LA VALIDACIÓN DEL MÉTODO DE WESOLOWSKY, PARA LA DETECCIÓN DE ANOMALIAS:

UN ESTUDIO SIMULADO

GERARDO H. TERRAZAS GONZALEZ

HÉCTOR V. ROBLES VASQUEZ

DIANA YANIRA CAAMAL PAT

JAIME MARTINEZ SANCHES

ÍNDICE GENERAL

RESÚMEN

Este reporte tiene como objetivo validar la metodología que aplica el INEE en la detección de la excesiva similitud en los trenes de respuesta de los alumnos que presentan la prueba Planea, la cual usa el método propuesto por Wesolowsky (2000). Para dicho fin, se generaron 900,000 simulaciones bajo dos escenarios fundamentales: el primero (420 000 simulaciones) considera la independencia en las respuestas de los estudiantes, y la segunda (480 000) bajo un patrón de copia que se describe en el cuerpo del documento. Los resultados de las simulaciones, muestran evidencia favorable y confiable para la aplicación de este método en la detección de anomalías, y particularmente, muestran que el método de Wesolowsky no excede la significancia fijada, para acusar falsamente a estudiantes que atienden la prueba.

Palabras clave: Método de Wesolowsky, análisis estadístico para detección de irregularidades, anomalías en exámenes con opción múltiple, Simulaciones.

I. INTRODUCCIÓN

A partir de enero del 2015, el Instituto Nacional para la Evaluación de la Educación (INEE), diseño una nueva prueba de aprendizajes, denominada **Plan Nacional para las Evaluaciones de los Aprendizajes (PLANEA).** A través de ella se evaluarán, inicialmente, competencias disciplinares básicas de los campos de Lenguaje y Comunicación y Matemáticas y progresivamente se extenderá al esquema de evaluación otras áreas de aprendizaje como Ciencias y Formación Ciudadana. Planea se aplicó por primera vez en junio del 2015 en sexto de primaria, tercero de secundaria.

En su contenido, **Planea** está conformado por un conjunto de instrumentos para evaluar aprendizajes cognitivos y no cognitivos. Los primeros derivan del currículum nacional vigente de la educación básica y del marco curricular común. Los segundos están orientados a evaluar aspectos sociales y emocionales, así como valores y actitudes. Los instrumentos evaluativos de educación básica serán diseñados y elaborados por el INEE, además de que tendrá a su cargo la supervisión de la aplicación de todas las pruebas; asimismo, aplicará evaluaciones muestrales, en tanto que a la SEP le corresponderá la aplicación de las evaluaciones censales.

La principal preocupación de **Planea** es que las evaluaciones deben cumplir con los criterios técnicos que aseguren sus resultados sean útiles y confiables para los niveles de dominio que fueron diseñados. Particularmente debe comentarse que estas no están diseñadas para evaluar la calidad educativa de los planteles o del desempeño de sus docentes, ni tampoco deberá usarse para compensar a los docentes o escuelas, o castigar a los alumnos (tomando fundamentalmente las experiencias de ENALCE y Excale).

Los propósitos que se establecen para las pruebas **Planea** son:

- Conocer la medida en que los estudiantes logran el dominio de un conjunto de aprendizajes esenciales al término de los distintos niveles de la educación obligatoria.
- Ofrecer información contextualizada para la mejora de los procesos de enseñanza en los centros escolares.
- Informar a la sociedad sobre el estado que guarda la educación, en términos del logro de aprendizaje de los estudiantes.
- Aportar a las autoridades educativas información relevante y utilizable para el monitoreo, la planeación, programación y operación del sistema educativo y sus centros escolares.

Para evitar el fenómeno conocido como inflación de resultados de las pruebas censales, se tienen contemplados tres mecanismos de control, donde el INEE:

1. Definirá criterios específicos para la aplicación de las pruebas censales y uso de los resultados. Uno de estos criterios indica que los aplicadores de las pruebas deberán ser externos al centro educativo correspondiente.
2. Supervisará los procesos asociados a la aplicación de las pruebas.
3. Realizará una verificación estadística de la consistencia de los resultados.

El tercer punto se relaciona directamente con proponer metodologías estadísticas que ayuden a detectar anomalías en las pruebas, las cuales son el tema fundamental del presente escrito; Las anomalías deben considerarse a nivel de grupo escolar y a nivel de dominios de análisis que permitan ayudar a dar la certeza de la validez de los datos de logro de las pruebas.

Existen métodos estadísticos existentes para la detección de irregularidades, para comparar las respuestas de una prueba estandarizada desde los años 20´s como los desarrollados por Gundlanch (1925), Bird (1929), Dickenson (1945). Más recientemente se reportan los desarrollados por Aiken (1991), Angoff (1974), Cizek (1999, 2006), Frary (1993), Haney, Wesolowsky (2000) y Clarke (2007), por mencionar algunos de ellos. A la fecha se siguen desarrollando métodos estadísticos que pretenden tener mejores resultados en la detección de irregularidades.

Wesolowsky (2000) propone un modelo estadístico que compara las respuestas por pares de estudiantes y procede a identificar excesivas respuestas coincidentes. El énfasis del método es la de prevenir falsas acusaciones (controlando el error tipo I). El método se aplicó a los resultados de las pruebas PLANEA 2015 para analizar si posibles patrones anómalos en los trenes de respuesta de los estudiantes. Los resultados, muestran un porcentaje de alumnos involucrados en irregularidad hasta el 7.2% en las pruebas aplicadas a alumnos de 6º de primaria y del 11.8% en 3º de secundaria en la prueba de matemáticas. Los porcentajes de irregularidad encontrados para Lenguaje y Comunicación son más bajos en ambos casos.

Ante la preocupación del INEE por prevenir falsas acusaciones, se toma la iniciativa de investigar la validez del método de Wesolowsky, ya que es el que se aplicó en el análisis de las pruebas y por ende la validez de los resultados.

II. El método de Wesolowsky.

A partir de este capítulo y a lo largo de este reporte, cuando se mencione *el método*, se hará referencia al método de Wesolowsky (2000).

En este capítulo se proporciona un resumen del *método*, así como algunas definiciones necesarias. Para mayores detalles, véase Wesolowsky (2000).

II.1 Notación:

En la discusión del escrito se asume la siguiente notación:

$n :=$ número de estudiantes en el grupo.

$q :=$ número de ítems en el examen.

$m_{jki} :=$ probabilidad de que los estudiantes j y k, coincidan en la respuesta del ítem i.

$M_{jk} :=$ número de coincidencias en las respuestas de los estudiantes j y k.

$p_{ji} :=$ probabilidad de que el estudiante j conteste correctamente el ítem i.

$r_i :=$ proporción del grupo que contestó correctamente el ítem i.

$c_j :=$ proporción de ítems que contestó correctamente el estudiante j.

$v_i :=$ número de opciones incorrectas correspondientes al ítem i.

$w_{ti} :=$ probabilidad de que, dado que la respuesta es incorrecta, el estudiante seleccione la opción incorrecta t en el ítem i.

La probabilidad de que dos estudiantes j y k coincidan en la respuesta del ítem i, considera la probabilidad de que ambos tengan la respuesta correcta, además de la probabilidad de que ambos coincidan en la respuesta incorrecta. Wesolowsky propone como forma funcional de la probabilidad p_{ji} a:

$$p_{ji=[1-(1-r_i)^{a_j}]^{1/a_j}} \tag{1}$$

para $0 \le r_i \le 1$, $a_j > 0$, con $j = 1,...,n$, $i = 1,...,q$.

