A New Approach To Detect Answer Copying

Dmitry Belov

Law School Admission Council

Problem statement

Given a set of responses, identify pairs of examinees involved in answer copying

Structure of response vector

operational part (scored) has items identical for all examinees

`1111101111101110111111111` **`1011010000001011001 0`**

ABCDEABCDEABCDEABCDEABCDE **ABCDEABCDEABCDEABCDE**

variable part (unscored) has items different for adjacent examinees

This is relevant to PPT and CBT

Two effects: blind copy effect and shift copy effect

Pairs found in real data from LSAT

…10111111101011011… 100110000100000000001000000
…53135532122315543… 255254322454242423122254142
…53135532122315543… 445254322454242423125254142

…11100110101… 010000000100000000000100000
…25312243312… 545322243141243332244534423
…25312243312… 54332224312124333224453442

…00911010110100… 000100000000000000001000000
…12445111122215… 323531215351541334523214245
…12445111122215… 232532215114421252132314245

…11011011110110… 000000000010000001000000100
…35122413433333… 121335232115132142254145441
…35122413433333… 121332523211513214254145441

A new statistic VM-Index

Given two examinees c and s taking a linear test partitioned into T and V, consider the following random variable:

$$\eta_i \equiv \eta_i(w_c, w_s) = \sum_{j=l}^{u} \begin{cases} 1 & \begin{array}{l} c \text{ selects incorrect answer to item } i \in V_c \text{ and} \\ s \text{ selects the same answer to item } i + j \in V_s \end{array} \\ 0 & \text{otherwise} \end{cases},$$

where w_c [w_s] is the number of incorrect responses of c [s] to T.

The conditional number of agreements

$$\xi \equiv \xi(w_c, w_s) = \sum_{i \in V_c} \eta_i$$

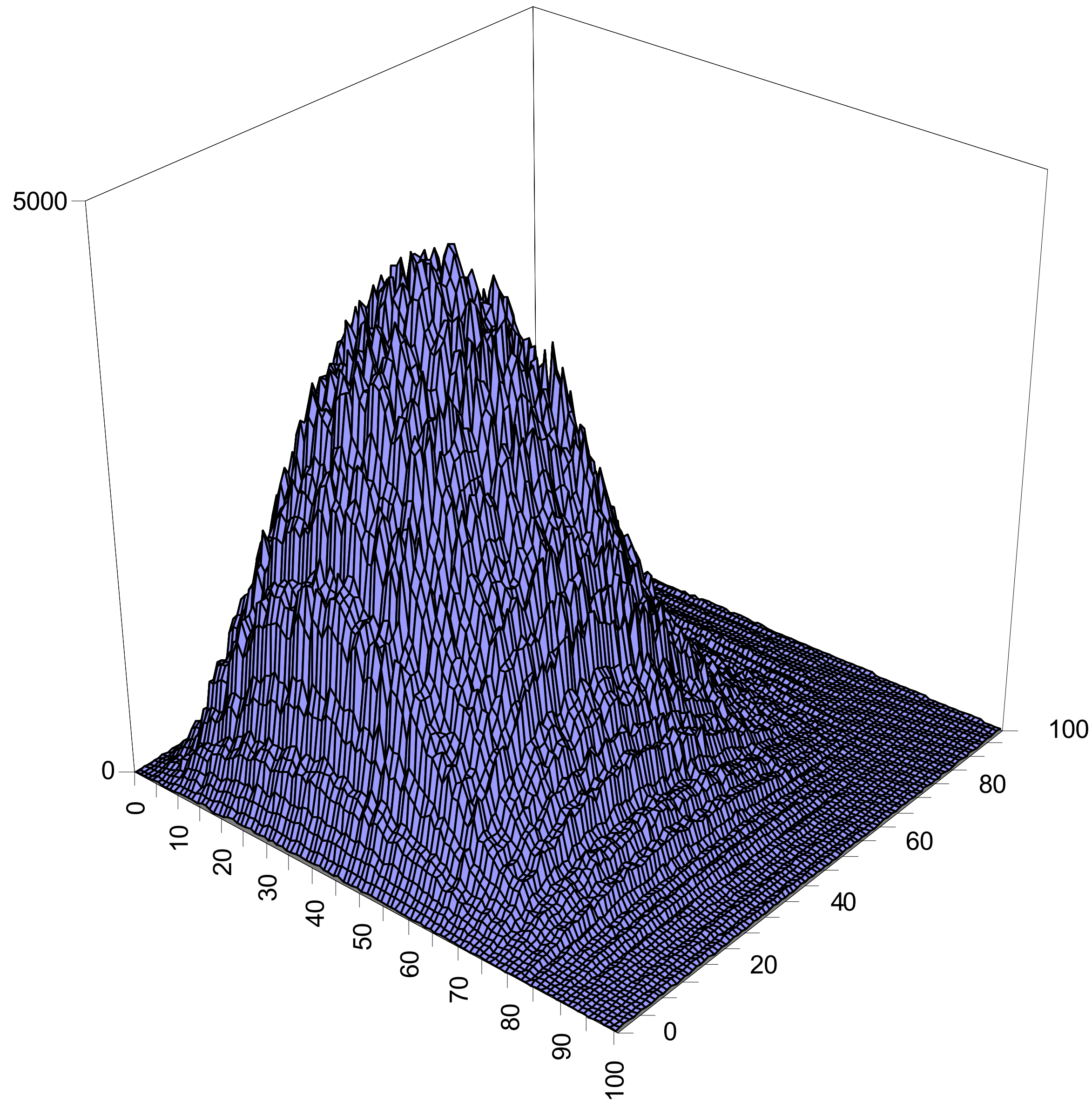
1. If $l = u = 0$, the VM-Index detects a blind-copy effect.
2. If $l \le u < 0$, the VM-Index detects a negative shift-copy effect.
3. If $0 < l \le u$, the VM-Index detects a positive shift-copy effect.
4. If $l < 0 < u$, the VM-Index detects all of the above effects.

Monte Carlo method to compute empirical distributions of ξ

Step 1: Create $(|T|+1)\times(|T|+1)$ matrix $\mathbf{A}$, where each element $\mathbf{A}[x][y]$ contains a frequency distribution of the corresponding random variable $\xi(x,y)\in\{0,1,2,\ldots,n\}$, $n=(u-l+1)\max_{i=1}^{m}|V_i|$, where m is the number of different variable parts. Initialize $\mathbf{A}$ to zero.

Step 2: Select a random pair of examinees (c,s). Compute w_c, w_s, and ξ. Increase $\mathbf{A}[w_c][w_s][\xi]$ by 1. Repeat Step 2 multiple times.

Real data from LSAT (22000). Distribution of random variable (w_c, w_s):



Monte Carlo method to compute empirical distributions of ξ

Step 3: Compute matrix $\mathbf{P}$ of empirical distributions of ξ by the following:

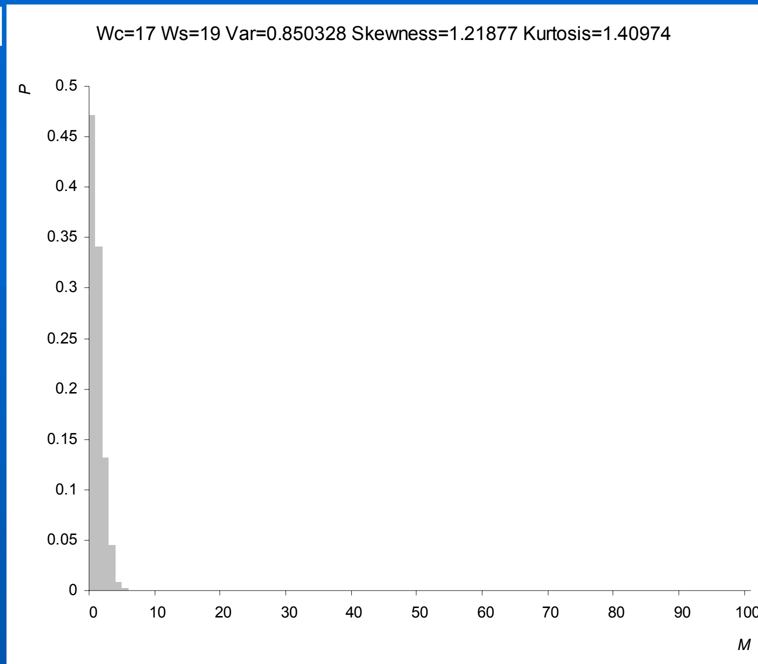
$$\mathbf{P}[x][y][k] = \mathbf{A}[x][y][k] / \sum_{j=0}^{n} \mathbf{A}[x][y][j],$$

where $\mathbf{P}[x][y]$ represents the distribution of $\xi \equiv \xi(x,y)$.

