

quantil

Matemáticas Aplicadas

INFORME FINAL

Algoritmo para la detección de copia en exámenes estandarizados de múltiples opciones: Modelos de respuesta nominal y pruebas múltiples de copia.

PRESENTADO A:

Instituto Colombiano para la  
Evaluación de la Educación  
ICFES

Enero 11, 2011

Todo material presentado en este reporte, a menos que se indique lo contrario, es de derecho reservado de Quantil S.A.S («Quantil»). Ninguna parte del material presentado puede ser alterado, transmitido, copiado o distribuido sin el previo consentimiento escrito de Quantil.

Quantil S.A.S ®, 2010. Derechos reservados.

## Índice

<b>1. Introducción</b>	<b>3</b>
<b>2. Estado del arte</b>	<b>5</b>
2.1. Índice K . . . . .	6
2.2. Índice $S_1$ . . . . .	6
2.3. Índice B . . . . .	7
2.4. Índice H . . . . .	7
2.5. Comparación error tipo I y potencia . . . . .	8
2.6. Avances recientes . . . . .	9
2.7. Pruebas de múltiples hipótesis . . . . .	10
<b>3. Índice <math>\kappa</math></b>	<b>12</b>
3.1. Índice $\kappa$ : recodificación . . . . .	14
3.2. Índice $\kappa$ : recodificación condicional . . . . .	15
3.3. Resultados . . . . .	15
<b>4. Índice <math>\omega</math></b>	<b>17</b>
<b>5. Metodología ICFES: Índice C</b>	<b>22</b>
<b>6. Simulaciones</b>	<b>24</b>
6.1. Metodología para los índices $\omega$ y $\kappa$ . . . . .	24
6.2. Metodología para el índice C . . . . .	25
6.3. Resultados sin condicionar . . . . .	26
6.3.1. Error Tipo I . . . . .	26
6.3.2. Potencia . . . . .	27
<b>7. Análisis de las Pruebas Saber 5° y 9°</b>	<b>34</b>
7.1. Análisis Descriptivo . . . . .	34

7.2. Análisis Pruebas Controladas y Estadísticas con el Índice $\omega$ .	37
7.3. Análisis Pruebas Censales con el Índice $\omega$ . . . . .	38
7.4. Copia Masiva con el Índice $\omega$ . . . . .	47
7.5. Análisis Pruebas Censales con el Índice C . . . . .	50
7.6. Copia Masiva con el Índice C . . . . .	66
<b>8. Conclusiones</b>	<b>76</b>
<b>9. Bibliografía</b>	<b>78</b>
<b>A. Resultados condicionando por habilidad</b>	<b>80</b>
A.1. Error Tipo I . . . . .	80
A.2. Potencia . . . . .	87

## 1. Introducción

En un proyecto anterior contratado por el ICFES el equipo asesor de Quantil construyó un índice para detección de copia en exámenes de múltiples opciones y presentó los resultados del ejercicio para todas las pruebas del núcleo común del Examen de Estado para el ingreso a la educación superior, actualmente Saber 11, 2009 así como para las pruebas de docentes. Este estudio se basó en la implementación del índice  $\kappa$  con una técnica de recodificación que arroja resultados prometedores como herramienta cuantitativa expedita para generar alertas de fraude; sin embargo una de las principales conclusiones del estudio fue que, para tener mayor confianza en la utilización del índice construido, era necesario contrastarlo contra otras metodologías existentes en la literatura especializada.

En particular, el estudio concluyó que era necesario intentar otras metodologías que mitigaran un problema de especificación del modelo probabilístico de respuesta de cada ítem que se propuso. Concretamente, la distribución de probabilidad con la que se responde a las diferentes opciones varía con las preguntas. Que la distribución no variara era una suposición necesaria para la determinación de la distribución a la que converge el índice  $\kappa$ . De hecho, se estimó que la media del índice para formas independientes no era cero, como se esperaría si no hubiera variación.

Este documento presenta tres metodologías para detectar copia. Un índice utilizando por el ICFES en el pasado, el índice  $\kappa$  y el índice  $\omega$ . Estas tres metodologías se comparan en términos del error tipo I utilizando muestreos independientes de datos reales. También se compara la potencia incorporando copia de manera controlada a estos muestreos. De acuerdo a estas simulaciones, el índice  $\omega$  es preferible al índice  $\kappa$  en términos del error tipo I pues en raras ocasiones acusa a inocentes de copia; sin embargo a primera vista no es claro qué índice tiene una mayor potencia. Lo anterior justifica el uso del índice  $\omega$ .

Se aplica la metodología del índice  $\omega$  para detectar las parejas sospechosas de copia en las pruebas SABER 5° y 9° de mayo y octubre de 2009. Con este resultado se determina en qué sedes-jornada hay indicios de copia masiva, controlando la tasa de falsos positivos al realizar pruebas de hipótesis conjuntas.

El documento se organiza de la siguiente manera. Primero se realiza una breve revisión de la literatura, seguido de una explicación del índice  $\kappa$  y el índice  $\omega$  y de la metodología usada en la actualidad por el ICFES para

detectar copia. La siguiente sección compara el error tipo I y la potencia de estas tres metodologías. La última sección contiene los resultados de analizar las pruebas SABER con el índice  $\omega$  y de detectar copia masiva controlando la tasa de falsos positivos.

## 2. Estado del arte

Los exámenes de opción múltiple se usan frecuentemente pues presentan una herramienta eficiente y objetiva de evaluar los conocimientos de un estudiante al simplificar la logística administrativa y eliminar los elementos subjetivos en la clasificación. Sin embargo este tipo de exámenes es más vulnerable a la copia que aquellos de pregunta abierta. Los índices de copia presentan una herramienta estadística expedita para detectar la copia en este tipo de exámenes buscando patrones de respuesta sospechosamente similares.

Los índices de copia adolecen de tres problemas. El primero es que pueden existir explicaciones alternativas para que dos individuos presenten un patrón de respuestas sospechosamente similar. El más evidente es que ningún estadístico puede eliminar por completo la posibilidad de que dos individuos tengan respuestas idénticas por azar. Adicionalmente dos estudiantes que tuvieron la misma educación tienen una probabilidad alta de tener respuestas similares.

El segundo problema consiste en que una prueba estadística no puede ser usada como un argumento contundente para acusar a alguien de copia. Dado que es imposible eliminar el error tipo I por completo, nunca se puede estar 100 % de que alguien incurrió en copia así el estadístico lo indique.

Finalmente, todo índice de copia estadístico supone que existe una distribución de probabilidad con la cual los estudiantes responden a las preguntas. Aproximaciones incorrectas de esta distribución pueden llevar a conclusiones igualmente erróneas.

La mayoría de los modelos de detección de copia existentes (Van der Linden & Sotaridona 2006) se basan en calcular diferentes estadísticos que cuentan el número de respuestas comunes entre el individuo sospechoso de hacer copia y aquel de quien copia (la fuente). Estos estadísticos difieren principalmente en: (1) El conjunto de respuestas sobre las que se calcula (todas las respuestas, correctas e incorrectas, solo las respuestas incorrectas, solo las respuestas incorrectas condicional a la habilidad de los dos individuos, etc.); (2) La distribución que se supone tiene este estadístico y si se usa o no alguna estandarización del estadístico relevante. En esta sección hablaremos de los índices  $K$ ,  $S_1$ ,  $S_2$ ,  $B$  y  $H$ . El índice  $\kappa$  y el índice  $\omega$  se explican detalladamente más adelante.

Los avances más recientes se enfocan en utilizar múltiples índices al mis-

mo tiempo (Wollack 2006) y utilizar la asignación de puestos y de temas - exámenes visualmente similares pero con preguntas o respuestas diferentes - para enriquecer los modelos (Van der Ark, Emons & Sijtsma 2008).

A continuación se presenta una breve descripción de cinco índices que se encuentran en la literatura. La explicación del índice  $\kappa$  y el índice  $\omega$  se encuentra en los capítulos 3 y 4.

## 2.1. Índice K

El índice  $K$  (Holland 1996) y su modificación el  $K_2$  (Sotaridona & Meijer 2001), suponen que la probabilidad de que dos individuos, el que copia (C) y la fuente de copia (S), tengan al menos  $K$  respuestas incorrectas iguales o más, dado que S respondió  $N$  preguntas incorrectamente, es:

$$K = \sum_{i=K}^N \binom{N}{i} p^i (1-p)^{N-i}, \quad (1)$$

en otras palabras se supone que el número de respuestas incorrectas iguales sigue una distribución binomial. Si  $K$  es menor que  $1 - \alpha$ , donde  $\alpha$  es el nivel de significancia, se sospecha que C copió de S. La diferencia entre  $K$  y  $K_2$  radica en la forma de calcular  $p$ . En el primero, éste se calcula como la proporción de respuestas iguales incorrectas promedio que tiene S con todas las personas que tienen el mismo número de respuestas correctas. En el segundo es el resultado de una regresión lineal. Para más detalles ver (Holland 1996) y (Sotaridona & Meijer 2001).

## 2.2. Índice $S_1$

El índice  $S_1$  (Sotaridona & Meijer 2003), es conceptualmente similar al índice K. La principal diferencia radica en que este índice utiliza una distribución de Poisson para determinar la probabilidad de tener  $n$  coincidencias en las respuestas incorrectas, mientras que el índice K utiliza una distribución binomial.

El que copia (C) y la fuente de la copia (S) tienen una probabilidad de tener al menos  $K$  respuestas incorrectas iguales, dado que S respondió  $N$

preguntas incorrectamente, de:

$$S_1 = \sum_{i=K}^N \frac{e^{-\hat{\mu}} - \hat{\mu}^i}{i!}, \quad (2)$$

donde  $\mu$  se modela por medio de una regresión log-lineal. El índice  $S_2$  es una extensión de  $S_1$  que incluye información de coincidencias en el número de respuestas correctas. Si  $S_1$  es menor que  $1 - \alpha$  donde  $\alpha$  es el nivel de significancia se sospecha que C copió de S. Análogo para  $S_2$ . Para más detalles ver (Sotaridona & Meijer 2003).

### 2.3. Índice B

Este índice desarrollado por Angoff (1974) condiciona el número de respuestas incorrectas idénticas entre C y S al producto del número de respuestas incorrectas de C y de S suponiendo una distribución bi-normal. Si denotamos la primera cantidad como  $K_{cs}$  y la segunda como  $W_{cs}$ , entonces se define  $B$  como:

$$B = \frac{K_{cs} - \bar{K}_{W_{cs}}}{S_{K_{W_{cs}}}}, \quad (3)$$

donde  $\bar{K}_{W_{cs}}$  es el promedio del número de respuestas incorrectas idénticas para todas las parejas con valores de  $W_{ij}$  similares a  $W_{cs}$ .  $S_{K_{W_{cs}}}$  es la desviación estándar del número de respuestas incorrectas idénticas entre todas las parejas con valores de  $W_{ij}$  parecidos a  $W_{cs}$ .  $B$  se compara con un valor crítico  $z^*$  extraído de una distribución normal estándar. En otras palabras, si se tiene un nivel de significancia de  $\alpha$ , entonces el valor crítico  $z^*$  es tal que  $P(Z \leq z^*) = \alpha$  (si  $B$  es mayor a  $z^*$  se sospecha copia). Para más detalles ver (Angoff 1974).

### 2.4. Índice H

Este índice desarrollado por Angoff (1974) condiciona el número de respuestas incorrectas idénticas entre C y S al número de respuestas incorrectas del que tenga el mayor puntaje suponiendo una distribución bi-normal. Se

define  $H$  como:

$$H = \frac{K_{cs} - K_+}{S_+}, \quad (4)$$

donde  $K_+$  y  $S_+$  son el promedio y la desviación estándar del número de respuestas incorrectas idénticas de individuos con un número de preguntas correctas similar al de aquél con el mayor puntaje entre C y S.  $H$  se compara con un valor crítico  $z$  extraído de una distribución normal estándar. En otras palabras si se tiene un nivel de significancia de  $\alpha$ , entonces el valor crítico  $z^*$  es tal que  $P(Z \leq z^*) = \alpha$  (si  $H$  es mayor a  $z^*$  se sospecha copia). Para más detalles ver (Angoff 1974).

## 2.5. Comparación error tipo I y potencia

En la literatura se encuentran dos métodos para estimar el error tipo I. El primero consiste en simular vectores de respuesta, de acuerdo a algún modelo como el modelo de respuesta nominal. Se empareja un número  $n$  de respuestas simuladas. Se aplica la metodología del índice que se desea estudiar y se calcula la proporción de parejas sospechosas de copia. Esta proporción es el error tipo I, pues los vectores de respuesta fueron simulados sin copia alguna (Sotaridona, van der Linden & Meijer 2006). El segundo método consiste en tomar un examen real, y tomar  $n$  parejas de individuos de tal manera que los miembros de una misma pareja no hayan podido realizar copia. Estas parejas se pueden seleccionar tomando individuos que hayan realiza la prueba en años diferentes o en salones diferentes. Al igual que con el primer método, se aplica la metodología del índice que se desea estudiar y se calcula la proporción de parejas sospechosas de copia. Esta proporción es el error tipo I (Wollack 2006).

Para calcular la potencia de un índice existen tres metodologías. Las primeras dos se derivan de los métodos mencionadas en el párrafo anterior para estimar el error tipo I. La idea consiste en modificar los vectores de respuesta copiando  $k$  preguntas de un miembro de una pareja a el otro. Se hace esto para las  $n$  parejas y se calcula la proporción de parejas acusados de copia. Esta proporción es la potencia para un nivel de copia  $k$ . El tercer método consiste en determinar la distribución del índice bajo la hipótesis alternativa y encontrar la probabilidad de detección de copia para un nivel de copia dado.

Wollack (2006) compara los índices mencionados en la seccion anterior

y el índice  $\omega$  en términos del error tipo I, utilizando exámenes reales contestados por 10,000 individuos en dos años diferentes. Se emparejan individuos de un año con otro lo que asegura que no puedo haber copia entre ellos. Se encuentra que todos los índices preservan el error tipo I.

Para medir el poder de los índices se copian respuestas de un individuo a otro y se mide la tasa de detección. En general los resultados de Wollack (2006) encuentran que entre los índices individuales el índice  $\omega$  es el que más potencia tiene. Adicionalmente en este estudio se estudian los beneficios de utilizar varios índices simultáneamente y se encuentra que el índice  $\omega$  en conjunto con un nuevo estadístico denominado  $H^{*1}$  es el que tiene la mayor potencia. Sin embargo en este artículo no se compara con el índice  $\kappa$ .

Sotaridona et al. (2006) muestran, utilizando simulaciones, que la versión del índice  $\kappa$  que se utilizó en el anterior estudio que Quantil realizó para el ICFES (Jara, Riascos & Romero 2010), no preserva el error tipo I, sobre todo cuando la habilidad del individuo que se copia y de la fuente son similares.

El presente estudio es un aporte a la literatura en esta dirección pues compara el índice  $\omega$ , el que más potencia tiene, entre aquellos índices que respetan el error tipo I (Wollack 2006), con el índice  $\kappa$  utilizando la metodología propuesta por Wollack (2006).

## 2.6. Avances recientes

Belov & Armstrong (2010) describen una metodología que combina dos aproximaciones estadísticas diferentes para detectar copia. Primero se utiliza la divergencia de Kullback-Leibler para detectar individuos con patrones de respuesta inconsistentes y después se utiliza el índice K para detectar la posible fuente de la copia. Los principales beneficios de esta técnica son: reducción en el tiempo computacional y un error tipo I menor. Sin embargo, esta metodología sólo se puede aplicar cuando un examen tiene particiones claras, donde una parte del examen es idéntica para ambos individuos de una pareja y una parte es diferente. En el caso de las pruebas de estado realizadas por el ICFES estas particiones pueden corresponder a las diferentes áreas que se examinan, siempre y cuando una pareja tenga algunos formatos en común y otros diferentes.

La divergencia de Kullback-Leibler se utiliza para detectar cambios en los

---

<sup>1</sup>Este índice es similar a  $H$  pero con valores críticos sacados de una distribución empírica, mientras que  $H$  los obtiene de una distribución teórica.

patrones de respuesta entre diferentes particiones del examen. Por ejemplo, si una pareja realiza matemáticas al mismo tiempo y se copia, pero después una persona debe responder lenguaje y la otra ciencias, es de esperar que la habilidad de los individuos caiga en la segunda prueba comparada con la primera. Aquellos individuos que demuestren cambios sustanciales en sus patrones de respuestas, son examinados con el índice K. En este paso se compara cualquier individuo detectado en la primera etapa contra todos los demás, pues es posible que la copia solo haya ocurrido en una dirección. En (Belov & Armstrong 2010) se muestra que el error tipo I es prácticamente inexistente, pero se pierde cierto poder en comparación con el índice K.

Van der Ark et al. (2008) describen una metodología que explota el lugar designado de los estudiantes así como el uso de pruebas de apariencia similar para diferentes estudiantes. Primero, y esto en general es posible aplicarlo con cualquier metodología siempre y cuando se tenga un mapa con los lugares designados para los estudiantes, solo se consideran parejas de personas sentadas una al lado de la otra. Para aplicar esta metodología se necesita que la mitad de las parejas tengan la misma versión del examen mientras que la otra mitad tenga versiones diferentes. La ventaja de los estadísticos desarrollados es que no requieren una muestra grande para determinar la distribución de las respuestas; sin embargo no tienen un poder muy alto y no se ha estudiado a profundidad su error tipo I.

## 2.7. Pruebas de múltiples hipótesis

Cuando se quiere determinar si un salón incurrió en copia masiva o no, es necesario realizar una prueba de múltiples hipótesis. Es deseable mantener un error tipo I bajo pues no se quiere acusar a inocentes de copia, sin embargo tampoco se quiere ser muy estricto pues la potencia de la prueba se vería comprometida.

Controlar pruebas de múltiples hipótesis por el error tipo I presenta dificultades teóricas que impiden llevar el procedimiento a la práctica en la mayoría de los casos. En los pocos casos donde se puede aplicar, la potencia de la prueba se ve seriamente disminuida. A la luz de estos resultados, la literatura ha intentado controlar medidas similares al error tipo I, pero cuya aplicación sea generalizada, buscando al mismo tiempo no comprometer la potencia de la prueba.

La mayoría de estas técnicas son procedimientos tipo Bonferroni, que permiten controlar la tasa de falsos positivos (hipótesis nulas rechazadas

erróneamente como proporción de las hipótesis nulas rechazadas) a un nivel de confianza deseado. La principal contribución de Benjamini & Hochberg (1995) es un procedimiento que tiene una potencia mayor. Considere que se van a probar  $m$  hipótesis nulas (el número de estudiantes en una sede-jornada). De estas  $m_0$  son verdaderas y  $R$  la cantidad rechazada, entonces la siguiente tabla representa los errores cometidos:

Tabla 1: Errores cometidos al probar  $m$  hipótesis nulas

	Test NO Significativo	Test Significativo	Total
Hipótesis Nulas Verdaderas	$U$	$V$	$m_0$
Hipótesis Nulas NO Verdaderas	$T$	$S$	$m - m_0$
Total	$m - R$	$R$	$m$

La tasa de falsos positivos esperada es  $E[V/(V+S)]$ . Como se muestra en Benjamini & Hochberg (1995), es posible controlar esta tasa; sin embargo lo ideal sería controlar el error tipo I de la prueba conjunta, que es  $P(V \geq 1)$ . La tasa de falsos positivos y el error tipo I son idénticos si todas las hipótesis son verdaderas; de lo contrario, la tasa de falsos positivos es menor. Raras veces se controla por el error tipo I en la práctica debido a las dificultades teóricas que presenta.

Suponga que se tienen  $H_1, \dots, H_m$  hipótesis a ser probadas, ordenadas de tal manera que sus p-valores cumplen  $P_1 \leq P_2 \leq \dots \leq P_m$ , donde  $P_i$  es el  $p$ -valor de la hipótesis  $H_i$ . Sea  $k$  el número  $i$  más grande tal que:

$$P_i \leq \frac{i}{m}q^* \tag{5}$$

Entonces se rechaza  $H_i$  para todo  $i \in 1, \dots, k$ . Esto controla la tasa de falsos positivos a un nivel máximo de  $q^*$ . La proposición anterior supone independencia entre las hipótesis nulas verdaderas. En el contexto de copia, este supuesto indica que el hecho de que un individuo no haya copiado es independiente del hecho de que otro individuo no haya copiado, lo que es razonable.

### 3. Índice $\kappa$

Esta sección está basada en (Jara et al. 2010). Para comenzar, se supone que tenemos un cuestionario con  $N$  preguntas,  $n$  posibles respuestas a cada pregunta y dos estudiantes  $c$  (sospechoso de copia) y  $s$  (sospechoso de ser la fuente) que responden de forma independiente cada pregunta (no se copian). Los resultados de las respuestas de cada estudiante se pueden representar por una matriz  $n \times n$  de la siguiente forma. Supongamos que las filas corresponden a respuestas de  $c$  y las columnas a respuestas de  $s$ . La fila  $i$  se llena de la siguiente forma. Se seleccionan todas las preguntas en las que  $c$  respondió la opción  $i$  ( $i$  es una de las posibles respuestas a una pregunta). En esas preguntas se determina el número de veces que  $s$  respondió cada una de las  $n$  alternativas y se llenan las respectivas celdas. Luego, en la celda  $(i, i)$  (en la diagonal) va a aparecer el número de respuestas coincidentes entre  $c$  y  $s$  con respuesta común  $i$  y en una celda  $(i, j)$  (por fuera de la diagonal), va a aparecer el número de veces que  $s$  respondió la opción  $j$  en las preguntas en que  $c$  respondió la opción  $i$ .

Sea  $\pi_{vv}$  la probabilidad de que una pregunta sea clasificada en la celda  $(v, v)$  y  $\pi_{v+}$ ,  $\pi_{+v}$  la probabilidad de que la pregunta sea clasificada en la fila o columna  $v$  respectivamente. La probabilidad observada de que los individuos coincidan en por lo menos una pregunta,  $\pi_0$  es:

$$\pi_0 = \sum_v \pi_{vv}. \quad (6)$$

Ahora, bajo la hipótesis de que la distribución conjunta de  $c$  y  $s$  de responder las diferentes opciones es igual para todas las preguntas se puede demostrar que la probabilidad de que coincidan en por lo menos una pregunta  $\pi_e$  es:

$$\pi_e = \sum_v \pi_{v+} \pi_{+v} \quad (7)$$

El estadístico  $\kappa$  (Kappa) se define como:

$$\kappa = \frac{\pi_0 - \pi_e}{1 - \pi_e} \quad (8)$$

que es un indicador del grado de coincidencia en las respuestas de dos individuos controlando por el número esperado de coincidencias si no hubiera

copia. Obsérvese que el índice  $\kappa$  es creciente en ambas variables. Entre más cercano a cero, menor es la probabilidad de que  $c$  se haya copiado de  $s$ .<sup>2</sup>

La prueba estadística para detectar sospechosos de copia se basa en la siguiente hipótesis:

$$\begin{aligned} H_0 &: \pi_0 \leq \pi_e \\ H_1 &: \pi_0 > \pi_e \end{aligned}$$

o de forma equivalente:

$$\begin{aligned} H_0 &: \kappa \leq 0 \\ H_1 &: \kappa > 0 \end{aligned}$$

Con las hipótesis mencionadas hasta el momento se puede obtener la distribución asintótica de  $\kappa$ , en el número de preguntas  $N$ : sea  $\hat{\kappa}$  el estadístico que se obtiene de sustituir las probabilidades por las frecuencias observadas (proporciones en la diagonal, fila y columna). Entonces asintóticamente  $\hat{\kappa}$  se distribuye  $N(\mu, \sigma^2)$  donde:

$$\begin{aligned} \mu &= 0 \\ \sigma^2 &= \frac{1}{N} \left( \frac{\pi_0(1 - \pi_0)}{(1 - \pi_e)^2} + a + b \right) \end{aligned}$$

y  $a$  y  $b$  son las siguientes constantes:

$$\begin{aligned} a &= \frac{2(1 - \pi_0)(2\pi_0\pi_e - \sum_v \pi_{vv}(\pi_{v+} + \pi_{+v}))}{(1 - \pi_e)^3} \\ b &= \frac{(1 - \pi_0)^2(\sum_v \sum_{v'} \pi_{vv'}(\pi_{v+} + \pi_{+v'})^2 - 4\pi_e^2)}{(1 - \pi_e)^4} \end{aligned}$$

Ahora, en el caso específico de la prueba de hipótesis mencionada arriba,  $\hat{\kappa}$  se distribuye asintóticamente  $N(0, \sigma^2)$ , donde las ecuaciones anteriores se reducen a:

$$\begin{aligned} \mu &= \kappa \\ \sigma^2 &= \frac{1}{N} \frac{1}{(1 - \pi_e)^2} (\pi_e(1 - \pi_e) + \sum_v \sum_{v'} \pi_{vv'}(\pi_{v+} + \pi_{+v'})^2 - 2 \sum_v \pi_{v+}\pi_{+v}(\pi_{v+} + \pi_{+v})) \end{aligned}$$

<sup>2</sup>Con las hipótesis hasta ahora mencionadas el índice de la pareja ordenada  $(c, s)$  es igual al índice de la pareja ordenada  $(s, c)$ . La diferencia entre los dos se hace evidente más adelante que se condicione el cálculo del índice a la habilidad de  $c$ .

Por simplicidad se estandariza la prueba. Sea  $Z = \frac{\hat{\kappa} - \mu}{\sigma}$ ; luego con la hipótesis nula se sigue que  $Z$  se distribuye asintóticamente  $N(0, 1)$ . Dada una significancia  $\alpha$  para la prueba, el valor crítico,  $z$  se determina a partir de (por ejemplo para un nivel de significancia del 95 % da un valor crítico de 1.645):

$$P(Z \geq z) = \alpha$$

### 3.1. Índice $\kappa$ : recodificación

La idea de la recodificación se basa en la observación de que una permutación de las opciones de las preguntas no afectaría la suma de la diagonal de la matriz mencionada anteriormente. Sin embargo, es en la diagonal que se registran las respuestas coincidentes. Luego, en principio, existe cierta libertad para modificar las celdas por fuera de la diagonal (las marginales). Utilizando esta idea, el objetivo es recodificar las respuestas de tal forma que la probabilidad conjunta de respuesta sea igual para todas las preguntas. El procedimiento propuesto es el siguiente.

Primero se calcula la distribución empírica de las respuestas a cada pregunta y se ordenan las respuestas a cada pregunta según su frecuencia (esto es lo que se denomina  $valor_a$  más abajo en la explicación del código). Luego se toman cada uno de los individuos de la base de datos y se cambian sus respuestas (A, B, C, etc.) por (1, 2, 3, etc.) respetando el ordenamiento de la distribución empírica mencionada anteriormente. El único cambio hecho es el de forzar que la respuesta 1 corresponda a la correcta. Las demás respuestas siguen el ordenamiento empírico entre las que no son correctas. Una vez recodificada toda la base de datos de respuesta de los individuos para todo  $c$  y  $s$ , se calcula la matriz de respuestas mencionada anteriormente pero utilizando como cadena de repuestas la cadena recodificada.

Obsérvese que la recodificación no iguala la distribución de las respuestas entre preguntas pero reduce sus diferencias manteniendo la consistencia con los datos observados. Ahora, la diferencia de la distribución entre preguntas puede depender de la habilidad de los individuos. Por esta razón en la tercera versión de índice Kappa se condiciona la distribución empírica de las respuestas a la habilidad medida como el número de respuestas correctas.

### 3.2. Índice $\kappa$ : recodificación condicional

Esta es la versión del índice Kappa que se ha implementado como un algoritmo de detección de copia. La diferencia principal con el método anterior es que ahora se calcula la distribución empírica de las respuestas por pregunta condicional al número de respuestas correctas. De esta forma para cada número posible de respuestas correctas y para cada pregunta se calcula una distribución empírica de las respuestas. Una vez hecho esto se toma la totalidad de la base de datos y se determina, según su cadena de respuestas, la habilidad de cada individuo (número de respuestas correctas) y la distribución empírica correspondiente a ese número de respuestas correctas. Ahora, dados  $c$  y  $s$  se utiliza el procedimiento anterior para determinar la distribución empírica con la misma habilidad de  $c$  y se recodifican  $c$  y  $s$  con esa misma distribución empírica de  $c$  para cada una de las preguntas. El efecto de esta recodificación es uniformizar aún más la distribución de las respuestas entre individuos condicional a la habilidad. Ahora, con la recodificación condicional de  $c$  y  $s$  se sigue el mismo procedimiento de la versión simplificada del índice Kappa.