Según Wesolowsky (2000), la probabilidad de coincidencia en la respuesta de la i-esima pregunta entre los estudiantes j y k es:

$$m_{jki} = p_{ji}p_{ki} + (1-p_{ji})(1-p_{ki})\sum_{t=1}^{v_i} w_{ti}^2 \ , \tag{2}$$

para $j = 1,...,n-1$, $k = j+1,...,n$, $i = 1,...,q$.

La deducción de la expresión sigue una lógica simple donde le primer término se refiere a la coincidencia en la respuesta correcta, mientras que la segunda en la respuesta incorrecta.

El parámetro a_j, la habilidad del estudiante j, se estima resolviendo la ecuación:

$$\frac{\sum_{i=1}^{q} p_{ji}}{q} = c_j \quad , \quad j = 1,...,n, \tag{3}$$

la cual es independiente para cada estudiante, y por su estructura no lineal, debe resolverse numéricamente.

Adicionalmente, *el método* considera deseable que la estimación para las probabilidades $\hat{p}_{ji}$ sea consistente con la proporción de respuestas correctas en cada una de las preguntas del examen que se aplicó al grupo, es decir, que se cumplan (al menos de manera aproximada) la siguiente ecuación:

$$\frac{\sum_{j=1}^{n} p_{ji}}{n} = r_i \quad , \quad i = 1,\ldots,q \ . \tag{4}$$

II.2. Supuestos del método

Para desarrollar las expresiones que sustentan el modelo, se parte de los siguientes supuestos:

a) Los estudiantes responden de manera independiente las preguntas del examen.
b) w_{ti} es la misma para todos los estudiantes.
c) La variable aleatoria, *Xij* que indica si la respuesta se respondió correctamente, sigue una distribución Bernoulli con parámetro *p_ij*.

El modelo tiene las propiedades deseables de un modelo de respuesta, como la de una relación directa entre la probabilidad de respuesta correcta tanto con la habilidad como con la dificultad del ítem; es decir, un alumno con mayor habilidad tiene mayor probabilidad de contestar una pregunta de una dificultad dada que estudiante de menor habilidad; y una pregunta simple (*ri* mas grande) tiene mayor probabilidad de ser respondida correctamente por los estudiantes que una menos simple.

II.3. La distribución asintótica del número de coincidencias y estadístico de prueba.

Para deducir un resultado asintótico que sea la herramienta de prueba de la excesiva similitud de coincidencias en las respuestas entre pares de alumnos, se considera la siguiente estructura de los datos. Se asumió que *Xij* es una variable aleatoria Bernoulli con parámetro *p_ij*, por lo tanto E(*Xij*)= *p_ij* y la varianza es Var(*X_ij*) = *p_ij*(1- *p_ij*). Considerando la independencia entre los estudiantes en responder las preguntas, se dedujo que la probabilidad de coincidencia está dada por (2), por lo que una nueva variable aleatoria sobre la probabilidad de coincidencia entre dos estudiantes - digamos *j* y *k*- se puede definir como *M_ijk*, la cual puede considerarse como variable aleatoria Bernoulli con valor esperado dado por la expresión en (2) y varianza $m_{ijk}(1-m_{ijk})$. A partir de esta expresión, se deduce que la variable aleatoria que define el número total de coincidencias es:

$$M_{jk} = \sum_{i=1}^{q} M_{ijk} \ ,$$

con valor esperado, la suma de los valores esperados de las *M_ijk*; es decir,

$$E(M_{jk}) = \sum_{i=1}^{q} E(M_{ijk}) = \sum_{i=1}^{q} m_{ijk} = \mu_{jk} \ . \tag{5}$$

Por independencia, la varianza, sería:

$$Var(M_{jk}) = \sum_{i=1}^{q} m_{ijk}(1 - m_{ijk}) = \sigma^2_{jk} \ . \tag{6}$$

Usando el Teorema del Límite Central, se justifica que La variable aleatoria, M_{jk}, tiene como distribución asintótica a la Normal con valor esperado dado por (5) y varianza dada en (6).

Estimación

Una vez que se tienen los resultados (trenes de respuesta) de una prueba aplicada, se procede a la estimación de los parámetros del modelo. El método más común para llevar a cabo el proceso de estimación es el de Máxima Verosimilitud. Partiendo del supuesto de que las dificultades de los ítems son conocidas (dadas), se plantea que los estudiantes responden los reactivos de manera independiente las preguntas, pero con diferentes probabilidades que depende de la habilidad de cada uno de ellos (no son idénticamente distribuidos). Este aspecto, hace que el proceso de estimación sea un tanto complicado, ya que el interés se centra en la estimación de las probabilidades de responder correctamente los reactivos de cada estudiante. En este sentido, dado que se asume la independencia entre los alumnos, cada estudiante tiene su propia muestra de reactivos (independientes) con una dificultad dada, los cuales son respondidos con una probabilidad que depende de su dificultad.

Si se plantea el problema como un proceso de estimación sobre una **probabilidad promedio** de responder correctamente cada reactivo, digamos $\bar{p}_j$, para una variable aleatoria X_{ji} con distribución Bernoulli con dicho parámetro, la función de verosimilitud para cada estudiante, $j=1,2,\ldots,n$, está dada por:

$$\ell(\bar{p}_j) = \prod_{i=1}^{q} \bar{p}_j^{x_{ji}} (1-\bar{p}_j)^{1-x_{ji}},$$

$$\Rightarrow \quad \ln[\ell(\bar{p}_j)] = \sum_{i=1}^{q} [x_{ji} \ln(\bar{p}_j) + (1-x_{ji}) \ln(1-\bar{p}_j)],$$

donde $\bar{p}_j = \frac{1}{q}\sum_{i=1}^{q} p_{ji}$.

Se puede demostrar que el estimador de Máxima Verosimilitud de $\bar{p}_j$, está dado por:

$$\hat{\bar{p}}_j = c_j.$$

Igualando entonces las expresiones correspondientes $\bar{p}_j$, se llega a la expresión:

$$\frac{1}{q}\sum_{i=1}^{q} p_{ji} = c_j \Rightarrow \frac{1}{q}\sum_{i=1}^{q} [1-(1-r_i)^{a_j}]^{1/a_j} = c_j.$$

El objetivo ahora es resolver la expresión para el parámetro de habilidad, a_j; dicha expresión se resuelve a través de métodos numéricos, lo cual resulta en el MV de la habilidad a_j (por resultados conocidos de los estimadores de *MV*); además.

Una vez estimados los parámetros, el estimador de las probabilidades de responder correctamente cada reactivo, *i*, por el estudiante *j* es:

$$\hat{p}_{ji} = [1-(1-r_i)^{\hat{a}_j}]^{1/\hat{a}_j},$$

De la misma manera, se obtienen las probabilidades de coincidencia entre los estudiantes j y k, las cuales se estiman sustituyendo las respectivas $\hat{p}_{ji}$ y $\hat{p}_{ki}$ en la expresión (2). Ahora, se puede asumir a la variable que describe la probabilidad de coincidencia entre los estudiantes *j* y *k* en la pregunta *i* como *M_jki*, cuya distribución sería Bernoulli con parámetro m_{jki}, dada en la expresión (2). La estimación de la probabilidad de coincidencia en la respuesta del reactivo *i*, por parte de los estudiantes *j* y *k*, m_{jki}, resulta en:

$$\hat{m}_{jki} = \hat{p}_{ji}\hat{p}_{ki} + (1-\hat{p}_{ji})(1-\hat{p}_{ki})\sum_{t=1}^{v_i}\hat{\omega}_{ti}^2 . \qquad (7)$$

Para la variable aleatoria $M_{jk} = \sum_{i=1}^{q} M_{ijk}$, la cual define el número de coincidencias observadas, entre los estudiantes j y k, se puede considerar como un proceso de q variables aleatorias independientes Bernoulli pero con diferentes probabilidades de éxito (las cuales fueron estimadas usando las probabilidades $\hat{p}_{ji}$ y $\hat{p}_{ki}$ de cada estudiante). Por ende, el número esperado de coincidencias entre los estudiantes j y k se obtienen sumando los valores esperados de las q variables aleatorias, lo que deriva en:

$$EM_{jk} \equiv \mu_{jk} = \sum_{i=1}^{q} m_{jki} , \qquad (6)$$

para $j = 1,\ldots,n-1$, $k = j+1,\ldots,n$.