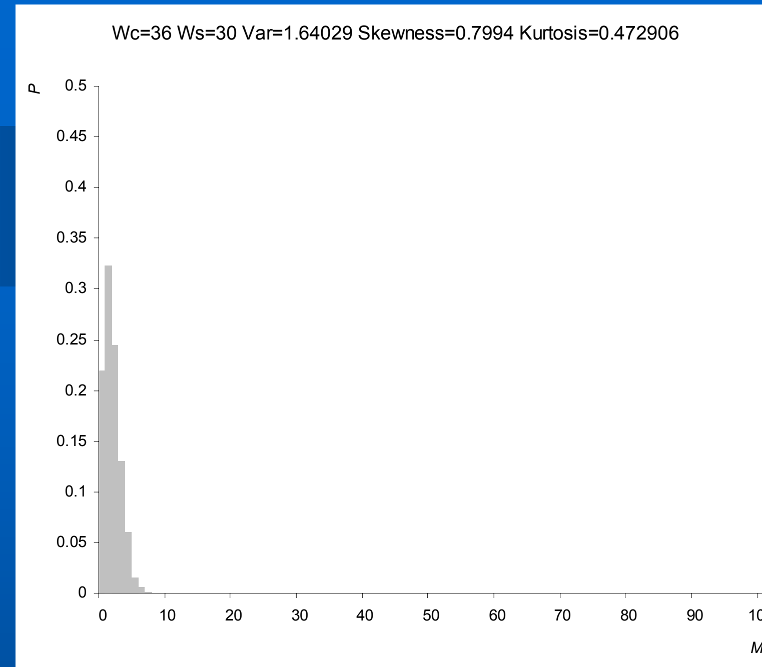
Optional smoothing of matrix $\mathbf{A}$:

$$\mathbf{B}[x][y][k] = \sum_{x_1=x-1}^{x+1} \sum_{y_1=y-1}^{y+1} \mathbf{A}[x_1][y_1][k]$$

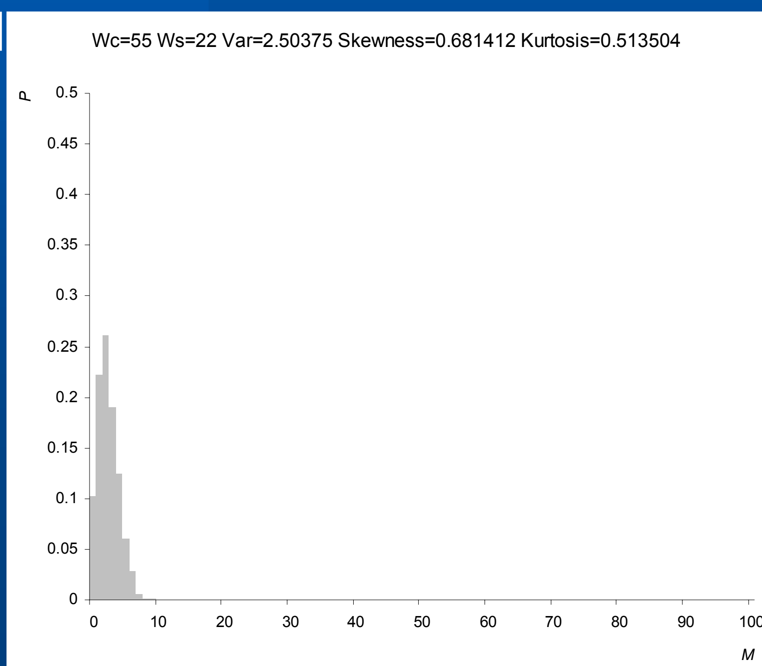
P[17][19]



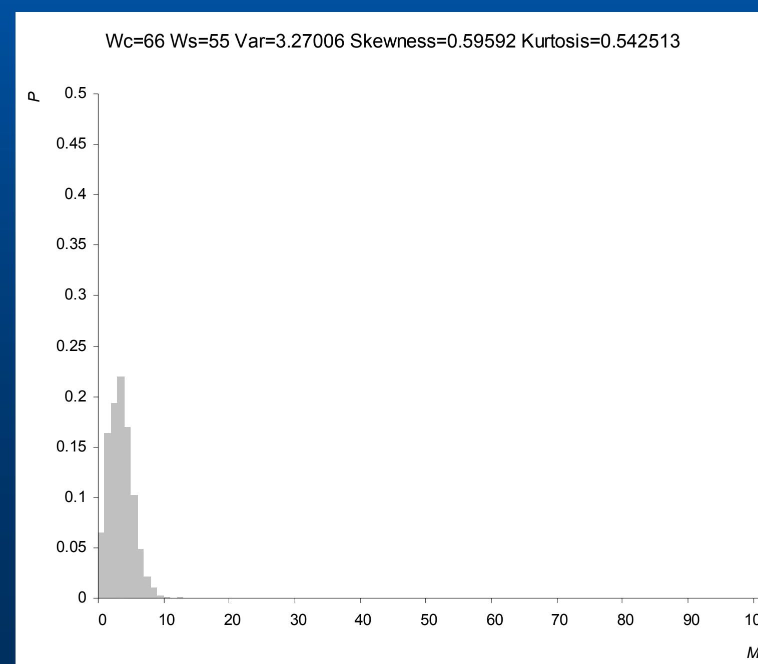
P[36][30]



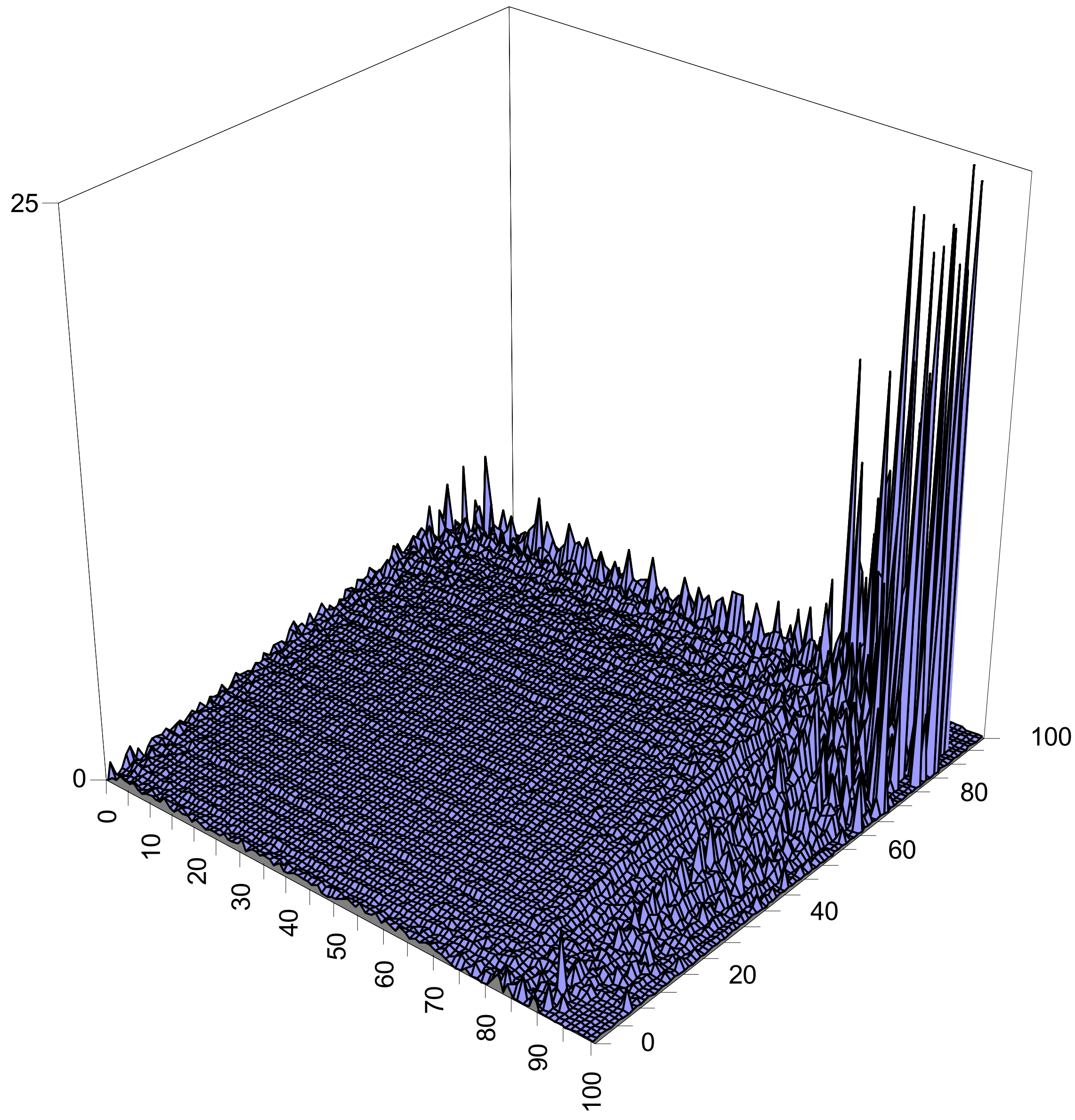
P[55][22]



