

Efectos de la Multiplicidad en una Situación de Pruebas Seriales de Asunto Único: ¿Cuándo una Prueba de Asunto Único no es una prueba de Asunto Único?

Raymond Nelson

Abstract

Se investigó el enfoque serial de asunto único (SSI) para pruebas exploratorias de asunto múltiple, con el objetivo de entender los efectos potenciales de la multiplicidad que podrían jugar un papel en los tamaños de efecto y en las ventajas esperadas asociadas a este método. Utilizando las tasas de sensibilidad, especificidad y de error de estudios poligráficos publicados previamente, se utilizaron métodos Monte Carlo para estudiar la situación del SSI. Se proporcionan ejemplos para el contexto de la exploración poligráfica de asuntos múltiples y de la repetición de pruebas en el contexto del polígrafo diagnóstico de asunto único. Para proporcionar información adicional entorno a las posibles ventajas de las estrategias de pruebas en serie, también se modelan estrategias de repetición de pruebas para el contexto del COVID-19. Los resultados indican que las pruebas en serie potencialmente podrían disminuir la precisión general de las pruebas y generar un aumento desproporcionado de errores falso positivo y de resultados inconclusos entre personas inocentes. Las pruebas en serie pueden proporcionar pequeñas mejoras en la sensibilidad de la prueba y en errores falso negativo. Las tasas de precisión y de error esperadas se muestran a manera de precaución para aquellos lectores que deseen una mejor comprensión de las posibles ventajas y desventajas asociadas al uso de una serie de pruebas exploratorias de asunto único. Se proporciona una discusión sobre la multiplicidad estadística (efectos de error acumulativos) y del uso de métodos analíticos ómnibus. El mayor efecto observado en el enfoque SSI fue una gran reducción en la especificidad de la prueba. Aunque pueden existir ciertas ventajas al probar una pregunta o hipótesis a la vez, en este momento la evidencia disponible no respalda el abandono total de los métodos de análisis ómnibus o de las técnicas poligráficas exploratorias de asuntos múltiples.



Introducción

En una publicación reciente, O'Burke (2022) describió el uso de cuatro asuntos-objetivo exploratorios, denominadas preguntas relevantes en la práctica poligráfica de campo, incluyó uso de drogas ilegales, participación en delitos graves, violencia doméstica y delitos sexuales, y ha sugerido el uso de una serie de exámenes de asunto único en lugar de la práctica común de utilizar pruebas de exploratorias de asuntos múltiples. Investigaciones publicadas previamente (Barland et al., 1989) no mostraron ventajas por el uso de una serie de exámenes de asunto único en comparación con los efectos de un solo examen de asunto múltiple. Sin embargo, O'Burke ha planteado la hipótesis de un enfoque exploratorio de una serie de asunto único (SSI), utilizando el formato de prueba exploratorio de mentira dirigida (Handler et al., 2008; Research Division Staff, 1995a; 1995b) como formato de asunto único, como lo describieron previamente Prado et al. (2015a; 2015b)¹, puede proporcionar un diagnóstico más preciso del comportamiento problemático de un examinado y podría facilitar un uso más óptimo de los recursos disponibles de prueba.

O'Burke (2022) no proporcionó un análisis para respaldar la hipótesis del SSI, y la única evidencia discutida fue una descripción anecdótica de casos no confirmados² para los cuales se observó que los totales de las pruebas de asunto único eran mayores que las puntuaciones subtotales reportadas para pruebas exploratorias de asuntos múltiples³, y que fueron reportados como anecdóticamente satisfactorios para la agencia y los examinadores. Se informaron las medias y las desviaciones estándar de los resultados observados en veraces y engañosos. Sin embargo, los gráficos incluidos mostraron que los datos tenían una distribución muy anormal, de modo

que cualquier intento de hacer un uso práctico o analítico de estas estadísticas descriptivas utilizando métodos paramétricos tradicionales llevaría a conclusiones inestables e irreproducibles.

Además de la ausencia de datos y análisis de respaldo, no se proporcionó información sobre los tamaños del efecto del SSI ni se proporcionó una discusión sobre el hecho central de que el enfoque de SSI estará sujeto a los mismos fenómenos estadísticos conocidos que todas las demás actividades científicas y estadísticas que involucran las actividades de clasificación e inferencia múltiples. Finalmente, aunque pueden existir efectos económicos en términos de tiempo, financiamiento, espacio físico, recursos de personal o mayor precisión de clasificaciones de engaño, O'Burke no proporcionó datos o análisis econométricos para respaldar la hipótesis de estos efectos.

Una forma útil de obtener perspectiva y objetividad para una hipótesis o idea científica es encontrar otros ejemplos que puedan utilizarse para hacer cambios en el contexto aplicado. Se espera que los constructos válidos exhiban consistencia de dominio, lo que significa que podemos esperar observar fenómenos y efectos similares al basarnos en el concepto sin importar el contexto.⁴ Los constructos válidos rara vez son válidos en un paradigma de aplicación única. ¿Qué otros ejemplos se pueden encontrar para el uso de una serie de pruebas asunto único? En otras palabras, si la hipótesis del SSI es válida y útil, ¿podemos identificar otros ejemplos del uso de estrategias de pruebas en serie?

La pandemia mundial reciente proporciona un ejemplo potencial del concepto de coherencia de dominio. Entre la infinidad de reglas y directrices que se ofrecieron durante la pandemia de COVID-19 hubo una sugerencia de que múltiples resultados negativos consecutivos de pruebas podrían usarse como base para ciertas decisiones, como poner fin al período de cuarentena o aislamiento⁵

¹Prado et al. informaron el uso de un formato DLST en un examen de asunto único con tres repeticiones de 2 RQ en un solo registro. Encontraron tasas de precisión similares a otras técnicas de polígrafo de un solo tema con 2 RQ, y ninguna ventaja en el uso de un solo registro gráfico.

²La presunta culpabilidad e inocencia no se justifica para los sujetos de esta muestra de conveniencia porque la evidencia de tal presunción no es independiente del resultado de la prueba poligráfica. El conocimiento publicado sobre las tasas de error para la técnica poligráfica validada indica que se debe suponer razonablemente que los resultados observados incluyen alguna proporción de errores de prueba.

³Este hallazgo no es sorprendente si se considera que las puntuaciones manuales de los exámenes poligráficos se agregan mediante sumatoria. Dada la teoría de la prueba del polígrafo, es axiomático esperar que las puntuaciones de gran total sean mayores que las puntuaciones subtotales.

⁴Un ejemplo de esto se puede ver en las pruebas de personalidad e inteligencia en el campo de la psicología, aunque se están llevando a cabo discusiones profundas e interesantes sobre la naturaleza real de estos constructos - un constructo es una idea abstracta que se formula para ser considerada algo tangible - puede aceptarse como válido cuando diferentes aplicaciones contextuales conducen a observaciones, efectos y resultados similares. Aunque algunas diferencias son inevitables, las puntuaciones de inteligencia y personalidad tienden a correlacionarse fuertemente con diferentes métodos de medición y análisis.

⁵Esta recomendación se enfatizó con más fuerza durante el período anterior a la introducción de vacunas que redujeron sustancialmente la probabilidad de enfermedad grave, hospitalización y muerte. Se ha hablado menos de ello a medida que el público se ha acostumbrado a una forma más endémica de enfermedad viral.



El requisito de resultados de pruebas múltiples de COVID-19 - pruebas múltiples de asunto único - fue un intento de hacer uso deliberado del fenómeno de la *multiplicidad estadística*, también conocido simplemente como *multiplicidad*, y del *problema de las comparaciones múltiples*. Cuando se hace referencia a resultados positivos de las pruebas, el fenómeno se denomina *alfa inflado* o efecto de *error acumulativo* porque describe la inflación de las tasas de error falso positivo más allá de una tolerancia planificada. Cuando se describen resultados negativos de las pruebas, el fenómeno de la multiplicidad estadística resulta en lo contrario - un umbral *alfa reducido* o *desinflado*. En el contexto de las pruebas de COVID-19, se espera que el requisito de múltiples resultados negativos reduzca la aparición de errores falsos negativos.