Mientras que la varianza, se estima como:

$$VM_{jk} \equiv \sigma_{jk}^2 = \sum_{i=1}^{q} m_{jki}(1-m_{jki}) . \qquad (7)$$

Haciendo uso de la distribución asintótica (definida anteriormente por el Teorema de Límite Central), la variable aleatoria M_{jk}, es aproximada con la siguiente variable normal estándar:

$$Z_{jk} = \frac{M_{jk}-\mu_{jk}}{\sigma_{jk}} \sim N(0,1) , \qquad \text{para } j = 1,\ldots,n-1 , \; k = j+1,\ldots,n , \qquad (8)$$

Para efectuar el proceso de inferencia sobre las coincidencias, se debe estimar tanto μ_{jk} como σ_{jk}. Usando este resultado, se puede llevar a cabo la prueba de la hipótesis sobre la independencia en las

respuestas de los alumnos. Por la in-varianza de los estimadores, la expresión para el estimador de μ_{jk} está dada por:

$$\hat{\mu}_{jk} = \sum_{i=1}^{q} \hat{m}_{jki} ;$$

y para σ^2_{jk}, se logra con:

$$\hat{\sigma}^2_{jk} = \sum_{i=1}^{q} \hat{m}_{jki}(1 - \hat{m}_{jki}).$$

Así es estadístico de prueba sería el equivalente a la expresión (8) usando los estimadores correspondiente de μ_{jk} y σ_{jk}.

$$Z_{jk,Calculada} = \frac{\tilde{M}_{jk} - \hat{\mu}_{jk} - 1/2}{\hat{\sigma}_{jk}}, \qquad (9)$$

donde $\tilde{M}_{jk}$ son las coincidencias observadas. La constante "½" (conocida como factor de continuidad) se justifica dado que la variable aleatoria M_{jk}, es discreta (ver Rosner 2011).

De acuerdo, con Wesolowky (2000), se considera la prueba simultanea de todos los pares de estudiantes, lo que debe considerar que a un nivel de significancia α, se debe corregir por el número de comparaciones realizadas (Bonferroni); Así, la hipótesis de independencia, implica que el estadístico de compararse con un valor teórico de la distribución normal $Z_{(1-\alpha/[n(n-1)/2]}$. El rechazo de la prueba, sugiere una posible sospecha de un excesivo número de coincidencias en los trenes de respuesta, lo cual puede deberse a factores no controlados en la aplicación de la prueba. Wesolowky (2000), sugiere usar niveles de significancia del orden 0.001 con el fin de evitar falsas acusaciones.

III. METODOLOGÍA.

El objetivo del reporte es evaluar la validez y que tan robusto es el *método de Wesolowsky* en la detección de irregularidades en pruebas de opción múltiple. Para ello, se realizan simulaciones bajo dos escenarios diferentes. El primero considera patrones de simulación en un contexto de independencia en las respuestas de los alumnos; mientras que el segundo, considera las simulaciones imputando un patrón de copia usando un "grupo control". Los detalles específicos de las simulaciones se dan enseguida.

III.1. Proceso de simulación bajo independencia.

Los objetivos que se persiguen bajo este escenario son:

1) Determinar la eficiencia y robustez del método. Se evalúa que tanto método falla en la detección de posibles falsos positivos; lo deseables es que falle lo menos posible, es decir, que se minimicen los falsos positivos. Por construcción, lo esperado es que *el método* no detecte anomalías en los trenes de respuesta simulados.
2) Determinar una cota superior sobre el número de coincidencias permitidas para sospechar que existe anomalía en la prueba. Es decir, se responde la pregunta: ¿qué tanto se pueden desviar las coincidencias por arriba de lo esperado, dado en (6), y deberse estrictamente al azar?
3) Se consideran diferentes tamaños de grupo (n) y número de preguntas (q), para evaluar la sensibilidad del método.

Los trenes de respuesta de los alumnos, asumiendo que contestan de manera autónoma, siguió el siguiente proceso de simulación:

- Se generaron aleatoriamente las q dificultades de los reactivos a través de una variable aleatoria Uniforme(0,1).

$$r_i \sim unif(0,1), \quad \text{para toda } i = 1,\ldots,q. \tag{9}$$

- Las n habilidades de los estudiantes, se simularon usando una distribución Gamma(16.32, 15.87). Dicha distribución proviene de los resultados obtenido de la prueba planea, la cual se sometió al software SAS para determinar los valores de los parámetros.

$$a_j \sim Gam(16.32,15.87), \quad \text{para toda } j = 1,\ldots,n. \tag{10}$$

  Cabe destacar que solo se buscaba una distribución aproximada que diese valores adecuados de las habilidades más no una distribución que de un ajuste a la distribución de las habilidades. Los valores de los parámetros de la distribución se tomaron de estimaciones de las pruebas de Matemáticas de Planea para tercero de secundaria.
- Conocidas las dificultades y las habilidades, se procedió a calcular las probabilidades p_{ji}, usando la expresión (1).
- Con las probabilidades estimadas, se generaron los trenes de respuesta de las q preguntas para los n sustentantes.
- Los trenes simulados se sometieron al proceso de detección de las irregularidades con el *método de Wesolowsky*.
- Se consideraron valores de $q =$ 10, 15, 20, 25, 30, 40 y 50 (número de reactivos) y de $n =$ 10, 20, 30, 40, 50 y 60 (número de sustentantes).
- Se realizaron 10,000 simulaciones para cada combinación (n,q) dando un total de 420,000 simulaciones bajo el escenario de respuestas independientes. Luego, los trenes de respuesta simulados para cada combinación de (n,q) se sometieron al *método*. El nivel de significancia usada fue de α = 0.001.

Para dar una idea de las posibles fallas del método, se contaron el número de falsos positivos de cada combinación y se estimaron, tanto los promedios como los errores estándar de las coincidencias. Con esto, se puede aplicar la distribución normal y determinar las cotas de tolerancia del método.

III.2. Proceso de simulación bajo patrones de copia.

Los objetivos de la simulación bajo las condiciones de copia, son:

1. Determinar la eficiencia del método cuando se somete a un patrón de copia conocido.
2. La idea de la simulación es que detecte los patrones de copia simulados y siga sin detectar falsos positivos.