### 3.3. Resultados

En esta sección se resumen los resultados de aplicar el índice  $\kappa$  a las pruebas del núcleo común del Examen de Estado para el ingreso a la educación superior, actualmente Saber 11, 2009 II. Con el 99.9% de confianza, el porcentaje de personas sospechosas de copia según el algoritmo es 48.29%. El porcentaje de parejas sospechosas de copia es de 2.94%. Este valor corresponde a las parejas ordenadas que son detectadas como sospechosas. Considerablemente mayor, es el porcentaje de salones en donde se sospecha hubo copia. Este es 96.94%.

Si se toman parejas compuestas por individuos que presentan pruebas en salones distintos, es de esperarse que el índice delate sospecha de copia en uno de cada mil casos, si las suposiciones de construcción del modelo son satisfechas en la práctica. Para verificar esto, se construyen 50,000 parejas de forma aleatoria e independiente (sobre distintos salones) para cada forma. Los resultados se incluyen en la tabla 2.

Tabla 2: Casos de copia por mil parejas independientes, para las 36 formas analizadas.

Forma	Tasa de copia	Forma	Tasa de copia	Forma	Tasa de copia
2	3.208333	106	1.041667	130	4.775
3	3.283333	108	2.141667	131	5.266667
4	5.9	120	5.233333	132	3.583333
8	3.258333	121	5.333333	133	3.625
10	3.075	122	5.483333	134	3.383333
41	4.15	123	3.033333	135	3.125
42	5.791667	124	2.691667	136	3.316667
43	2.341667	125	2.883333	137	3.241667
44	1.016667	126	3.225	138	3.266667
51	16.533333	127	2.9	139	3.541667
100	0.833333	128	2.883333	140	3.141667
101	1.216667	129	6.166667	149	19.575

Los resultados sugieren una fuerte diferencia entre las probabilidades de respuesta en las preguntas, que se suponían idénticas para la definición del índice. En efecto, los patrones de respuesta varían bastante con las preguntas. Es precisamente esta variación y la tasa de inocentes acusados de copia lo que motiva la utilización del índice  $\omega$ .

## 4. Índice $\omega$

Sea  $N$  el número de preguntas y  $n$  el número de alternativas para responder cada pregunta. Supongamos que un individuo con habilidad  $\theta_j$  que no ha copiado responde con probabilidad  $\pi$  la opción  $v$  en la pregunta  $i$  de acuerdo a un modelo de respuesta nominal. En otras palabras:

$$\pi_{iv}(\theta_j) = \frac{e^{\xi_{iv} + \lambda_{iv}\theta_j}}{\sum_{h=1}^m e^{\xi_{ih} + \lambda_{ih}\theta_j}} \quad (9)$$

$\xi_{iv}$  y  $\lambda_{iv}$  son parámetros del modelo y se conocen como el intercepto y la pendiente respectivamente. El intercepto y la pendiente pueden variar a través de las diferentes preguntas. Los parámetros de las preguntas ( $\xi_{iv}$  y  $\lambda_{iv}$ ) se estiman por máxima-verosimilitud marginal y la habilidad se estima usando el método EAP (Expected A Posteriori). Para la estimación se utilizó el paquete *virt*<sup>3</sup> de R.

El método de máxima-verosimilitud marginal, es una extensión del método de máxima-verosimilitud donde no se observan todas las variables de interés. En este caso no podemos observar la habilidad cuando queremos estimar  $\xi_{iv}$  y  $\lambda_{iv}$ . La función de maximó-verosimilitud marginal de un individuo  $j$  se puede escribir como:

$$L_j(\xi, \lambda | v_1^j, v_2^j, \dots, v_N^j) = \prod_{i=1}^N \pi_{iv_i^j}(\theta) f(\theta) d\theta, \quad (10)$$

donde  $f$  es la función de densidad de  $\theta$ ,  $v_1^j, v_2^j, \dots, v_N^j$  son las respuestas del individuo  $j$  a cada una de las preguntas, y  $\pi_{iv_i^j}(\theta)$  la probabilidad de que un individuo de habilidad  $\theta$  tenga el mismo patrón de respuesta del individuo  $j$ . Dado que no conocemos la habilidad del individuo  $j$  multiplicamos por la probabilidad de que tenga una habilidad  $\theta$  e integramos sobre todos los posibles valores de la habilidad. Se supone que  $f(\theta)$  es la función de densidad normal estándar.

La función de máxima-verosimilitud marginal total es:

$$L(\xi, \lambda | V) = \prod_{j=1}^T L_j, \quad (11)$$

---

<sup>3</sup>El paquete *virt* se puede encontrar en: <http://libirt.sourceforge.net/>.

donde  $T$  es el número de estudiantes que responde el examen,  $V$  es la matriz que contiene los patrones de respuesta de estos individuos y  $(\xi, \lambda)$  son los vectores de parámetros. Entonces los estimadores que buscamos son:

$$(\hat{\xi}, \hat{\lambda}) = \arg \max_{(\xi, \lambda)} L(\xi, \lambda | V) \quad (12)$$

Es necesario estimar la habilidad pues la proporción de respuestas correctas puede no reflejar el conocimiento de un individuo. Contestar una pregunta difícil correctamente es un indicador de habilidad superior a contestar una pregunta fácil correctamente.

El método de máxima probabilidad a posteriori para estimar la habilidad se basa en técnicas Bayesianas. Suponiendo que conocemos los parámetros  $\xi$  y  $\lambda$ , de acuerdo a la regla de Bayes tenemos que:

$$f(\theta | u, \xi, \lambda) \propto f(u | \theta, \xi, \lambda) f(\theta), \quad (13)$$

donde  $u$  es el patrón de respuestas de un individuo y  $\theta$  es la habilidad.  $f(u | \theta) = \prod_{i=1}^N \pi_{iv_i}(\theta)$  donde  $v_i$  es la opción que el individuo respondió a la pregunta  $i$ , y se supone que  $f(\theta)$  es la distribución normal estándar. Con esta información obtenemos la distribución a posteriori; sin embargo se necesita un estimador puntual de  $\theta$ , por lo que se toma su media. En resumidas cuentas:

$$\hat{\theta} = E[f(\theta | u)] \quad (14)$$

Dado que sólo obtenemos  $f(u | \theta) f(\theta)$ , que es proporcional, más no necesariamente igual a  $f(\theta | u)$ , es necesario calcular una esperanza normalizada. Si dividimos la distribución a posteriori de  $\theta$  en  $k$  intervalos podemos obtener la media como:

$$\hat{\theta} = E[f(\theta | u)] \approx \frac{\sum_{h=1}^k \theta_h f(u | \theta) f(\theta)}{\sum_{h=1}^k f(u | \theta) f(\theta)} \quad (15)$$

En (Van der Linden & Hambleton 1997) y (Hambleton, Swaminatha & Jane Rogers 1991) se puede encontrar más información de cómo se utiliza el método de máxima-verosimilitud marginal y máxima probabilidad a posteriori.

Una vez se estima el modelo de respuesta nominal, se intenta buscar patrones de respuestas que se desvíen del mismo para detectar la copia. La

variable de interés en este estudio es el número de respuestas que  $c$  le copió a  $s$ . Sea  $\gamma_{cs}$  este número.

El objetivo es probar las siguientes hipótesis:

$$H_0 : \gamma_{cs} = 0$$

$$H_1 : \gamma_{cs} > 0$$

Sea  $U_{ji}$  la respuesta del individuo  $j$  a la pregunta  $i$  y  $I_{csi}$  la función indicadora del conjunto  $U_{ci} = U_{si}$ . Es decir, toma el valor de 1 cuando el individuo  $c$  y  $s$  coinciden en su respuesta a la pregunta  $i$  y cero de lo contrario. Luego el número de respuestas en común entre  $c$  y  $s$  se puede expresar como:

$$M_{cs} = \sum_{i=1}^N I_{csi}. \quad (16)$$

Si los individuos no se copian (responden de forma independiente), la probabilidad de que  $c$  y  $s$  escojan la alternativa  $v$  en la pregunta  $i$  es

$$\pi_{csiv} = \pi_{iv}(\theta_c)\pi_{iv}(\theta_s) \quad (17)$$

Luego la probabilidad de una coincidencia entre  $c$  y  $s$  en la pregunta  $i$  es

$$P(I_{csi} = 1) = \pi_{csi} = \sum_{v=1}^n \pi_{csiv} \quad (18)$$

Teniendo en cuenta esta información, una forma equivalente de nuestra hipótesis es:

$$H_0 : \pi_{csi} = \sum_{a=1}^n \pi_{ia}(\theta_c)\pi_{ia}(\theta_s), \forall i \quad (19)$$

$$H_1 : \pi_{csi} = \begin{cases} \sum_{a=1}^n \pi_{ia}(\theta_c)\pi_{ia}(\theta_s) & \text{para } n - \gamma_{cs} \text{ preguntas} \\ 1 & \text{para } \gamma_{cs} \text{ preguntas} \end{cases} \quad (20)$$

Bajo la hipótesis nula el número de respuestas coincidentes,  $M_{cs} = \sum_{i=1}^N I_{csi}$  es la suma de  $N$  variables aleatorias de Bernoulli independientes cada una con una probabilidad distinta de coincidencia. La suma de

estas variables aleatorias sigue una distribución binomial generalizada que se puede calcular mediante un algoritmo recursivo. El valor crítico  $m^*$  que queremos encontrar es aquel tal que:

$$P(M_{cs} \geq m^*) \leq \alpha$$

Sea  $f_n(m)$  probabilidad de  $m$  éxitos en  $n$  intentos (éxitos entendidos como coincidencias en las respuestas). Ésta se puede calcular como el coeficiente de  $z^m$  en el polinomio:

$$\prod_{i=1}^n [1 - \pi_{csi} + z\pi_{csi}] \quad (22)$$

También es posible calcular esta probabilidad de manera recursiva:

$$f_n(m) = (1 - \pi_{csn})f_{n-1}(m) + \pi_{csn}f_{n-1}(m - 1) \quad (23)$$

Si  $n = 1$  entonces  $f_m(0) = Q$  y  $f_m(1) = P$ . Además  $f_n(m) = 0$  si  $m < 0$  o  $m > n$ .

Van der Linden & Sotaridona (2006) muestran que la distribución acumulada  $F_n(m)$  se puede determinar mediante:

$$F_n(m) = (1 - \pi_i)f_{n-1}(m) + F_{n-1}(m - 1), \quad (24)$$

donde  $\pi_i$  es la probabilidad de éxito en una pregunta  $i$  cualquiera,  $f_{n-1}(m)$  y  $F_{n-1}(m - 1)$  son la función de distribución marginal y acumulada de las  $n - 1$  preguntas que quedan al quitar  $i$ . Es con esta distribución acumulada que se calcula  $m^*$ .

Esta recursión sin embargo requiere un tiempo computacional que no permite realizar el ejercicio con las pruebas de estado que realiza el ICFES. Por ejemplo, para el caso de las pruebas Saber 5° y 9° de 2009 se necesitaría de un año<sup>4</sup> aproximadamente para ejecutar el análisis completo. Esto se debe a que la distribución binomial generalizada cambia de pareja a pareja, dado que la habilidad de ambos individuos, y por ende la probabilidad de éxito cambia con cada pareja. Este problema no puede ser solventado con

---

<sup>4</sup>En un computador con procesador Intel® Core 2 Duo 2.53 GHz y 4 gigas de memoria RAM

una matriz de búsqueda debido a que la habilidad estimada toma tantos valores como patrones de respuesta diferentes se observen.

Debido a esto se calcula un índice condicional que se comporta asintóticamente como una normal estándar. Si denotamos como  $U_s$  el vector de respuestas del individuo  $s$  y  $\pi_{ivs}(\theta_c)$  la probabilidad de responder la opción que  $s$  eligió en la pregunta  $i$  para un individuo de habilidad  $\theta_c$ , entonces el índice  $\omega$  se define de acuerdo a la siguiente ecuación:

$$\omega = \frac{M_{cs} - E[M_{cs}|U_s]}{\text{Var}[M_{cs}|U_s]^{1/2}}, \quad (25)$$

donde el valor esperado y la varianza se definen como:

$$E[M_{cs}|U_s] = \sum_{i=1}^N \pi_{ivs}(\theta_c) \quad (26)$$

$$\text{Var}[M_{cs}|U_s] = \sum_{i=1}^N \pi_{ivs}(\theta_c)(1 - \pi_{ivs}(\theta_c)) \quad (27)$$

El objetivo es probar las siguientes hipótesis:

$$H_0 : \omega \leq 0$$

$$H_1 : \omega > 0$$

Este índice se debe comparar con un valor crítico  $\omega^*$  obtenido a partir de una distribución normal estándar (Van der Linden & Sotaridona 2006). Para el análisis de las pruebas SABER 5° y 9° se utiliza un nivel de confianza del 99.9 %.

Nótese que el índice  $\omega$  es un test condicional. En otras palabras trata asimétricamente al individuo que se copia y a la fuente. Esto puede ser visto como una ventaja si se considera que la copia puede haber ocurrido en una sola dirección, o como una desventaja si se considera que dos individuos colaboraron mutuamente, y puede que por tener habilidades diferentes uno sea señalado como sospechoso de copia y otro no. Esto es especialmente preocupante al considerar que la potencia del test varía de acuerdo a las habilidades de los individuos.

## 5. Metodología ICFES: Índice C

El índice C cuantifica la similitud de las respuestas entre un individuo y todas las personas en su mismo salón. Si la similitud es improbable se dice que la persona es sospechosa de copia. En términos formales:

**Definición 1** (Distancia). *Se define la **distancia**,  $d$ , entre dos cadenas de respuestas como el número de respuestas diferentes entre ellas.*

**Definición 2** (Vecindad). *Sea  $d_i^A(j)$  el número de individuos a una distancia igual a  $i$  del individuo  $j$  en el conjunto  $A$ , y  $d_{min}^A(j)$  la distancia mínima entre  $j$  y los demás individuos del conjunto  $A$ .*

Se desea obtener la distribución empírica de la probabilidad de que en un salón haya  $x$  individuos a una distancia menor a  $d$ . Esta situación se modela mediante una distribución hipergeométrica donde tenemos una población total de tamaño  $N$ , de los cuales hay un grupo de individuos de tamaño  $D$  que tienen una característica especial (distancia  $d$  o menos) y deseamos ver cuál es la probabilidad de extraer  $x$  individuos con esa característica de un grupo de tamaño  $n$  (el tamaño del salón). Para obtener el parámetro  $D$  de esta distribución sería necesario calcular la distancia entre todas las parejas de individuos que presentan la prueba. Dado que esta prueba busca ser una alarma con bajo costo computacional, se decide aproximar la distribución hipergeométrica por una binomial con parámetros  $n$  y  $p \approx D/N$ , estimando  $p$  de una muestra aleatoria de tamaño  $M$ .

Sea  $Q$  el número de preguntas que contiene una forma. Sea  $A$  el número de individuos en una sede-jornada cualquiera que presetan la forma. Sea  $j$  un individuo de esa sede-jornada. Los pasos que se siguen en esta metodología son:

1. Sea  $q = \lceil Q * 0.15 \rceil$
2. Sea  $G^A(j) = d_0^A(j) + \dots + d_q^A(j)$  para todo  $j \in A$ .
3. Aquellos individuos para los que  $G^A(j) > 0$  son considerados sospechosos iniciales de copia.
4. Se toma una muestra de tamaño  $B$  de los individuos que no son sospechosos iniciales de copia entre todas las sedes-jornada. Para cada individuo,  $j$ , Se calcula  $d_0^B(j), \dots, d_{10}^B(j), \dots$

5. Sea  $\hat{p} = \frac{d_0^B(j) + \dots + d_{d_{min}^A + 2}^B(j)}{|B|}$
6. Se calcula  $P(j) = 1 - F(x_a, \hat{p}, |A|)$ , donde  $P$  denota la probabilidad de fraude,  $x_a = d_0(j)^A + \dots + d_{d_{min}^A + 2}^A(j)$  y  $F$  es la función de distribución acumulada de una binomial con parámetros  $\hat{p}$  y  $|A|$ .
7. Se escala esta probabilidad para crear un índice:  $I(j) = \left\| \log\left(\frac{1}{P(j)}\right) \right\|$
8. En los casos en que  $P(j)$  es nula el índice  $I(j)$  es igualado a 14.
9. Puesto que el procedimiento consideró todas las respuestas, se hace un ajuste por la proporción de respuestas correctas:

$$C(j) = I(j) - 14 \left( \frac{RC}{T} \right)^3$$

Donde  $RC$  es el número de respuestas correctas, y  $T$  el número de preguntas totales.

10. Si  $C(j) \geq 6$  se considera que el individuo  $j$  es sospechoso copia.

La metodología anterior; sin embargo tiende a sospechar de los colegios con buen rendimiento. La siguiente modificación se realiza para mitigar este efecto.

Sea  $NC$  el número de respuestas correctas de  $j$ . Sea  $B$  el conjunto de colegios con al menos un estudiante que tiene entre  $NC - 3$  y  $NC + 3$  respuestas correctas. De estos colegios se extrae una muestra aleatoria de tamaño 3000. De cada colegio es seleccionado un estudiante que tiene entre  $NC - 3$  y  $NC + 3$  respuestas correctas. A partir de este conjunto de estudiantes se calcula  $\hat{p}$  como en los pasos anteriores.  $P(j)$  e  $I(j)$  se calculan de manera idéntica. Sin embargo,  $C(j) = I(j) - \left\| \frac{6 * (PRC - \overline{PRC})}{SRC} \right\|$ , donde  $PRC$  es la proporción de respuestas correctas del individuo  $j$ ,  $\overline{PRC}$  es el promedio de la proporción de respuestas correctas de todos los estudiantes que presentan el examen, y  $SRC$  es la desviación estándar.

Nótese que el índice  $C$  carece de nivel de confianza. Esto implica que no se conoce la cantidad de inocentes acusados de copia teóricamente. Las simulaciones que se realizan a continuación son el primer paso para determinar este error empíricamente. Por otro lado, la filosofía de esta metodología es diferente a la de los índices  $\kappa$  y  $\omega$ : no compara parejas de individuos, sino un individuo en un salón (sede-jornada) con todos aquellos en el mismo lugar simultáneamente.

## 6. Simulaciones

### 6.1. Metodología para los índices $\omega$ y $\kappa$

La metodología que se utiliza para determinar los errores tipo I y tipo II de las pruebas consiste en utilizar la distribución empírica de los vectores de respuesta para generar parejas que no se hayan copiado. Con estas parejas se mide el error tipo I de la prueba. Después se introduce un factor de copia para medir la potencia de la prueba. Esto se hace para cada formato en general y condicionando por el nivel de habilidad de la fuente y del individuo que copia.

En términos formales, los pasos que se siguen para el caso en que se condiciona al nivel de habilidad son:

1. Se genera una malla de habilidades  $(\theta_f, \theta_c)$ , con ambos valores variando entre -2 y 2, con escalones de 0.5. En total se generan 36 parejas de habilidades diferentes. En realidad  $\theta_f$  y  $\theta_c$  representan intervalos de tamaño 0.5, donde se tiene individuos con habilidades entre  $[\theta_i, \theta_i + 0.5)$  con  $i = \{f, c\}$ . El estimador de la habilidad se obtiene al ajustar un modelo de respuesta nominal.
2. Para cada pareja ordenada de niveles de habilidad  $(\theta_f, \theta_c)$  se seleccionan 2,000 parejas, de tal manera que estén en sedes-jornadas diferentes y por ende no pueda haber copia entre ellas.
3. Sobre estas parejas se corre cada uno de los modelos de copia seleccionados y se calcula la proporción identificada con copia. Esta proporción es el error tipo I.
4. Para calcular el error tipo II, para el individuo que se copia se modifican sus cadenas de respuesta para introducir el fraude.
  - a) Se selecciona un nivel de copia: número de respuestas en los que hay copia.
  - b) Se selecciona aleatoriamente un grupo de preguntas del mismo tamaño del nivel de la copia.
  - c) Se cambian las respuestas del individuo que se copia por las de la fuente.<sup>5</sup>

---

<sup>5</sup>Si inicialmente alguna respuesta era igual, se deja como esta. Esto refleja el hecho de que los individuos pueden tener respuestas iguales sin que estas fueran parte del fraude.

5. Sobre estas simulaciones se corre cada uno de los modelos de detección de copia seleccionados y se calcula la proporción identificada con copia. Esta proporción es la potencia de la prueba. Se calcula la potencia para diferentes niveles de copia.

En el caso del índice  $\kappa$  es necesario calcular el *valor a* y para el índice  $\omega$  se necesita ajustar el modelo de respuesta nominal. Esto se hace utilizando una muestra independiente de observaciones<sup>6</sup>. Para la recodificación condicional del índice  $\kappa$  se utiliza la proporción de respuestas correctas como estimador de la habilidad.

Para el caso en que no se condiciona al nivel de habilidad, se seleccionan 100,000 parejas, de tal manera que estén en sedes-jornadas diferentes y por ende no pueda haber copia entre ellas, sin importar la habilidad de los individuos. Después se siguen los pasos 3-5 descritos anteriormente.

## 6.2. Metodología para el índice C

En el caso del Índice C es necesario modificar las simulaciones para incorporar el hecho de que este índice no compara parejas de individuos sino una persona contra todos aquellos en su mismo salón simultáneamente. Se genera una malla de habilidades  $(\theta_f, \theta_c)$  al igual que en la metodología para los índices  $\omega$  y  $\kappa$ .

Primero se fija un tamaño de salón,  $X$ , igual al tamaño promedio de cada forma. Se selecciona un individuo de habilidad  $\theta_c$  y  $X$  individuos de sede-jornadas diferentes de habilidad  $\theta_f$ . Se calcula el índice C para el individuo de habilidad  $\theta_c$ . Se repite este ejercicio 2,000 veces y se calcula el error tipo I.

Para calcular la potencia se supone que el individuo de habilidad  $\theta_c$  se copia de solamente uno de los otros individuos en el salón creado artificialmente. El número de preguntas que copia se designa el nivel de copia. Se repite este ejercicio 2,000 veces y se calcula la potencia del índice.

Cuando no se condiciona por la habilidad se toman 100,000 personas, y para cada persona se toman  $X$  personas que estén en sede-jornadas diferentes, sin importar su habilidad, y se realiza el mismo ejercicio.

---

<sup>6</sup>Es de esperar que los resultados varíen poco con la selección de la muestra.

### 6.3. Resultados sin condicionar

Para todas las pruebas se utilizó un nivel de confianza de 99.9% y se calculó el poder de los estadísticos utilizando niveles de copia de: 1,5,10,15,20,...,N, donde  $N$  es el número de preguntas que tiene el examen. Los resultados que se presentan a continuación no son condicionados a un nivel de habilidad particular ni de la fuente ni del individuo que se copia. Los resultados condicionados por habilidad se encuentran en el anexo A.

#### 6.3.1. Error Tipo I

Las simulaciones del índice  $\kappa$  muestran que esta prueba en general excede el error tipo I. En la siguiente tabla se ve cómo la proporción de inocentes acusados de copia es mucho mayor a la esperada<sup>7</sup>. En contraposición el índice  $\omega$  preserva el error tipo I.

El índice C tiene un error tipo I sustancialmente mayor. Esto se debe, en parte, a que no es posible seleccionar un nivel de confianza en este índice.

Tabla 3: Error tipo I. Inocentes acusados de copia (por cada mil personas).

Formato	$\omega$	$\kappa$	$C_{ICFES}$
PBA5041F1	0.74	11.12	214.89
PBA5041F2	0.93	7.74	93.73
PBA5042F1	0.87	4.95	154.39
PBA5042F2	0.93	4.22	95.8
PBA5043F1	0.77	6.88	164.24
PBA5043F2	1.28	3.8	97.43
PBA9041F1	0.85	22.3	215.87
PBA9041F2	0.93	20.66	104.4
PBA9042F1	0.6	13.61	103.65
PBA9042F2	0.75	7.53	92.8
PBA9043F1	0.86	9.06	180.46
PBA9043F2	0.97	5.52	106.38
Promedio	0.873333	9.7825	135.3367

<sup>7</sup>De 1 caso entre mil de acuerdo al nivel de confianza de 99.9% utilizado.

### 6.3.2. Potencia

En las figuras 1-12 se observa que el índice C es el que tiene una potencia mayor para niveles de copia bajos, pero que ésta no aumenta considerablemente a medida que aumenta el nivel de copia. La elevada potencia en niveles bajos se puede explicar por el comportamiento del error tipo I (Ver tabla 3). El índice  $\kappa$  tiene una potencia superior para niveles bajos de copia que el índice  $\omega$ , pero éste último toma la delantera, aproximadamente, cuando se copian más de la mitad de las preguntas del examen. Es posible concluir que la mayor potencia del índice  $\kappa$  no se debe a que éste sea más eficiente a la hora de identificar aquellos individuos que copian, sino a que identifica a más individuos como sospechosos de copia, sin importar si copiaron o no. La evidencia que soporta esta conclusión se encuentra en el anexo A.