Los exámenes poligráficos exploratorios de asunto múltiple son una alegoría contextual del uso de métodos de comparación múltiples en la investigación científica. Las reglas de decisión para estos exámenes pueden considerarse alegóricas a los métodos de análisis estadísticos *ómnibus* en los que en un solo experimento y análisis se prueban múltiples hipótesis. Los resultados de actividades científicas que utilizan métodos *ómnibus* son a menudo una solución útil y preferida en la investigación científica, aunque pueden requerir análisis *post hoc* para comprender mejor la información resultante cuando se observan diferencias significativas.

Una limitación conocida de los métodos *ómnibus*, como el del ANOVA, es que aunque pueden identificar la presencia de diferencias significativas dentro de una serie de condiciones experimentales, no señalan las áreas exactas donde ocurren las diferencias. Se ha encontrado que los exámenes poligráficos de asuntos múltiples presentan limitaciones similares (Barland et al., 1989; Podlesney & Truslow, 1993; Raskin et al., 1988). En la práctica poligráfica de campo, el análisis *post hoc* a menudo se denomina *prueba de ruptura* o *prueba de barreras sucesivas*, en las que se pueden realizar exámenes de asunto único después de un resultado positivo en una prueba *ómnibus* o de asuntos múltiples.

En ocasiones, los investigadores, estadistas y profesionales de campo preferirían evitar por completo el problema de la multiplicidad utilizando metodologías estadísticas y de procedimiento que no involucren comparaciones múltiples. Los efectos de la multiplicidad también pueden mitigarse mediante métodos de análisis *ómnibus* como los de la familia ANOVA. También se pueden reducir mediante el uso de correcciones estadísticas/matemáticas como las correcciones de Bonferroni y Sidak (Abdi, 2007; Nelson, 2015; Sidak, 1967) que se pueden aplicar tanto a los umbrales *alfa* deseados o a resultados estadísticos computarizados. Lo más importante es que se sabe que los resultados analíticos de la investigación científica y de las pruebas poligráficas de campo involucran efectos de multiplicidad - una distorsión de la precisión analítica - siempre que se utilizan comparaciones estadísticas múltiples para obtener conclusiones acerca de los resultados de una prueba o experimento científico.

Hay razones para hipotetizar que los resultados analíticos de una serie de exámenes poligráficos de asunto único - la hipótesis del SSI - también pueden estar sujetos a los efectos de multiplicidad, al igual que en otros exámenes poligráficos de asuntos múltiples. Investigaciones previas de Barland, Honts y Barger (1989) respaldan esta posibilidad y la necesidad de este proyecto. Una diferencia entre la hipótesis de SSI - que se mejorará la precisión o exactitud de las decisiones - y el contexto de COVID-19 es que mientras el estado de criterio de caso para el contexto de COVID-19 es uniforme para la serie de pruebas dentro de un caso, el enfoque de SSI en el contexto del examen poligráfico implica asuntos-objetivo independientes dentro de cada caso. Raskin et al. (1988) informaron que la precisión general de la prueba se puede optimizar mediante el uso de preguntas con estados de criterio uniformes.⁶

Este proyecto emplea métodos estadísticos y analíticos simples y comunes para examinar los tamaños del efecto para una serie de pruebas poligráficas exploratorias de asunto único utilizando el conocimiento existente sobre la precisión de los formatos de prueba de asunto único utilizados en la práctica poligráfica de campo contemporánea.

⁶ En la práctica del polígrafo de campo, el uso de preguntas con estados de criterio uniformes significará que un examinado miente a todas o dice la verdad a todas las preguntas objetivo de la investigación.



Método

Se obtuvieron tasas de sensibilidad, especificidad y de error de prueba para exámenes poligráficos de asunto único de un estudio meta analítico de técnicas poligráficas validadas (APA, 2011). La sensibilidad agregada de prueba para formatos de asuntos múltiples incluidos en el estudio meta analítico fue de .771 y la especificidad de prueba para estos formatos fue de .719. Las tasas de falso negativo y falso positivo para técnicas poligráficas de asuntos múltiples fueron reportadas con .113 y .144, respectivamente. Para las técnicas de asunto único, la sensibilidad agregada de la prueba fue de .741 y la especificidad fue de .702. La tasa agregada de falsos negativos para técnicas de asunto único fue de .062, y la tasa agregada de falsos positivos para estas técnicas fue de .091. El porcentaje de decisiones correctas fue de .904 para técnicas de asunto único y de .850 para técnicas poligráficas de asuntos múltiples. Estas métricas descriptivas se muestran en el Apéndice A y se utilizaron como valores iniciales para un análisis Monte Carlo simple de los tamaños de efecto de precisión para los exámenes SSI, cuando cada caso consta de una serie de cuatro pruebas poligráficas de asunto único.

El análisis tuvo 1000 iteraciones de un espacio Monte Carlo de $n=1000$ casos, cada uno de los cuales constó de una serie de cuatro exámenes de asunto único⁷ utilizados para investigar cuatro objetivos comportamentales diferentes. Para cada iteración del espacio de Monte Carlo se realizó una clasificación única [positiva, negativa, inconclusa] para cada uno de los exámenes SSI dentro de cada caso. El resultado general de cada caso se codificó como resultado verdadero positivo, verdadero negativo, falso negativo, falso positivo e inconcluso según la serie de los resultados de las cuatro pruebas.

Se hizo una clasificación de cada caso en base a los resultados de los exámenes SSI. Los casos se clasificaron como veraces si los resultados de todos los objetivos de asunto único eran veraces y se

clasificaron como engañosos si los resultados de uno o más objetivos de asunto único eran engañosos. Este esquema de clasificación de todas-o-cualquiera es una alegórica de la regla de decisión utilizada por los examinadores de polígrafo de campo que realizan exámenes exploratorios de asuntos múltiples y permite una comparación directa e intuitiva de los tamaños del efecto entre los exámenes de asuntos múltiples y la hipótesis SSI. [Vea Nelson (2018) para una discusión sobre las reglas de decisión del polígrafo.]

Debido a que cada uno de los exámenes SSI aborda diferentes asuntos-objetivo de comportamiento, para lo que se concibe que una persona podría participar en ninguno, algunos o todos ellos, el estado del criterio para cada objetivo comportamental fue independiente para cada caso. El estado de criterio de cada caso fue inocente⁸ cuando el estado de criterio de todas las conductas objetivo fue veraz, y el estado de criterio del caso fue culpable si el estado de criterio de uno o más objetivos fue culpable. La tasa base previa de culpabilidad para todo el experimento se fijó en 0.5. Esto se hizo estableciendo la tasa base en $1-(1-.5)^{(1/4)} = .159$ para cada objetivo individual. Para cada pregunta de cada caso en cada iteración del espacio de Monte Carlo, se comparó un número aleatorio con esta previa y el estado del criterio de la pregunta se estableció como inocente si excedía el previo y culpable si no lo superaba. El estado de criterio del caso se estableció como culpable si cualquier criterio de pregunta era culpable y se estableció como inocente si todos los estados del criterio de pregunta eran inocentes. El resultado de este proceso fue que el estado del criterio fue inocente para todos los asuntos-objetivo - todos los exámenes SSI para la serie de exámenes dentro de cada caso para aproximadamente la mitad de los casos en el espacio Monte Carlo. Con numerosas iteraciones del espacio Monte Carlo, la tasa de incidencia media de casos de culpables convergió a 0.5.