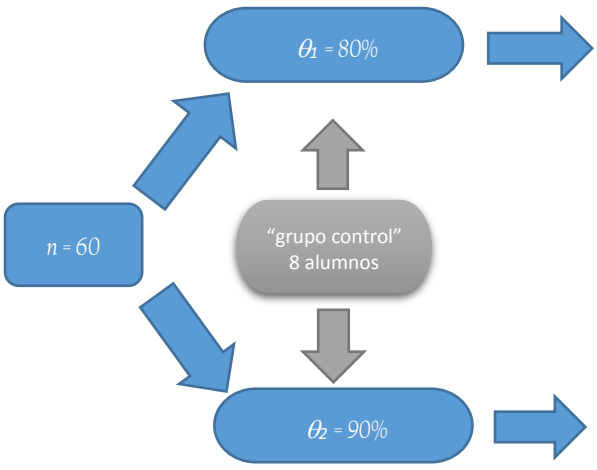
La principal diferencia entre la simulación con el patrón de copia y el patrón de independencia, fue la resignación de respuestas en los reactivos usando la siguiente estrategia:

- Para cada grupo de tamaño n, se identificaron al estudiante más hábil y al menos hábil, y a cada uno de ellos se les "ancló" de forma aleatoria a 3 estudiantes, formando un "grupo control" compuesto de 8 estudiantes.
- Para la prueba que consta de q ítems en cada uno de los grupos de tamaño n, se seleccionaron aleatoriamente dos porcentajes de similitud en las respuestas, el 80% y el 90%
- De esta forma, el "grupo control" tendrá un tren de respuestas muy similar
- El proceso seguido se muestra en la Figural, tomado como ejemplo un tamaño de grupo $n = 60$.
- Igual que en el caso de las simulaciones bajo independencia, se consideraron simulaciones para las siguientes combinaciones entre tamaño de grupo (n) y número de ítems (q):

  Para $q =$ 10, 20, 25, 30, 40 y 50 preguntas.

  Para $n =$ 10, 20, 30, 40, 50 y 60 estudiantes.
- Para cada combinación de (n, q, θ), se realizaron 10,000 simulaciones lo que arroja un total de 720,000 simulaciones.
- Finalmente, los trenes de respuesta simulados de cada combinación de se sometieron al *método de Wesolowsky* con la finalidad de evaluar la eficiencia en la detección de las irregularidades. El nivel de significancia usado fue de $\alpha = 0.001$.

Figural.- Proceso de simulación en patrones de copia para tamaño de grupo n=60 y todos los valores de q.



Total de preguntas en la prueba	10	20	25	30	40	50
Respuestas_copiadas	8	16	23	27	36	45

Total de preguntas en la prueba	10	20	25	30	40	50
Respuestas_copiadas	9	18	20	24	32	40

Esto es, se desea que el método tenga tolerancia contra quienes pudieron haber tenido irregularidades, en vez de acusar falsamente a quienes no incurrieron en las irregularidades.

Para determinar la eficacia del método, se contó el número de casos donde se identificaban al menos un par de sustentantes con anomalía de los que se indujo el patrón de copia. Además se cuentan todos los casos en donde se detectó el total de las anomalías simuladas. Cabe destacar que el objetivo no es detectar el total de las irregularidades, sino detectar las más evidentes y evitar que haya falsos positivos. Esto es porque se busca que el método sea tolerante a las irregularidades en vez de acusar falsamente. Este objetivo se logra, parcialmente, con el nivel de significancia del 0.001.

De acuerdo con Wesolowsky (2000) una condición deseable de un modelo es:

$$r_i \approx \frac{1}{n}\sum_{j=1}^{n}\hat{p}_{ji}\ .$$

Es decir, el promedio de las probabilidades de contestar correctamente cada una de las preguntas, debe ser similar a la proporción de estudiantes que la contestaron correctamente. Este argumento se dice que tiene mejor consistencia estructural. Gráficamente, puede visualizarse como un diagrama de dispersión entre las desviaciones entre las r_i y el promedio de las probabilidades en cada reactivo, contra el valor de la r_i, lo que debe generar una gráfica con puntos cercanos al valor de cero. Usando este criterio, Wesolowsky (2000) demostró que su modelo tiene mejor estructura que el modelo de Frary *et al.* (1977), donde se cumple el criterio solo en valores de r_i cercanos a $\bar{r}$. Ver Wesolowsky (2000) para mayores detalles.

IV. RESULTADOS.

A continuación se presentan los resultaos obtenidos del estudio de simulación. Se presentan por separado, los casos de independencia y grupos de control, ya que en ambos casos se responde a preguntas específicas.

IV.1. Resultados de las simulaciones: Bajo condiciones de independencia.

Una vez que se realizaron las 10,000 simulaciones de los trenes de respuesta para cada combinación (n, q), se sometieron al *método* para que detectara posibles anomalías en estos trenes de respuesta. Por la construcción de las simulaciones, lo esperado es que no se detecte anomalías. Los resultados obtenidos, usando $\alpha = 0.001$, son los siguientes:

Uno de los objetivos del ejercicio es determinar el número máximo de coincidencias permitidas entre dos estudiantes debidas a causas puramente aleatorias. Para ello, y en base a las simulaciones generadas se obtuvo el número esperado de coincidencias debidas puramente al azar, entre pares de estudiantes.

Tabla 1. Resumen de simulaciones de patrones de respuesta para diferentes tamaños de grupo y número de reactivos (α=0.001).

		Tamaño de la prueba (q)						
Tamaño de grupo (n)		*10*	*15*	*20*	*25*	*30*	*40*	*50*
	10	cero	cero	cero	cero	cero	cero	un par
	20	cero	cero	cero	cero	cero	un par	Cero
	30	cero	cero	cero	un par	un par	un par	Cero
	40	cero	cero	cero	cero	cero	un par	Cero
	50	cero	cero	cero	un par	cero	cero	Cero
	60	cero	cero	cero	cero	cero	cero	Cero

Para cada una de las celdas, se generaron 10,000 simulaciones.

Las casillas con "cero" significa que el *método* no detectó anomalías en las 10,000 simulaciones, mientras que en las casillas con "un par" implica que el *método* detectó un falso-positivo, *¡¡ un falso positivo en 10,000 simulaciones !!*.

Dado el nivel de significancia de $\alpha = 0.001 = P\left\{ \text{Re}\, chazar\, H_0 \mid H_0 \quad es \quad verdadera \right\} = 1/1000$, se admite el riesgo de que el *método* falle en 1 de 1000 veces que se aplique, es decir, dadas las 10,000 simulaciones en cada una de las combinaciones (n, q) se acepta el riesgo de que *el método* detecte hasta 8 falsos-positivos. Por lo que las simulaciones realizadas aportan evidencia favorable al método ya que detecta sólo 7 anomalías en las 420,000 simulaciones realizadas. La Tabla2, muestra los resultados de las simulaciones realizadas bajo independencia, sobre el valor esperado de las coincidencia y su desviación estándar (entre paréntesis). Por ejemplo cuando se le simulan 25 reactivos el número esperado de coincidencias entre pares de alumnos es de 12.6; es decir entre 12 y 13, con una desviación estándar de alrededor de 2.1. Es muy interesante notar que estos valores no dependen del tamaño del grupo, ya que se mantienen relativamente estables. Obsérvese además que el número esperado de coincidencias debidas al azar con el esquema de simulación descrito, es de aproximadamente la mitad del número del tamaño de la prueba.

Tabla 2. Número esperado de coincidencias, $E(M_{jk})$ y sus desviaciones estándar (entre paréntesis) por pares de alumnos.