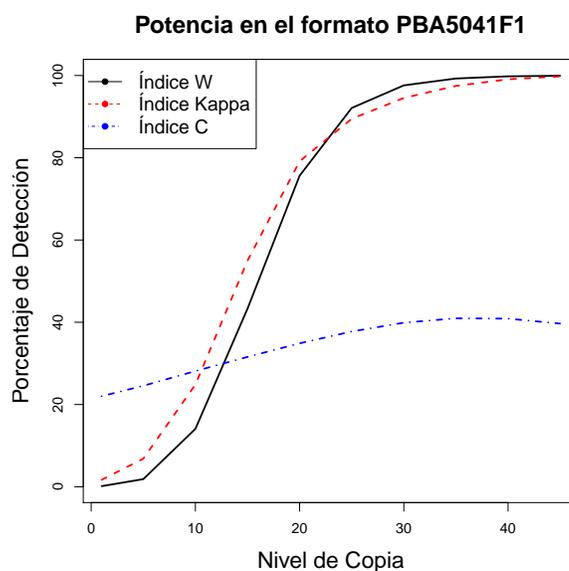


Figura 1: Comparación de índices en la forma 5º-Mayo-Matemáticas

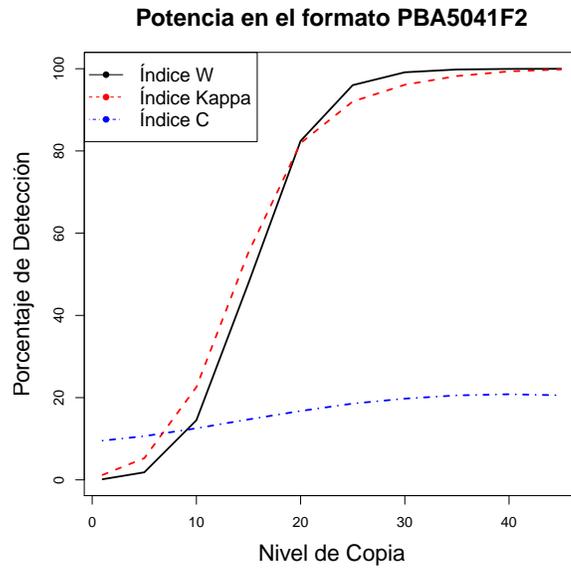


Figura 2: Comparación de índices en la forma 5°-Octubre-Matemáticas

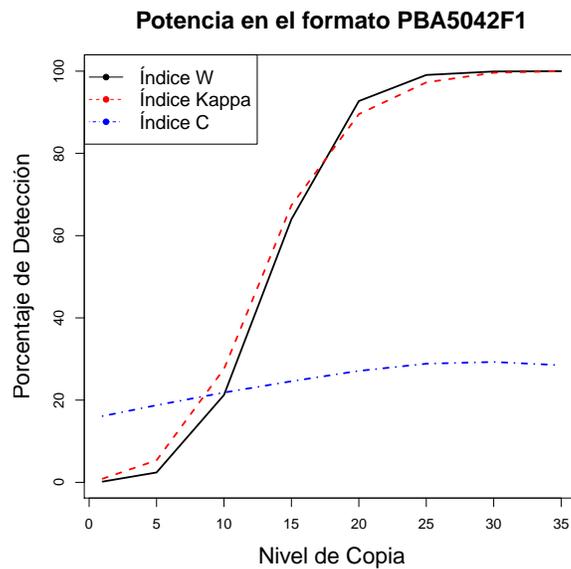


Figura 3: Comparación de índices en la forma 5°-Mayo-Lenguaje

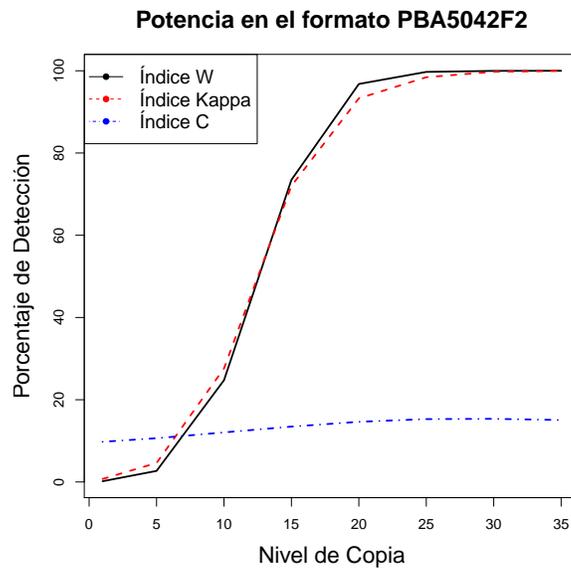


Figura 4: Comparación de índices en la forma 5°-Octubre-Lenguaje

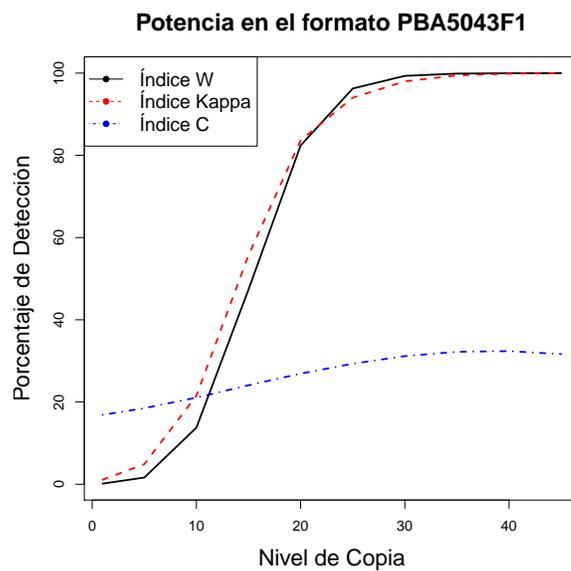


Figura 5: Comparación de índices en la forma 5°-Mayo-Ciencias

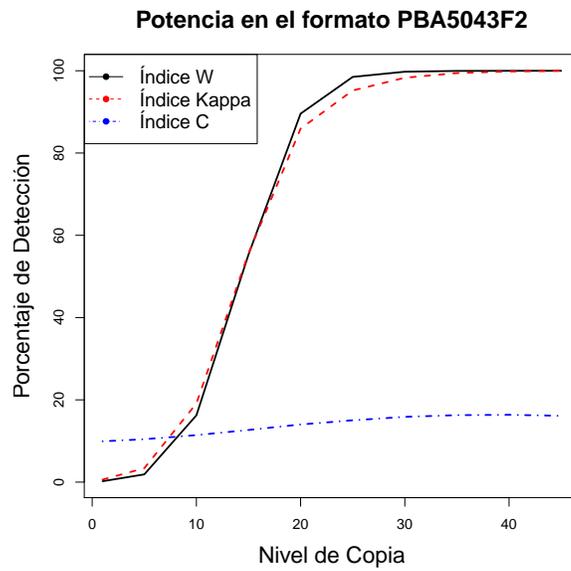


Figura 6: Comparación de índices en la forma 5°-Octubre-Ciencias

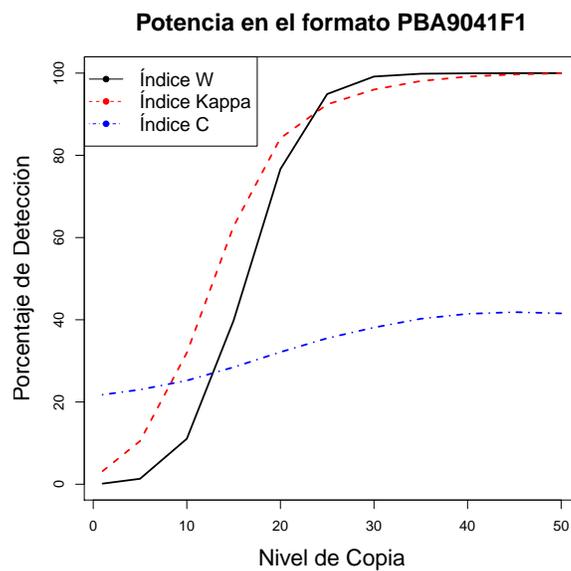


Figura 7: Comparación de índices en la forma 9°-Mayo-Matemáticas

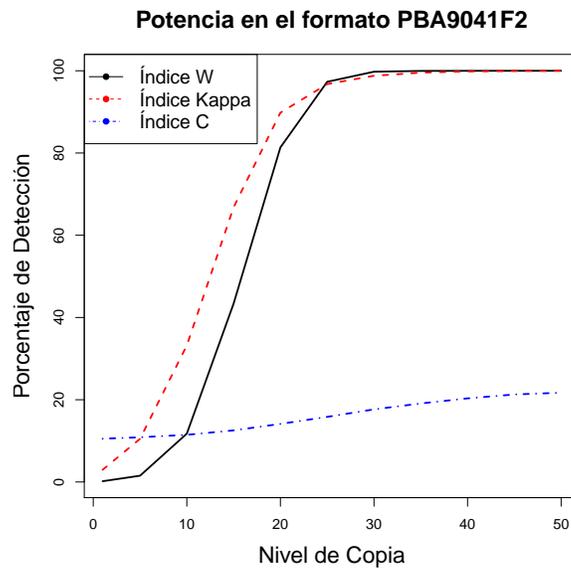


Figura 8: Comparación de índices en la forma 9°-Octubre-Matemáticas

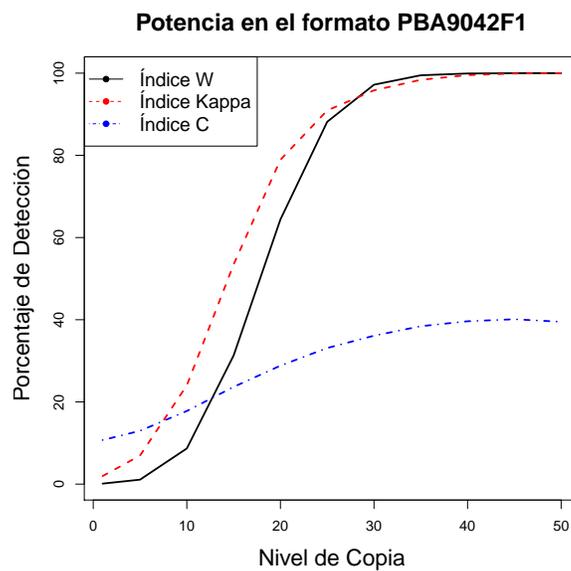


Figura 9: Comparación de índices en la forma 9°-Mayo-Lenguaje

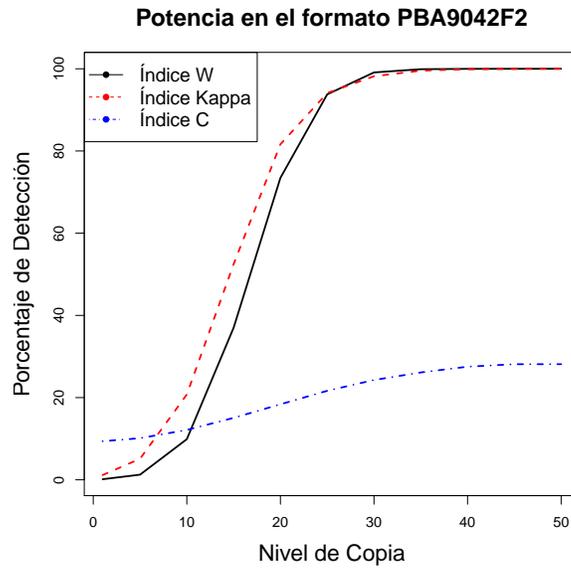


Figura 10: Comparación de índices en la forma 9°-Octubre-Lenguaje

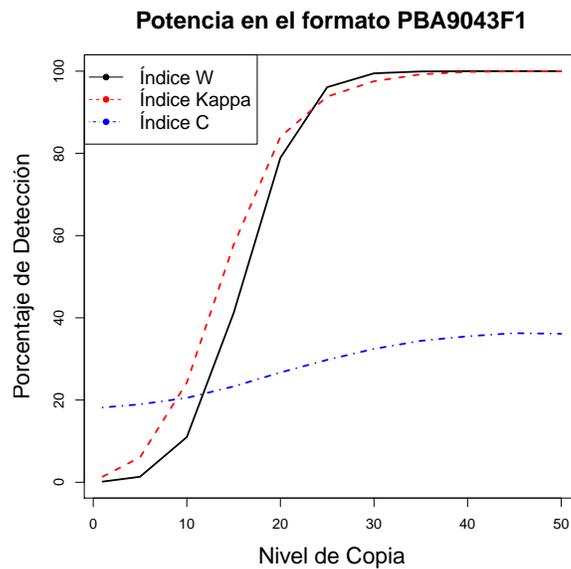


Figura 11: Comparación de índices en la forma 9°-Mayo-Ciencias

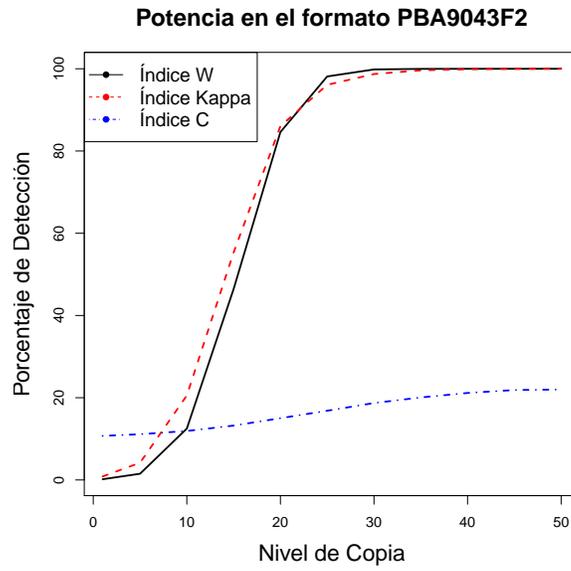


Figura 12: Comparación de índices en la forma 9º-Octubre-Ciencias

## 7. Análisis de las Pruebas Saber 5° y 9°

### 7.1. Análisis Descriptivo

Se consideran las doce formas como exámenes separados. Esto se debe a que hay tres áreas que se examinan, dos fechas en las que se realiza el examen y hay pruebas para los grados 5° y 9°. Las siguientes abreviaciones se utilizan de ahora en adelante para referirse a las distintas formas. Por área: 041 para matemáticas, 042 para lenguaje y 043 para ciencias. Por fecha: F1 para mayo y F2 para octubre. Por grado: 50 para 5° y 90 para 9°.

La información completa de cada uno de estos exámenes se encuentra en un archivo «raw». Cada uno de estos archivos contiene la siguiente información para todos los individuos que contestaron el examen: SNP, Formato, Sitio, String. El **SNP** corresponde a un identificador del estudiante. El **Formato** corresponde al área, fecha y grado del examen. El **Sitio** corresponde a la sede-jornada donde se realiza el examen, y el **String** a la cadena de respuestas del individuo. Adicionalmente, se cuenta con un archivo que contiene la cadena de respuestas correctas para el examen.

En la figura 13 se muestra la distribución de las habilidades estimadas usando el modelo de respuesta nominal para cada formato.

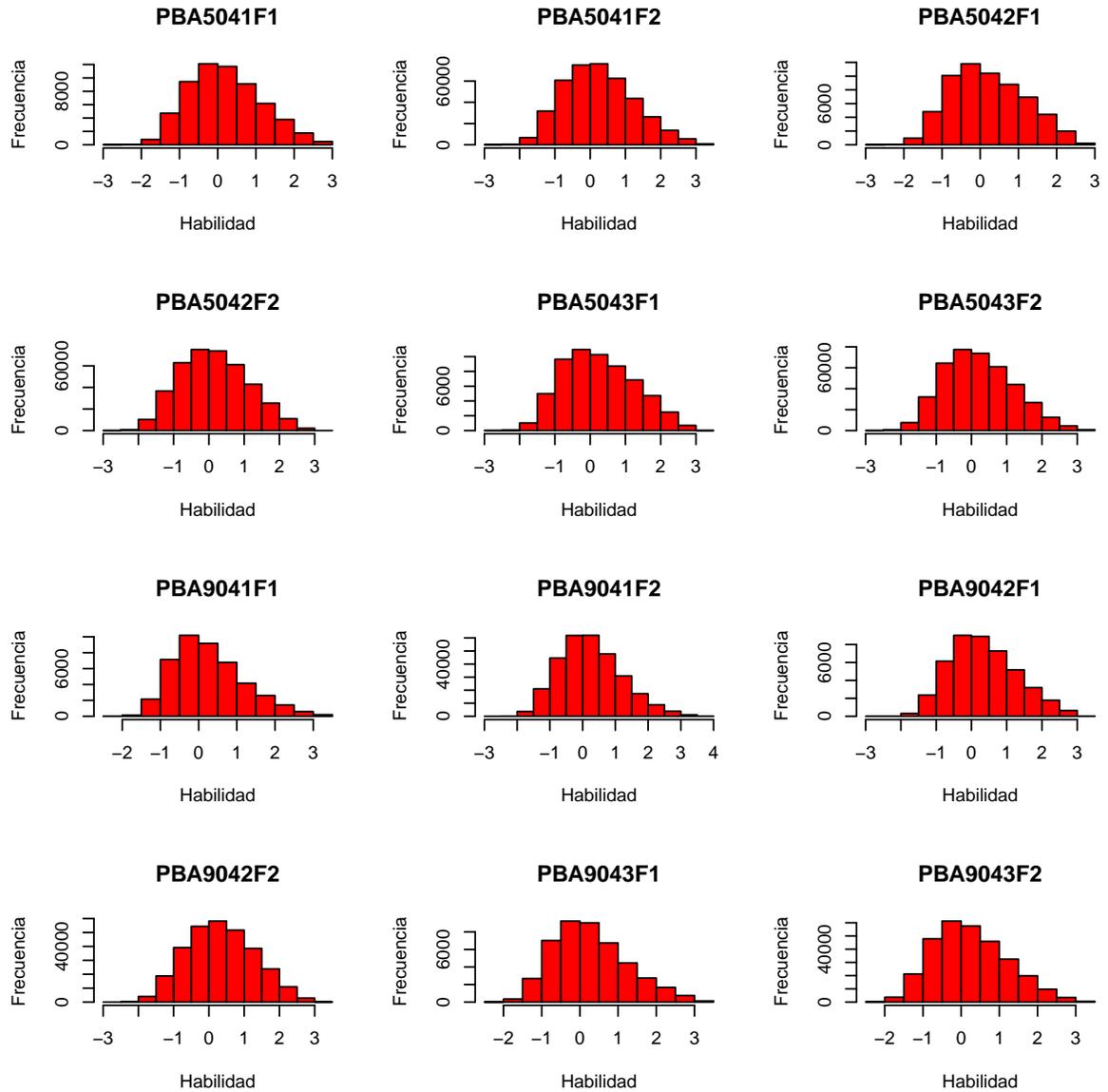


Figura 13: Habilidad estimada por formato.

Es posible extraer información útil del modelo de respuesta nominal. Por ejemplo se puede calcular la función de respuesta de cada pregunta. Esta

función indica la probabilidad de responder la pregunta correctamente en función de la habilidad. En la figura 14 se muestra la función de respuesta para cuatro preguntas diferentes de la prueba de matemáticas de 5° en mayo. Con estas funciones se puede ver qué preguntas eran las más sencillas, qué tan discriminatoria es cada pregunta, y en qué habilidad es discriminatoria. Por ejemplo, preguntas con una función de respuesta con una pendiente suave son poco discriminatorias, mientras que preguntas con se empinan rápidamente son discriminatorias. Aquella vecindad de habilidades donde se empina rápidamente la función es donde la pregunta es más discriminatoria.

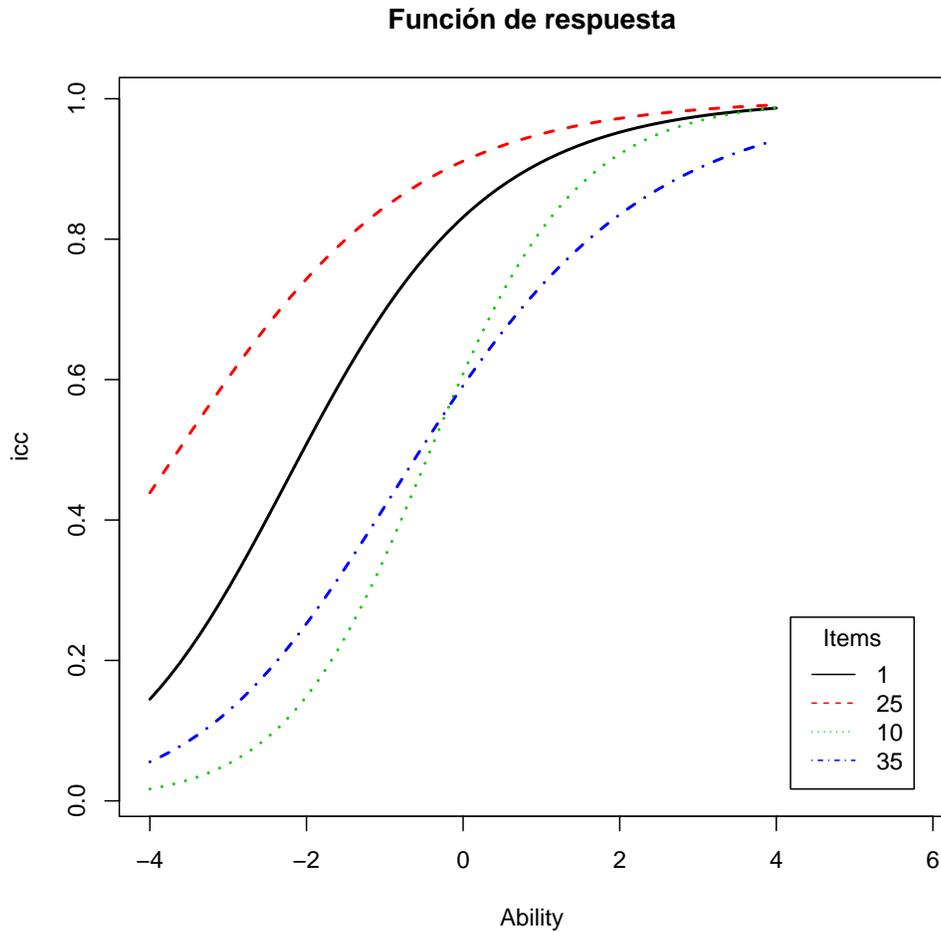


Figura 14: Función de respuesta del formato PBA5041F1.

## 7.2. Análisis Pruebas Controladas y Estadísticas con el Índice $\omega$

Las aplicaciones controladas y estadísticas se realizan con un control más cercano del ICFES, por lo que es de esperar que el porcentaje de parejas sospechosas de copias sea cercano al error tipo I. El nivel de confianza utilizado es del 99.9%.

Tabla 4: Parejas sospechosas de copia (por mil)

Formato	Estadístico	Controlado
PBA5041F1	15.643	10.834
PBA5041F2	2.637	2.01
PBA5042F1	7.349	4.348
PBA5042F2	2.482	2.846
PBA5043F1	7.702	10.145
PBA5043F2	2.953	3.155
PBA9041F1	14.082	20.785
PBA9041F2	4.225	2.683
PBA9042F1	3.482	0
PBA9042F2	2.911	3.027
PBA9043F1	7.165	4.866
PBA9043F2	3.96	3.285

Como se puede ver, en todos los casos la tasa de copia es superior al error tipo I, lo que podría tomarse como una señal de que existe copia en este tipo de pruebas, especialmente a la luz de los resultados encontrados en la sección 6.3.1. Sin embargo, tal y como se mencionó en la sección 2, estudiantes con educación similar tienen una probabilidad mayor de tener respuestas similares. Dado que en la aplicación de las pruebas SABER 5° y 9° se evalúa en un mismo sitio a todos los estudiantes de un colegio, este efecto puede sesgar los resultados. Si se considera que la prueba controlada es libre de copia, los resultados de la tabla 4 pueden sugerir qué tan fuerte es este efecto e insinuar el verdadero error tipo I de la prueba.

### 7.3. Análisis Pruebas Censales con el Índice $\omega$

Todos los resultados que se muestran a continuación utilizan un nivel de confianza del 99.9%. Sin embargo, es inmediato extraer el subconjunto de sospechosos con un nivel de confianza superior, usando los respectivos valores del estadístico.

En un computador con procesador Intel® Core 2 Duo 2.53 GHz y 4 gigas de memoria RAM, el análisis de cada forma demora entre 6 y 48 horas. En total las 12 formas analizadas tomaron aproximadamente 8 días. El resultado se encuentra en los archivos adjuntos dentro de “consolidado.rar”.

Con el nivel de confianza seleccionado, el porcentaje de individuos sos-

pechosos de copia según el algoritmo es 36.68 %. El porcentaje de parejas sospechosas de copia es 1.35 %. Este valor corresponde a las parejas *ordenadas* que son detectadas como sospechosas. El porcentaje de salones en donde se sospecha hubo copia es del 57.91 %.

La figura 15 presenta el número de copias que realiza una persona, considerando distintas formas y fuentes de copia.

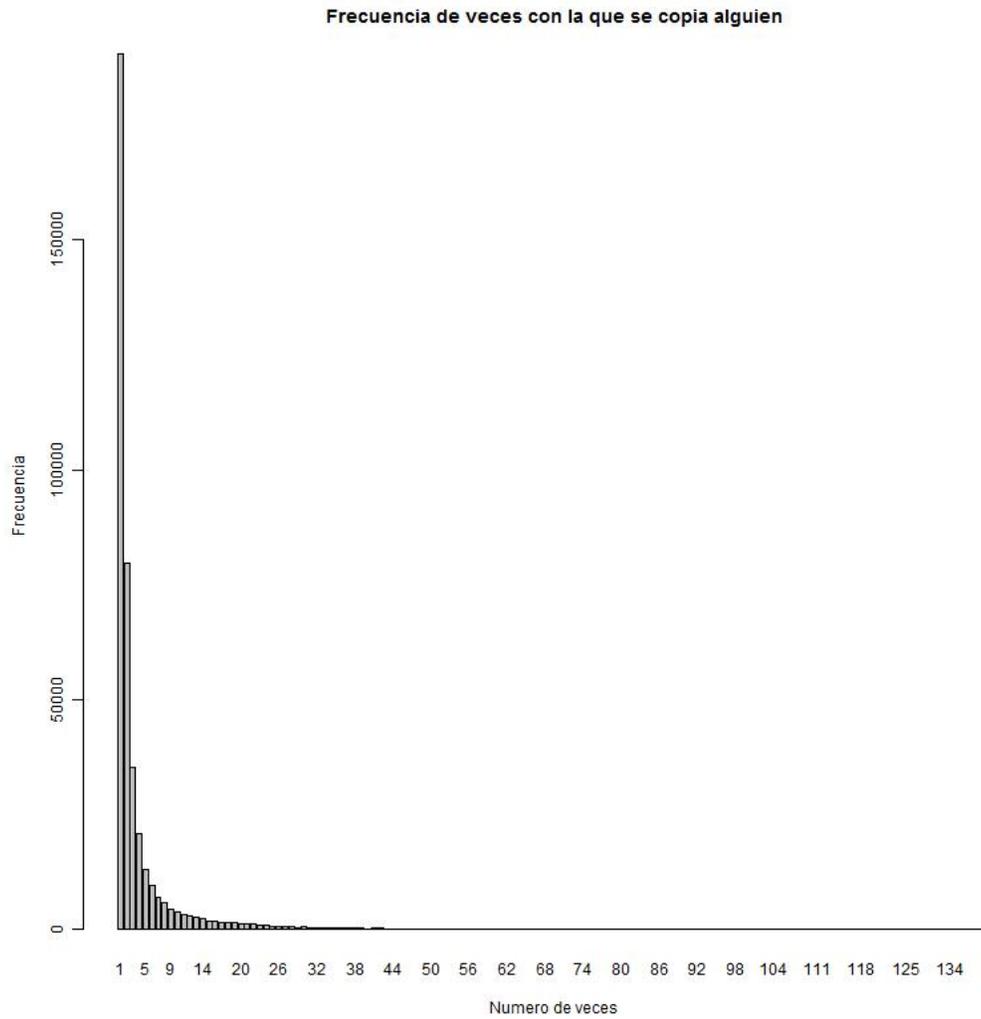


Figura 15: Número de veces que copia una persona (variando por salón, forma y fuente de copia).

En la figura 16 se exhibe la relación y la frecuencia que existe entre el número de respuestas correctas que tiene la fuente de la copia y el que copia. Como era de esperar, la proporción de respuestas correctas entre la fuente y la persona que copia suele ser similar.

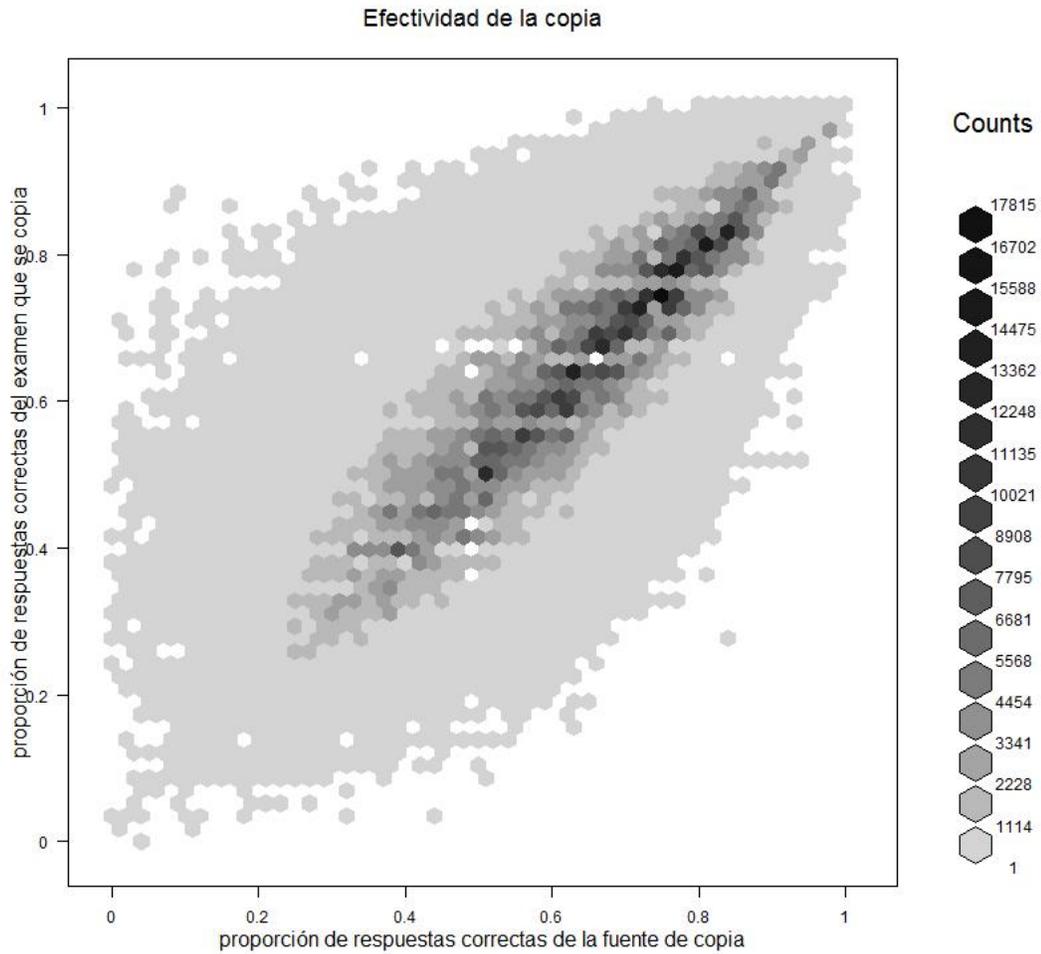


Figura 16: Relación del número de respuestas correctas entre individuo que copia y fuente, tomando las parejas sospechosas de copia.