Algunos métodos de análisis de datos poligráficos emplean una corrección estadística para los efectos de multiplicidad que ocurren al realizar clasificaciones de casos veraces con polígrafos exploratorios de asuntos múltiples.

⁷ O'Burke (2022) se refiere al enfoque propuesto de prueba poligráfica como prueba exploratoria de asunto único (SIST), aunque este término es potencialmente confuso porque en realidad cada caso consta de una multiplicidad de cuatro asuntos-objetivo. O'Burke intenta capitalizar la precisión de las técnicas poligráficas diagnósticas de asunto único, lo que puede ser razonable si se considera que la precisión de la técnica surge del diseño de un solo tema para el cual hay más datos disponibles para respaldar un solo resultado. Sin embargo, O'Burke pasa por alto que el enfoque de asunto único en serie sigue siendo probabilístico y todavía está sujeto a los efectos del error acumulativo bajo el requisito de múltiples resultados probabilísticos para lograr una clasificación general de caso.

⁸ En este uso, los términos "inocente" y "culpable" no se refieren a una sentencia judicial, sino que se utilizan para denotar el estado del criterio real de una manera distinta y menos fácil de combinar con la clasificación "engañoso" o "veraz", de los resultados de las pruebas.



El efecto esperado es reducir la tasa de resultados inconclusos para los casos inocentes, además de evitar una reducción de la especificidad de la prueba como resultado del requerimiento de resultados negativos para todos los objetivos de investigación para clasificar un caso como negativo⁹. Sin embargo, los tamaños de efecto agregados informados por APA (2011) no separan los resultados con y sin corrección estadística. O'Burke (2022) no describió el uso de una corrección estadística para el enfoque SSI y, por esta razón, no se utilizó ninguna corrección estadística al clasificar los casos Monte Carlo.

Además de estudiar los tamaños del efecto para los exámenes SSI con estados de criterio independientes, también se estudiaron los efectos de precisión para las pruebas poligráficas en serie en las que el estado del criterio era uniforme para la serie de exámenes dentro de cada caso. Esta

condición se parece más al uso de estrategias de pruebas en serie en otros contextos (por ejemplo, pruebas en serie para COVID-19).

Análisis

Se computarizaron varios tamaños de efecto de precisión para la hipótesis del SSI utilizando un procedimiento de arranque Monte Carlo que se muestran en la Tabla 1. Estos incluyen tasas de sensibilidad y especificidad de la prueba, además de las tasas de error de falsos negativos y falsos positivos. Se calcularon los errores estándar y los límites superior e inferior del intervalo de confianza (IC) del 90%. [Consulte Nelson (2020) para obtener una descripción de las métricas de precisión de las pruebas.] Todos los cálculos se completaron utilizando el lenguaje de computación estadística R (R Core Team, 2022).

Tabla 1. Métricas de precisión de la prueba

Métrica.	Significado
Sensibilidad (VP)	Tasa de verdaderos positivos (casos de culpabilidad que se clasifican correctamente como engañosos)
Especificidad (VN)	Tasa de verdaderos negativos (casos inocentes que se clasifican correctamente como veraces)
Falso Negativo (FN)	Tasa de falsos negativos (casos inocentes que se clasifican incorrectamente como engañosos)
Falso Positivo (FP)	Tasa de falsos positivos (casos culpables que se clasifican incorrectamente como veraces)
Culpable Inconcluso (C-INC)	Tasa de Culpables Inconclusos
Inocente Inconcluso (I-INC)	Tasa de Inocentes Inconclusos
Inconclusos No Ponderados (INC)	Tasa de Inconclusos no ponderados = $(C-INC + I-INC) / 2$
Porcentaje de Culpables Correctos (PCC)	Proporción de casos culpables correctos sin resultados inconclusos = $VP / (VP + FN)$
Porcentaje de Inocentes Correctos (PIC)	Proporción de casos inocentes correctos sin resultados inconclusos = $VN / (VN + FP)$
Precisión no ponderada (PNP)	Promedio del porcentaje de culpables correctos y del porcentaje de inocentes correctos = $(PCC + PIC) / 2$
Valor Predictivo Negativo (VPN)	Proporción de VN y todos los casos negativos = $VN / (VN + FN)$
Valor Predictivo positivo (VPP)	Proporción de VP y todos los casos positivos = $VP / (VP + FP)$

Resultados

Se analizaron dos versiones del experimento de Montecarlo. El primer experimento involucró la hipótesis SSI con cuatro estados de criterio independientes. Esta condición se asemeja al contexto exploratorio descrito por O'Burke (2022) que involucra preguntas poligráficas exploratorias sobre la posible participación de un examinado en el uso de drogas ilegales, delitos graves,

violencia doméstica y delitos sexuales. Para los examinadores de poligrafo de campo, es concebible que un examinado pueda participar en ninguna, algunas o todas estas actividades. En la práctica, los resultados de las pruebas poligráficas de asuntos múltiples se interpretan con el supuesto de que los estados de los criterios varían de forma independiente para diferentes preguntas objetivo.

⁹ No se utiliza una corrección estadística para clasificaciones engañosas con polígrafos exploratorios de asuntos múltiples porque hacerlo puede reducir la sensibilidad exploratoria. Por el contrario, se utiliza una corrección estadística cuando se hacen clasificaciones engañosas basadas en puntuaciones subtotales de polígrafos de asunto único, con el fin de reducir la inflación de errores falsos positivos. Para polígrafos de asunto único, la sensibilidad de la prueba se mantiene mediante el uso de la puntuación de gran total, para la cual no se utiliza ninguna corrección.



Esto significa que los resultados se realizan primero al nivel de cada asunto-objetivo, y luego se analizan para lograr un resultado general de la prueba. La heurística procesal es que el resultado general del caso se clasifica como veraz cuando el resultado para todos los asuntos objetivo se clasifica como veraces, y el resultado del caso general se clasifica como engaño cuando el resultado para cualquier asunto objetivo se clasifica como engaño.¹⁰

El segundo experimento involucró la hipótesis SSI con estados de criterio uniformes. Esta condición se parece más a un proceso de pruebas repetidas con el mismo objetivo de investigación. Esto es más similar a la estrategia de pruebas en serie con COVID-19, con la importante diferencia de que el objetivo aquí es observar los efectos después de una serie de cuatro pruebas.

Investigación SSI de múltiples conductas-objetivo independientes.

Los resultados del análisis Monte Carlo y Bootstrap se muestran más adelante en la Tabla 2. La aplicación de los exámenes SSI a cuatro objetivos independientes dentro de cada caso produjo una tasa de error de falsos negativos de $<.001$ y una tasa de error de falsos positivos fue de $.371$. La tasa de inconclusos para casos culpables fue de $.126$ y para casos inocentes se cargó con una tasa de $.270$.¹¹ El enfoque del SSI para cuatro objetivos independientes produjo una tasa alta de sensibilidad de prueba - la proporción de casos culpables que se clasifican correctamente - de $.874$ con un IC del 90% de $.851$ a $.897$. La especificidad de la prueba - la proporción de casos inocentes que se clasificaron correctamente - fue menor, con una media de $.350$ y un IC del 90% de $.326$ a $.394$.

La precisión no ponderada - la media no ponderada del porcentaje correcto para casos de culpables e inocentes, excluyendo resultados inconclusos - fue de $.746$, con un error estándar de $.013$ y un IC del 90% de $.725$ a $.768$ para el enfoque SSI con cuatro objetivos independientes. La tasa no ponderada de inconclusos fue de $.198$. El límite superior de especificidad de la prueba no superó el nivel de $.5$ para los exámenes SSI con

cualquiera de los estados de criterios mixtos.