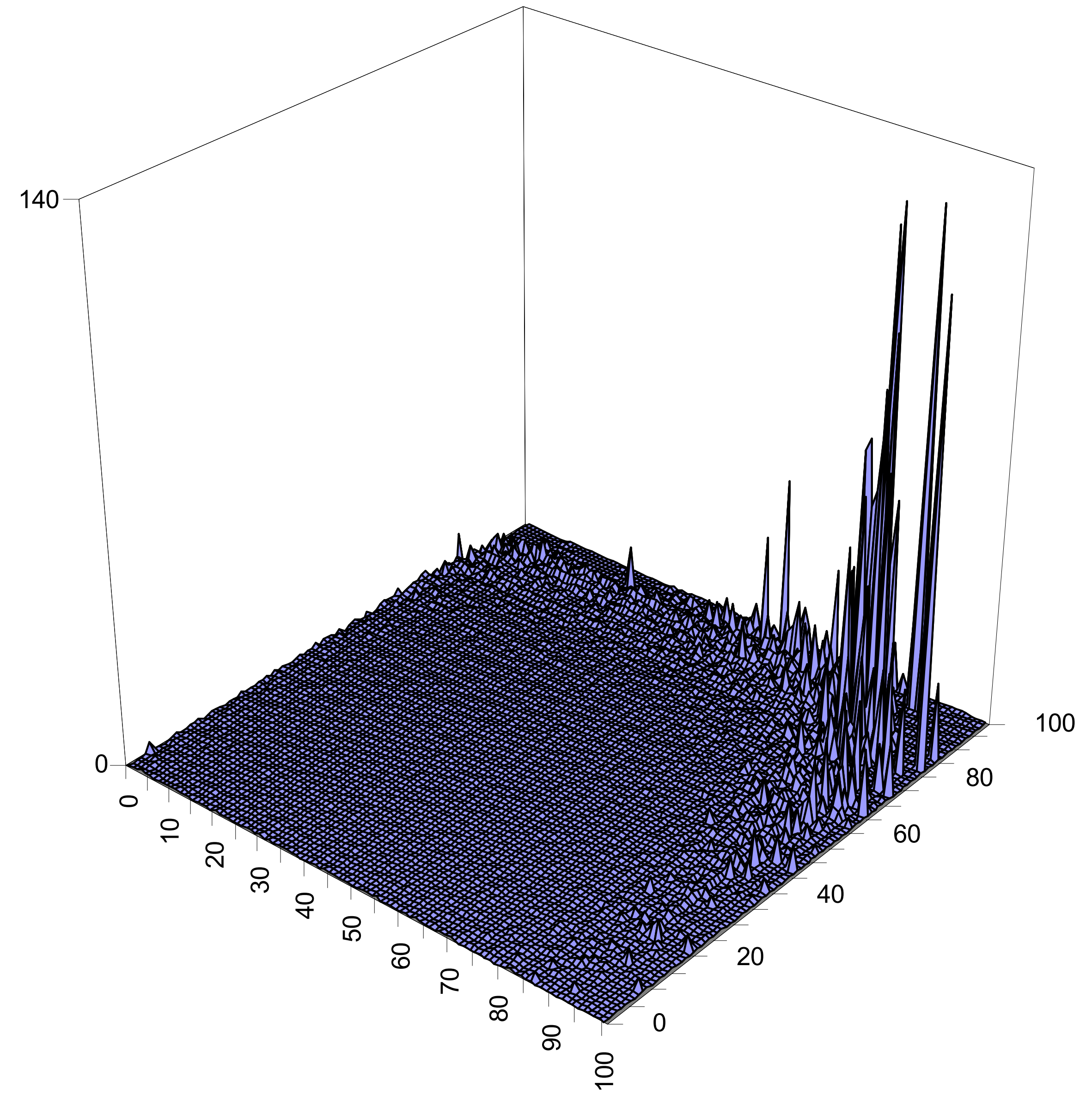
P[66][55]



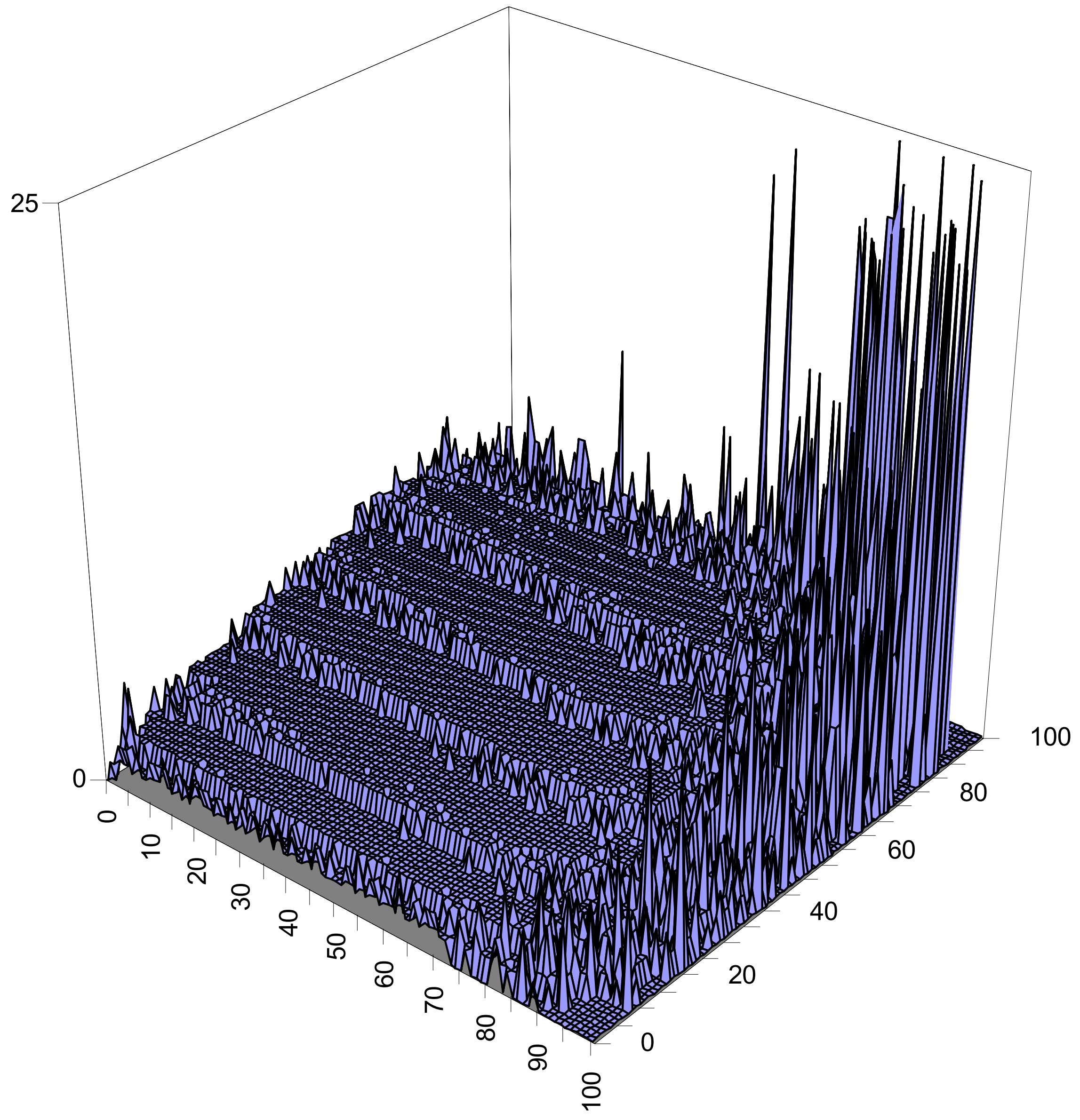
Matrix of means:



Matrix of variances:



Matrix of critical values for ξ (α=0.05):



Comparison Study: Setup

Real LSAT form (operational part with 100 items and 10 variable parts with about 25 items each uniformly distributed among examinees)

10000 examinees from N(0,1)