En la figura 17 se relaciona el número de respuestas correctas que tiene el que copia, con el número de respuestas iguales que tiene con la fuente. Como se ve, aunque el número de respuestas incrementa con el número de preguntas iguales, la copia no es del todo efectiva, pues el número de respuestas correctas es menor que el número de preguntas iguales.

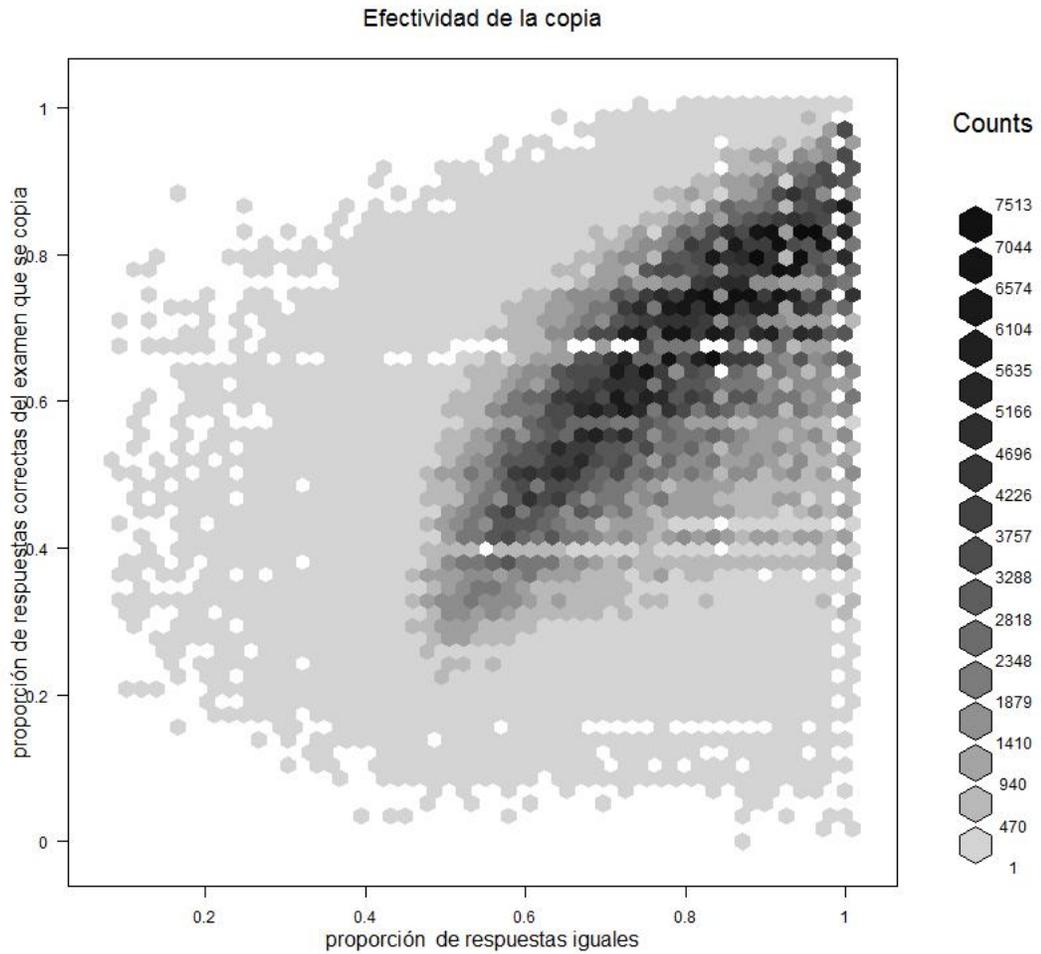


Figura 17: Relación entre número de respuestas correctas del sospechoso de copia y número de respuestas iguales con la fuente.

En la figura 18 se presenta la distribución de la habilidad entre aquéllos que se copian. Como se puede ver la mayoría de la gente sospechosa de copia presenta una habilidad por debajo del promedio<sup>8</sup>.

<sup>8</sup>El promedio es cero pues se supone que la habilidad sigue una distribución normal estándar.

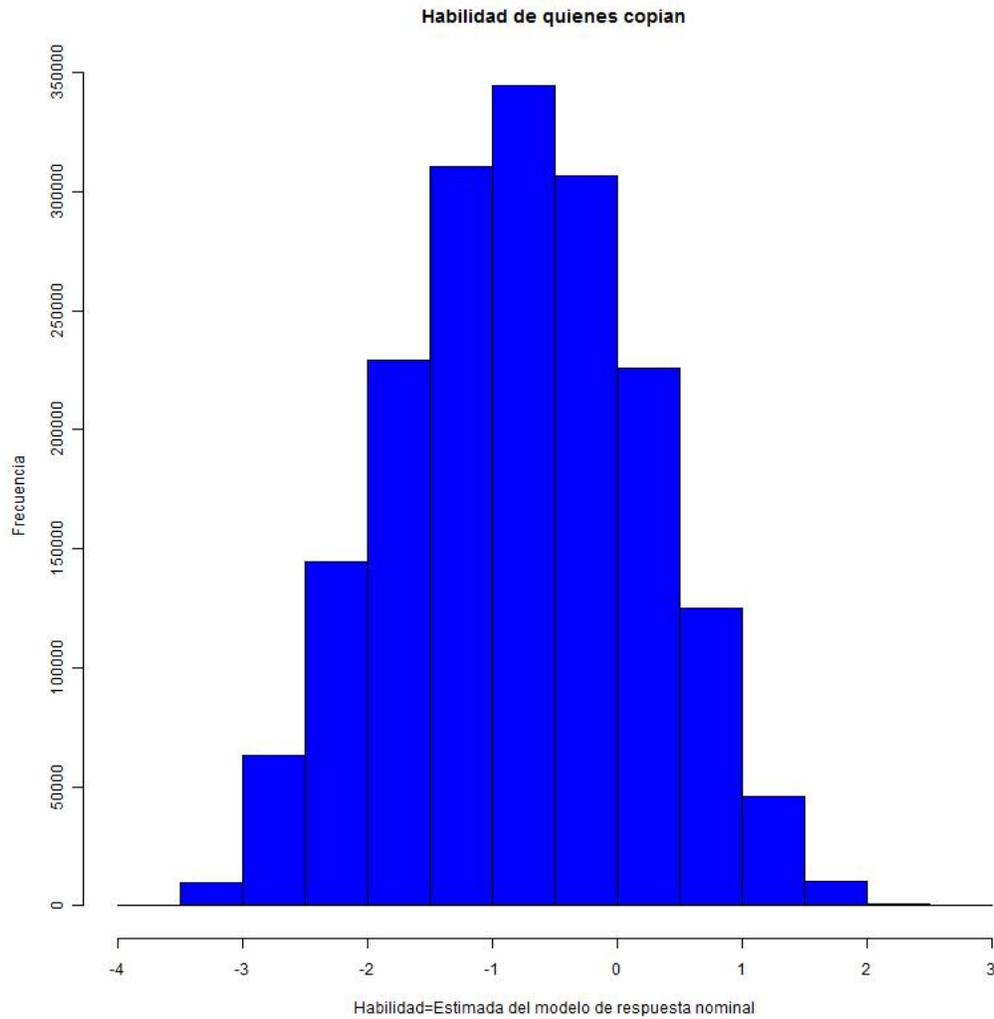


Figura 18: Distribución de la habilidad, calculada con el modelo de respuesta nominal, entre aquéllos que se copian.

En la figura 19 (y tabla 5) se encuentra la proporción de copia por forma. Esta proporción se calcula como el número de parejas sospechosas de copia, sobre el número de parejas que podrían copiarse<sup>9</sup> y presentan la copia. Un

<sup>9</sup>Se tiene en cuenta que solo puede haber copia entre las parejas dentro de un mismo

patrón consistente es que la copia en mayo (F1) es considerablemente mayor que en octubre (F2).

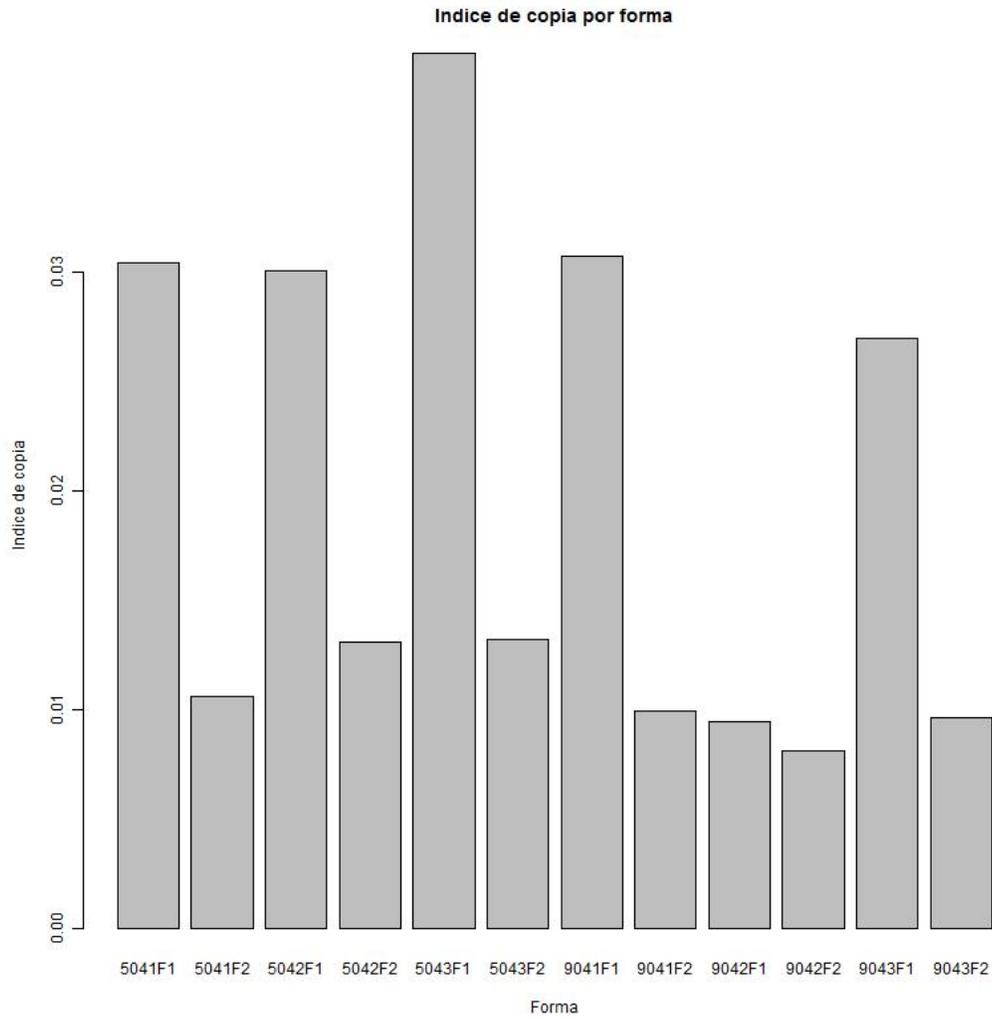


Figura 19: Proporción de parejas sospechosas de copia en cada forma.

---

salón.

Tabla 5: Proporción de parejas sospechosas de copia en cada forma.

Forma	Tasa de Copia (por mil)
PBA5041F1	30.377
PBA5041F2	10.632
PBA5042F1	30.028
PBA5042F2	13.113
PBA5043F1	39.984
PBA5043F2	13.191
PBA9041F1	30.693
PBA9041F2	9.949
PBA9042F1	9.446
PBA9042F2	8.149
PBA9043F1	26.948
PBA9043F2	9.671

Al comparar los resultados de copia en las pruebas censales con las pruebas estadísticas y controladas (Tabla 6 y figura 20), se puede ver que en general la copia es más extendida en la aplicación censal, seguido de la estadística y finalmente la controlada. Una vez más es importante señalar, como se mencionó en las secciones 2 y , que estudiantes con educación similar tienen una probabilidad mayor de tener respuestas similares. Dado que en la aplicación de las pruebas SABER 5° y 9° se evalúa en un mismo sitio a todos los estudiantes de un colegio, este efecto puede sesgar los resultados. Si se considera que la prueba controlada es libre de copia los resultados de la tabla 4 pueden sugerir qué tan fuerte es este efecto y sugerir el verdadero error tipo I de la prueba.

Tabla 6: Proporción de parejas sospechosas de copia en cada forma (por mil).

Forma	Censal	Controlada	Estadística
PBA5041F1	30.377	10.834	15.643
PBA5041F2	10.632	2.01	2.637
PBA5042F1	30.028	4.348	7.349
PBA5042F2	13.113	2.846	2.482
PBA5043F1	39.984	10.145	7.702
PBA5043F2	13.191	3.155	2.953
PBA9041F1	30.693	20.785	14.082
PBA9041F2	9.949	2.683	4.225
PBA9042F1	9.446	0	3.482
PBA9042F2	8.149	3.027	2.911
PBA9043F1	26.948	4.866	7.165
PBA9043F2	9.671	3.285	3.96

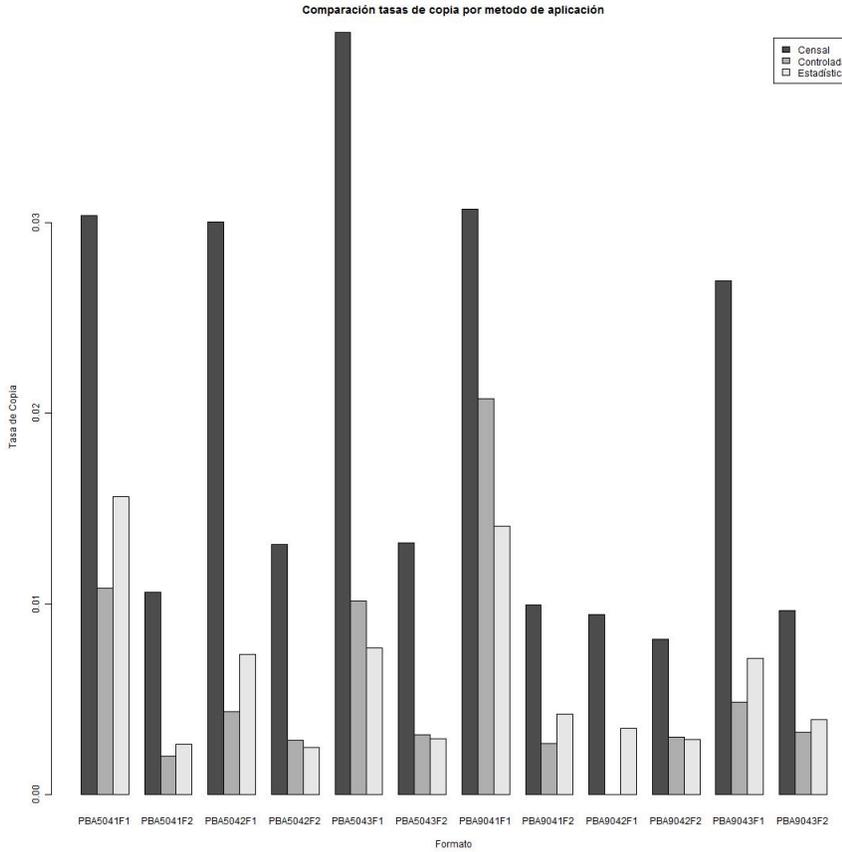


Figura 20: Comparación tasa de copia por aplicación.

### 7.4. Copia Masiva con el Índice $\omega$

Se dice que en un plantel hay copia masiva si más del 60% de los estudiantes es sospechoso de copia. Suponga que se rechaza  $H_i$  para todo  $i \in 1, \dots, k$ , controlando la tasa de falsos positivos a un nivel máximo de  $q^*$ , siguiendo la metodología de (Benjamini & Hochberg 1995). Para poder aplicar esta metodología es necesario tener un  $p$ -valor por cada individuo. Para cada individuo en cada sede-jornada y para cada prueba, se toma el menor  $p$ -valor, entre todas las parejas señaladas como sospechosas de copia donde este individuo es quien copia.

Si  $\frac{k}{m}$  es mayor al 60%, donde  $m$  es el número de estudiantes en un salón,

se dice que hubo copia masiva en el salón con un nivel de confianza de  $1 - q^*$ . Se toma  $q^* = 0.1\%$  en las pruebas realizadas. Los resultados se encuentran en el archivo adjunto “consolidado.rar”.

La figura 21 presenta la proporción de salones con copia masiva utilizando diferentes definiciones de copia masiva: más del 60%, 70% y 80% de los estudiantes.

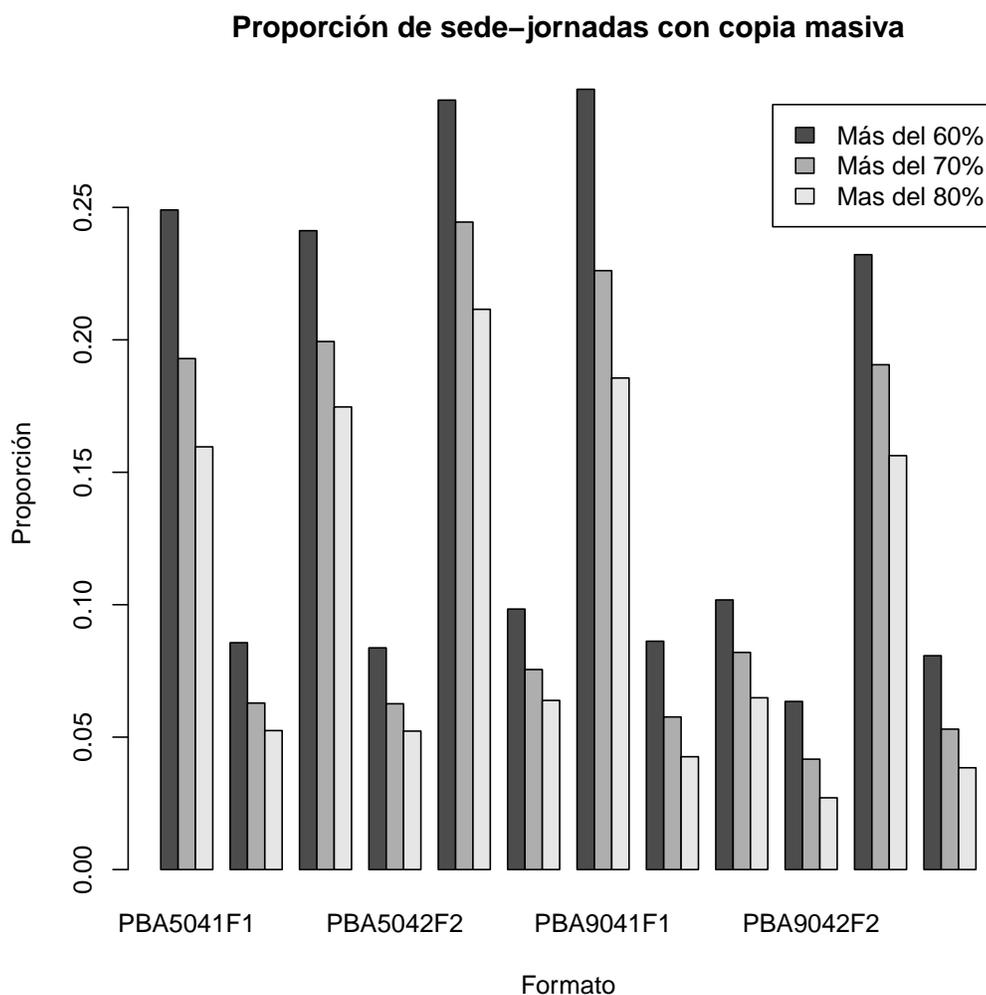
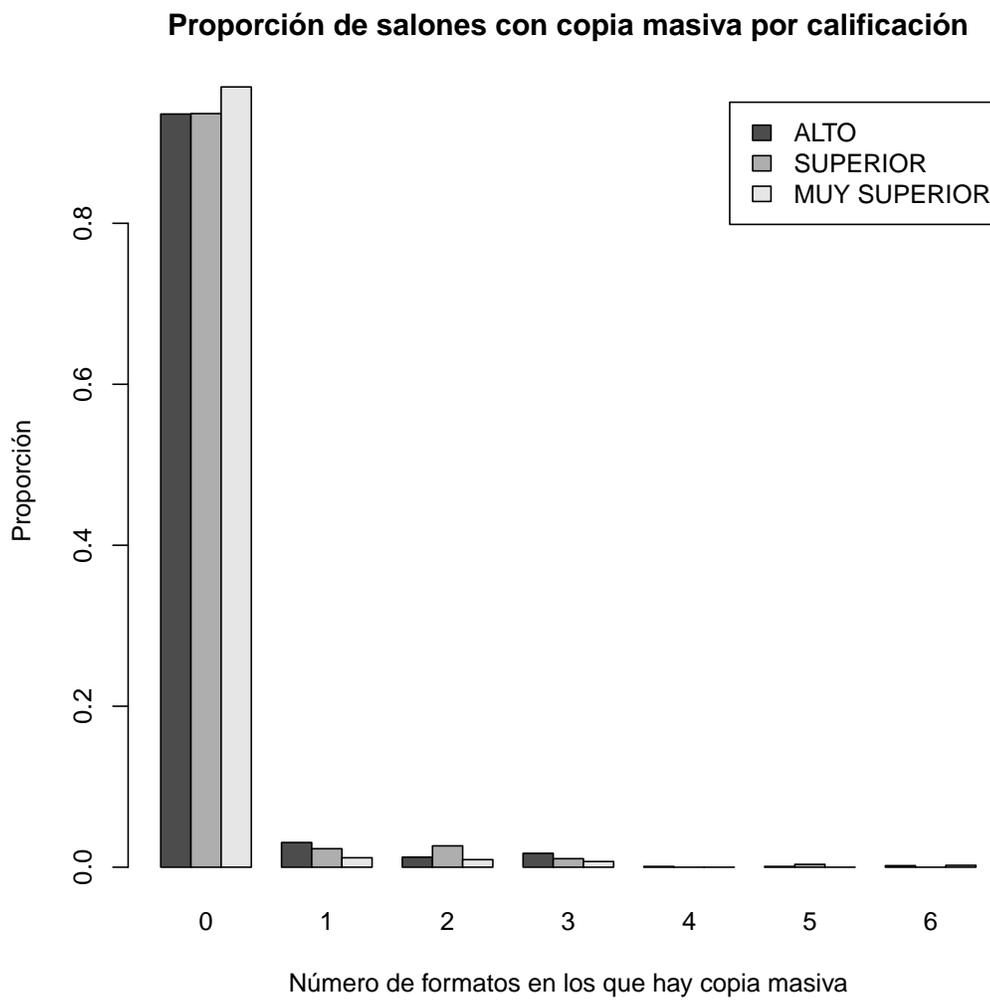
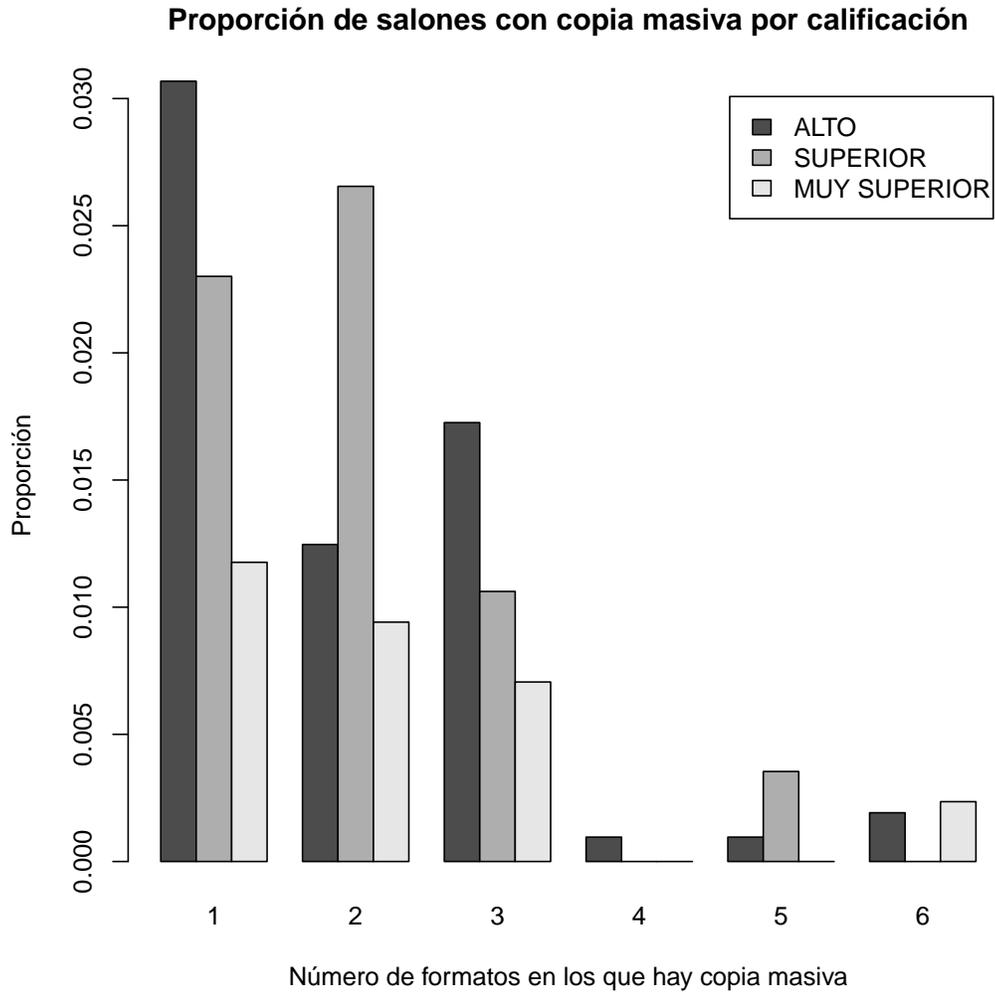


Figura 21: Copia masiva por forma

Las figuras 22 y 23 presentan la proporción de salones con copia masiva para diferentes categorías. Cada columna representa la proporción de sedes-jornada con copia masiva en  $X$  número de formas, entre la totalidad de sedes-jornada con esa calificación. La figura 23 excluye aquellas sedes-jornada donde no se presentó copia en ningún formato.

Figura 22: Número de formas con copia por calificación





### 7.5. Análisis Pruebas Censales con el Índice C

Al analizar las pruebas censales con el índice C, el porcentaje de individuos sospechosos de copia es 32.7%. El porcentaje de salones en donde se sospecha hubo copia es del 43.5%.

La figura 23 presenta el número de copias que realiza una persona, considerando distintas formas.

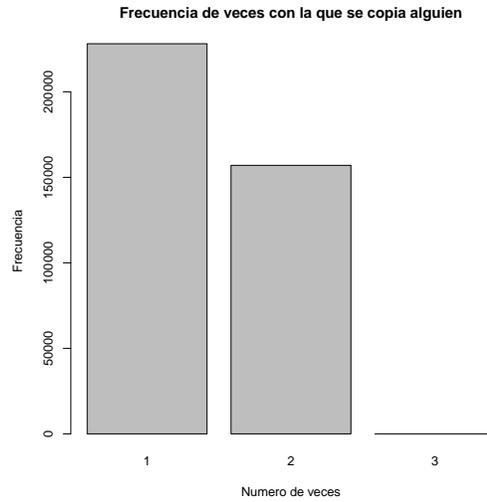


Figura 23: Número de veces que copia una persona (variando por salón y forma).