Investigación SSI de conductas-objetivo uniformes

Para proporcionar más información y conocimientos sobre los tamaños del efecto de una estrategia de pruebas en serie, se completó un segundo análisis Monte Carlo, que involucró el uso de exámenes SSI para los cuales los temas objetivo de la prueba, y por lo tanto los estados de criterio, fueron uniformes para una serie de cuatro exámenes para cada caso. Es decir, se estableció un único estado de criterio para cada caso y todas las pruebas dentro de cada caso tenían el mismo estado de criterio. La tasa base previa de casos culpables se fijó para alcanzar una media de $.5$ para todas las iteraciones del modelo Monte Carlo. Los resultados analíticos de exámenes seriados de un solo tema con objetivos de prueba uniformes se incluyen en la Tabla 2. Esta condición se parece más a un proceso de pruebas repetidas con el mismo objetivo de investigación.

La aplicación del enfoque SSI para el cual el asunto-objetivo y el estado del criterio eran uniformes para todos los objetivos de la investigación dentro de cada caso, produjo una tasa de precisión no ponderada - la media no ponderada del porcentaje correcto para casos culpables e inocentes, excluyendo resultados inconclusos - de $.747$ ($.726$ a $.768$), con una tasa de inconclusos no ponderada de $.135$ ($.119$ a $.152$). Se cargaron los resultados inconclusos para los casos inocentes, con una tasa de $.001$ para casos culpables y $.269$ para casos inocentes. También se cargaron errores de clasificación para casos inocentes, con una tasa de $<.001$ para errores falsos negativos y de $.370$ para errores falsos positivos. La sensibilidad de la prueba al engaño aumentó a $.998$ para la serie de cuatro exámenes con un objetivo uniforme. La especificidad de la prueba se redujo a $.359$. Curiosamente, el límite superior del intervalo de confianza para los exámenes SSI con estados de criterio uniforme no superó el nivel de $.5$.

Discusión

La validez de la hipótesis del SSI podría indicar cierta mejora sobre el efecto de los exámenes poligráficos de asuntos múltiples. La Tabla 2

¹⁰ De manera similar a los métodos omnibus en otras áreas del análisis de datos, la interpretación de resultados negativos (es decir, veraces) no está garantizada cuando se observan resultados estadísticamente significativos o positivos (es decir, engañosos) en cualquier lugar dentro de una prueba o experimento general. En el análisis de datos científicos, análisis posterior, post hoc, cuando hay interés o necesidad de más información acerca de la localización de un resultado estadísticamente significativo. [Consulte Nelson (2018) para obtener más información sobre las reglas de decisión del polígrafo.]

¹¹ Se sabe que en la práctica de campo las tasas de inconclusos son más bajas que las reportadas en los estudios publicados. Esto se debe a que a los examinadores de polígrafo de campo se les permite utilizar métodos estandarizados y basados en evidencia para resolver resultados inconclusos, mientras que es más probable que la información de los estudios científicos se reporte sin intentar manipular o resolver resultados inconclusos.



muestra un resumen de los resultados de este estudio utilizando una serie de exámenes de asunto único con estados de criterio tanto uniformes como no uniformes. También se muestra en la Tabla 2 un resumen de los efectos de precisión para exámenes de asuntos múltiples (que se muestran en el Apéndice A). Los estados de criterio uniforme de los exámenes SSI son

ejemplos de situaciones de pruebas repetidas en las que se desean múltiples resultados de pruebas para el mismo objetivo de investigación. Los exámenes SSI con estados de criterio mixtos son un ejemplo del uso de una serie de exámenes poligráficos exploratorios de asunto único en lugar de un examen poligráfico exploratorio de asuntos múltiples.

Tabla 2. Resumen de precisión para exámenes seriados de asunto único con estados de criterios independientes (mixtos), incluida la media (SE) [IC del 90 %].

Métricas	4 Exámenes SSI (estados de criterio independiente)	4 Exámenes SSI (estados de criterio uniforme)	Criterio independiente (temas-múltiples) Técnicas PDD†
Sensibilidad (VP)	.874 (.015) [.851 to .897]	.998 (.006) [.978 to >.999]	.771 (.072) [.630 to .911]
Especificidad (VN)	.359 (.022) [.326 to .394]	.359 (.065) [.255 to .477]	.719 (.047) [.626 to .811]
Falso Negativo (FN)	<.001 (<.001) [min=<.001, Max=<.001] *	<.001 (<.001) [min = <.001, Max=.004] *	.113 (.058) [.001 to .226]
Falso positivo (FP)	.371 (.022) [.336 to .404]	.370 (.022) [.334 to .404]	.144 (.039) [.066 to .221]
Culpable Inconcluso (C-INC)	.126 (.015) [.103 to .149]	.001 (.001) [min=<.001, Max=.008] *	.112 (.051) [.013 to .212]
Inocente inconcluso (I-INC)	.270 (.020) [.235 to .304]	.269 (.020) [.235 to .302]	.136 (.031) [.076 to .196]
Inconclusos No Ponderados (INC)	.198 (.013) [.177 to .218]	.135 (.010) [.119 to .152]	.125 (.029) [.068 to .183]
Porcentaje de Culpables Correctos (PCC)	.999 (<.001) [min=>.999, Max=>.999] *	>.999 (<.001) [min=.996, Max=>.999] *	.873 (.066) [.744 to .999]
Porcentaje de Inocentes Correctos (PIC)	.492 (.026) [.450 to .536]	.493 (.026) [.452 to .536]	.831 (.043) [.746 to .915]
Precisión no ponderada (PNP)	.746 (.013) [.725 to .768]	.747 (.013) [.726 to .768]	.850 (.039) [.773 to .926]
Valor Predictivo Positivo (VPP)	.705 (.013) [.685 to .725]	.732 (.011) [.714 to .751]	.828 (.059) [.712 to .943]
Valor Predictivo Negativo (VPN)	>.999 (<.001) [min=>.999, Max=>.999] *	>.999 (.001) [min=>.999, Max=>.999] *	.878 (.049) [.782 to .973]

* Los valores mínimos y máximos observados se muestran porque los valores observados están tan cerca de los límites que los datos no se distribuyen normalmente, lo que hace que las funciones cuantiles sean algo inestables y poco informativas en este contexto.

† Mostrado anteriormente en la Tabla 1. Agregado aquí para facilitar la comparación de las medias y los IC's

Los tamaños del efecto de las técnicas poligráficas de asuntos múltiples, descritas por APA (2011) como aquellas interpretadas bajo un supuesto de varianza de criterio independiente, se muestran en el Apéndice A e incluyen una tasa de precisión no ponderada de .850 con una tasa de inconclusos de .125.¹² La sensibilidad de la prueba para exámenes de asuntos múltiples se

muestra en .771 y la sensibilidad en .719. Se informa que las tasas de inconclusos para exámenes de asuntos múltiples son .112 para casos culpables y .136 para casos inocentes. La tasa de error de falsos negativos para polígrafos de asuntos múltiples se informó en .113 y la tasa de falsos positivos se informó en .144.

¹² Al igual que con el informe de la APA (2011), los exámenes de asuntos múltiples para este análisis se caracterizan por suposiciones pragmáticas de que los estados de criterio pueden variar de forma independiente para los asuntos-objetivo dentro de cada examen. Esta suposición se manifiesta en la selección de reglas de decisión utilizadas para analizar el resultado del caso categórico a partir de los datos de la prueba. No se han demostrado diferencias en los tamaños del efecto para diferentes conductas objetivo, y no se han publicado estudios de validación del polígrafo para diferentes asuntos operacionales objetivo.