		Tamaño de la prueba						
		q = 10	*q = 15*	*q = 20*	*q = 25*	*q = 30*	*q = 40*	*q = 50*
Tamaño del grupo	*n = 10*	4.9 (1.44)	7.7 (1.77)	10.7 (2.11)	12.7 (2.34)	15.3 (2.57)	19.9 (2.94)	26.7 (3.20)
	n = 20	4.9 (1.47)	7.8 (1.76)	10.9 (1.99)	12.7 (2.29)	15.6 (2.55)	20.1 (2.95)	25.6 (3.28)
	n = 30	4.8 (1.46)	7.6 (1.77)	10.0 (2.13)	12.3 (2.34)	14.3 (2.55)	19.8 (2.96)	25.6 (3.27)
	n = 40	4.7 (1.49)	7.6 (1.84)	10.4 (2.02)	12.6 (2.34)	14.8 (2.59)	20.0 (2.94)	24.8 (3.31)
	n = 50	5.0 (1.46)	7.6 (1.79)	10.2 (2.01)	12.6 (2.34)	15.5 (2.52)	19.5 (2.95)	24.8 (3.28)
	n = 60	5.1 (1.46)	7.7 (1.76)	9.6 (2.13)	12.6 (2.32)	14.7 (2.63)	19.0 (3.00)	23.7 (3.34)

A partir de la tabla se pueden responder preguntas como:

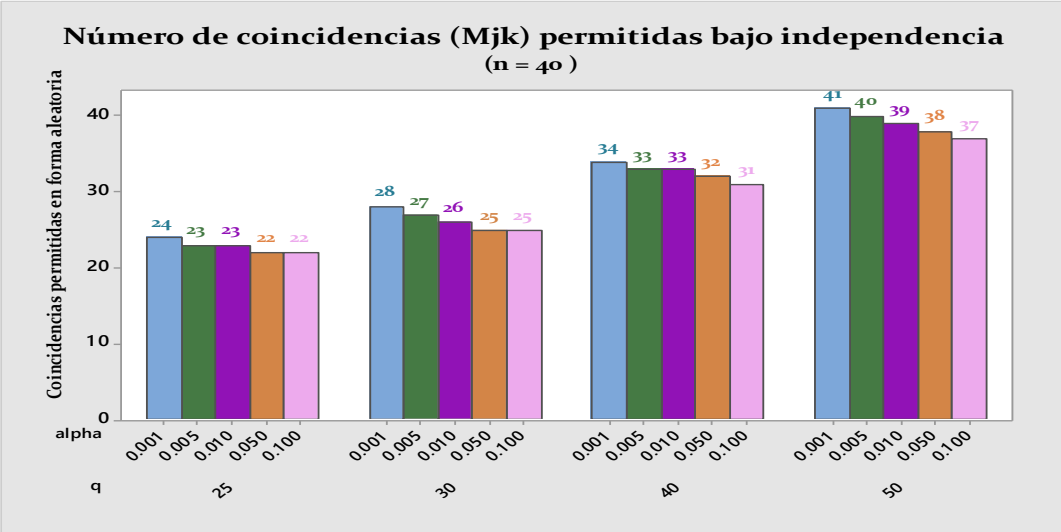
- ¿Qué tanto se pueden desviar las coincidencias por encima del número esperado (dado en la tabla anterior) y deberse estrictamente a causas aleatorias?
- ¿Cuál debe de ser la cota mínima en las coincidencias para poder concluir sobre la existencia de anomalías?

Aunque el análisis no es totalmente concluyente, si se da una panorámica sobre el tipo de resultados y/o análisis que se pueden realizar en las pruebas estandarizadas. Para que este sea más concluyente, se deben considerar varios factores que resulten en más escenarios de simulación, ya que los factores que tienen influencia en la respuesta son:

a) El nivel de significancia utilizado (α),
b) La habilidad de los estudiantes (a_j) y
c) El número de ítems de la prueba (q);

La figura 2 y la Tabla3, muestran un resumen de los casos que estamos analizando. Por ejemplo, en la figura se hace notar que, para un nivel de significancia de $\alpha = 0.001$, se pueden presentar hasta 24

Figura2. Representación gráfica de las cotas de tolerancia de coincidencias entre pares de estudiantes de grupos de 40 estudiantes, para cinco niveles de significancia y cuatro tamaños de la prueba. $\alpha = 0.001$.



coincidencias en total (11 coincidencias adicionales a las 13 esperadas), para concluir con un 99.9% de confianza que existe evidencia de irregularidades en las respuestas de los estudiantes (es decir, no contestaron de manera independiente), cuando el número reactivos es de 25. ¡Esto es, aún con 24 respuestas iguales, es probable que no se detecte la existencia de anomalías, al 99.9% de confianza, cuando la prueba de solo 25 preguntas!!! A un nivel de significancia del 90% la detección de respuestas anómalas entre pares de estudiantes se presenta a partir de 22 coincidencias, siendo aún muy similar al tamaño de la prueba.

Por otro lado, la Tabla 3 presenta las cotas para todos los casos simulados. Lo interesante de la tabla es que para tamaños de prueba muy pequeños (10, 15 y 20 reactivos), el número de coincidencias que pueden tolerarse, sin que se pueda asegurar la existencia de irregularidades, es muy similar al tamaño de la prueba; lo que sugiere usar tamaños de prueba suficientemente extensos para poder realizar el análisis de las posibles anomalías y no caer en la determinación de falsos positivos. Por ejemplo, para los casos de 50 reactivos, se permiten desde de 35 hasta 40 (aproximadamente) para detectar anomalías lo que ya es un tanto menor que el tamaño de la prueba y disminuiría el riesgo de los falsos positivos.

Tabla3. Número de coincidencias permitidas bajo independencia, para combinaciones de tamaños de grupo y número de reactivos, a cinco diferentes niveles de significancia, α.

	Nivel de significancia α	$q = 10$	$q = 15$	$q = 20$	$q = 25$	$q = 30$	$q = 40$	$q = 50$
$n = 10$	0.001	10	15	19	22	25	32	39
	0.005	10	14	18	21	24	30	38
	0.010	10	13	18	21	24	30	37
	0.050	10	13	17	19	23	28	36
	0.100	10	12	16	19	22	28	35
$n = 20$	0.001	10	15	20	22	26	33	40
	0.005	10	15	18	21	25	32	38
	0.010	10	15	18	21	25	31	38
	0.050	10	13	17	20	24	30	36
	0.100	10	13	17	20	24	29	36
$n = 30$	0.001	10	15	20	23	26	33	40
	0.005	10	15	19	22	25	32	39
	0.010	10	15	18	21	24	31	39
	0.050	10	14	17	20	23	30	37
	0.100	10	13	17	20	23	30	37
$n = 40$	0.001	**10**	**15**	**20**	*24*	*28*	*34*	*41*
	0.005	**10**	**15**	**19**	*23*	*27*	*33*	*40*
	0.010	**10**	**15**	**18**	*23*	*26*	*33*	*39*
	0.050	**10**	**15**	**18**	*22*	*25*	*32*	*38*
	0.100	**10**	**14**	**17**	*22*	*25*	*31*	*37*
$n = 50$	0.001	10	15	20	23	27	33	40
	0.005	10	15	19	23	26	32	39
	0.010	10	15	19	22	26	32	39
	0.050	10	15	18	21	25	31	37
	0.100	10	14	18	21	25	30	38
$n = 60$	0.001	10	15	20	23	26	33	40
	0.005	10	15	19	23	25	32	38
	0.010	10	15	19	22	25	32	38
	0.050	10	15	18	21	24	31	37
	0.100	10	15	17	21	23	30	36

NOTA: Las cotas para n=40 y q=25, 30, 40 y 50 de esta tabla, se presentan en la **Gráfica 3**.