100 test centers, 100 examinees per test center

1, 5, 10, 15 aberrant pairs (subject *c*, source *s*) per test center

100, 500, 1000, 1500 aberrant pairs

10%, 20%, 30%, 40% of answer copying

In each aberrant pair subject and source have different variable parts

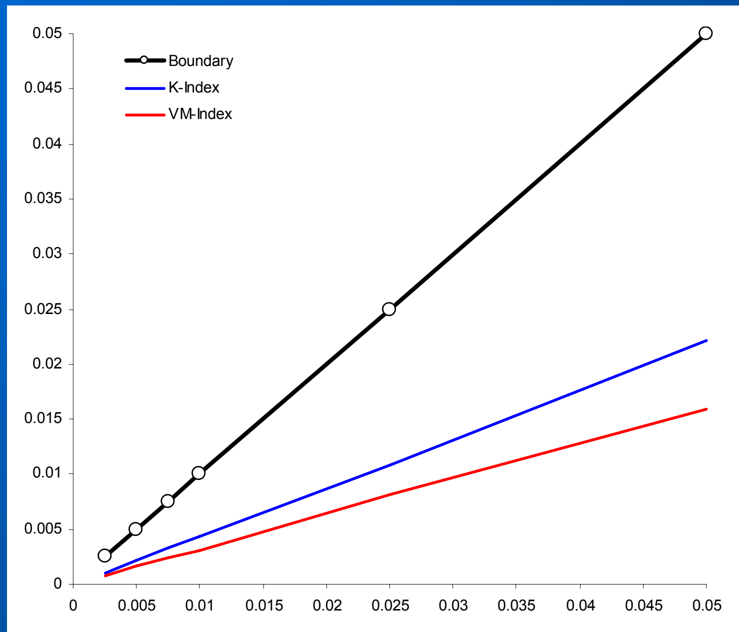
$\theta s \sim U(0,3)$, $\theta c = \theta s - 1$

Distribution of incorrect answers for each item was computed from real data

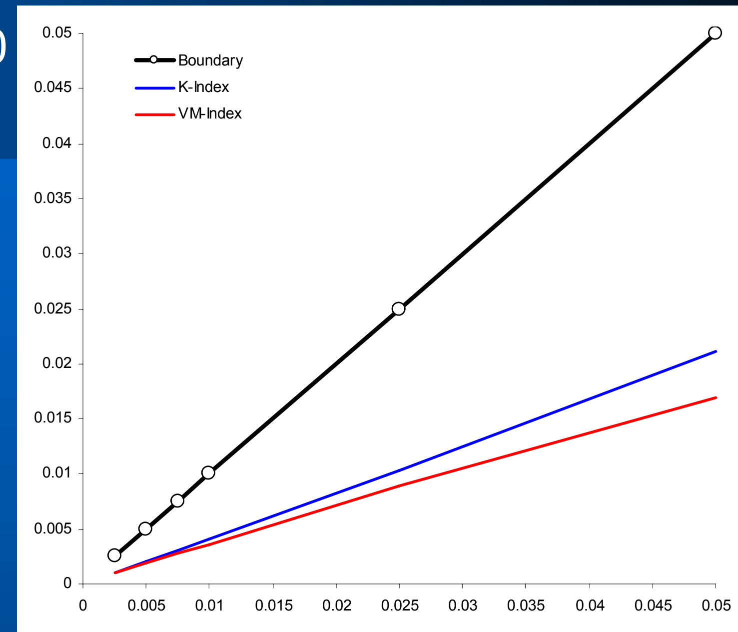
K-Index vs. VM-Index

Type I error

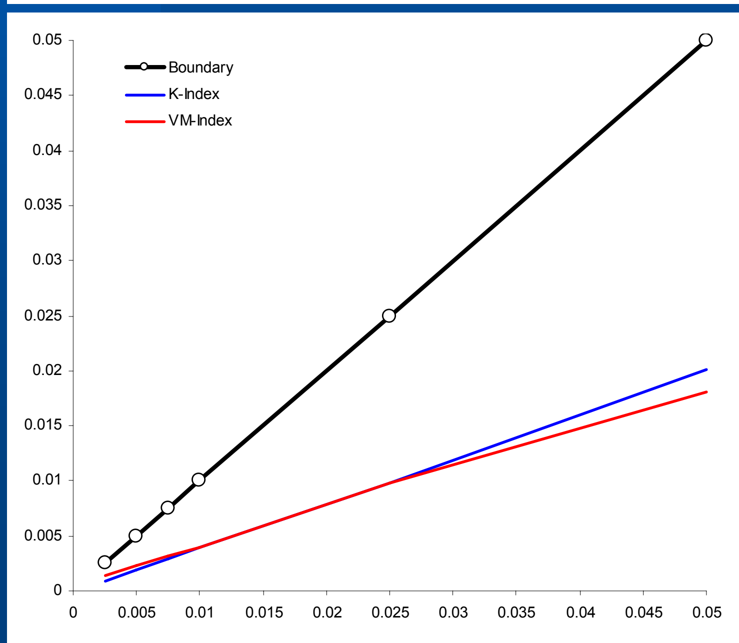
100



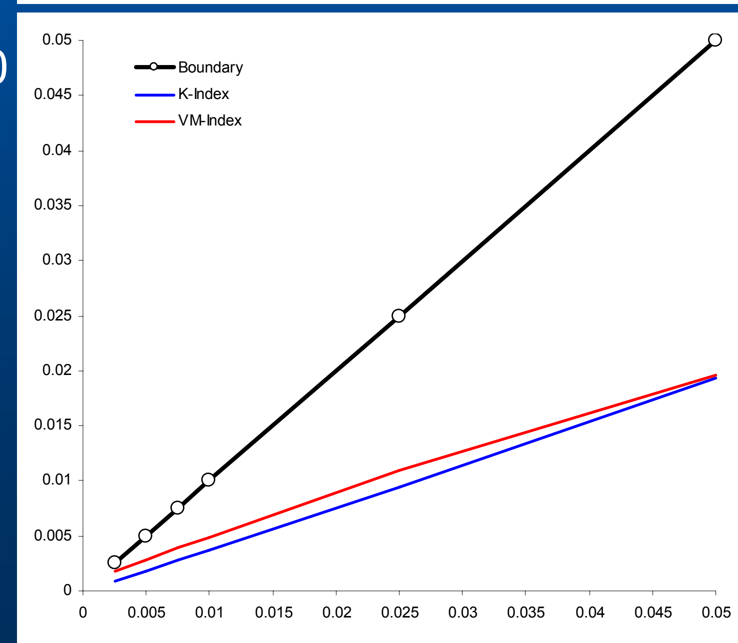
500



1000

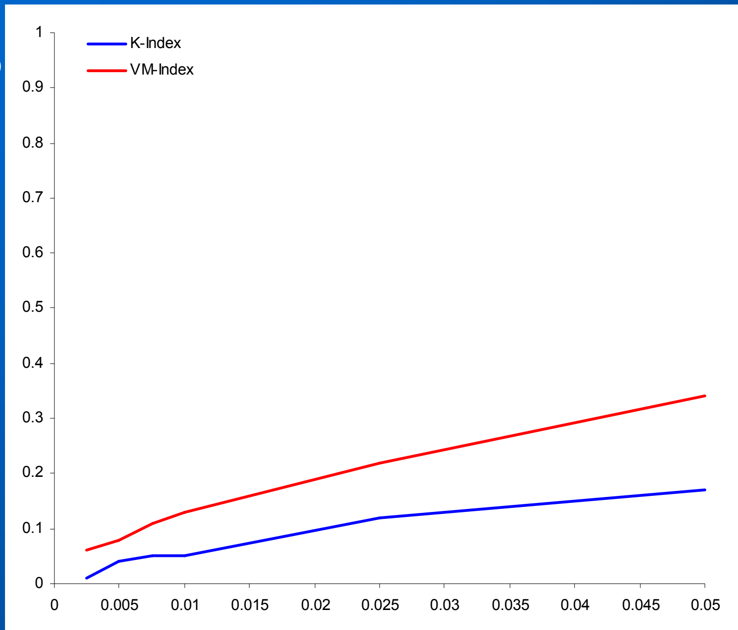


1500

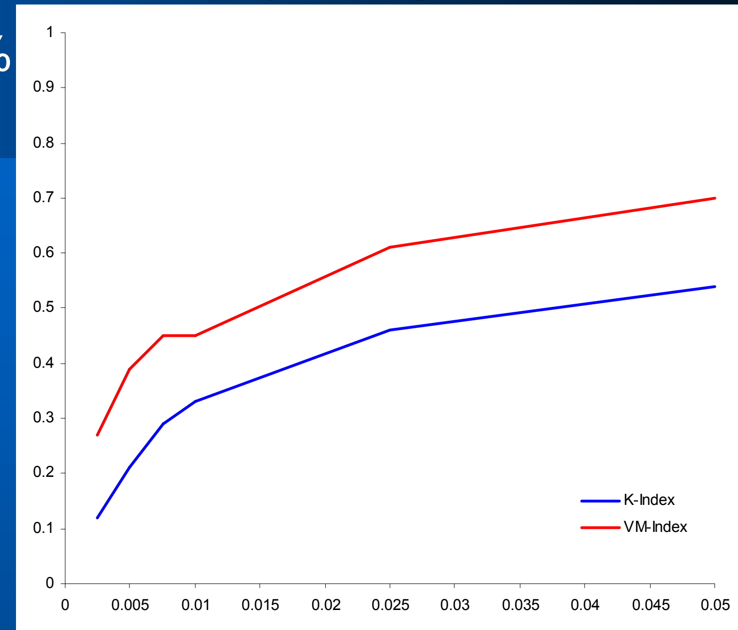


Detection rate (random copying)