En la figura 24 (y tabla 7) se encuentra la proporción de copia por forma. Esta proporción se calcula como el número de personas sospechosas de copia, sobre el número de personas que presentan el examen.

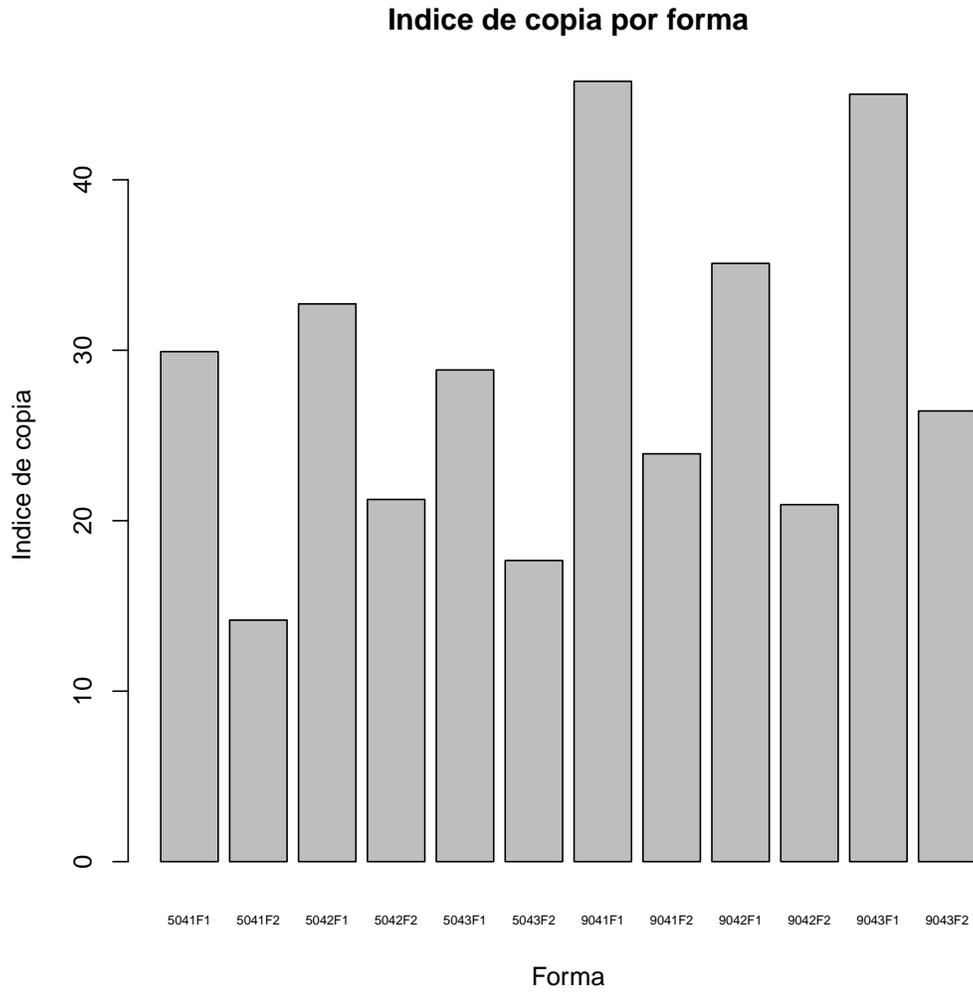


Figura 24: Proporción de parejas sospechosas de copia en cada forma.

Tabla 7: Porcentaje de individuos sospechosos de copia en cada forma.

Forma	Porcentaje de Copia
PBA5041F1	29.91 %
PBA5041F2	14.16 %
PBA5042F1	32.71 %
PBA5042F2	21.24 %
PBA5043F1	28.84 %
PBA5043F2	17.66 %
PBA9041F1	45.77 %
PBA9041F2	23.92 %
PBA9042F1	35.09 %
PBA9042F2	20.94 %
PBA9043F1	45.01 %
PBA9043F2	26.43 %

A continuación se encuentra una comparación de los resultados arrojados por el índice C y el índice  $\omega$ . Las figuras 26-37 muestran la concordancia entre los índices. La figura 25 explica las figuras de comparación.

Índice w	No Copia	Número de individuos que no copian según ambos índices.	Número de individuos que no copian según el índice W, pero sí según el índice C.	Individuos que se copian según el índice C como proporción de los que no se copian según el índice W.	Individuos que se no copian según el índice C como proporción de los que no se copian según el índice W.
	Copia	Número de individuos que no copian según el índice C, pero sí según el índice W.	Número de individuos que copian según ambos índices.	Individuos que no se copian según el índice C como proporción de los que se copian según el índice W.	Individuos que se copian según el índice C como proporción de los que se copian según el índice W.
		Individuos que se copian según el índice W como proporción de los que no se copian según el índice C.	Individuos que no se copian según el índice W como proporción de los que se copian según el índice C.	La proporción de individuos donde ambos índices coinciden.	
		Individuos que se no copian según el índice W como proporción de los que no se copian según el índice C.	Individuos que se copian según el índice W como proporción de los que se copian según el índice C.	La Proporción de individuos donde los índices no coinciden.	
		No Copia	Copia		
		Índice C			

Figura 25: Explicación de las figuras de comparación entre los índices.

Índice w	No Copia	32042	5861	0.15	0.85
	Copia	10076	12120	0.45	0.55
		0.24	0.33	0.734820879	
		0.76	0.67	0.265179121	
		No Copia	Copia		
Índice C					

Figura 26: Comparación en el formato PBA5041F1

Índice w	No Copia	304627	25781	0.08	0.92
	Copia	41809	31407	0.57	0.43
		0.12	0.45	0.832542168	
		0.88	0.55	0.167457832	
		No Copia	Copia		
		Índice C			

Figura 27: Comparación en el formato PBA5041F2

Índice w	No Copia	32001	9300	0.23	0.77
	Copia	8673	10481	0.45	0.55
		0.21	0.47	0.702704491	
		0.79	0.53	0.297295509	
		No Copia	Copia		
Índice C					

Figura 28: Comparación en el formato PBA5042F1

Índice w	No Copia	279108	50055	0.15	0.85
	Copia	37878	35467	0.52	0.48
		0.12	0.59	0.781537261	
		0.88	0.41	0.218462739	
		No Copia	Copia		
		Índice C			

Figura 29: Comparación en el formato PBA5042F2

Índice w	No Copia	31564	6578	0.17	0.83
	Copia	11418	10844	0.51	0.49
		0.27	0.38	0.70207271	
		0.73	0.62	0.29792729	
		No Copia	Copia		
		Índice C			

Figura 30: Comparación en el formato PBA5043F1

Índice w	No Copia	288666	36351	0.11	0.89
	Copia	45233	35287	0.56	0.44
		0.14	0.51	0.798824768	
		0.86	0.49	0.201175232	
		No Copia	Copia		
		Índice C			

Figura 31: Comparación en el formato PBA5043F2

Índice w	No Copia	16678	7239	0.30	0.70
	Copia	7494	13166	0.36	0.64
		0.31	0.35	0.669493236	
		0.69	0.65	0.330506764	
		No Copia	Copia		
Índice C					

Figura 32: Comparación en el formato PBA9041F1

Índice w	No Copia	188844	33083	0.15	0.85
	Copia	41839	39467	0.51	0.49
		0.18	0.46	0.75292267	
		0.82	0.54	0.24707733	
		No Copia	Copia		
		Índice C			

Figura 33: Comparación en el formato PBA9041F2

Índice w	No Copia	24777	9445	0.28	0.72
	Copia	4349	6305	0.41	0.59
		0.15	0.60	0.692619663	
		0.85	0.40	0.307380337	
		No Copia	Copia		
		Índice C			

Figura 34: Comparación en el formato PBA9042F1

Índice w	No Copia	206352	30637	0.13	0.87
	Copia	33017	32775	0.50	0.50
		0.14	0.48	0.789768843	
		0.86	0.52	0.210231157	
		No Copia	Copia		
		Índice C			

Figura 35: Comparación en el formato PBA9042F2

Índice w	No Copia	18413	8781	0.32	0.68
	Copia	6230	11396	0.35	0.65
		0.25	0.44	0.665082552	
		0.75	0.56	0.334917448	
		No Copia	Copia		
		Índice C			

Figura 36: Comparación en el formato PBA9043F1

Índice $\omega$	No Copia	184895	38686	0.17	0.83
	Copia	38528	41614	0.48	0.52
		0.17	0.48	0.745774933	
		0.83	0.52	0.254225067	
		No Copia	Copia		
Índice C					

Figura 37: Comparación en el formato PBA9043F2

### 7.6. Copia Masiva con el Índice C

En general, se dice que en un plantel hay indicios de copia masiva si más del 60% de los estudiantes es sospechoso de copia. Se comparan los resultados obtenidos al utilizar el índice  $\omega$  con los del índice C. Las figuras 38,39 y 40 presentan la proporción de sede-jornadas con copia masiva para ambos índices, utilizando diferentes definiciones de copia masiva.

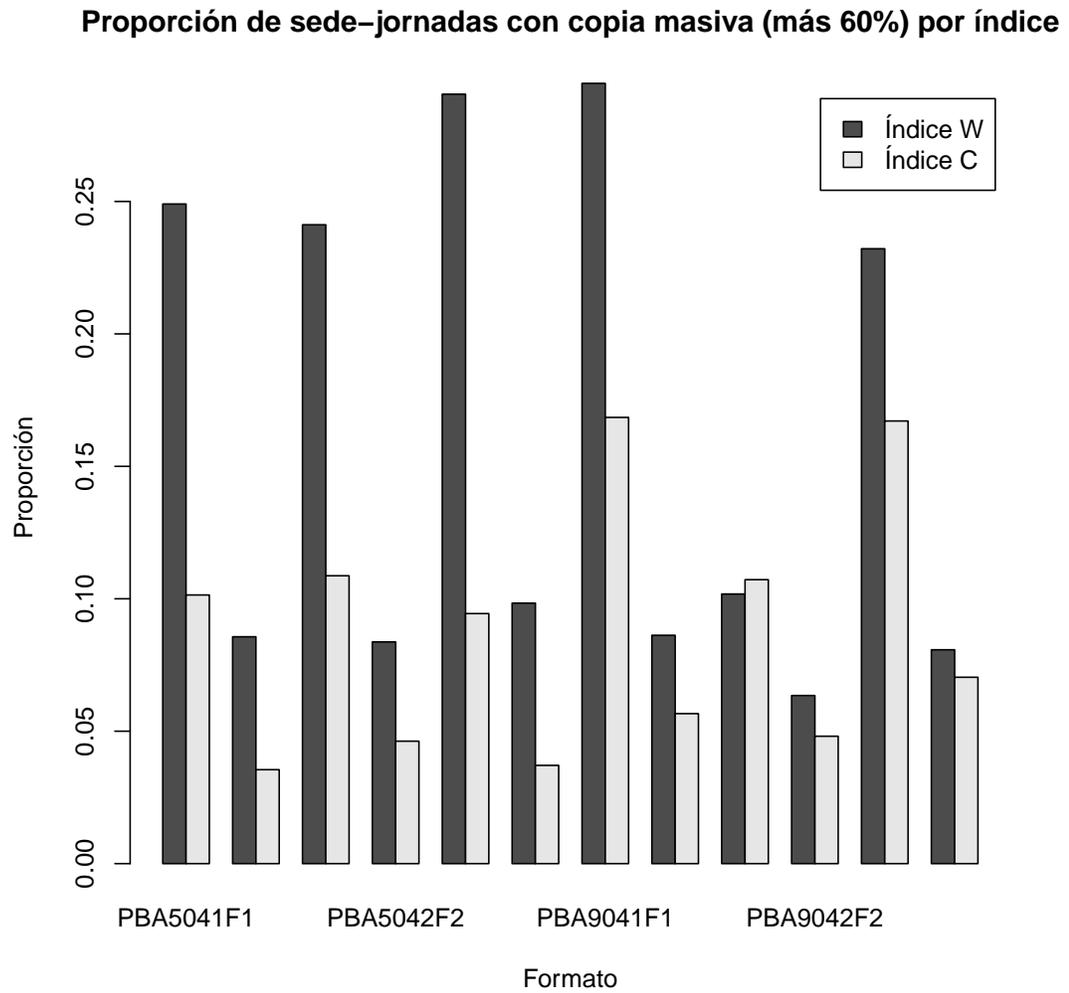


Figura 38: Comparación copia masiva (más 60%) por índice

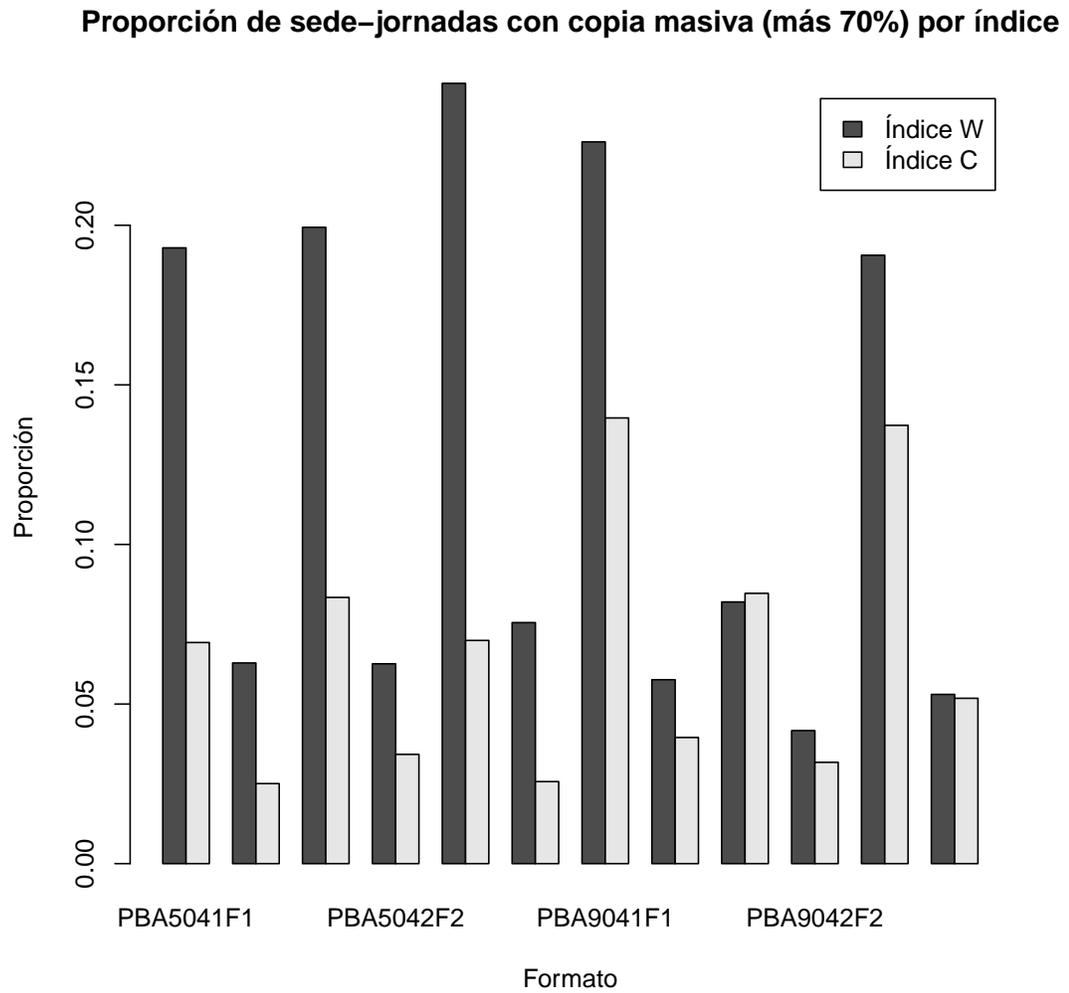


Figura 39: Comparación copia masiva (más 70%) por índice

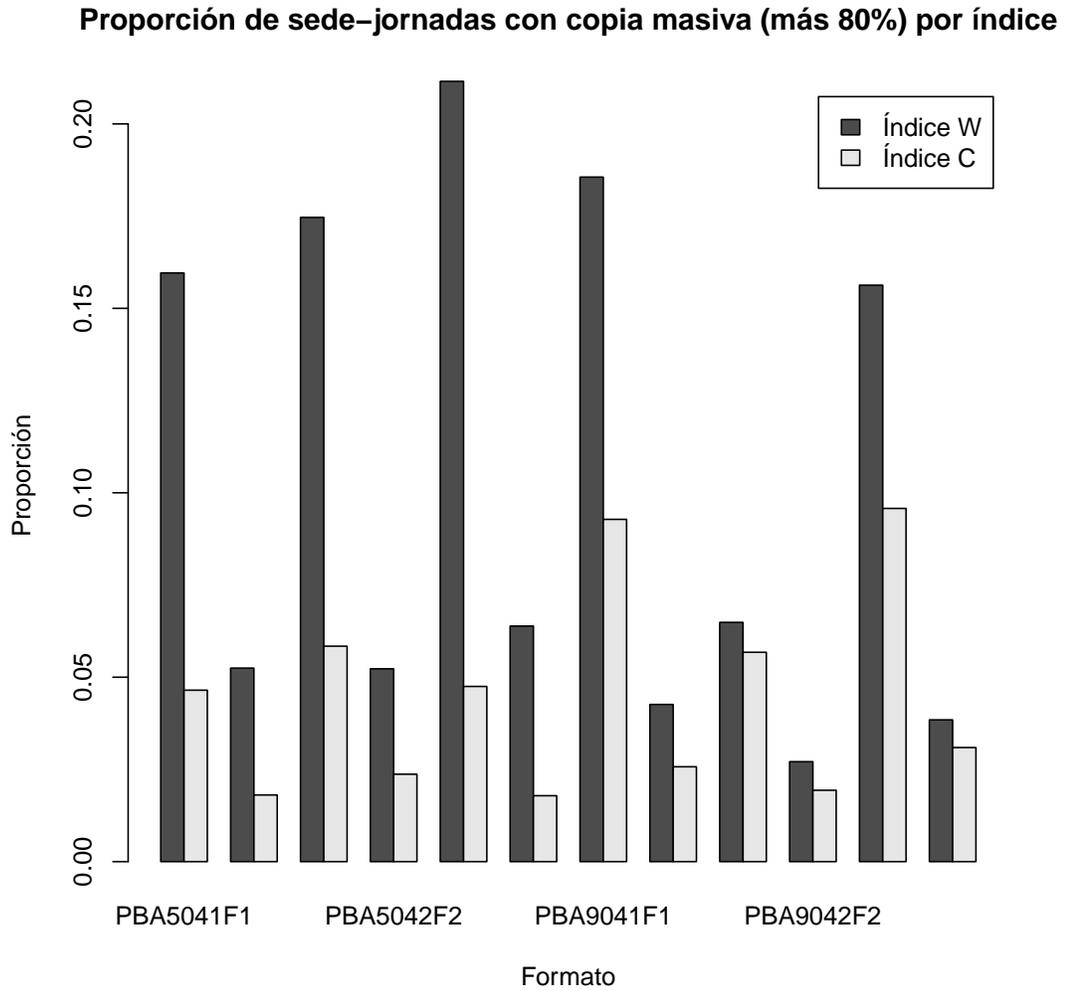


Figura 40: Comparación copia masiva (más 80 %) por índice

Las figuras 41, 42 y 43 presentan la proporción de salones con copia masiva, según el índice C y el índice  $\omega$ , para diferentes categorías. Cada columna representa la proporción de sedes-jornada con copia masiva en  $X$  número de formas, entre la totalidad de sedes-jornada con esa calificación.

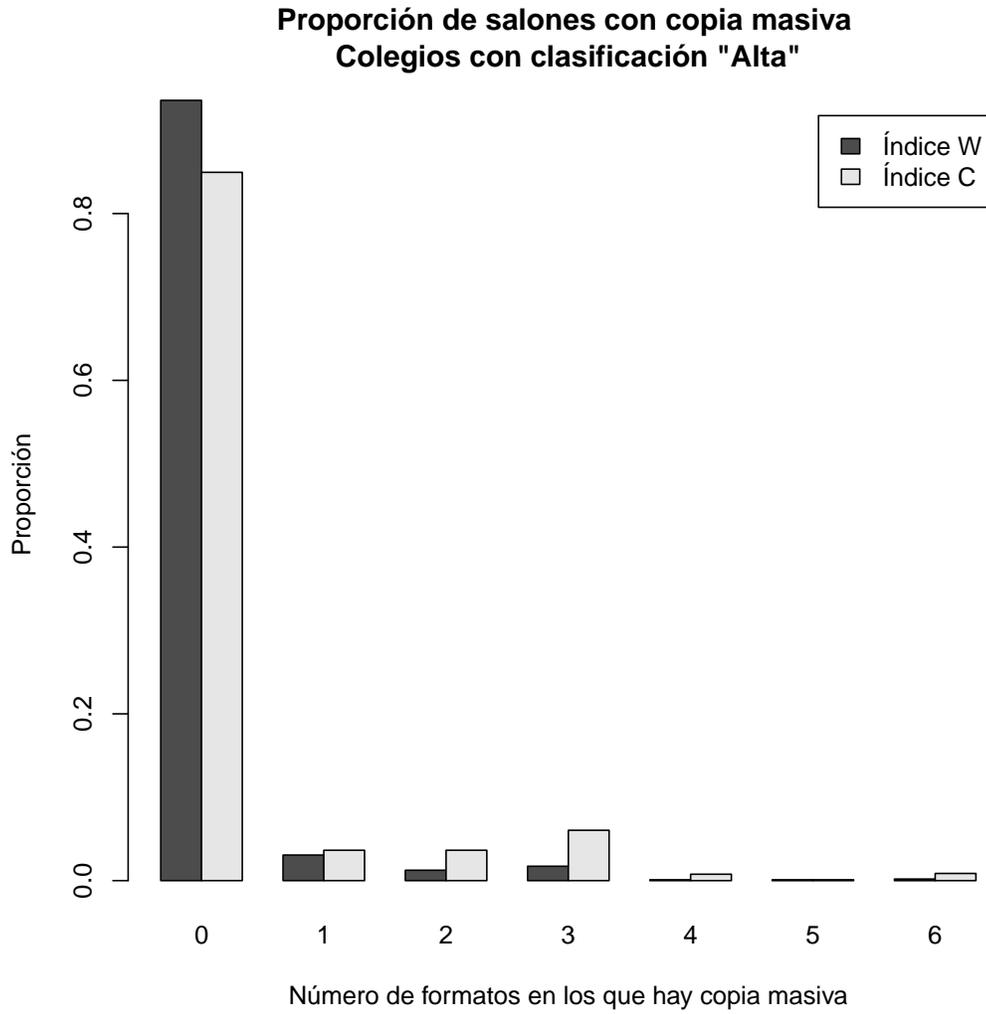


Figura 41: Comparación copia masiva por índice para colegios con clasificación "alta"

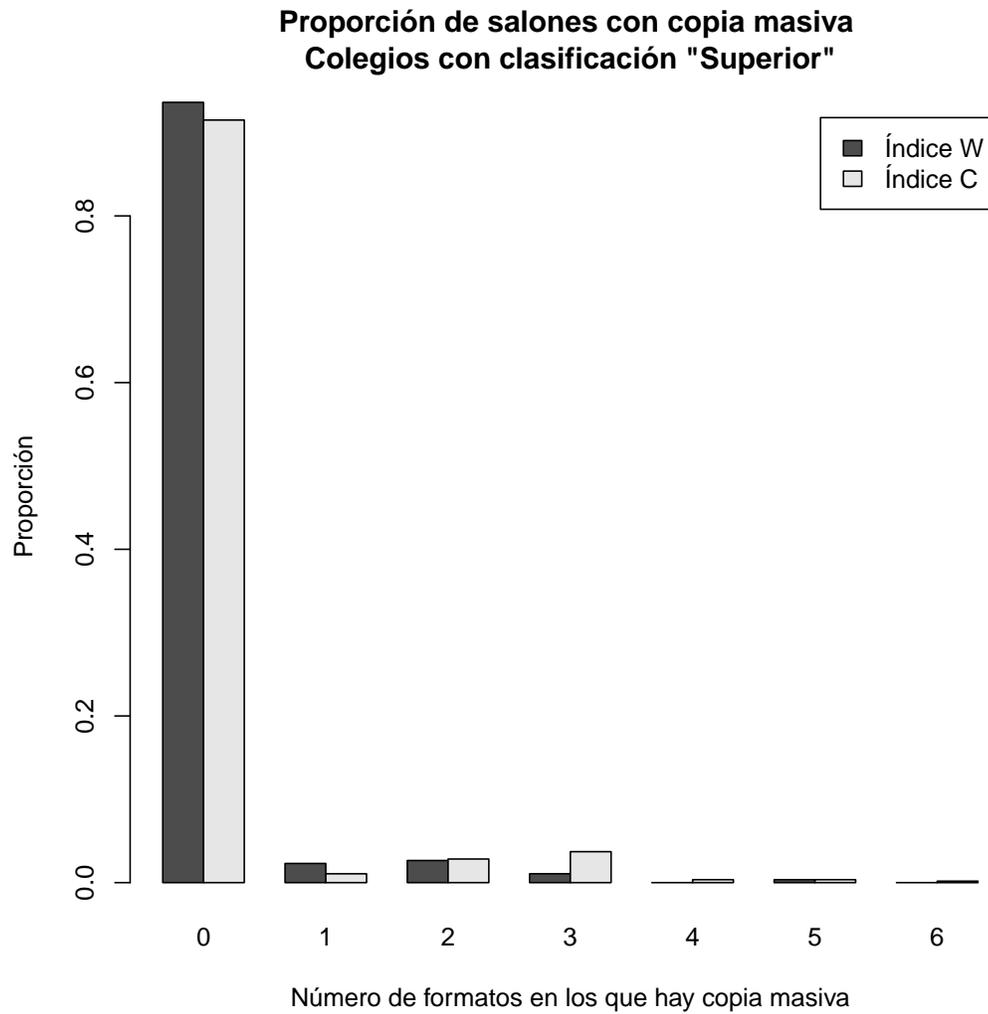


Figura 42: Comparación copia masiva por índice para colegios con clasificación "superior"

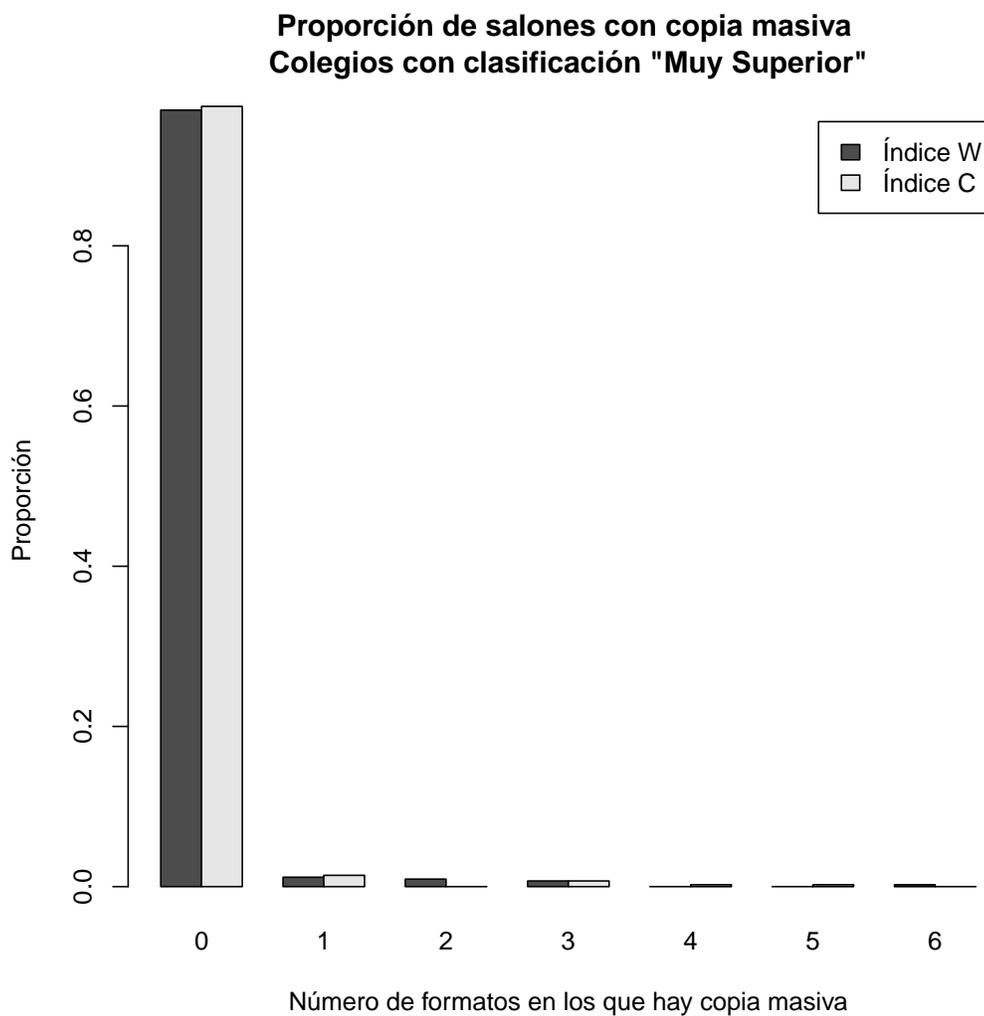


Figura 43: Comparación copia masiva por índice para colegios con clasificación “muy superior”

Las figuras 44,45 y 46 excluyen aquellas sedes-jornada donde no se presentó copia en ningún formato.