Para tener una idea inicial del efecto de las habilidades de los estudiantes, se realizó como ejercicio adicional, una simulación de 10,000 para los casos de alumnos y reactivos de (60,50) y (25,10). Con este ejercicio se quiere responder las preguntas:

¿Qué pasaría con estas coincidencias si se comparan los trenes de respuestas de dos estudiantes que son *muy habilidosos*? y, ¿si se comparan dos alumnos con no tan *hábiles*? *¿Cambiaría la cota de tolerancia, y en caso positivo en cuantas*?

La idea es notar, la diferencia en el número de coincidencias permitidas entre alumnos habilidosos y entre alumnos poco habilidosos. Para ello, se crearon dos clases de alumnos, estudiantes habilidosos y estudiantes poco habilidosos. Para crear la "clase" de alumnos habilidosos, (clase A) y no habilidosos (clase B), se usó como valor crítico de $a_j = 1$. Esto es porque de acuerdo al modelo de Wesolowsky, $a_j = 1$, corresponde a una probabilidad de responder correctamente el reactivo i igual a

su dificultad r_i. Así, los estudiantes con habilidad $a_j > 1$ pertenecerían a la clase A (dado que la probabilidad de responder correctamente es mayor a la dificultas); y los estudiantes con habilidad $a_j \leq 1$ formaron la clase B. Bajo este esquema, al tomar las 10,000 simulaciones de cada grupo (60x50 y 10x25), los resultados obtenidos bajo esta división, se muestra en la Tabla 4.

Tabla 4. Número de coincidencias permitidas bajo independencia, dado un nivel de significancia para dos niveles de habilidad (A y B). $\alpha = 0.001$.

NIVEL DE SIGNIFICANCIA	CLASE	GRUPO 10X25 (Por habilidad)	GRUPO 10X25 Cota y Diferencia entre Clases vs. General (A,B)	GRUPO 60x50 (Por habilidad)	GRUPO 60x50 Cota y Diferencia de Clases vs. General (A,B)
	A ($a_{j>1}$)	22		39	
	B ($a_{j<=1}$)	21	19 (3,2)	36	36 (3,0)
α = 0.100	Diferencia A-B	1		3	
	A ($a_{j>1}$)	23		40	
	B ($a_{j<=1}$)	21	19 (4,2)	37	37 (3,0)
α = 0.050	Diferencia A-B	2		3	
	A ($a_{j>1}$)	24		41	
	B ($a_{j<=1}$)	22	21 (3,2)	38	38 (3,0)
α = 0.010	Diferencia A-B	2		3	
	A ($a_{j>1}$)	25		43	
	B ($a_{j<=1}$)	23	22 (3,1)	40	40 (3,0)
α = 0.001	Diferencia A-B	2		3	

La comparación entre las cotas de la Tabla 4 y la Tabla3, es muy ilustrativa sobre el efecto, tanto de la habilidad como del tamaño de grupo. Por ejemplo, para el grupo de *60x50* en base a la tabla 3, se permiten 40 coincidencias en un par de estudiantes para sospechar de la anomalía con un nivel de significancia de $\alpha = 0.001$; sin embargo, para estudiantes suficientemente hábiles, al mismo nivel de significancia, se deben tener 43 coincidencia para sospechar de la anomalía. La tabla, también ilustra el efecto del nivel de significancia, lo cual se nota revisando las diferencias a través de las columnas. Es interesante notar que las diferencias, para los tamaños de exámenes simulados se mantienen relativamente estables para todos los niveles de significancia; para el grupo *10x25* es generalmente de dos; mientras que para el grupo *60x50* es de tres.

Para obtener una herramienta mucho más robusta sobre la determinación de las cotas de tolerancia, sería interesante plantear diferentes simulaciones que combinaran las habilidades de los estudiantes, e incluir las dificultades como factor de análisis de las anomalías. Así, se tendrían casos de alumnos hábiles con exámenes sencillos y complicados, y casos con alumnos poco hábiles con exámenes sencillos y complicados. El análisis de estas simulaciones, daría la pauta para determinar las cotas de tolerancia bajo diferentes circunstancias que, además, pudiese ayudar a resolver los casos más comunes de aplicación en las pruebas estandarizadas.

IV.2. Resultados de las simulaciones: Bajo patrones de copia.

Por la construcción de las simulaciones bajo este escenario, se espera que *el método* detecte las anomalías que se "anclaron" al "grupo control" compuesto por 8 estudiantes. Luego, para cada combinación de (n, q, θ) se espera que en las 10,000 simulaciones realizadas, *el método* detecte a los *mismos* alumnos que componen al "grupo control", idealmente.

Existen dos situaciones bajo las cuales se puede decir que *el método* falle:

1) Si existen simulaciones en las que se detecte 8 anomalías, pero entre alumnos que **no** pertenecen al grupo control.
2) Si se detecta un número excesivo de pares de estudiantes con respuestas anómalas y que estos estudiantes **no** pertenezcan al "grupo control" (existan más de 8 anomalía detectadas).

Los resultados encontrados para evaluar el primer punto, se presentan en la Tabla5. Obsérvese que, independientemente del tamaño del grupo (n), para pruebas con un número pequeño de reactivos (q), el método es poco eficiente para la detección de anomalías del "grupo control". Esto es, tolera que varios estudiantes que han cometido anomalías no sean detectados de esperarse, lo cual se justifica dado que, para exámenes pequeños, el método requiere de "muchas coincidencias" para tener la posibilidad de detectar anomalías; en varios casos de todo el examen. A partir de esto, es justificable que solo en pocos grupos se hayan detectado las ocho coincidencias generadas con el grupo control.

Tabla 5. Número de grupos simulados donde se detectaron las 8 anomalías generadas con el "grupo control"; $\alpha = 0.001$.

Tamaño de Grupo	% De reactivos	*Tamaño de la prueba*			
		q = 25	*q = 30*	*q = 40*	*q = 50*
n = 10	$\theta_1 = 90\%$	0	4	575	4575
	$\theta_2 = 80\%$	0	0	39	724
n = 20	$\theta_1 = 90\%$	4	0	3329	0
	$\theta_2 = 80\%$	0	0	708	0
n = 30	$\theta_1 = 90\%$	0	2026	6972	7004
	$\theta_2 = 80\%$	0	2027	4722	4037
n = 40	$\theta_1 = 90\%$	4	4953	8105	8181
	$\theta_2 = 80\%$	0	832	4785	5457
n = 50	$\theta_1 = 90\%$	5	886	6171	9076
	$\theta_2 = 80\%$	0	21	2405	6853
n = 60	$\theta_1 = 90\%$	0	2194	5101	9214
	$\theta_2 = 80\%$	0	95	2662	6435

Cada casilla es resultado de 10,000 simulaciones.

Por otro lado, la tabla muestra que el método es más eficiente mientras más estudiantes tenga el grupo y mientras más preguntas contenga el examen. Si bien el cuadro muestra números relativamente bajos respecto a número de simulaciones realizadas, no quiere decir, que el método no detecte ningún

caso de anomalía. La Tabla 6 muestra el número de simulaciones en las que el método detectó al menos una anomalía (no exactamente las ocho generadas). Las casillas que tienen *, identifican los casos donde el método detectó falsos positivos; es decir, detectó anomalías en estudiantes que no pertenecen al "grupo control". En casillas en las que no hay un *, implica que en las anomalías detectadas están involucrados los elementos (tal vez no todos) que forman el "grupo control", es decir, el método no cae en el riesgo de acusar falsamente. Claramente, lo esperado es que en cada casilla aparezca un 10,000, implicando que en las 10,000 simulaciones se detectaron anomalías y que las detectadas pertenecen a los elementos que forman el "grupo control".