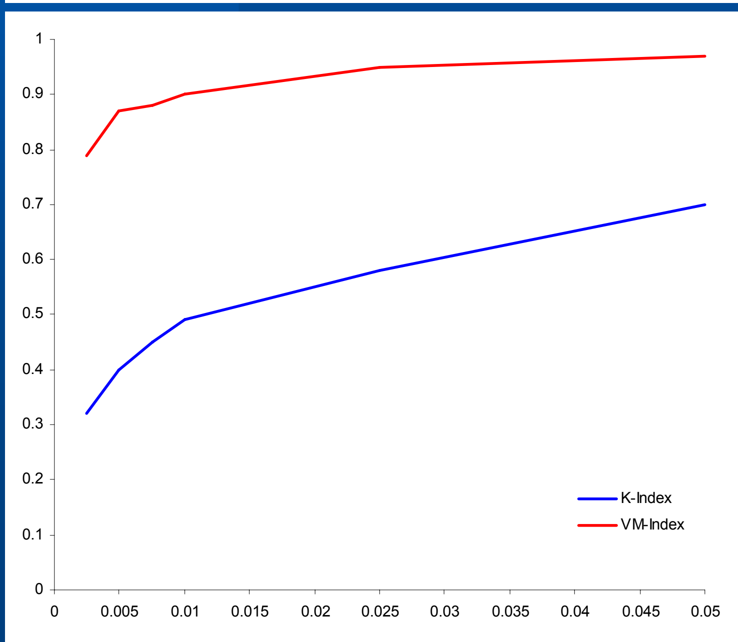
10%



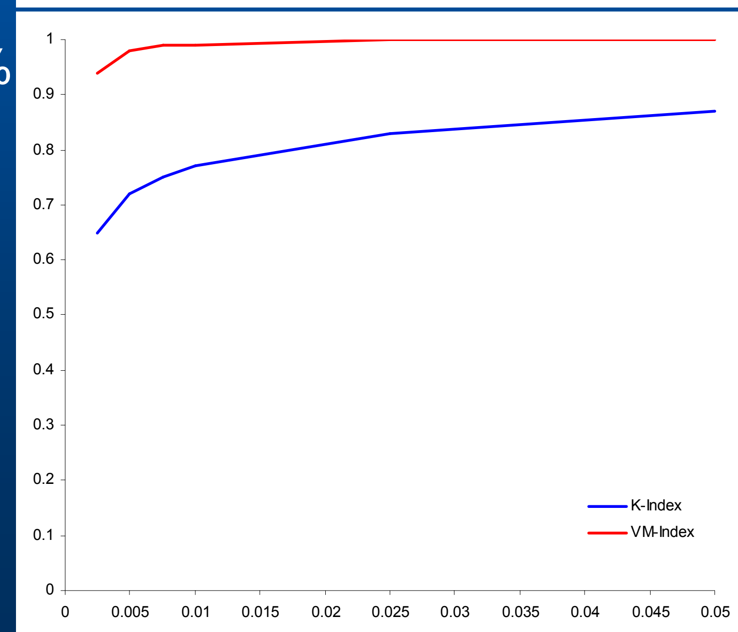
20%



30%

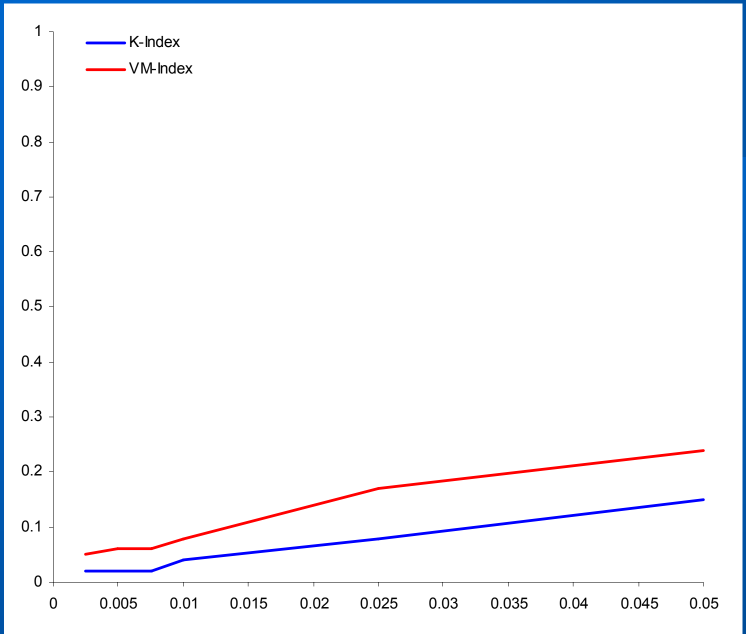


40%

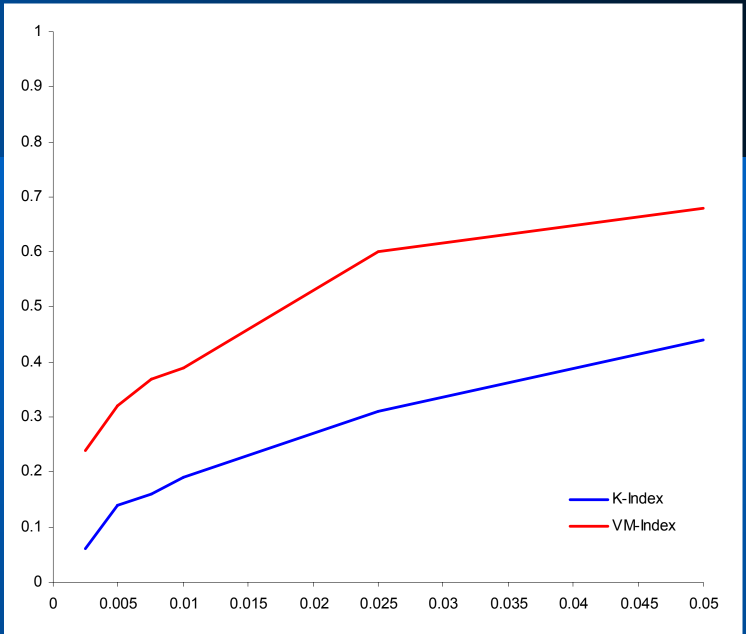


Detection rate (sequential copying)