Los resultados de los experimentos del SSI mostraron un aumento en la sensibilidad de la prueba para los exámenes de SSI con estados de criterio uniforme, junto con una disminución en los errores falsos negativos para los exámenes de SSI con criterios uniformes y mixtos. La mayor diferencia observada fue una fuerte disminución en la especificidad de la prueba para el enfoque SSI. La tasa de precisión media fue de .746 para los exámenes SSI con estados de criterios mixtos y de .747 para los exámenes SSI con estados de criterios uniformes, que fue inferior a la precisión media informada para los exámenes de polígrafo formulados con preguntas destinadas a ser interpretadas con el supuesto de que los estados de criterio de las RQ's varían de forma independiente.

Todos los intervalos de confianza previamente informados que se muestran en el Apéndice A, para exámenes de asunto único y de asuntos múltiples, tienen una superposición sustancial, lo que indica que las diferencias reportadas no han sido estadísticamente significativas. El significado de esto es que cualquier intento de afirmar que los exámenes de asuntos múltiples son inherentemente menos precisos que los de asunto único, no está respaldado por evidencia científica publicada. Sin embargo, las medias o ubicaciones (estimaciones puntuales) de las distribuciones de los tamaños del efecto para los exámenes de asuntos múltiples a veces parecen ser más débiles que para los exámenes de asunto único. Varios factores pueden influir en las diferencias observadas. Estos pueden incluir factores psicológicos como la atención dividida durante las pruebas, además de los efectos de la multiplicidad estadística que agravan las tasas de error y de los resultados inclusivos cuando las clasificaciones de casos se basan en cálculos estadísticos múltiples dentro de cada caso. Otro aspecto interesante del polígrafo exploratorio de asuntos múltiples es que diferentes temas objetivo pueden tener diferentes antecedentes. Por el contrario, las pruebas de diagnóstico, en respuesta a un problema, alegato o incidente conocido, se caracterizan por preguntas que no son independientes y cuyos antecedentes se caracterizan más fácilmente como uniformes.

En ocasiones, las analogías son útiles para ilustrar un punto abstracto. Las analogías de otro contexto son especialmente útiles porque pueden ayudar a crear una perspectiva más abstracta y menos miope. Consideremos el contexto de las armas (que a veces invoca el mismo tipo de compromiso práctico y

pragmático como ciencia y tecnología). ¿Cuáles son mejores: rifles o escopetas? Intuitivamente sabemos que esto no es tan sencillo como podría parecer a primera vista. Puede resultar tentador intentar sugerir que los rifles son más precisos y, por lo tanto, mejores. Sin embargo, si mejor se define como más precisa, y si precisión se define como la capacidad de alcanzar los objetivos deseados, entonces podemos pensar en situaciones en las que las escopetas son a veces más efectivas para alcanzar un objetivo y, por lo tanto, mejores. También podemos pensar en situaciones en las que los rifles pueden ser más eficaces. El punto aquí es que la noción de mejor es a menudo una cuestión de contexto.

Para este análisis, las escopetas pueden considerarse análogas a un análisis ómnibus en la investigación científica, o a una prueba de asuntos múltiples en las pruebas poligráficas de campo, en las que se evalúa todo al mismo tiempo. Esto tiene ventajas si esperamos muchas variaciones no controladas y si existen desafíos inherentes a la repetición de la prueba *en vivo*. Las escopetas podrían no ser la solución ideal cuando hay un problema con un solo objetivo y se necesita precisión suficiente para evitar la observación de efectos no deseados. Por otro lado, los rifles podrían ser la solución preferida cuando se identifica un solo objetivo y cuando tenemos una confianza razonable de que podemos lograr un nivel de control experimental suficiente para reducir la variación aleatoria a niveles aceptablemente bajos. Aunque no es imposible, podría haber complicaciones inherentes al uso de una solución para asuntos únicos (por ejemplo, un rifle) con objetivos múltiples. Por ejemplo, las limitaciones de tiempo, efectos prácticos, los efectos del observador y otros efectos no controlados que podrían jugar un papel en los resultados después del inicio de las actividades.

En las pruebas con interacción humana y de desempeño, hay pocas pruebas y actividades en general, que puedan repetirse sin experimentar algún tipo de efecto de reexaminación o efecto de práctica. Y si bien las metáforas son útiles para presentar y socializar nuevas construcciones abstractas, será importante evitar tomar cualquier metáfora de forma literal. El objetivo de esta metáfora es crear una mejor comprensión contextual de los aspectos de la multiplicidad en las pruebas científicas. El significado práctico de esto es que las sugerencias para el uso de un solo tipo de solución en todos los contextos no están respaldadas por evidencia científica.

Se cree que los polígrafos exploratorios de asuntos múltiples tienen la ventaja de ser prácticos y económicos.



Además, tienen la ventaja de una mayor sensibilidad para la exploración de una gama más amplia de asuntos exploratorios. Una mayor sensibilidad corresponderá a una menor tasa de error de falsos negativos. La principal desventaja de los polígrafos exploratorios de asuntos múltiples es su incapacidad de señalar de manera confiable la ubicación o la causa exacta cuando se observa un resultado estadísticamente significativo. Una característica de las pruebas exploratorias es que a veces tienen como objetivo predecir de manera excesiva los problemas. Las pruebas exploratorias no pretenden, por sí mismas, hacer un diagnóstico. A menos que se diseñen cuidadosamente para controlar o evitar los efectos de la multiplicidad, las pruebas exploratorias de asuntos múltiples pueden incrementar los errores falsos positivos a niveles inesperados. Nada de esto es exclusivo del contexto poligráfico y es bien conocido por todos los investigadores que están familiarizados con los métodos de investigación que involucren un análisis multivariado.

La principal ventaja de los polígrafos de asunto único es una mayor especificidad de la prueba, en comparación con los exámenes de asuntos múltiples. Para que los resultados diagnósticos sean útiles, se requieren dos factores: 1) suficiente sensibilidad al objetivo de la prueba para que las decisiones basadas en los resultados de la prueba sean significativamente mayores que las decisiones sin los resultados de la prueba, y 2) tasas de especificidad - capacidad de excluir los casos que no ejemplifican el tema específico de preocupación - significativamente mayores que las que se pueden lograr sin el resultado de la prueba. La sensibilidad de la prueba se refiere a la capacidad de una prueba para notar u observar los fenómenos de interés cuando están presentes. La especificidad de la prueba se refiere a la capacidad de aislar el tema de interés y excluir casos que no expresan o ejemplifican el tema específico de interés.

Una prueba ideal tendría métricas muy altas tanto de sensibilidad como de especificidad, junto con tasas de error e inconclusas muy bajas. Pero la mayoría de las veces nuestras pruebas científicas no son ideales. Y como son inherentemente probabilísticas, no se espera que las pruebas científicas sean infalibles. Sólo se requiere que las pruebas científicas proporcionen estimaciones conocidas y realistas de las tasas de sensibilidad, especificidad, de error y de inconclusos. En general, las pruebas científicas son útiles si de alguna manera mejoran nuestra toma de decisiones con valor práctico.

Nuestro conocimiento de las tasas de sensibilidad, especificidad y de error de las pruebas es lo que nos permite hacer un uso

informado y estratégico de los resultados de las pruebas, a pesar de las limitaciones conocidas. El uso racional y juicioso de los resultados de las pruebas es a veces más intuitivo cuando se equilibran las tasas de sensibilidad, especificidad y de error. Por el contrario, el uso estratégico de los resultados de las pruebas científicas puede ser menos intuitivo y más difícil cuando estos están desequilibrados. Esta dificultad puede mitigarse comprendiendo las características de precisión de la prueba, incluida la fuente del desequilibrio observado. Una fuente común de precisión desequilibrada en las pruebas es el problema de las comparaciones múltiples - efectos de multiplicidad.