Tabla 6. Número de simulaciones en las que se detectaron al menos una anomalía del grupo control. α = 0.001.

Tamaño de Grupo	*% De reactivos*	*Tamaño de la prueba*			
		q = 25	*q* = 30	*q* = 40	*q* = 50
n = 10	$\theta_1 = 90\%$	356	4,353	9,486	9,783
	$\theta_2 = 80\%$	169	1,642	7,367	8,509
n = 20	$\theta_1 = 90\%$	6,745	7,290	9,902	10,000
	$\theta_2 = 80\%$	5,514	4,516	9,545	10,000
n = 30	$\theta_1 = 90\%$	4,316	9,735	10,000	10,000
	$\theta_2 = 80\%$	28	9,206	9,995	10,000
n = 40	$\theta_1 = 90\%$	5,373	9,848	10,000	10,000
	$\theta_2 = 80\%$	1,755	7,832	9,995	10,000
n = 50	$\theta_1 = 90\%$	6,882	9,814	9,992	10,000
	$\theta_2 = 80\%$	1,779	8,138	9,848	9,998
n = 60	$\theta_1 = 90\%$	811	9,671	9,868	9,978
	$\theta_2 = 80\%$	258	7,614	9,487	9,979

Nota: Cada casilla es resultado de 10,000 simulaciones.

Cabe aclarar que, en todos los casos donde el método acusó falsamente, sólo de detectó un par de estudiantes, es decir, el *método* detectó solo un falso-positivo en cada caso *¡¡ un falso positivo en 10,000 simulaciones !!*. Así, los resultados dan evidencia de la confianza del método, ya que solo sólo detectó seis falsos positivos en 480,000 casos simulados.

Dado lo anterior, se puede concluir que si bien el método es permisivo en el proceso de detección de irregularidades (no acusa en casos donde se hubo irregularidad), cuando acusa casos irregulares es muy alta la confianza de que así sea; sobre todo cuando el número de reactivos es suficientemente grande. También es importante notar que cuando el método detecta irregularidades, lo hace porque hay una gran cantidad de reactivos donde las respuestas entre los estudiantes coinciden (80 y 90%). Si el porcentaje de coincidencia fuera menor, aunque si hubiera irregularidad, el método tendería a no detectarla ya que estaría cada vez más cerca de la cota de tolerancia. Por ejemplo, si en un grupo de 40 estudiantes se presume que hay irregularidades y esta ocurrió en el 70% de los reactivos (en 35 reactivos) para algunos pares de estudiantes en un examen de 50 preguntas, el método no detectaría

dicha anomalía dado que la cota de tolerancia es de 40 reactivos (α=0.001). Dependiendo de algunos factores, podría comenzar a detectar a partir de 36 con un nivel de significancia de α=0.1.

Finalmente, la Tabla 7 da el promedio de alumnos detectados con irregularidad para cada caso simulado. De la tabla es interesante notar que a medida que se incrementa el valore de *q* (Número de reactivos en la prueba), el promedio de irregularidades detectadas se acerca al ocho, que fueron los casos construidos con irregularidad (grupo control); es decir, cuando el tamaño de la prueba es suficientemente grande (50 o más), el método tendería a detectaría a la mayoría de los casos donde sí se presentaron irregularidades; lo cual es un aspecto muy deseable del método.

Tabla7. Número promedio de pares de estudiantes detectados con irregularidad del grupo de control. $\alpha = 0.001$.

Tamaño de grupo	*Parámetro de irregularidad*	*Número de reactivos (q)*			
		25	*30*	*40*	*50*
n = 10	θ= 90%	0.0826	1.2018	4.1063	6.3291
	θ= 80%	0.0340	0.3632	2.3295	3.8348
n = 20	θ= 90%	2.1146	2.1191	5.9970	6.9889
	θ= 80%	1.4132	1.1085	4.4385	6.8103
n = 30	θ= 90%	0.9826	5.0931	7.0257	7.2198
	θ= 80%	0.0057	4.8295	6.5835	6.3769
n = 40	θ= 90%	1.4136	6.3978	7.5162	7.5111
	θ= 80%	0.3915	3.2572	6.7598	6.7458
n = 50	θ= 90%	1.9830	4.7349	6.9906	7.7508
	θ= 80%	0.3900	2.6400	5.6022	7.1598
n = 60	θ= 90%	0.2053	5.3198	6.4064	7.7924
	θ= 80%	0.0536	2.6455	5.3084	7.1303

IV.3. Algunos casos extremos y algunas soluciones.

Vale la pena mostrar los casos extremos a los que se enfrenta el método y por ende este no funciona.

1) Para un grupo de n estudiantes y una prueba con q reactivos, <u>*todos*</u> los estudiantes tienen las respuesta correctas; es decir $r_i = 1$ para *i=1,2,...,q*,

$\Rightarrow$ El método **NO** detecta anomalías.

La razón es que siendo $r_i = 1$, la probabilidad de responder correctamente cualquier reactivo es igual a 1. Si la probabilidad es igual a 1, la probabilidad de coincidencia entre los estudiantes es también igual a 1 en todos los reactivos, por lo que el valor esperado de coincidencias es igual al número de

reactivos de la prueba. De aquí que el estadístico de prueba, quede indefinido, ya que al ser todo igual, la desviación estándar es cero; ya que tanto el denominador como el numerador son iguales a cero

Por lo tanto, no hay forma de detectar irregularidad.

2) Para un grupo de n estudiantes y una prueba con q reactivos, para la cual <u>*todos*</u> los estudiantes tienen la <u>*misma*</u> respuesta incorrecta $r_i = 0$ para toda *i=1,2,...,q*.

   $\Rightarrow$ El método **NO** detecta anomalías.

La razón es similar al caso anterior, solo que la probabilidad de responder correctamente es igual a cero; pero como todas las respuestas incorrectas son iguales, la probabilidad de coincidencia es otra vez igual a 1. Por lo tanto, el valor esperado de coincidencias vuelve a ser igual a número de reactivos y por igualdad en todas las respuestas la desviación estándar vuelve a ser cero.

Las razones por las que pueden ocurrir este tipo de casos, son muy diversas. Para el caso 1, algunas de ellas son:

- No hubo el control debido en la prueba lo que resultó en una copia generalizada;
- El maestro les ayudó a responder la prueba y lo sabía todo,
- Todos los estudiantes son extraordinariamente buenos que no se equivocaron en las respuestas de la prueba, o como caso muy inusual,
- Tanto los estudiantes como el profesor, pudieron tener acceso a la prueba con anterioridad;
- Otras posibilidades o combinaciones que se puedan ocurrir.

En el caso dos, pudo ser una falta de enseñanza adecuada (se les enseño mal TODO) y además hubo un acuerdo de responder la misma opción.

Algunos casos comunes a los que se les dio solución son los siguientes:

3) Solo en unas preguntas coinciden todos los estudiantes en la respuesta correcta; es decir en *k* de los *q* reactivos las respuestas son iguales para todos los estudiantes, por lo que para esos casos $r_i = 1$.

   Solución: La probabilidad de contestar correctamente es igual a 1, para todos los estudiantes, por lo que la probabilidad de coincidir en dicho reactivo para todos los pares de estudiantes es igual 1. Por lo tanto, el ítem aporta solo al valore esperado, con varianza cero para el estadístico de prueba.