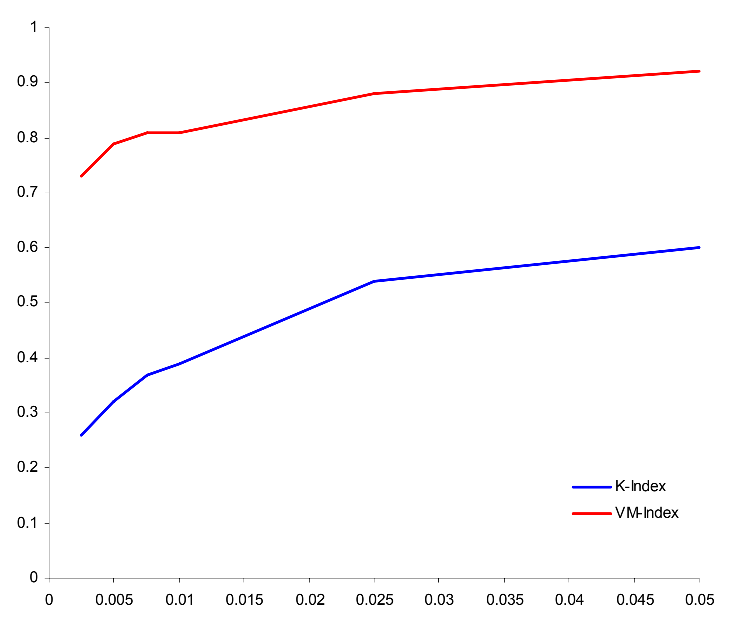
10%



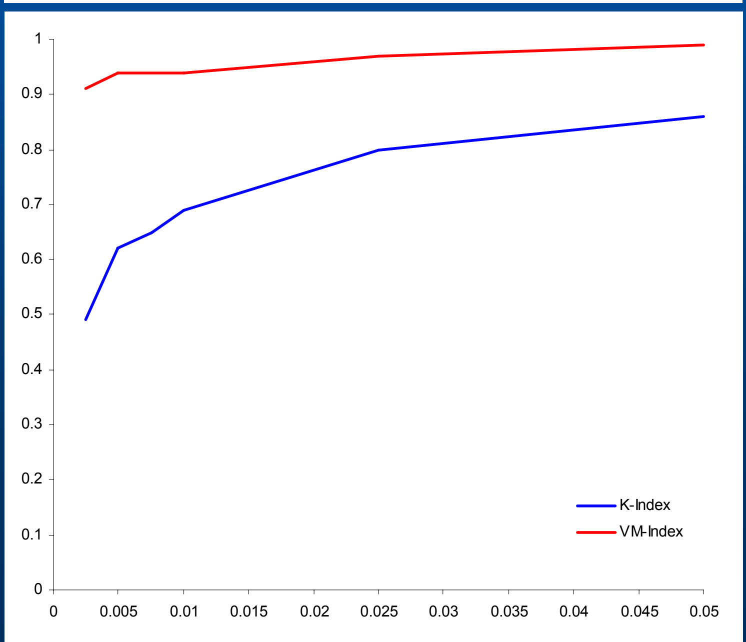
20%



30%

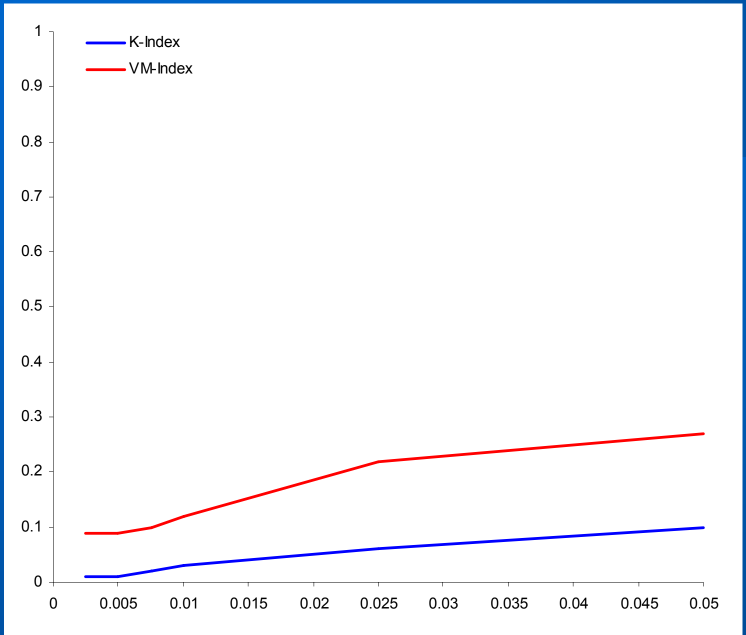


40%

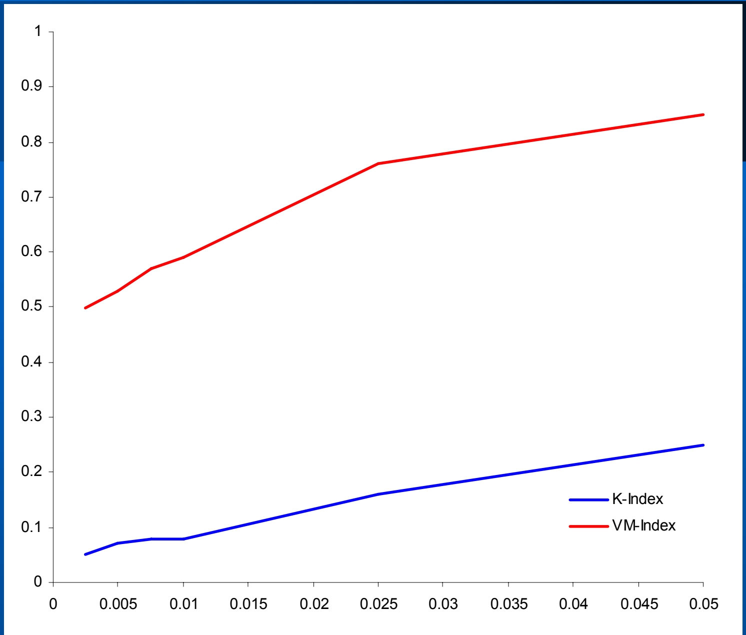


Detection rate (sequential copying, $\theta s \sim U(2,3)$)

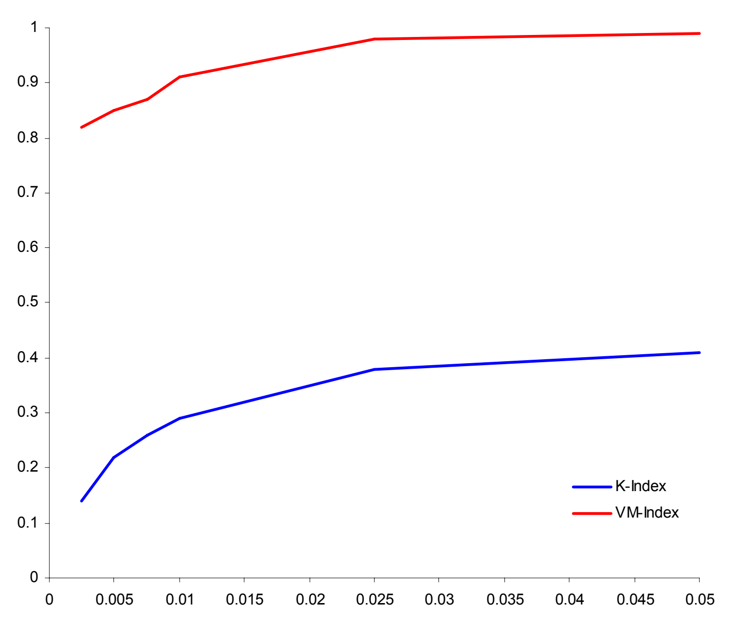
10%



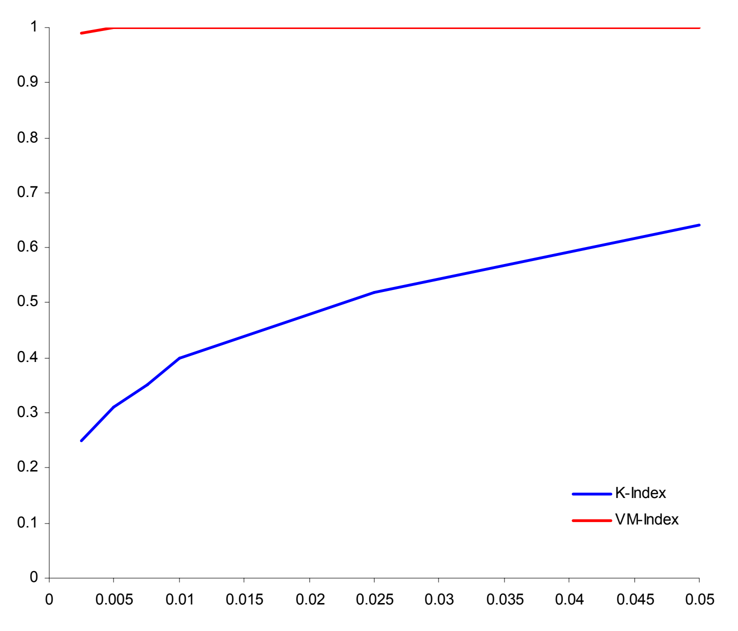
20%



30%



40%



Warning!

Number of reported pairs corresponding

to the previous slide (30% of copying, $\alpha = 0.05$).

Method	Number of reported pairs	Number of correctly reported pairs
K-Index	20,386	41
VM-Index	14,746	99

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement*, *34*, 379–392.

VM-Index

Given two examinees c and s taking a linear test partitioned into T and V, consider the following random variable:

$$\eta_i \equiv \eta_i(w_c, w_s) = \sum_{j=l}^{u} \begin{cases} 1 & \begin{array}{c} c \text{ selects incorrect answer to item } i \in V_c \text{ and} \\ s \text{ selects the same answer to item } i+j \in V_s \end{array} \\ 0 & \text{otherwise} \end{cases},$$

where w_c [w_s] is the number of incorrect responses of c [s] to T.