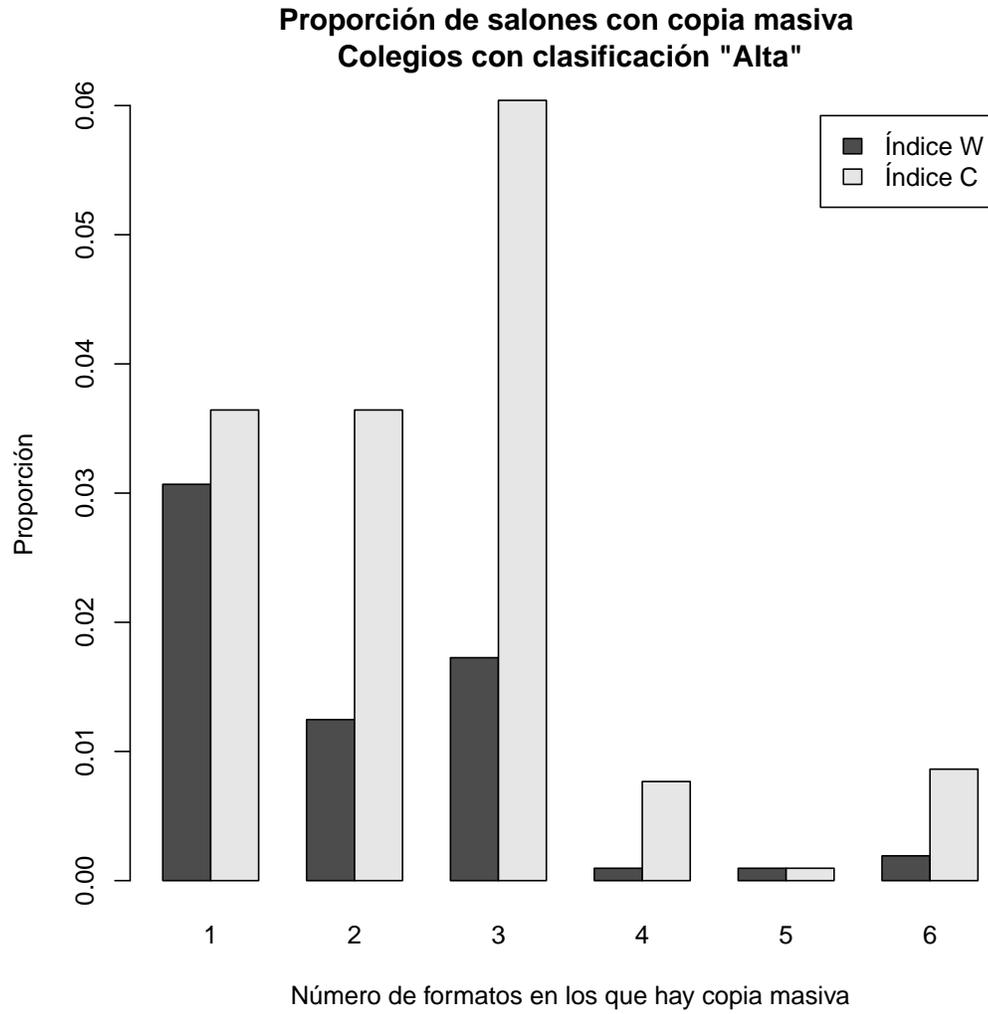


Figura 44: Comparación copia masiva por índice para colegios con clasificación "alta"

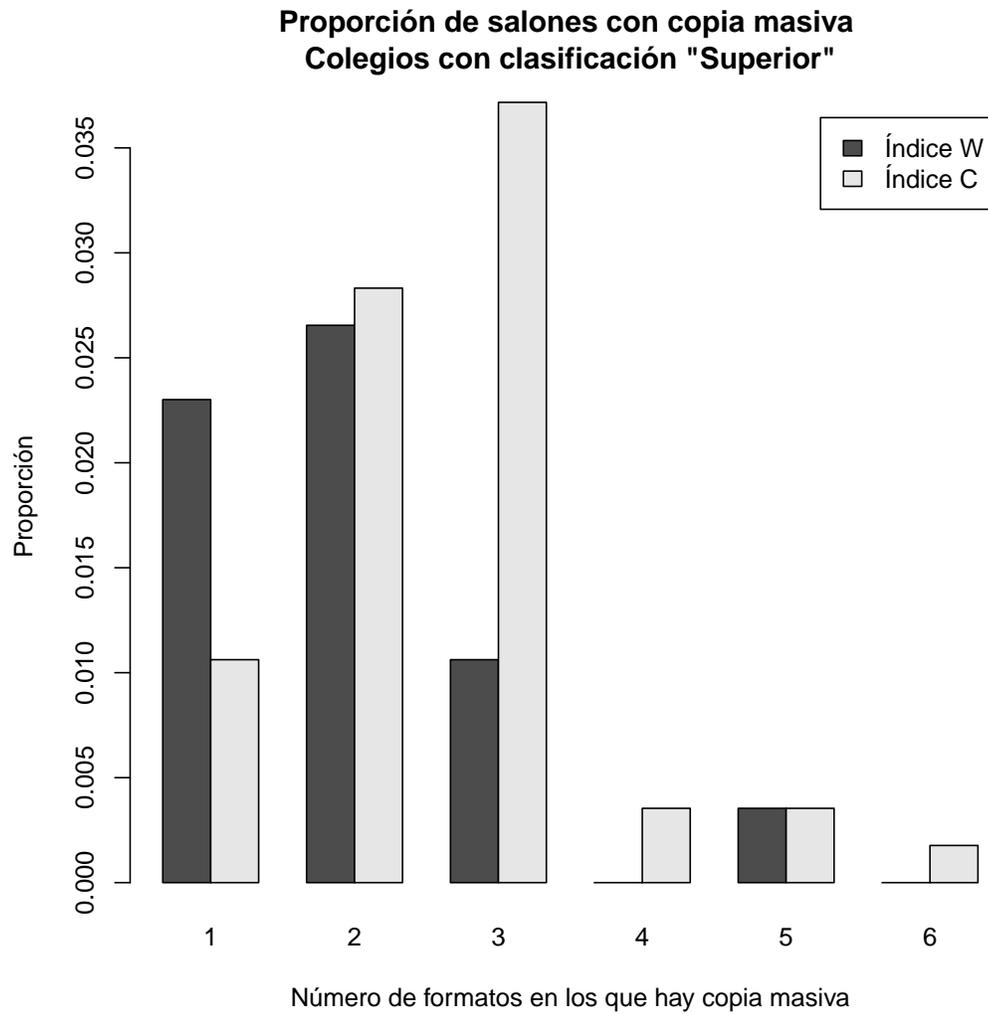


Figura 45: Comparación copia masiva por índice para colegios con clasificación "superior"

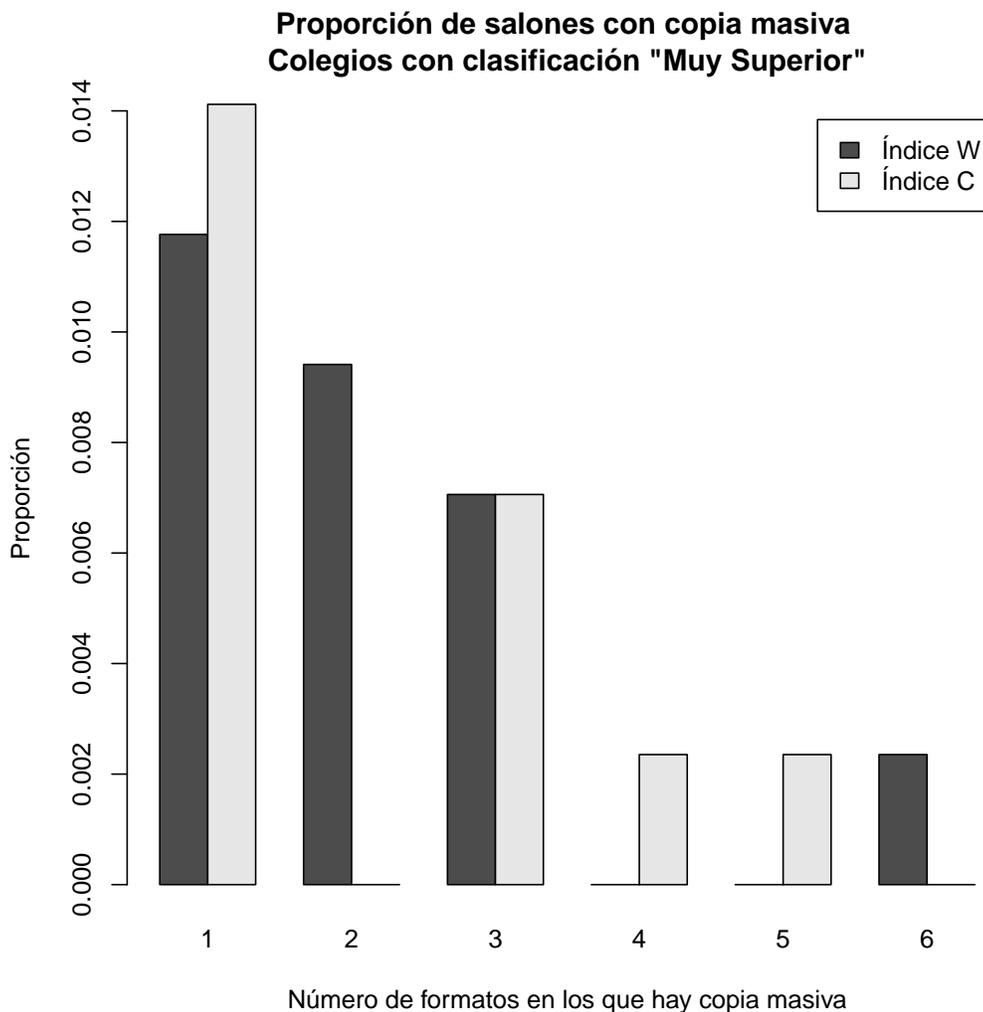


Figura 46: Comparación copia masiva por índice para colegios con clasificación “muy superior”

En general se puede ver que el índice  $\omega$  acusa a menos sede-jornadas de clasificaciones “alta”, “superior” y “muy superior” de copia masiva que el índice C. Este resultado es interesante, especialmente cuando se tiene en cuenta que al mirar la totalidad de sedes-jornada el índice  $\omega$  señala una proporción mayor de copia masiva (ver figura 38.).

## 8. Conclusiones

El presente documento describe la metodología utilizada para implementar el índice  $\omega$  a las pruebas SABER 5° y 9° realizadas por el ICFES en el 2009. Para validar la utilización de este índice se compara, usando muestreos independientes y simulaciones, con el índice  $\kappa$  y el índice utilizado por el ICFES en la actualidad.

En general la metodología utilizada por el ICFES presenta un error tipo I muy alto y una potencia relativamente baja. El índice  $\kappa$  a primera vista parece ser el índice más potente; sin embargo no preserva el error tipo I. El índice  $\omega$  es el único que preserva el error tipo I y no presenta grandes sacrificios en términos de potencia, lo que justifica su uso. En particular, tanto el índice  $\kappa$  como el índice  $\omega$  presentan avances importantes en comparación con el índice C utilizado por el ICFES en la actualidad.

Al analizar las pruebas se observa que la proporción de parejas sospechosas de copia excede el error tipo I en las tres formas de aplicación del examen. Dado que se presume que la aplicación controlada está libre de copia se cree que existe un problema de especificación del modelo de respuesta nominal. Estudiantes con educación similar pueden estar sesgados hacia alguna respuesta en particular. Dado que en la aplicación de las pruebas SABER 5° y 9° se evalúa en un mismo sitio a todos los estudiantes de un colegio, este efecto puede sesgar los resultados. En el modelo esto se vería reflejado en unos parámetros,  $\xi$  y  $\lambda$ , propios de cada sede-jornada; sin embargo, dado el pequeño tamaño de cada sede-jornada, sería imposible tener una estimación confiable de los mismos.

De cualquier manera, el índice  $\omega$  es un avance importante pues tiene un error tipo I menor que la metodología actual y una potencia mayor. El análisis de cada forma requiere entre 6 y 48 horas, en un computador con procesador Intel® Core 2 Duo 2.53 GHz y 4 gigas de memoria RAM. Con un nivel de confianza del 99.9% el porcentaje de individuos sospechosos de copia según el algoritmo es 36.68%. El porcentaje de parejas sospechosas de copia es 1.35%. Este valor corresponde a las parejas *ordenadas* que son detectadas como sospechosas. El porcentaje de salones en donde se sospecha hubo copia es del 57.91%. Finalmente, se detectaron las sede-jornada con copia masiva controlando el valor esperado de la tasa de falsos positivos mediante un procedimiento de tipo Bonferroni. Es importante anotar que aunque el índice  $\omega$  señala una proporción mayor de sedes-jornada con copia masiva que el índice C, cuando se analiza solamente el grupo de colegios de

alto desempeño señala una proporción menor.

Como comentario final, es importante tener en cuenta que es posible aprovechar más la información que se puede extraer del modelo de respuesta nominal como se mencionó en la sección 7.1. Se puede determinar qué preguntas son las más fáciles, cuales son las más discriminatorias y a qué nivel de habilidad. Esta información puede ser utilizada para apoyar el diseño de las pruebas del estado.

## 9. Bibliografía

- Angoff, W. (1974), 'The development of statistical indices for detecting cheaters', *Journal of the American Statistical Association* .
- Baird, L.S., J. (1980), 'Current trends in college cheating', *Psychology in the schools* .
- Belleza, F. & Belleza, S. (1989), 'Detection of cheating on multiple-choice test by using error-similarity analysis', *Teaching of Psychology* .
- 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.
- Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)*. **57**, 289-300.
- Cohen, J. (1960), 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement* pp. 37-46.
- Hambelton, R. K., Swaminatha, H. & Jane Rogers, H. (1991), *Fundamental of Item Response Theory*, SAGE Publications.
- Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.
- Holland, P. (1996), 'Assesing unusual agreement between the incorrect answers of two examinees using the k index: statistical theory and empirical support', *ETS technical report* .
- Jara, D., Riascos, A. & Romero, M. (2010), 'Detección de copia en pruebas del estado', *Documento CEDE* **15**.
- Lord, F. M. & Wingersky, M. S. (1984), 'Comparison of irt true-score and equipercentile observed-score equatings.', *Applied Psychological Measurement* **8**, 452-461.
- Sotaridona, L. & Meijer, R. (2001), 'Statistical properties of the k-index for detecting answer copying', *Research Report Educational Science and Technology* .

- Sotaridona, L. S. & Meijer, R. R. (2003), 'Two new statistics to detect answer copying', *Journal of Educational Measurement* **40**, 53–69.
- Sotaridona, L., van der Linden, W. & Meijer, R. (2006), 'Detecting answer copying using kappa statistic', *Applied Psychology Measurement* pp. 412–431.
- Stark, R. B. (2001), Statistical methods of detecting copying on multiple choice tests: a review and an application, Technical report, Integrative Statistics.
- Van der Ark, L. A., Emons, W. H. M. & Sijtsma, K. (2008), 'Detecting answer copying using alternate test forms and seat locations in small-scale examinations', *Journal of Educational Measurement* **45**, 991–17.
- Van der Linden, W. & Hambleton, R. (1997), *Handbook of Modern Response Theory*, Springer.
- Van der Linden, W. J. & Sotaridona, L. (2004), 'A statistical test for detecting answer copying on multiple-choice tests', *Journal of Educational Measurement* **41**, 361–377.
- Van der Linden, W. & Sotaridona, L. (2006), 'Detecting answer copying when the regular response process follows a known response model', *Journal of Educational and Behavioral Statistics* .
- Wollack, J. A. (2003), 'Comparison of answer copying indices with real data', *Journal of Educational Measurement* **40**, 189–205.
- Wollack, J. A. (2006), 'Simultaneous use of multiple answer copying indexes to improve detection rates', *Applied Measurement in Education* **19**, 265 – 288.

## A. Resultados condicionando por habilidad

Para todas las pruebas se utilizó un nivel de confianza de 99.9% y se calculó el poder de los estadísticos utilizando niveles de copia de: 1, 5, 10, 15, 20,  $\dots$ ,  $N$ , donde  $N$  es el número de preguntas que tiene el examen. Los resultados que se presentan a continuación son condicionados a un nivel de habilidad particular tanto para la fuente como para el individuo que se copia.

### A.1. Error Tipo I

Se presentan los resultados de ambos índices en la misma tabla por brevedad. El primero reportado, de izquierda a derecha, es el índice  $\omega$ , seguido del índice  $\kappa$  y finalmente se encuentra el índice  $C$ . En general se ve que el índice  $\kappa$  no preserva el error tipo I, especialmente cuando se comparan individuos con habilidades similares. Esto corrobora los hallazgos de Sotaridona et al. (2006). El índice  $C$  es el peor librado, acusando hasta el 50% de inocentes de copia cuando los demás individuos de salón presentan una habilidad similar a aquélla del individuo que se copia.

Tabla 8: Matemáticas-Mayo-5°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(0.5;55.5;111)	(0;28.5;116)	(0;5.5;106)	(0;0;95)	(0;0;109.5)	(0;0;117.5)	(0;0;102.5)	(0;0;102.5)
	(-1.5,-1)	(0.5;27;651)	(1;34.5;544)	(0;22;536)	(0;5.5;542)	(0;0;533.5)	(0;0;544)	(0;0;543.5)	(0;0;538)
	(-1,-0.5)	(1.5;4.5;544.5)	(1;22.5;387.5)	(0;34;344.5)	(1;26;348.5)	(0;8.5;349.5)	(0;0;358)	(0;0;348)	(0;0;359)
	(-0.5,0)	(0.5;0;548.5)	(1.5;6.5;335.5)	(2.5;21;270.5)	(0.5;42;228.5)	(0.5;25;247.5)	(0.5;5.5;247.5)	(0;0.5;244)	(0;0;248.5)
	(0,0.5)	(0.5;0;389.5)	(1;0;263.5)	(1;2;217.5)	(1.5;21;190.5)	(1.5;42.5;197)	(1;21.5;195)	(0.5;4;190)	(0;0;188.5)
	(0.5,1)	(0.5;0;188.5)	(0.5;0;173)	(2;1;155)	(1.5;6.5;159.5)	(1;16.5;162)	(1.5;25.5;157.5)	(0.5;15;178)	(0.5;6;152)
	(1,1.5)	(0.5;0;65.5)	(0;0;71.5)	(2;0;73.5)	(1;0;61)	(2;2;64)	(2;14;59)	(2;16;66)	(1;12.5;64)
	(1.5,2)	(0;0;1)	(0.5;0;0.5)	(1;0;0.5)	(3;0;1.5)	(1.5;0.5;0)	(2;5.5;0)	(3.5;10;2)	(2;17;1)

18

Tabla 9: Matemáticas-October-5°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(0;32;43.5)	(0.5;21.5;40.5)	(0;2.5;39)	(0;1;35.5)	(0;0;35)	(0;0;31.5)	(0;0;38)	(0;0;36.5)
	(-1.5,-1)	(0.5;21;223)	(0;30;184)	(1;18;169.5)	(0;7;141.5)	(0;1;160.5)	(0;0;142)	(0;0;135)	(0;0;135.5)
	(-1,-0.5)	(1.5;3;237)	(1.5;18.5;171.5)	(1;30.5;164.5)	(0.5;22;141.5)	(0;5.5;127)	(0;0;128.5)	(0;0;143)	(0;0;134)
	(-0.5,0)	(1;1;243)	(1.5;5.5;165.5)	(2;15.5;136.5)	(1;26.5;114.5)	(0.5;14.5;118.5)	(0;6;119.5)	(0;1;115)	(0;0;123.5)
	(0,0.5)	(1;0;179.5)	(1;0.5;140.5)	(1;5;109)	(3;18;100)	(4;14.5;116)	(0;10.5;97)	(0;2;106)	(0.5;1;112)
	(0.5,1)	(0;0;118.5)	(1;0;98.5)	(1;1;97.5)	(2.5;4.5;85.5)	(2.5;12;88.5)	(3;12.5;100.5)	(1;4.5;82)	(1.5;0;76.5)
	(1,1.5)	(0;0;39.5)	(0.5;0;38)	(2;0;30.5)	(2;0;25.5)	(2;2.5;35)	(3.5;6.5;31)	(3.5;5;33)	(2.5;2;32)
	(1.5,2)	(1;0;0)	(1;0;0)	(1;0;0)	(1.5;0;0.5)	(2;0;0)	(1.5;0.5;0.5)	(1.5;5;0.5)	(4.5;7;0)

Tabla 10: Lenguaje-Mayo-5°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(0.5;27.5;167.5)	(0;15.5;157.5)	(0;5;167)	(0;0;173.5)	(0;0;165.5)	(0;0;159.5)	(0;0;153)	(0;0;166)
	(-1.5,-1)	(1.5;17.5;362.5)	(1;20.5;339.5)	(1;13;344.5)	(0;2.5;351)	(0;0;327)	(0;0;346)	(0;0;321.5)	(0;0;333)
	(-1,-0.5)	(0.5;2.5;327.5)	(2;14.5;268.5)	(1.5;14.5;252)	(0.5;8;263)	(0;1.5;245)	(0;0.5;243.5)	(0;0;256.5)	(0;0;241)
	(-0.5,0)	(1;0;274)	(1;1.5;202)	(1;7.5;168)	(1;10;157.5)	(0;11.5;160)	(0;4.5;144)	(0;0.5;135)	(0;0;134.5)
	(0,0.5)	(0;0;178.5)	(0;0;143)	(1.5;2;136.5)	(1;6;117.5)	(0.5;10.5;121)	(1;8.5;120)	(0.5;2.5;121.5)	(0;0.5;130.5)
	(0.5,1)	(0;0;100)	(0;0;84.5)	(2.5;0;84)	(1.5;2;94)	(1.5;9.5;89.5)	(2;5;90.5)	(0.5;5.5;110)	(1;1.5;82)
	(1,1.5)	(0;0;35.5)	(0;0;24)	(0;0;27)	(3;0;26)	(2;2;32)	(2;4.5;33)	(2;6.5;33)	(2.5;3.5;32.5)
	(1.5,2)	(0.5;0;0)	(0.5;0;0)	(0;0;0)	(1;0;0)	(1.5;0;0)	(2.5;1.5;0)	(2;3;0)	(2;3;0)

82

Tabla 11: Lenguaje-October-5°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(1.5;29.5;26.5)	(0.5;13;21.5)	(0;6;20)	(0;0.5;24)	(0;0;23)	(0;0;16.5)	(0;0;17.5)	(0;0;15)
	(-1.5,-1)	(2;15;169.5)	(0.5;15.5;140.5)	(0.5;14;114)	(0.5;1.5;113.5)	(0;1;121.5)	(0;0;106.5)	(0;0;111.5)	(0;0;105.5)
	(-1,-0.5)	(0.5;5.5;207.5)	(1.5;11;163)	(1;10.5;157)	(0;8.5;143.5)	(0.5;3;142.5)	(0;0;119)	(0;0;139.5)	(0;0;132.5)
	(-0.5,0)	(1.5;1.5;183)	(1.5;2;149.5)	(1;10;137.5)	(1;10;130)	(0.5;8.5;122)	(0.5;1.5;108.5)	(0.5;0.5;114)	(0;0;121)
	(0,0.5)	(0.5;0;167.5)	(0;0;118.5)	(2;5.5;129.5)	(2;9;115.5)	(0.5;5;118.5)	(1.5;5.5;110)	(1;2;109)	(0;0.5;112)
	(0.5,1)	(0;0;113)	(0.5;0;103)	(1;0.5;104.5)	(2;1;99.5)	(4;4;94.5)	(2.5;5;97)	(2;2.5;103)	(2.5;0.5;95)
	(1,1.5)	(1;0;56)	(0;0;54)	(2;0;52)	(0.5;0;50.5)	(4;1.5;44.5)	(4;1;40)	(2.5;3;43)	(0.5;3;37.5)
	(1.5,2)	(0;0;3)	(0.5;0;6)	(0.5;0;5.5)	(2.5;0;6)	(1;0;7)	(3;1;4.5)	(1.5;2;7.5)	(5;1.5;4.5)

Tabla 12: Ciencias-Mayo-5°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(1.5;50.5;113)	(0;23.5;109)	(0;11.5;120)	(0;1;111)	(0;0;105.5)	(0;0;113)	(0;0;103)	(0;0;119.5)
	(-1.5,-1)	(2;28.5;278.5)	(0.5;31;266.5)	(0;18;289.5)	(0;7.5;268)	(0;1;276)	(0;0;278.5)	(0;0;271.5)	(0;0;261.5)
	(-1,-0.5)	(0;8.5;351.5)	(1;17;250.5)	(0.5;23;254.5)	(1;10;246.5)	(0;5;249.5)	(0;0.5;245)	(0;0;246)	(0;0;256)
	(-0.5,0)	(1;0.5;361.5)	(1;6.5;260.5)	(1;21.5;230)	(1.5;16.5;212.5)	(0;12;225.5)	(0;2;212.5)	(0;0;200)	(0;0;224)
	(0,0.5)	(0.5;0.5;279.5)	(1.5;1.5;206)	(0.5;4;198.5)	(4;10.5;165)	(2;13;175)	(0.5;5.5;159)	(0;1.5;176.5)	(0;0;180)
	(0.5,1)	(1;0;129)	(1;0;132)	(2;0.5;114.5)	(1.5;2.5;122.5)	(1;11;136)	(4;6;139.5)	(1;6.5;134.5)	(1;1.5;134)
	(1,1.5)	(0;0;55)	(0.5;0;49.5)	(0;0;53.5)	(2.5;0.5;54)	(1;2.5;56.5)	(0.5;6.5;62.5)	(0;6;60.5)	(1.5;3;55)
	(1.5,2)	(0;0;4)	(0;0;6.5)	(0.5;0;5)	(2.5;0;6)	(1;0;5.5)	(0.5;1.5;5)	(3.5;5;5.5)	(2.5;4.5;5)

83

Tabla 13: Ciencias-October-5°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(1.5;23.5;20)	(0;18.5;11.5)	(0;4.5;11)	(0;1;15.5)	(0;1;11)	(0;0;9.5)	(0;0;14.5)	(0;0;15)
	(-1.5,-1)	(2;18;168)	(0.5;17;145)	(0.5;12.5;117.5)	(0;2.5;126.5)	(0.5;0;133)	(0;0;125)	(0;0;123.5)	(0;0;123)
	(-1,-0.5)	(2.5;4;218)	(0.5;13;171.5)	(0;13.5;155.5)	(0.5;9;150.5)	(0.5;3;145.5)	(0.5;0;149)	(0;0;154)	(0;0;146.5)
	(-0.5,0)	(2.5;1;222)	(3;3;160.5)	(1.5;6.5;154)	(1.5;12.5;154)	(0.5;8.5;141.5)	(0;1;138)	(0.5;0;129.5)	(0;0;129.5)
	(0,0.5)	(0.5;0;163)	(3.5;1;126)	(0.5;2.5;131.5)	(3;5;118)	(0.5;7;112.5)	(1;1.5;106.5)	(0.5;2.5;106.5)	(0.5;0;105)
	(0.5,1)	(0.5;0;108.5)	(0.5;0;99.5)	(0.5;0.5;81.5)	(2;2;95)	(2.5;1.5;89.5)	(2.5;4;92)	(1;3;73)	(0.5;0.5;81)
	(1,1.5)	(1.5;0;37)	(1.5;0;35.5)	(1.5;0;35)	(1.5;0.5;35.5)	(2.5;1.5;41.5)	(3.5;3.5;37)	(2.5;4.5;38.5)	(2.5;3.5;43.5)
	(1.5,2)	(0;0;1)	(2;0;0)	(1;0;1.5)	(1;0;3.5)	(2;0.5;2.5)	(1.5;2.5;0)	(3;4;3.5)	(7.5;3.5;1.5)