4) Solo en unas preguntas coinciden todos los estudiantes en la respuesta incorrecta; es decir en *k* de los *q* reactivos las respuestas incorrectas son iguales para todos los estudiantes, por lo que para esos casos $r_i = 0$.

   Solución: La probabilidad de contestar correctamente es igual a cero, para todos los estudiantes, por lo que, la probabilidad de coincidir en dicho reactivo para todos los pares de estudiantes es igual 1. Por lo tanto, el ítem aporta solo al valor esperado, con varianza cero para el estadístico de prueba.

5) El número de respuestas correctas para algunos estudiantes es cero.

Solución: Todas las probabilidades de contestar correctamente los reactivos para esos estudiantes son iguales a cero. Por lo tanto, la ecuación para resolver la habilidad queda igualada a cero, por la expresión para la habilidad queda igualada a cero. Analíticamente, se pierde la expresión en función de la habilidad y la solución implica una $r_i = 0$, lo cual no es necesariamente cierto. Nuestra propuesta es asignar desde un inicio el valor mínimo de habilidad que en este caso es cero y evitar la solución analítica.

6) Dos estudiantes coinciden en el mismo tren de respuesta, y en algunos reactivos TODOS los estudiantes tienen la misma respuesta.
Solución: Se asignó la habilidad para las dos estudiantes igual al número de reactivos correctos entre *q*.

V. CONCLUSIONES

Este reporte es un resumen de los resultados después de haber realizado 900,000 simulaciones de trenes de respuesta usando el modelo de Wesolowky (2000). Los resultados muestran suficiente evidencia de que el método es eficiente para detectar irregularidades. En general, se puede concluir que el método propuesto por Wesolowsky para la detección de un excesivo número de coincidencias en los patrones de respuesta de las pruebas estandarizadas, por pares de estudiantes, es eficiente.

En base a estos resultados se admite que, aunque *el método* puede acusar falsamente, la probabilidad de que esto suceda no excede el nivel de significancia (error tipo I de la prueba).

En el artículo origina, Wesolowsky no determina cotas de tolerancia máxima sobre el número de coincidencias permitidas entre pares de estudiantes debidas al azar; este estudio si considera importante conocer las cotas de tolerancia, las cuales son debidas al azar bajo las condiciones en las que se desarrolló el estudio. El objetivo fue solo conocer el orden de magnitud del número de coincidencias arriba del promedio de coincidencias, sobre las cuales se espera una posible irregularidad en los trenes de respuesta de las pruebas (Ver Tabla3).

Para visualizar un efecto esperado de las habilidades, el estudio arrojó que:

1) El número esperado de coincidencias observadas y debidas al azar es de aproximadamente la mitad del número de ítems del examen. Además, es independiente del tamaño del grupo, véase tabla 2.

2) Se notó que a medida que se aumenta el número de reactivos, así como el tamaño del grupo, cuando se conoce a los estudiantes que participaron en las irregularidades, el método tiende a detectar a "prácticamente" a todos los involucrados.

3) A menor número de reactivos, la cota de tolerancia en el número de coincidencias esperadas se acerca al número de preguntas de la prueba. Por lo tanto, el método aumenta su eficiencia, sobre todo cuando el tamaño de la prueba es mayor a 25 reactivos.

4) El número de coincidencias observadas, puede exceder a la cota de tolerancia, debido fundamentalmente a:

- El nivel de ***significancia*** utilizado (α);
- las ***habilidades*** de los estudiantes involucrados (a_j);
- del ***número de reactivos de la prueba*** (q);

VI. INVESTIGACIONES FUTURAS PLANEADAS

En base a las conclusiones dadas en la sección anterior, se propone:

- Comparación del método con ***otros métodos existentes***, presumiblemente eficientes, para medir la bondad del método y considerar posibles extensiones del método (por ejemplo, el método de Wollack el cual se basa en modelos logísticos).
- Desarrollar estrategias de detección de irregularidades usando estrategias Bayesianas, como el considerar la dificultad de los reactivos con una distribución a-priori en el proceso de detección de las irregularidades.
- Aunque, se da la pauta del efecto de la habilidad de los estudiantes, se podría estudiar con mayor detalle, además de la dificultad de la prueba, en el valor esperado de coincidencias.

VII. REFERENCIAS

Dwyer, D. J. and Hecht, J.B. (1996). Using statistics to catch cheaters: methodological and legal issues for Student Personnel Administrators, NASPA Journal, 33(2), pp. 125-135.

Frary, R., Tideman, T. and Watts, T. (1977). Indices of cheating on multiple-choice test. Journal of Educational Statistics, 4, pp. 235-256.

Frary, R. (1992). Statistical Detection of Multiple-Choice Test Answer copying: State of the Art. Paper presented at the Annual Meeting of the Measurement Services Association (San Francisco, CA, April 1992).

Frary, R. (1993). Statistical Detection of Multiple-Choice Test Answer copying: Review and commentary. Applied Measurement in Education, 6(2), 153-165.

Fray, R. and Tideman, T. (1994). The evaluation of two indices of answer copying and the development of a spliced index. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Fray, R. and Tideman, T. (1997). Comparison of two indices of answer copying and development of a spliced index. Educational and Psychological Measurement, 57, 20-32.

Hanson, B. A., Harris D. J., and Brennan, R. L. (1987). A comparasion of several statistical methods for examining allegations of copying, ACT Research Report Series 87-15. Iowa City, IA: American College Testing Program.

Post, G.V. (1996). A quantal choice model for the detection of copying on multiple choice exams, Decision Sciences, 25(1), pp. 123-142.

Rosner B., (2010) Fundamentals of Biostatistics. 7th edition, Brooks/Cole, Boston MA, USA.

Wesolowsky, G. (2000). Detecting excesive similarity in answers on multiple choice exams. Journal of Applied Statistics, Vol. 27, pp. 909-921.

REFERENCIAS COMPLEMENTARIAS

Bakhtiyari, K., Patel, A., Taghavi, M. (2011). Evaluation of Cheating Detection Methods in Academic Writings. Journal of Library Hi Tech, Vol. 29, Issue 4.

Bay, L. (1995). Detection of Cheating on Multiple-Choice Examinatons. Paper presented at the Annual meeting of the American Educational Research Association.

Belov, D. and Armstrong R. (2010). Automatic Detection of Answer Copying via Kullback-Leibler Divergence and K-index. Applied Psychological Measurement, 34(6) 379-392.

Drasgow, F. and Guo, J. (2010). Identifying Cheating on Unproctored Internet Test: The Z-test and the likelihood ratio test. International Journal of Selection and Assessment, Vol. 18, No. 4.

Harpp, D. and Hogan, J. (1996). Crime in the classroom- Part II, an update. Journal of Chemical Education, 73, 4, pp. 349-351.

Love, R., Morris, J., and wesolowsky, G. (1988). Facilities location: Models and Methods (New York, Elsevier Publishing).

Nathanson, C., Paulhus, D., Williams, K. (2006). Predictors of a behavioral measure of scholastic cheating: Personality and competence but not demographics. Contemporary Educational psychology, 37, 97-122.

Nelson, L. (2006). Using selected índices to monitor cheating on multiple-Choice exams. Journal of Educational Research and Measurement, Vol. 4, No. 4.

Richmond, P. and Roehner B. (2015). The detection of cheating in multiple choice examinations. Psychometrica, Vol 1, No 1.

Van der Linden, W. and Sotaridona, L. (2004). A Statistical Test for Detecting Answer Copying on Multiple-Choice Test. Journal of Educational Measurement, Vol. 41, No. 4., pp. 361-377.