The conditional number of agreements

$$\xi \equiv \xi(w_c, w_s) = \sum_{i \in V_c} \eta_i$$

1. If $l = u = 0$, the VM-Index detects a blind-copy effect.
2. If $l \le u < 0$, the VM-Index detects a negative shift-copy effect.
3. If $0 < l \le u$, the VM-Index detects a positive shift-copy effect.
4. If $l < 0 < u$, the VM-Index detects all of the above effects.

Reduction to Poisson trials

Reorganize equations for VM-Index into the following:

$$\eta_{i,j} \equiv \eta_{i,j}(w_c, w_s) = \begin{cases} 1 & c \text{ selects incorrect answer to item } i \in V_c \text{ and} \\ & s \text{ selects the same answer to item } i + j \in V_s \\ 0 & \text{otherwise} \end{cases}$$

$$\xi_j \equiv \xi_j(w_c, w_s) = \sum_{i \in V_c} \eta_{i,j}$$

$$\xi \equiv \xi(w_c, w_s) = \sum_{j=l}^{u} \xi_j(w_c, w_s)$$

Poisson trials

Random variables η_i , $i = 1, 2, ..., n$, $\begin{matrix} 1 & 0 \\ p_i & q_i = 1 - p_i \end{matrix}$, mean p_i , variance $q_i p_i$

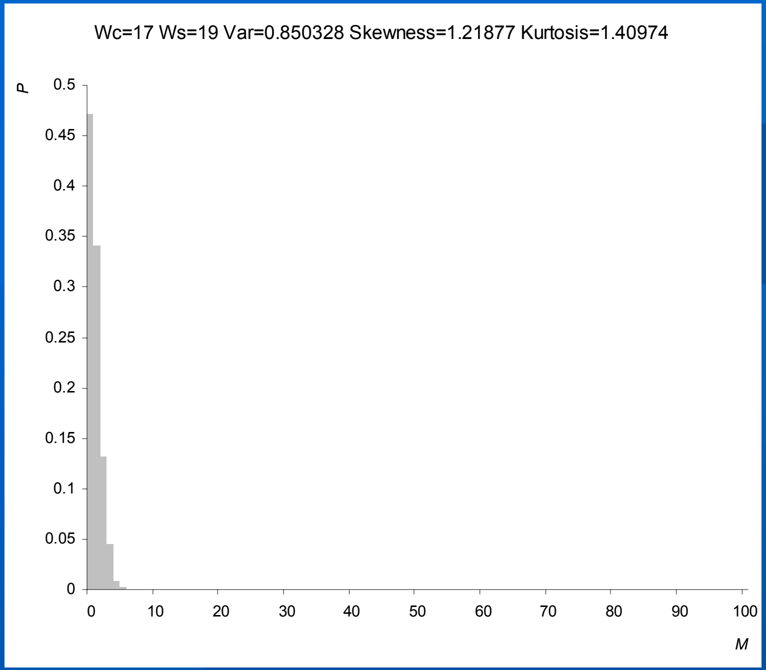
$\xi = \sum_{i=1}^{n} \eta_i$, mean $E(\xi) = \sum_{i=1}^{n} p_i$, variance $Var(\xi) = \sum_{i=1}^{n} q_i p_i$

Proposition: When random variables η_i are independent, $n \to \infty$, and

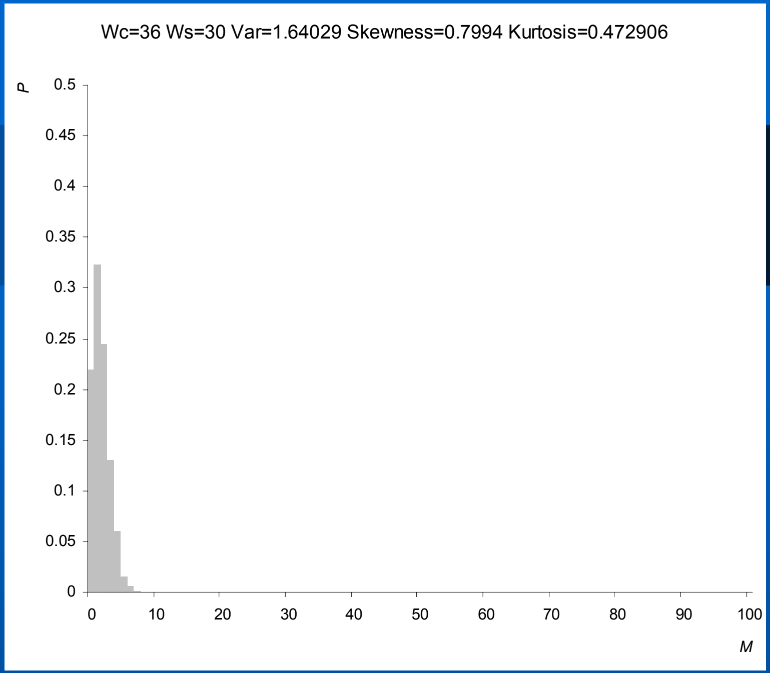
$\lim_{n\to\infty} \sum_{i=1}^{n} q_i p_i = \infty$ then distribution of $\lambda = \left(\xi - E(\xi)\right) / \sqrt{Var(\xi)}$ converges to N(0,1).

Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.

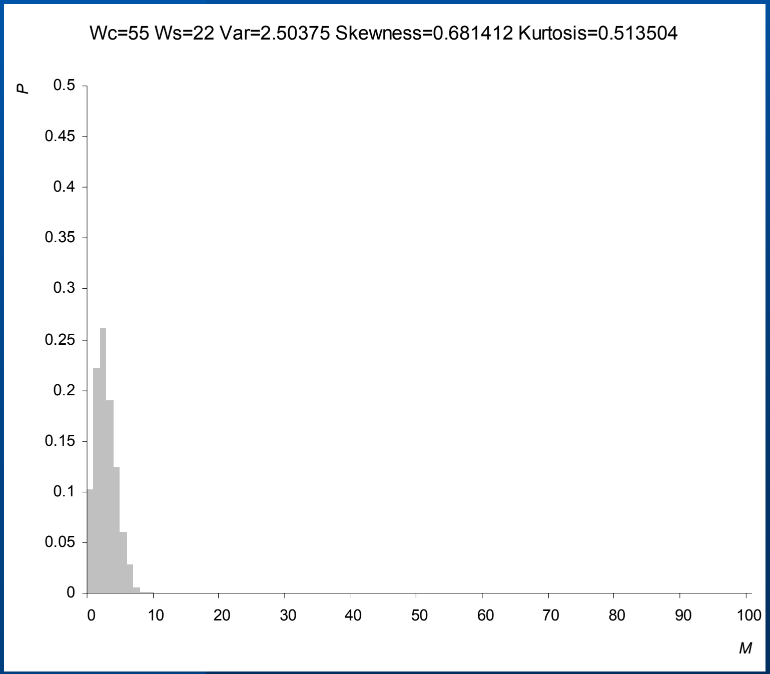
P[17][19]



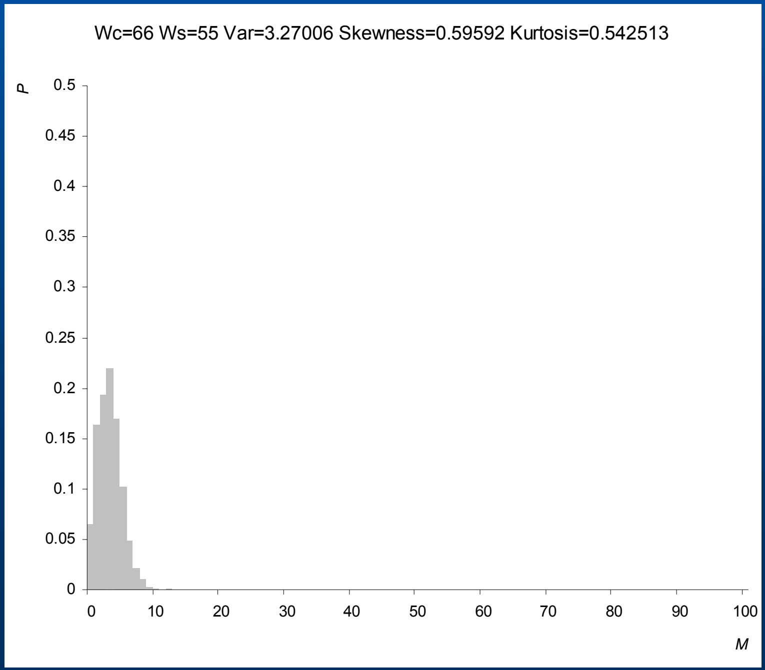
P[36][30]