Tabla 14: Matemáticas-Mayo-9°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(0.5;52.5;8.5)	(1;45.5;9)	(0;12.5;8)	(0;3.5;12.5)	(0;0;7.5)	(0;0;7.5)	(0;0;9.5)	(0;0;10.5)
	(-1.5,-1)	(1;31.5;232.5)	(0.5;51.5;227.5)	(0.5;47.5;214.5)	(0.5;15.5;212.5)	(0.5;6;224.5)	(0;0.5;209.5)	(0;0;224)	(0;0;196)
	(-1,-0.5)	(2;19.5;465)	(1;50;400)	(0;66;363)	(1.5;67.5;361.5)	(0.5;22.5;363)	(0;3;386.5)	(0;0;374)	(0;0;358.5)
	(-0.5,0)	(2;3;486)	(1.5;16;390.5)	(2;60.5;356)	(2.5;79;357)	(0;54;363.5)	(0;9;333.5)	(0;3.5;341.5)	(0;0.5;346)
	(0,0.5)	(0.5;0;319.5)	(1;2;299.5)	(1;22;245)	(3;57;275)	(3.5;56.5;246.5)	(1.5;38.5;271.5)	(0;12.5;253)	(0.5;2;276.5)
	(0.5,1)	(0.5;0;175)	(0.5;0.5;183)	(2;2.5;179)	(1.5;13;183)	(2;33.5;191)	(1;33;179.5)	(2;19;194.5)	(1;4;184)
	(1,1.5)	(1.5;0;52)	(0;0;51.5)	(2.5;0;47)	(1;0.5;50)	(2.5;11;48)	(2.5;20.5;48.5)	(3.5;18;56)	(0.5;13;55.5)
	(1.5,2)	(0;0;0)	(1;0;0)	(1.5;0;0)	(1.5;0.5;0)	(0;4;0)	(2;13.5;0)	(1.5;17.5;0)	(2;21.5;0)

84

Tabla 15: Matemáticas-October-9°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(1;79.5;30.5)	(1;67.5;28)	(0;28;27.5)	(0;4.5;22.5)	(0;0.5;24.5)	(0;0;28)	(0;0;30.5)	(0;0;30)
	(-1.5,-1)	(3;62.5;136.5)	(2;91.5;131)	(0.5;67;124.5)	(0;31;123.5)	(0;8;117.5)	(0;0.5;124.5)	(0;0;129)	(0;0;128.5)
	(-1,-0.5)	(2;28.5;175)	(1.5;64.5;173.5)	(1;72.5;160)	(0;53.5;158)	(0.5;29.5;162)	(0;5.5;161)	(0;1;150)	(0;0.5;167)
	(-0.5,0)	(2;8.5;184)	(0.5;23.5;135)	(0.5;53;136.5)	(2;64.5;131)	(0.5;34.5;135.5)	(0;8.5;141.5)	(0;2;127.5)	(0.5;0.5;144)
	(0,0.5)	(1;1.5;124.5)	(0.5;8;121)	(4;20;117.5)	(3;28;128.5)	(1;38.5;119.5)	(0;25.5;115.5)	(0;7;123)	(0.5;0.5;124)
	(0.5,1)	(0.5;0;94.5)	(1.5;1.5;92.5)	(2.5;4.5;100)	(3.5;13.5;95)	(4;15.5;96.5)	(1.5;13;80)	(1;12;100.5)	(1;4.5;105)
	(1,1.5)	(1;0;23)	(0;0;29)	(1;0;32)	(2;0.5;28)	(1.5;6;27.5)	(1.5;8;34.5)	(4;7;29)	(1;5;35.5)
	(1.5,2)	(0.5;0;1.5)	(0;0;1)	(1;0;5.5)	(3;0;3.5)	(2;1;3.5)	(2.5;4.5;3)	(3.5;4;3)	(1;6.5;1.5)

Tabla 16: Lenguaje-Mayo-9°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(0;68.5;73.5)	(0;35;75.5)	(0;17.5;62.5)	(0;2;54)	(0;0;59)	(0;0;64)	(0;0;66.5)	(0;0;65.5)
	(-1.5,-1)	(1.5;40.5;198.5)	(0;52;148)	(0;44.5;137)	(0;12.5;133)	(0;1;133)	(0;0;139)	(0;0;140.5)	(0;0;140)
	(-1,-0.5)	(1;13;378.5)	(0.5;39.5;186.5)	(1.5;50;127.5)	(1;30;146)	(0.5;13;152.5)	(0;0.5;162.5)	(0;0;166.5)	(0;0;146.5)
	(-0.5,0)	(1;3;519.5)	(1;10;305.5)	(1.5;28;152)	(0;35.5;124)	(0.5;24.5;128)	(0;6;132)	(0;1;132.5)	(0;0;137)
	(0,0.5)	(0.5;0;373)	(1.5;1.5;245)	(1;9;177)	(1;21;136.5)	(0.5;26;125.5)	(0;16;127.5)	(0;3.5;131)	(0;0.5;131.5)
	(0.5,1)	(0;0;162.5)	(1;0;137)	(1.5;1;99.5)	(1.5;5;101)	(1.5;20.5;104)	(0.5;18.5;91)	(0.5;9.5;94.5)	(0.5;2;86.5)
	(1,1.5)	(0.5;0;35)	(0.5;0;25)	(1;0;28.5)	(1.5;1.5;21.5)	(0.5;6.5;23)	(1.5;8.5;22.5)	(3;13;24.5)	(1;4.5;22)
	(1.5,2)	(0.5;0;0)	(1.5;0;0)	(1;0;0)	(0.5;0;0)	(2;0;0)	(3.5;3;0)	(1.5;6.5;0)	(2;7.5;0)

85

Tabla 17: Lenguaje-October-9°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(1;48.5;49.5)	(1;26.5;41.5)	(0;11.5;48.5)	(0;3.5;43)	(0;0.5;47.5)	(0;0;45.5)	(0;0;39.5)	(0;0;46.5)
	(-1.5,-1)	(0.5;32.5;144.5)	(0;33;106.5)	(0.5;26.5;96)	(0.5;9;105.5)	(0;1.5;102.5)	(0;0;105.5)	(0;0;95.5)	(0;0;107.5)
	(-1,-0.5)	(1;12;272)	(1.5;23.5;143)	(0;22;122)	(1.5;14;126.5)	(0;6;127.5)	(0;2;120.5)	(0;0;140)	(0;0;106.5)
	(-0.5,0)	(1;2.5;353)	(0.5;4.5;195)	(1.5;15.5;122.5)	(0;18;120.5)	(1;11.5;130.5)	(0;2;120.5)	(0.5;1;117.5)	(0;0;127.5)
	(0,0.5)	(0;0.5;246)	(1.5;1.5;170)	(1;7;121)	(1;8;127)	(1.5;12;110.5)	(0;8.5;115)	(0;2;108.5)	(0;0;108)
	(0.5,1)	(0.5;0;122)	(1.5;0;90.5)	(1.5;1.5;88.5)	(2.5;5.5;93.5)	(2;7;105)	(1;8.5;96)	(3;7;91.5)	(1;1;88)
	(1,1.5)	(0;0;30)	(1;0;37.5)	(2;0;27.5)	(3.5;1.5;30.5)	(0.5;5;32.5)	(1.5;7;28)	(2;3.5;30.5)	(4.5;5;33.5)
	(1.5,2)	(0;0;1)	(0.5;0;0)	(1.5;0;0.5)	(1.5;0;0)	(3;1;0.5)	(2.5;3.5;0.5)	(2.5;4.5;0)	(4;3.5;0)

Tabla 18: Ciencias-Mayo-9°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(0;37.5;44)	(0;36;43)	(0;12;32.5)	(0;0.5;40.5)	(0;0;34.5)	(0;0;36)	(0;0;35)	(0;0;37)
	(-1.5,-1)	(0.5;33;214)	(0;38;210)	(0.5;29.5;196.5)	(0;8;192.5)	(0;1.5;190)	(0;0;189)	(0;0;196.5)	(0;0;193.5)
	(-1,-0.5)	(0.5;15.5;376)	(1;27.5;299.5)	(1.5;34.5;300.5)	(1.5;22;278)	(0.5;6.5;299.5)	(0;0.5;297.5)	(0;0;282.5)	(0;0;269)
	(-0.5,0)	(0.5;2;445.5)	(0;8.5;324)	(1;29.5;283)	(1.5;25.5;289.5)	(1;16.5;295)	(1;3;258.5)	(1.5;2;274.5)	(0;0;275.5)
	(0,0.5)	(1;0.5;286)	(1.5;0.5;242.5)	(2.5;7;238)	(2.5;18;197)	(1;16.5;215.5)	(0;9.5;214.5)	(0;3.5;207)	(0;1;208.5)
	(0.5,1)	(0;0;166.5)	(0.5;0;148)	(2;3;144.5)	(0.5;6;148)	(1;11;143.5)	(2.5;14;154.5)	(1.5;4;154.5)	(0.5;1.5;154)
	(1,1.5)	(0.5;0;49)	(2;0;45.5)	(1;0;49.5)	(2;1.5;46)	(3.5;5;50)	(1;6;50)	(0;3.5;46.5)	(2;3.5;46)
	(1.5,2)	(0;0;0)	(1;0;0)	(2.5;0;0)	(2.5;0;0)	(1;1.5;0)	(3.5;2;0)	(2.5;2.5;0)	(1.5;4;0)

86

Tabla 19: Ciencias-October-9°. Inocentes acusados de copia (por cada mil personas).

		$\theta_c$							
		(-2,-1.5)	(-1.5,-1)	(-1,-0.5)	(-0.5,0)	(0,0.5)	(0.5,1)	(1,1.5)	(1.5,2)
$\theta_s$	(-2,-1.5)	(0.5;27;28.5)	(0;19;22.5)	(0.5;5.5;20.5)	(0;3;28)	(0.5;0.5;26)	(0;0;23)	(0;0;25.5)	(0;0;22.5)
	(-1.5,-1)	(1.5;20.5;138.5)	(1;21;131.5)	(0;14;122.5)	(0;7;118)	(0;2;115.5)	(0.5;0.5;114)	(0.5;0;111)	(0;0;106)
	(-1,-0.5)	(2;10;236.5)	(1.5;12.5;166.5)	(1;20;165)	(1;14;156)	(0;2;150.5)	(0;0.5;153)	(0;0;150.5)	(0;0.5;151.5)
	(-0.5,0)	(1.5;2;285)	(2.5;7.5;191)	(0.5;9;152)	(1.5;13.5;160.5)	(1.5;9;151.5)	(0;4;154.5)	(0;0;142.5)	(1;0;149)
	(0,0.5)	(0.5;0.5;171)	(2;0.5;142)	(1.5;5.5;124)	(4;14;132)	(2.5;12;128.5)	(1;4.5;137)	(0;2.5;127.5)	(0.5;0.5;133.5)
	(0.5,1)	(0.5;0;95)	(0.5;0;89)	(1;1.5;86.5)	(1.5;3;84.5)	(1;5.5;99.5)	(2;10;82)	(0;3.5;82)	(0.5;2;86.5)
	(1,1.5)	(0.5;0;36.5)	(0;0;38.5)	(1;0;38.5)	(2;1.5;33)	(1.5;0.5;30)	(3;4;41)	(0.5;6;35)	(1.5;4.5;39)
	(1.5,2)	(0.5;0;2)	(0.5;0;1)	(1.5;0;1.5)	(1;0;3)	(2;1;1.5)	(2;1.5;4)	(1;2.5;5)	(2.5;3.5;1.5)

## A.2. Potencia

Se presentan 9 resultados por forma, para cuando la habilidad de la fuente o la persona que copia se encuentra en los intervalos  $(-2,-1.5)$ ,  $(-0.5,0)$ ,  $(1.5,2)$ , sin embargo dado que se tomaron 8 intervalos de habilidad, como se muestra en las tablas de arriba, en realidad existen 64 gráficas por forma.

De estas gráficas es posible extraer una conclusión. El índice  $\omega$  parece ser más poderoso que el índice  $\kappa$  exceptuando los casos en que la habilidad de ambos individuos es similar, pero es precisamente en estos casos que el error tipo I es grande. Parece que la mayor potencia del índice  $\kappa$  no se debe a que éste identifique a aquellos individuos que copian con una probabilidad mayor que el índice  $\omega$ , sino a que identifica a más individuos como sospechosos de copia, sin importar si copiaron o no.

El índice C presenta una potencia alta en algunos casos; sin embargo en otros casos la potencia es casi nula y en ocasiones presenta un comportamiento errático. Este comportamiento extraño se puede explicar por el ajuste final que se hace al calcular el índice. Éste reduce el valor de C cuando el individuo tiene un número de preguntas correctas diferente al promedio. Cuando un individuo copia de otra persona tendrá más preguntas correctas que el promedio si copio de alguien con habilidad alta, por el contrario tendrá menos cuando copia de un individuo con habilidad baja. El valor del índice solo se reduce cuando el número de preguntas correctas es muy diferente del promedio, lo cual se da en niveles de copia altos y cuando la habilidad de la fuente de copia es muy alta o muy baja y diferente a la del individuo que copia. Esto explica el comportamiento en “U” invertida.