En términos simples, los efectos de multiplicidad reducen la precisión de las estimaciones estadísticas y las conclusiones categóricas. Esta reducción en la precisión general se produce en función del aumento de la oportunidad del error de varianza. Se puede considerar que la variación diagnóstica influye en las clasificaciones correctas. Se puede considerar que los problemas de clasificación están demasiado influenciados por el error de varianza. El error de varianza puede adoptar varias formas, que incluye el error sistemático y el error no controlado o inexplicable. El error de varianza sistemático, a veces denominado *sesgo*, está influenciado por la imperfección o la mala comprensión del conocimiento o la información disponible que se utiliza para desarrollar una prueba o experimento científico. El error no controlado a veces se denomina *error aleatorio* o error inexplicable.

La precisión o exactitud de los resultados de las pruebas o de los experimentos se logra comprendiendo y reduciendo en la medida de lo posible las posibles fuentes de error. Los métodos de prueba y de investigación que usan clasificaciones o estimaciones basadas en comparaciones estadísticas múltiples tendrán de manera inevitable más oportunidades de error que aquellos que requieren una sola comparación estadística. Sin embargo, muchos contextos de investigación y evaluación se enfrentan a la necesidad de evaluar preguntas o hipótesis múltiples, lo que requiere que los profesionales que participan en pruebas e investigaciones científicas estén adecuadamente informados sobre los efectos de la multiplicidad.

Se puede considerar que los esfuerzos para mitigar los efectos de la multiplicidad estadística se dividen en tres áreas principales. El primero de ellos es evitar, siempre que sea posible, decisiones, estimaciones y clasificaciones basadas en comparaciones estadísticas múltiples.



Esto no siempre es posible dentro del contexto de la prueba poligráfica, cuando una agencia ha identificado múltiples conductas-objetivo de interés para las decisiones exploratorias. Una segunda estrategia para mitigar los efectos de la multiplicidad en las pruebas e investigaciones científicas es hacer uso de correcciones estadísticas de los límites alfa y los puntos de corte de decisión. Algunos métodos de análisis de datos poligráficos utilizan correcciones estadísticas para reducir los efectos de la multiplicidad estadística. Un tercer enfoque para mitigar los efectos de la multiplicidad estadística y la consiguiente reducción de la precisión de los resultados analíticos es hacer uso de métodos de análisis ómnibus como la familia de ANOVA que pueden evaluar múltiples pruebas o hipótesis de investigación en un solo procedimiento analítico.

Resultados Analíticos

El enfoque SSI dio como resultado una reducción en un nivel muy bajo de los errores FN. La inspección del intervalo de confianza estadístico indica que la diferencia observada es significativa. Sin embargo, hay un aumento correspondiente en los errores de FP (media = .371, IC del 90%: .336 a .404) junto con un aumento en los resultados inconclusos para personas inocentes (media = .270, IC del 90%: .235 a .304). La inspección de los intervalos de confianza también indica que las reducciones observadas en la especificidad de la prueba, de .719 a .359 (IC del 99%: .326 a .394), y la precisión no ponderada, de .850 a .746 (IC del 90%: .725 a .768) también fueron significativos.

Los resultados del segundo análisis, que utilizó una serie de cuatro exámenes con objetivos uniformes, revelaron un patrón de efectos similar. La especificidad de la prueba se reduce significativamente y los errores de FP aumentan significativamente, junto con un aumento significativo de resultados inconclusos entre personas inocentes. El aumento significativo en la sensibilidad de la prueba es aún mayor cuando los estados de criterio son uniformes (en comparación con el enfoque SSI con estados de criterio mixtos, como se muestra en la Tabla 3). La precisión no ponderada (.747, SE=.013, IC del 90%: .726 a .768) para una serie de polígrafos de asunto único con estados de criterio uniformes fue significativamente menor que las tasas de precisión reportadas para exámenes de evento único (asunto único) específico (que se muestra en el Apéndice A).

En el contexto del examen poligráfico de asuntos múltiples, los efectos de multiplicidad estadística están presentes independientemente de si se investigan asuntos múltiples utilizando un formato de prueba de asuntos múltiples o el enfoque SSI.

Para fines prácticos, las pruebas en serie o repetidas de asuntos-objetivo uniformes pueden considerarse como asunto de confiabilidad por la repetición de las pruebas. La mayoría de los estudios de confiabilidad del polígrafo en el pasado no han abordado la confiabilidad de las reevaluaciones y, en cambio, han abordado cuestiones de confiabilidad entre evaluadores (entre calificadores). Una excepción a esta tendencia se puede ver en el análogo de Honts et al. (2015), en el que no se observaron efectos en la repetición de prueba que redujeran la disponibilidad de datos registrados en la prueba de preguntas de comparación. Un mayor uso de algoritmos de análisis automatizados puede mitigar sustancialmente la mayoría de las preocupaciones restantes sobre la confiabilidad entre calificadores. Y aunque este análisis no fue diseñado para abordar la cuestión de la confiabilidad de la reevaluación de pruebas, estos resultados sugieren la necesidad de comprender mejor la repetición de pruebas y la confiabilidad de las nuevas pruebas tanto en contextos de exploración poligráfica como de diagnóstico. Una diferencia importante entre este proyecto, que utiliza una serie de cuatro exámenes, es que el contexto de reexamen más común puede involucrar una sola repetición del examen.

Análisis auxiliares de estrategias de pruebas en serie

Para comprender mejor los efectos de la repetición de pruebas asociados con estrategias de prueba repetidas con un solo asunto-objetivo, en contraste con la hipótesis SSI aplicada con asuntos-objetivo múltiples, se completaron dos análisis auxiliares. El primer análisis auxiliar implicó un cálculo de los efectos potenciales asociados a la recomendación o requisito de resultados negativos repetidos de las pruebas durante la pandemia de COVID-19. Este tipo de requisito puede estar asociado con el procedimiento de finalización de la cuarentena o de la atención hospitalaria. Un segundo análisis auxiliar involucra el cálculo de las tasas de sensibilidad, especificidad y de error para la reexaminación de una prueba poligráfica de diagnóstico de asunto único.

Pruebas en serie para COVID-19

La reciente pandemia de COVID-19 proporciona un punto interesante de yuxtaposición para una mejor ilustración y comprensión del uso potencial de las estrategias de pruebas en serie. El gobierno de los Estados Unidos (Administración de Alimentos y Medicamentos, 2022a) extendió una autorización de uso de emergencia y autorizó además el uso de pruebas rápidas de antígenos (AR) para pruebas seriadas de COVID-19 para personas sintomáticas y asintomáticas. Poco después se proporcionó orientación adicional (Administración de Alimentos y Medicamentos, 2022b) a manera de recomendación para realizar pruebas en serie - una segunda prueba de AR de COVID-19, uno o dos días después - después de un resultado negativo de la prueba, para reducir el riesgo de error falso negativo.



Los resultados del metaanálisis (Dinnes et al., 2022) mostraron que las pruebas de AR para COVID-19, así como las aprobadas para uso doméstico, en sitios de atención, para pacientes en cuarentena, y para la aprobación de viajes, tienen una tasa de sensibilidad de .73 con personas sintomáticas y de .55 con personas asintomáticas. Se informaron tasas de especificidad para estas pruebas de $>.99$ para personas sintomáticas y $>.99$ para personas asintomáticas. La tasa correspondiente de error falso negativo fue .27 y la tasa correspondiente de error falso positivo fue .01.

También se tuvo en cuenta la tasa base o prevalencia de COVID-19. El COVID-19 se consideró un peligro para la salud pública debido a la alta tasa de transmisión junto con el alto potencial de complicaciones graves de salud, incluyendo la muerte. A pesar de estas preocupaciones, la mayoría de las personas de la comunidad no tendrían COVID-19 en ningún momento. Para este proyecto, era necesario determinar una tasa base adecuada.