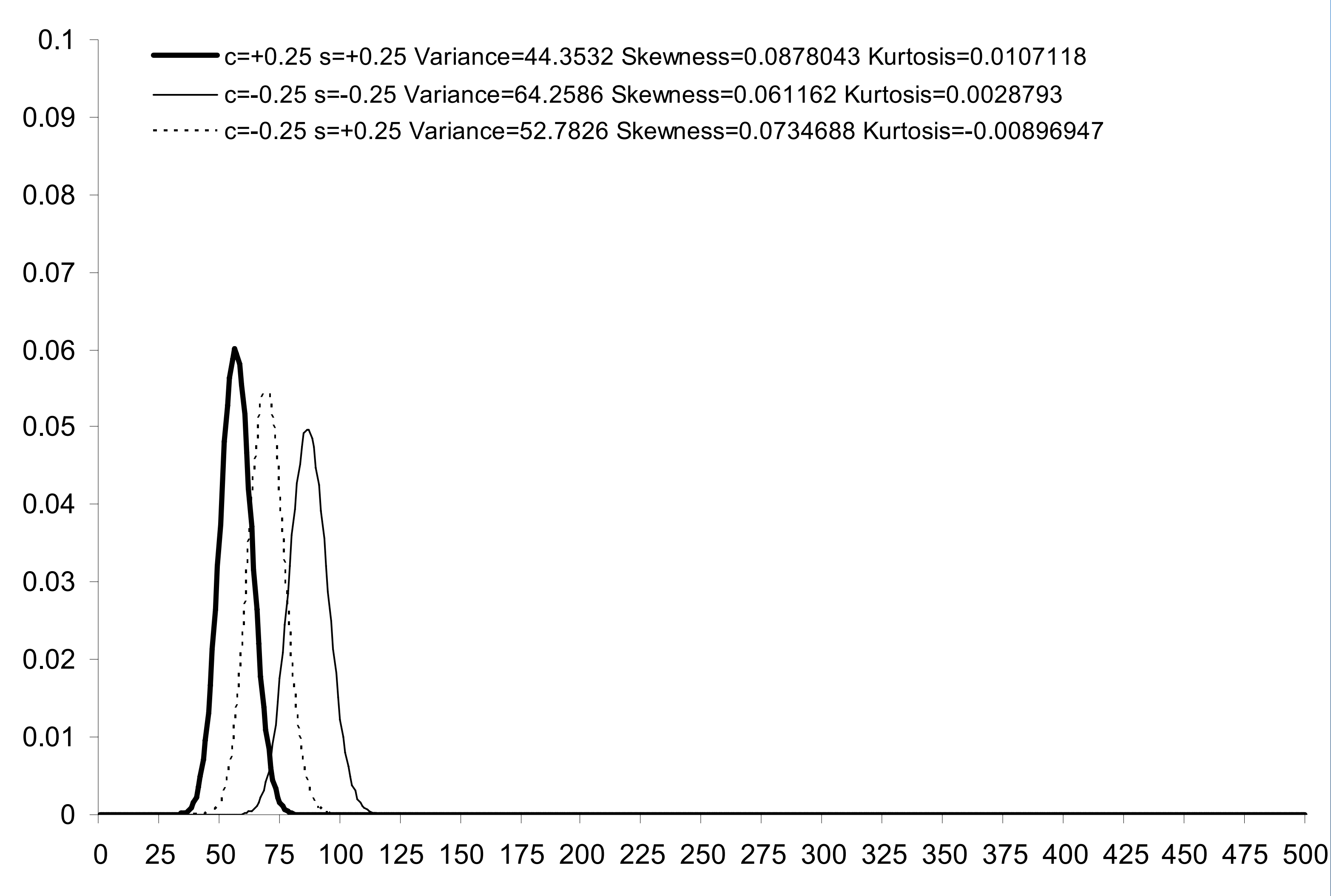
P[55][22]



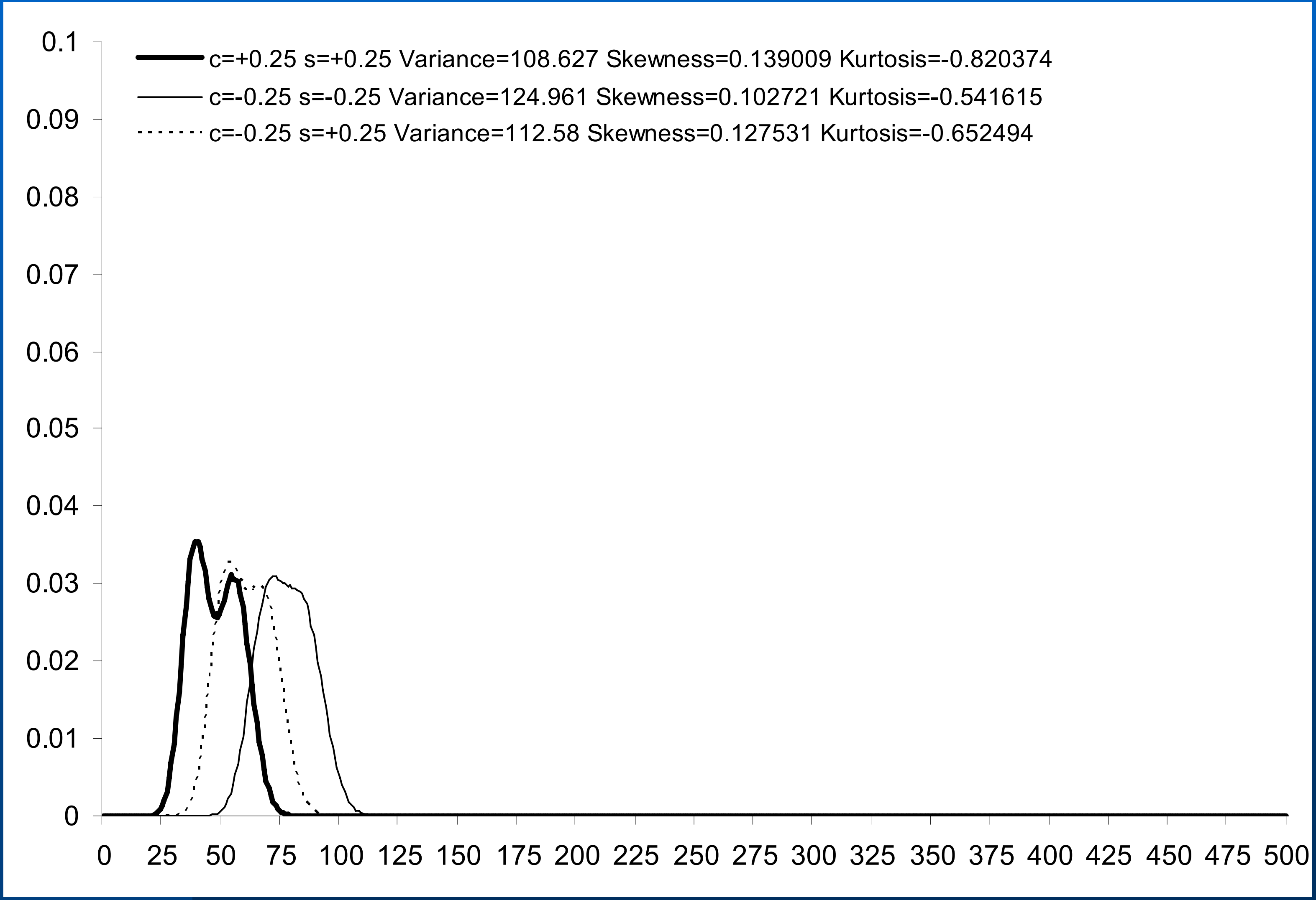
P[66][55]



One variable part



Two variable parts



Asymptotic distribution of VM-Index

In general, random variable $\xi_j(w_c, w_s)$ has a mixture density.

Assume that there are k different variable parts. We can condition on all pairs of variable parts which will result on k^2 matrices of normal distributions for $\xi(w_c, w_s)$.

Summary

- VM-Index assesses unusual agreement (blind-copy effect, shift-copy effect) between the answers of two examinees in the variable part.
- VM-Index provides Type I error rate lower than K-Index, when the number of examinees from aberrant pairs is below 9%.
- VM-Index provides higher detection rate than K-Index, especially when **a source has high ability**.
- VM-Index can be approximated by matrices of normal distributions.
- Pairwise statistics (K, M, VM, K*, K1, K2, ω, etc.) should not be used alone in practice.

Call for papers on

Applications of Operations Research in Educational Measurement

special volume of the journal

Annals of Operations Research

published by Springer

Submission deadline: February 1, 2011

dbelov@LSAC.org