17.31% | 16.90% |
| CCT3KE.A2L2S2 | 17.70% | 17.39% |

Note: The error rates for the ability interval from –1 to +1 were relative to 10000 examinees.