Doernberg et al. (2022) encuestaron a trabajadores de la salud en San Francisco durante los primeros meses de la pandemia, de mayo a septiembre de 2020. Se descubrió que los trabajadores de la salud, que se consideraba que tenían una tasa de exposición al virus superior al promedio, tenían una prevalencia inicial de aproximadamente el 1% en ese momento. Otra consideración igualmente importante es que, a pesar de la naturaleza pandémica del COVID-19, la mayoría de las personas no darían positivo al virus. Kalish et al. (2021) informaron un análisis de datos de los primeros seis meses de la pandemia, utilizando un muestreo por cuotas de más de 9,000 personas de un grupo de

voluntarios de más de 460,000 personas, y mostraron una tasa de aproximadamente cinco casos con COVID-19 no diagnosticados por cada diagnosticado. Una consideración práctica importante es que las tasas de prevalencia o tasas base – también denominadas tasas de incidentes – están siempre en función del tiempo de riesgo. Las tasas más altas tienden a observarse durante periodos de tiempo más largos. Haciendo caso omiso del aspecto del tiempo en riesgo, solo por simplicidad, la tasa base previa para este análisis auxiliar se fijó en .05 (5%) según la información publicada disponible.

Las métricas de precisión, junto con la tasa de prevalencia estimada, se utilizaron como parámetros de entrada para el modelo Monte Carlo que se utilizó para un mejor cálculo de la precisión de las pruebas de AR de COVID-19 únicas y en serie, incluyendo el rango de confianza esperado para VPP, VPN y exactitud general de la prueba. El espacio de Monte Carlo constaba de 1000 iteraciones de $n=1000$ casos. Cada caso constó de dos pruebas de AR, para las cuales los parámetros de sensibilidad, especificidad y error de entrada se obtuvieron del metaanálisis de Dinnes et al. (2022).

Los casos en el análisis seriado de AR de COVID-19 se clasificaron como positivos si alguna de las dos pruebas daba un resultado positivo. Los resultados de la prueba COVID-19 AR no positivos se clasificaron como negativos¹³, sin zona para inconclusos. Los casos se clasificaron como negativos cuando ambos resultados de las pruebas fueron negativos. El efecto de esta regla de decisión fue aumentar tanto la sensibilidad como la especificidad de la prueba. Los resultados se muestran en la Tabla 3.

¹³ En contraste, las pruebas poligráficas con preguntas comparativas se evalúan para determinar el nivel de significancia estadística tanto para el engaño como para la veracidad. Es posible que algunos resultados del polígrafo no sean estadísticamente significativos para ninguno de los dos, para lo cual los examinadores de polígrafo de campo utilizan los términos categóricos contextuales inconcluso y no opinión.



Tabla 3. Estimación de la precisión de las pruebas rápidas de antígenos de COVID-19 con previa = 5 %, utilizadas en contextos de pruebas únicas y en serie, incluida la media (SE) y el [IC del 90 %].

	Prueba Única de AR para COVID-19	Prueba Serial de AR para COVID-19
Precisión no ponderada†	.860 (.030) [.810 to .909]	.953 (.018) [.922 to .980]
Sensibilidad	.730 (.060) [.630 to .833]	.925 (.035) [.880 to .981]
Especificidad	.990 (.03) [.984 to .995]	.980(.005) [.973 to .987]
Errores Falso Negativo	.270 (.06) [.167 to .370]	.075 (.035) [.019 to .130]
Errores Falso-Positivos	.010 (.030) [min=.001, Max=.024] *	.020 (.005) [.013 to .027]
VPP	.809 (.054) [.720 to .898]	.730 (.047) [.654 to .810]
VPN	.985 (.003) [.979 to .990]	.996 (.002) [.992 to .999]

† La precisión no ponderada puede considerarse como la estimación de la precisión si los tamaños de los grupos seropositivos y seronegativos estuvieran equilibrados. En este análisis, el porcentaje real de aciertos es menos informativo porque el 95% de los casos eran inocentes.

* Los valores mínimos y máximos observados se muestran porque los valores observados están tan cerca de los límites que los datos no se distribuyen normalmente, lo que hace que las funciones cuantiles sean algo inestables y poco informativas en este contexto.

No es sorprendente que el VPN fuera alto (.985) para una sola prueba de AR de COVID-19, debido a la alta especificidad de la prueba en combinación con una alta probabilidad previa de ser negativo y se mantuvo alto para la situación de las pruebas en serie. Sin embargo, la tasa de error FN para una sola prueba de RA fue de .27, lo que podría ser excesivo para algunos fines en el manejo de riesgos. El objetivo de la estrategia de pruebas en serie era disminuir los errores de FN. La Tabla 5 muestra que la tasa de error FN se redujo a .075 mediante la administración de una segunda prueba de AR. El aumento de los errores de FP de .01 a .02, con la correspondiente disminución del VPP, de .809 a .730 no es sorprendente y puede ser de poca preocupación en este contexto. Sin embargo, en otros contextos de prueba se podría estar preocupado por estos efectos y, por lo tanto, podría requerir una estrategia de prueba diferente.

Pruebas en serie en el contexto del polígrafo diagnóstico

Volviendo a la prueba del polígrafo, se han observado estrategias de pruebas diagnósticas en serie (PDS) en el contexto de un nuevo examen o de la reexaminación o repetición de una prueba de

polígrafo diagnóstica de evento específico. La Tabla 4 muestra los resultados Monte Carlo de la PDS de un solo tema. El modelo de Monte Carlo constaba de 1000 iteraciones de $n=1000$ casos de asunto único. Cada caso constaba de dos exámenes para los cuales el criterio de culpabilidad o inocencia era idéntico en ambos exámenes. Los parámetros de entrada se obtuvieron de la información de la Tabla 1, para exámenes de polígrafo de un solo tema.

El enfoque PDS difiere de la estrategia de exploración de obstáculos sucesivos (EOS) (Krapohl y Stern, 2003). El EOS involucra el uso en serie de pruebas exploratorias de asuntos múltiples seguido del uso de pruebas de exploratorias de asuntos únicos más precisas antes de clasificar un caso como positivo.¹⁴ En el presente análisis PDS los casos se clasificaron como engaño cuando los resultados de ambas pruebas eran engaño y se clasificaron como veraces cuando los resultados de ambas pruebas eran veraces. Este esquema de clasificación es diferente del de la hipótesis SSI en la que un único resultado positivo se utilizaría para hacer clasificaciones de engaño. La PDS se puede utilizar en casos de alto interés o de valor elevado (como aquellos destinados a ser utilizados en un



procedimiento legal) e involucra la reexaminación de un resultado de prueba de un solo tema,

positivo o negativo.

Tabla 4. Resumen de precisión para polígrafos repetidos y concordantes de eventos específicos.

Métricas	Media (EE) [IC 90%]
Sensibilidad (VP)	.704 (.021) [.668 to .738]
Especificidad (VN)	.600 (.023) [.563 to .638]
Falso negativo (FN)	.006 (.003) [min = <.001, Max=.020] *
Falso positivo (FP)	.012 (.005) [.004 to .020]
Culpables Inconclusos (C-INC)	.291 (.021) [.258 to .326]
Inocentes – Inconclusos (I-INC)	.388 (.022) [.350 to .423]
No ponderados inconclusos (INC)	.339 (.015) [.314 to .364]
Porcentaje de Culpables Correctos (PCC)	.992 (.005) [.983 to >.999]
Porcentaje de Inocentes Correctos (PIC)	.981 (.008) [.966 to .993]
Precisión No Ponderada (PNP)	.986 (.005) [.978 to .994]
Valor Predictivo Positivo (VPP)	.984 (.007) [.972 to .994]
Valor Predictivo Negativo (VPN)	.991 (.006) [.981 to >.999]

* Los valores mínimos y máximos observados se muestran porque los valores observados están tan cerca de los límites que los datos no se distribuyen normalmente, lo que hace que las funciones cuantiles sean algo inestables y poco informativas en este contexto