Figura 47: Comparación de índices en la forma 5°-Mayo-Matemáticas

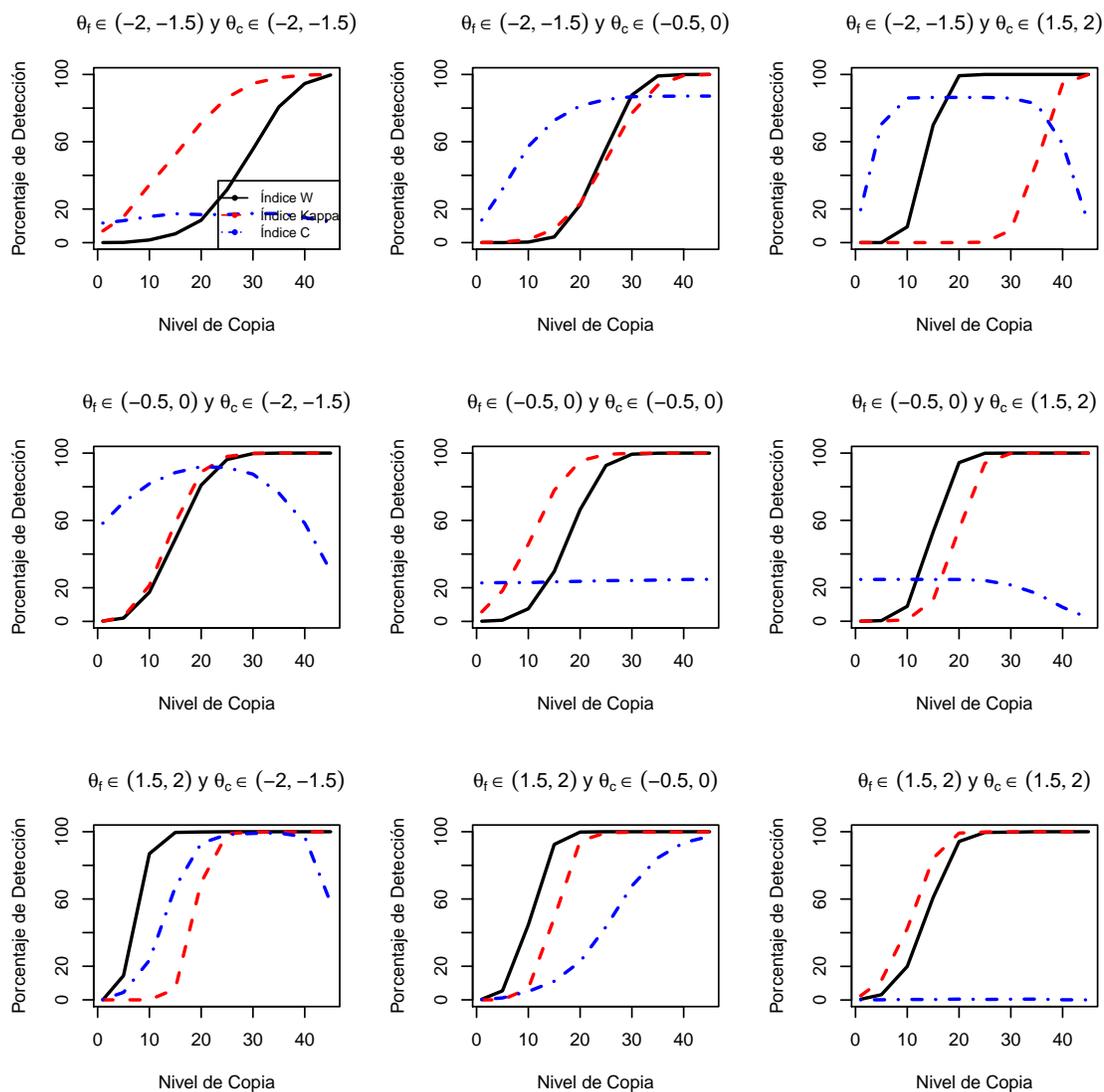


Figura 48: Comparación de índices en la forma 5°-Octubre-Matemáticas

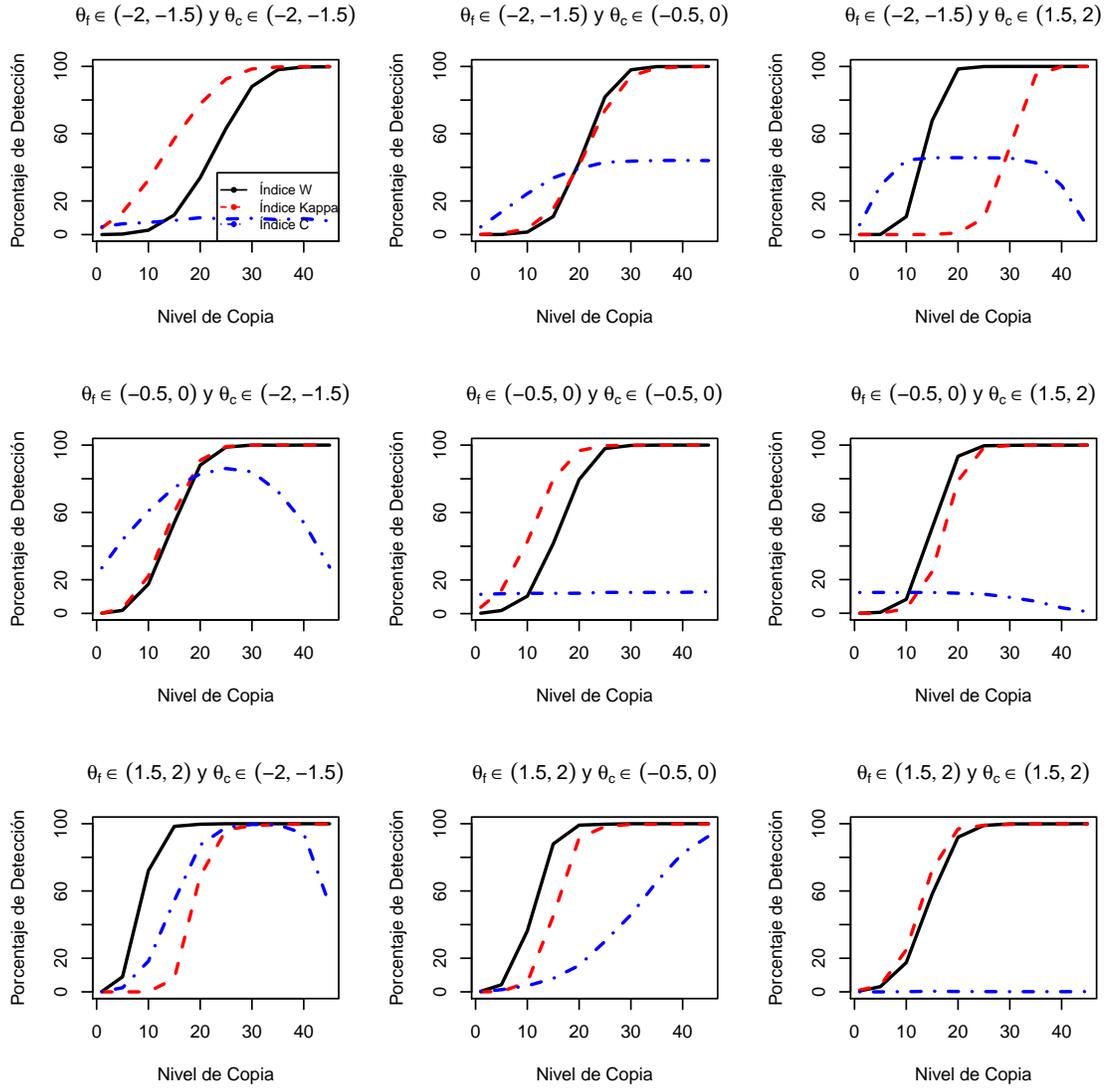


Figura 49: Comparación de índices en la forma 5°-Mayo-Lenguaje

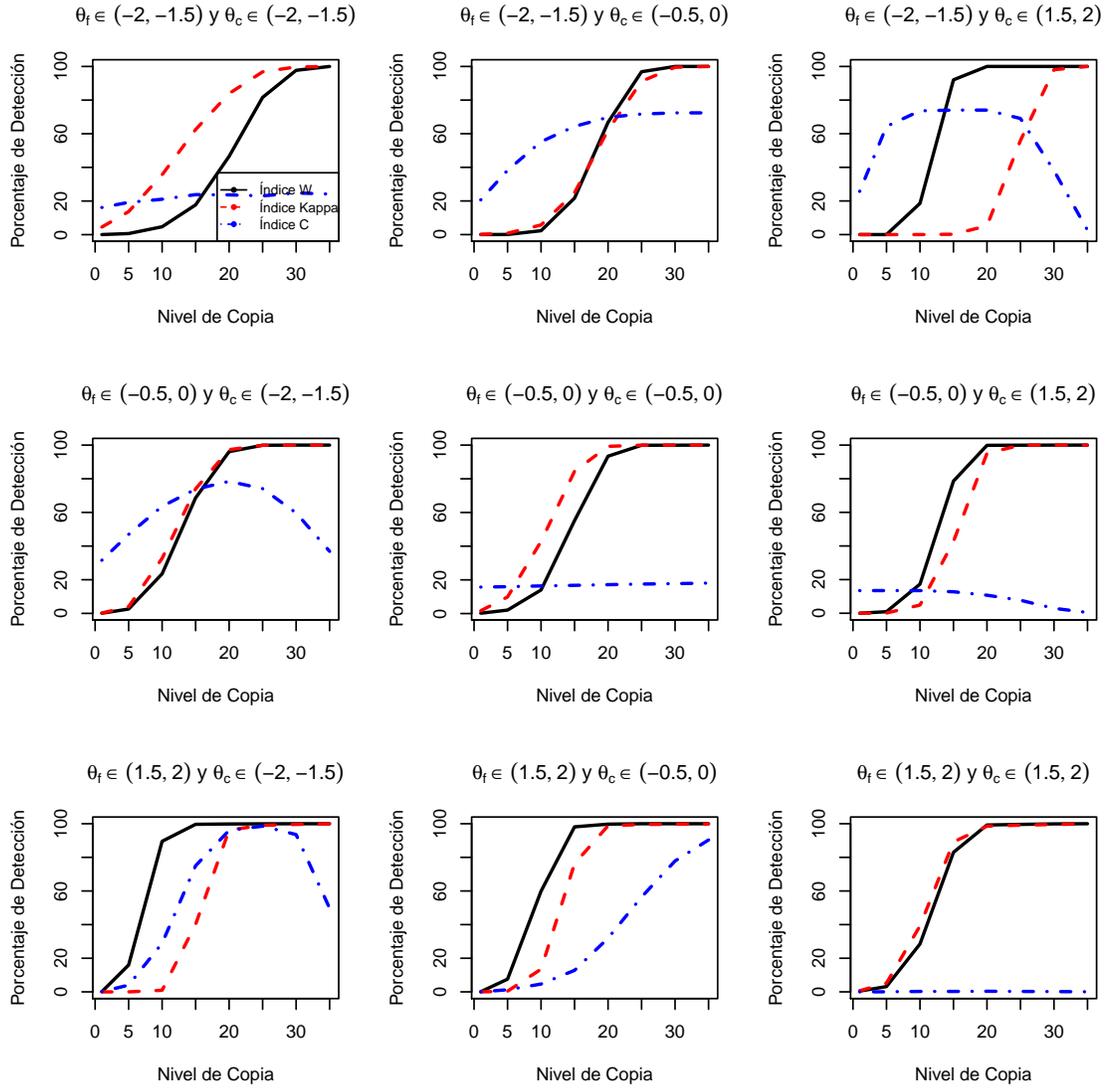


Figura 50: Comparación de índices en la forma 5°-Octubre-Lenguaje

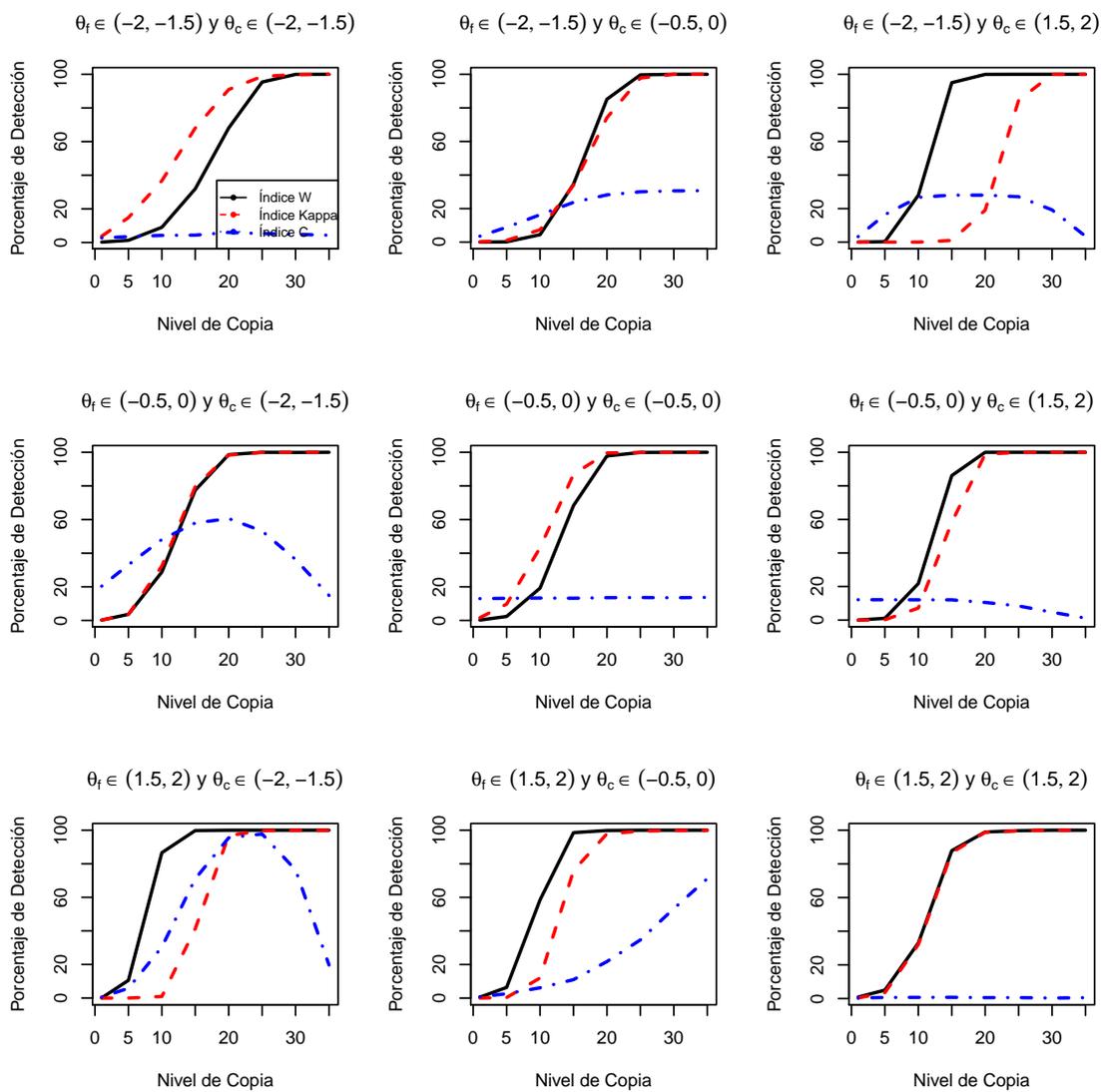


Figura 51: Comparación de índices en la forma 5º-Mayo-Ciencias

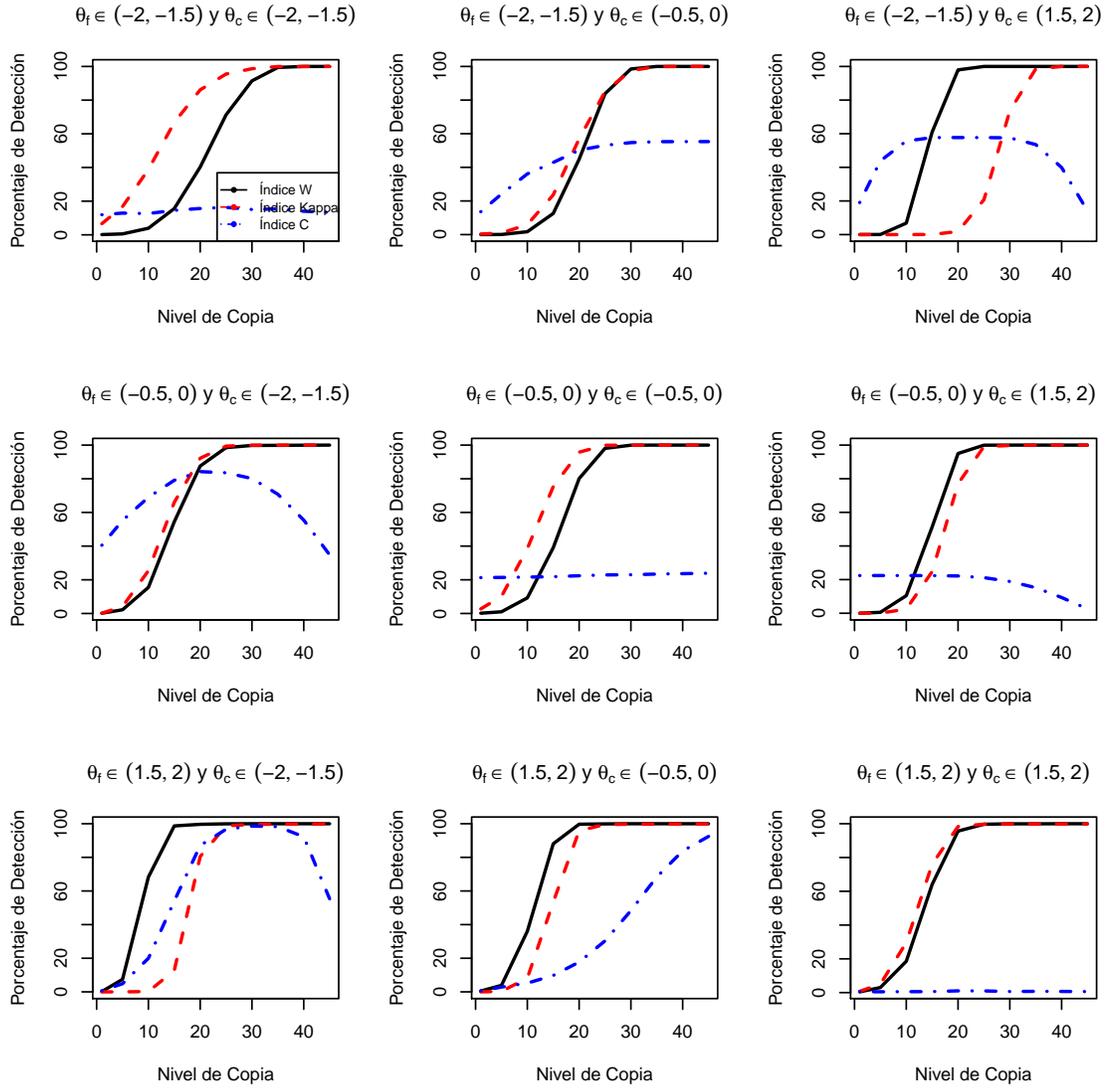


Figura 52: Comparación de índices en la forma 5<sup>o</sup>-Octubre-Ciencias

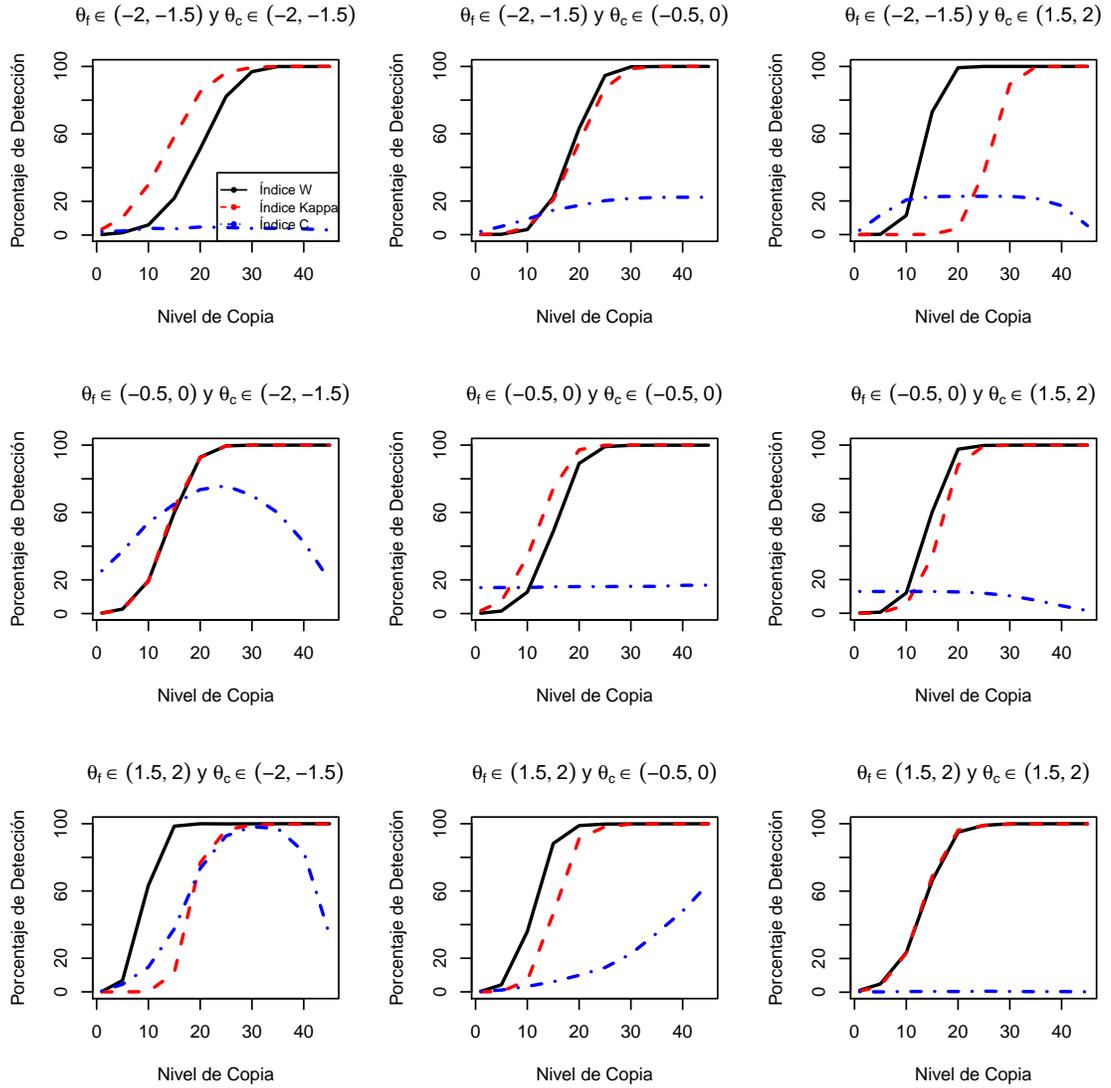


Figura 53: Comparación de índices en la forma 9º-Mayo-Matemáticas

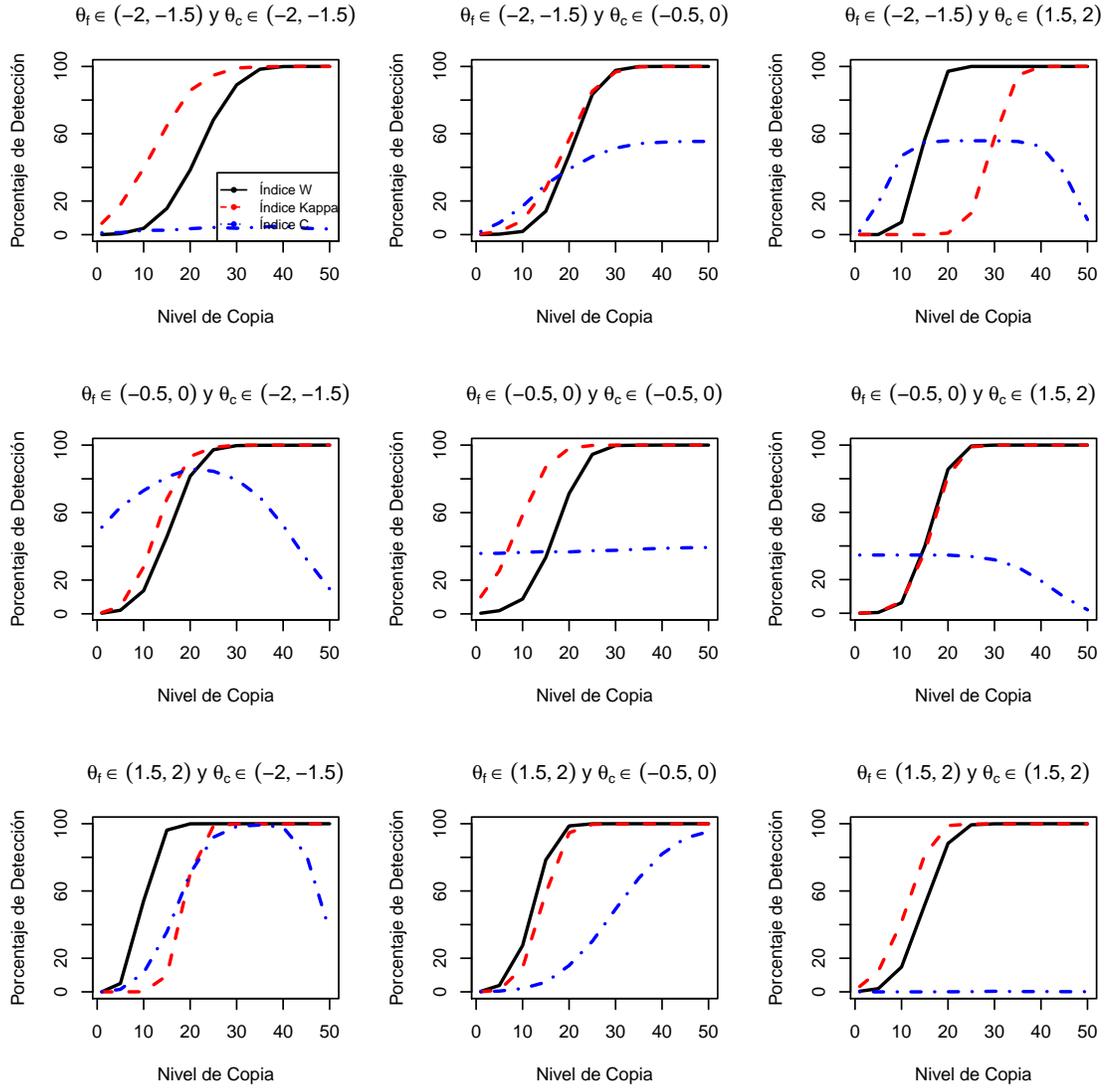


Figura 54: Comparación de índices en la forma 9°-Octubre-Matemáticas

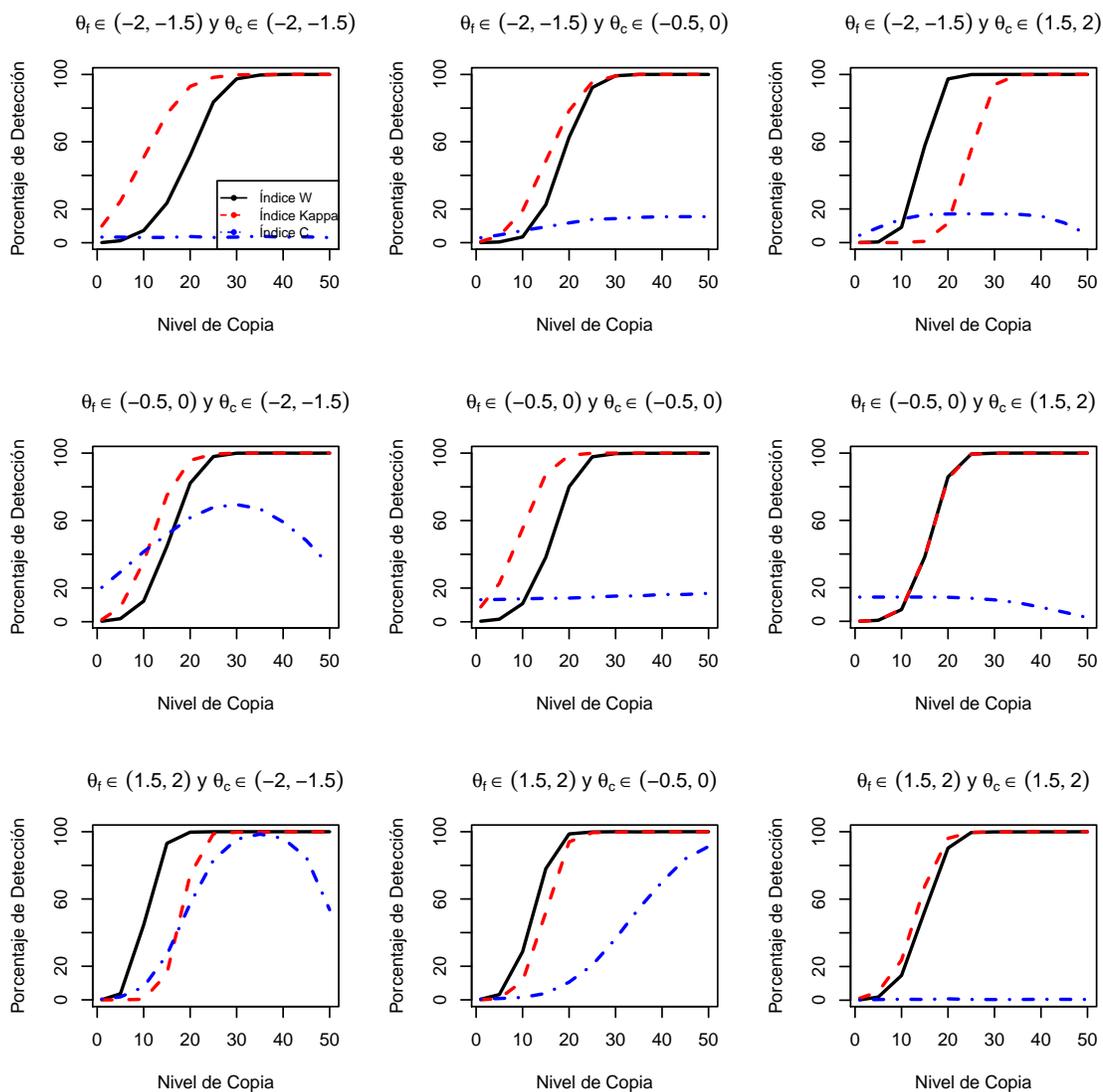


Figura 55: Comparación de índices en la forma 9°-Mayo-Lenguaje

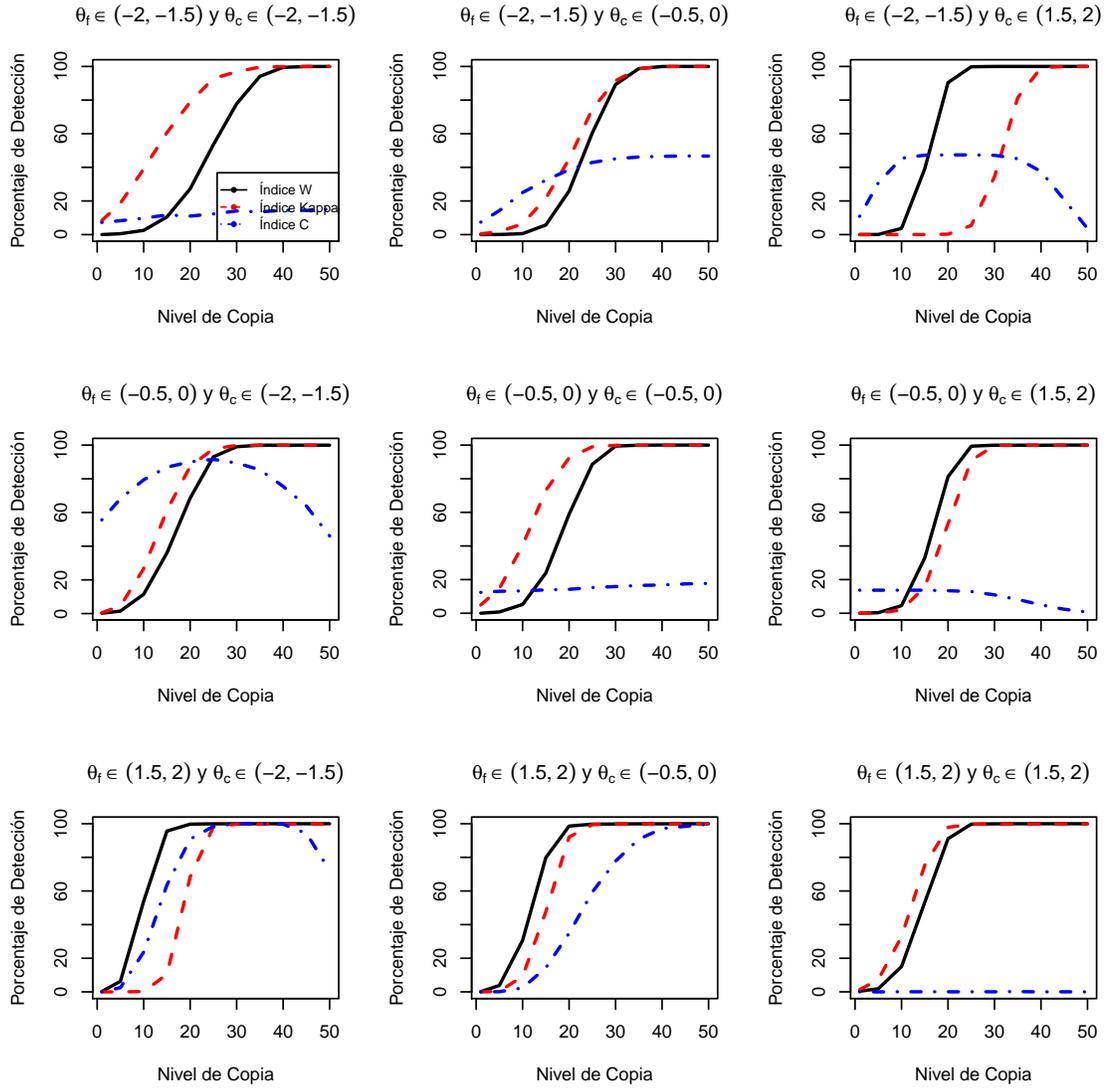


Figura 56: Comparación de índices en la forma 9°-Octubre-Lenguaje

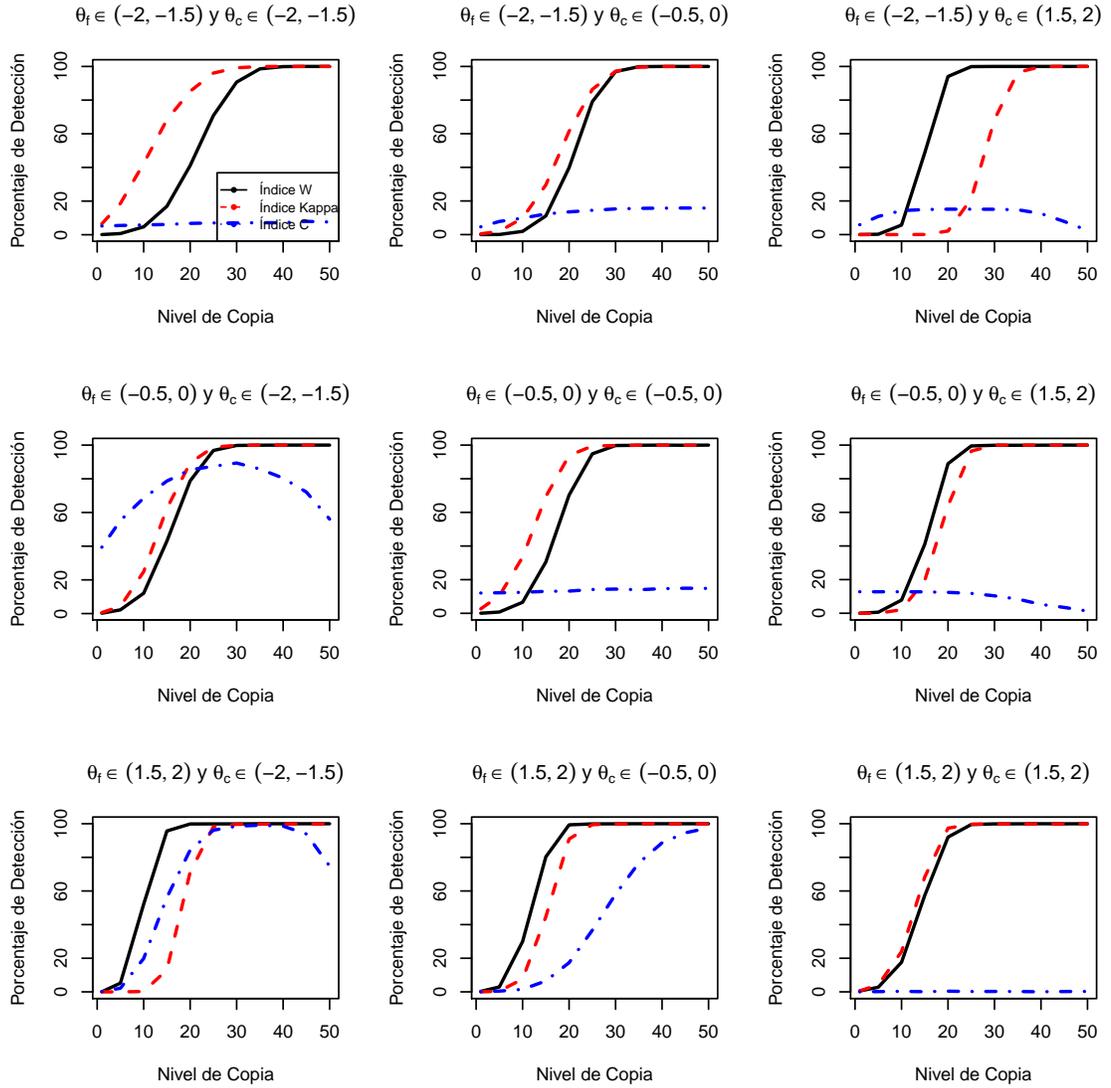


Figura 57: Comparación de índices en la forma 9º-Mayo-Ciencias

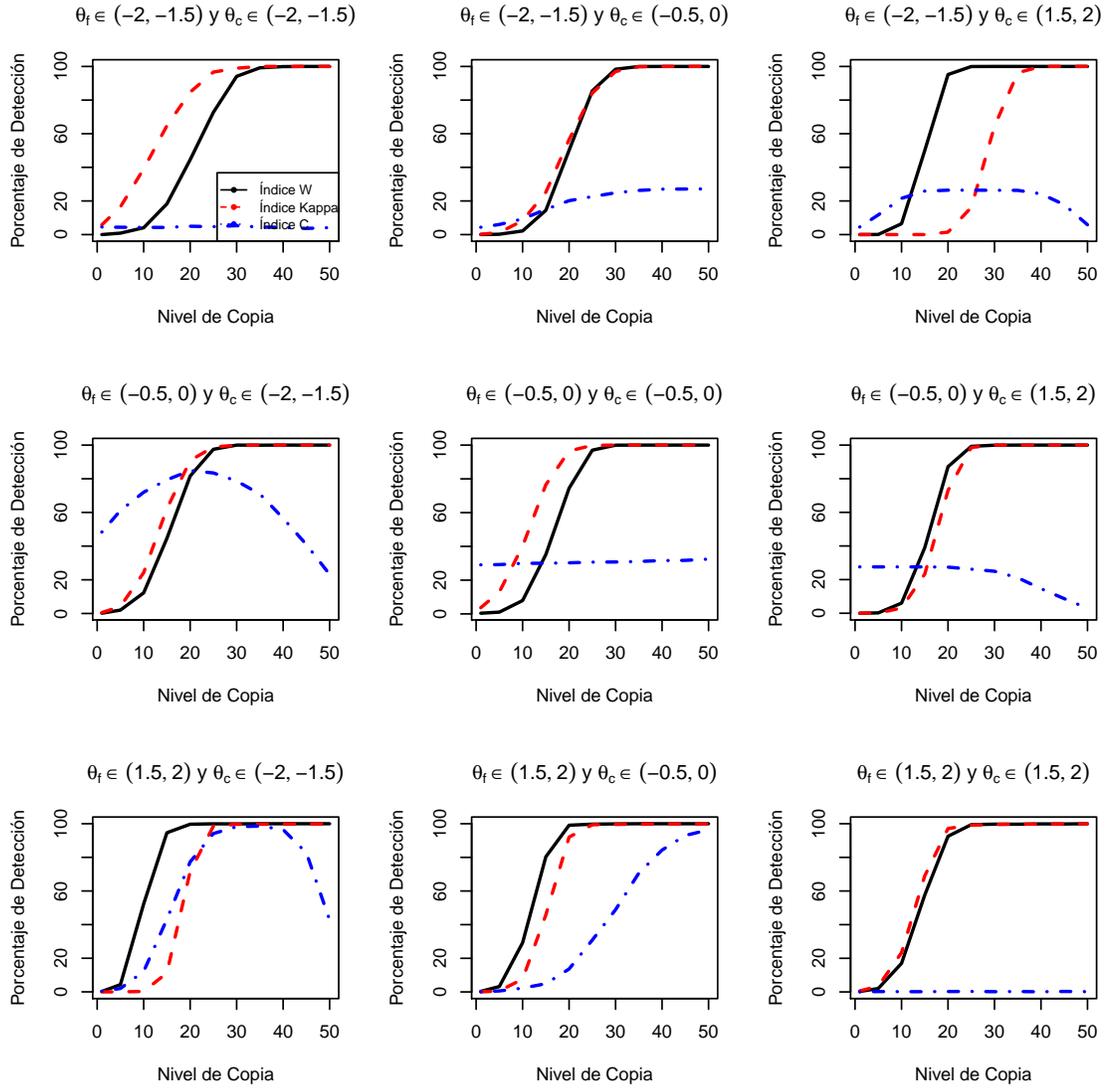


Figura 58: Comparación de índices en la forma 9º-October-Ciencias