Los resultados que se muestran en la Tabla 4 ilustran el valor potencial de la PDS y el uso de resultados de pruebas concordantes. La precisión general aumentó a .986 con VPP = .984 y VPN = .991, junto con la correspondiente reducción a niveles muy bajos de los errores FN y FP. Sin embargo, las tasas efectivas de sensibilidad y

especificidad para la PDS fueron menores que para los exámenes individuales. Esto se relacionó con un aumento sustancial de resultados inconclusos (más de un tercio de los casos) cuando los resultados de las dos pruebas no coincidieron [Consulte Nelson (2016) y Nelson y Turner (2017)

¹⁴El EOS es análogo al uso médico de pruebas exploratorias y de diagnóstico, y análogo al uso de pruebas ómnibus y post hoc en la investigación científica.



para información y discusión acerca de los efectos de las pruebas en serie en las que no se requieren resultados concordantes - en las que se selecciona un resultado de la prueba como base para la clasificación de engaño o veracidad^{15, 16]}

Las estrategias PDS fueron descritas previamente (Handler, 2016; Kircher y Raskin, 2017). Las ventajas potenciales de la PDS fueron descritas por primera vez por Meehl y Rosen (1955). El cálculo de la precisión de dos resultados de pruebas implica el uso del coeficiente de correlación phi, y fue descrito por Nelson, Kircher y Handler (2018). Un aspecto importante de la PDS es que es indispensable que las dos pruebas sean independientes, es decir, no debe haber una fuente de varianza compartida o contaminación entre las dos pruebas. En otras palabras, independencia significa que los resultados de una prueba no pueden influir en los resultados de la otra. En la práctica, esto puede ser difícil de lograr si el examinador de la segunda prueba está informado de los resultados de la primera. Otra fuente potencial de error es la posibilidad de que existan vulnerabilidades inherentes a la metodología o la tecnología de la prueba. Por esta razón, las pruebas en serie en el contexto médico a menudo van a utilizar dos pruebas que usan diferentes datos y diferentes métodos de análisis. En el contexto poligráfico, esto podría implicar el uso en serie de dos tecnologías diferentes de prueba de evaluación de credibilidad.

Conclusión

En general, estos resultados son consistentes con los hallazgos reportados previamente (Barland et al., 1989) donde no lograron respaldar la hipótesis de SSI como ventajoso en comparación con los resultados de los polígrafos exploratorios de asunto múltiple tradicionales. Aunque no se discutieron en ese momento como fenómenos estadísticos, los resultados de Barland et al. hoy se entienden fácilmente como una cuestión de multiplicidad o error estadístico acumulado. La hipótesis del SSI es equivalente al uso de clasificaciones estadísticas múltiples, aunque de una manera ligeramente diferente a las pruebas poligráficas de asuntos múltiples y está sujeta a distorsiones y reducciones potenciales similares de la precisión analítica.

La propuesta del SSI para la detección poligráfica de asuntos múltiples es alegórico a proceder inmediatamente al análisis post-hoc en la investigación científica, sin realizar primero un análisis ómnibus. El efecto sería aumentar las tasas de error con las que estamos familiarizados (a nivel experimental). En el contexto poligráfico, el aumento de errores y resultados inconclusos se cargan hacia los casos inocentes. Las soluciones analíticas ómnibus tienen el efecto de reducir las tasas de error en los experimentos. En el contexto poligráfico, esto puede proporcionar un mejor equilibrio entre las tasas de sensibilidad y especificidad de la prueba.

Actualmente, la evidencia no da respaldo a que la hipótesis SSI proporcione una solución simple y conveniente en el contexto de la exploración poligráfica de asuntos múltiples. Este enfoque puede reducir inadvertidamente la precisión general, con el correspondiente aumento de errores y resultados inconclusos que afectan a examinados inocentes. Sin embargo, esto no implica que la hipótesis SSI no agregue un valor potencial a los programas poligráficos que entienden correctamente el efecto de la multiplicidad discutido anteriormente. Para aquellos interesados en aumentar de manera estratégica la sensibilidad de la prueba y en reducir los errores FN, y que están correctamente informados sobre la reducción de la especificidad y precisión de la prueba, el valor potencial del enfoque SSI puede ser determinado por consideraciones económicas además de las métricas de precisión de la prueba.

Una limitación obvia de este proyecto es que no investigó el uso potencial de correcciones estadísticas para reducir los efectos de multiplicidad con ninguna de las pruebas exploratorias tradicionales de asuntos múltiples.

Otra limitación obvia de este proyecto es que no investigó el uso potencial de correcciones estadísticas para reducir los efectos de multiplicidad ni con la prueba poligráfica exploratoria tradicional de asuntos múltiples ni con el enfoque SSI. Esto se debe en parte al diseño del modelo de Monte Carlo, que utiliza parámetros de siembra para tasas de sensibilidad, especificidad y de error de la prueba que no son fácilmente sujetos a corrección estadística. Otra limitación es que, al realizarse un estudio

¹⁵ En general, la clasificación de una prueba como de engaño o veracidad basándose en un resultado único de prueba en un contexto de pruebas en serie puede aumentar la probabilidad de error.

¹⁶ Nelson (2016) y Nelson y Turner (2017) también describieron los efectos cuando se utiliza el resultado de una prueba inicial como un previo bayesiano para una prueba posterior. Esto también fue discutido por Handler (2016) y Kircher y Raskin (2016).



Monte Carlo, no se involucró el análisis de datos analógicos o datos *en vivo*. Esto se ve mitigado en cierta medida con la coherencia de estos resultados con los resultados de un estudio análogo previo (Barland et. al., 1989), que comparó los resultados de pruebas poligráficas de asuntos múltiples con una serie de exámenes de asunto único.

La limitación más importante de este proyecto es que no investigó el conjunto de factores económicos que deben tener en cuenta los administradores de programas y los que crean políticas al considerar el enfoque SSI para los polígrafos exploratorios. Los factores económicos incluyen tanto la economía fiscal como la física. Si bien los factores económicos fiscales se relacionan con los costos directos de realizar múltiples exámenes poligráficos de asunto único en lugar de un solo examen de asuntos múltiples, los factores económicos físicos pueden requerir un mayor tiempo de prueba y de programación de profesionales, instalaciones e instrumentación. Los factores económicos también incluyen mayores requisitos para el control de calidad necesario para garantizar el cumplimiento de las normas en una serie de cuatro exámenes. Los factores económicos, sociales y psicológicos podrían incluir situaciones como una mayor fatiga al realizar evaluaciones y entrevistas al tener varios conjuntos de RQs y CQs, mayor oportunidad para generar errores de procedimiento, efectos de habituación, efectos de la práctica y efectos del observador, junto con una mayor oportunidad de falta de cooperación conductual.

Los costos económicos podrían incluir elementos tanto a corto como a largo plazo que afectan los resultados de las pruebas. Puede ser que los costos a corto plazo sean inmediatamente muy obvios para las agencias y los profesionales poligráficos de campo. Los resultados inconclusos provocan la repetición de pruebas y un aumento de los interrogatorios de personas inocentes. También pueden dar lugar a una mayor oportunidad o riesgo de confesiones falsas. Otro costo económico obvio es la reducción potencial del grupo de solicitantes adecuados disponibles en las agencias de seguridad pública. Los costos potenciales de los errores FN pueden incluir un mayor riesgo de daño a una agencia o comunidad y la pérdida de reputación profesional de un examinador. Los costos potenciales por errores de FP los paga el examinado. Sin embargo, esto no implica que los errores de FP estén exentos de costos. Lo más probable es que los costos económicos de los errores de FP para las agencias y las comunidades, y para la profesión del polígrafo, puedan observarse durante períodos de tiempo más largos. Corresponde a los administradores de los programas estudiar y comprender las función de los costos asociados con los errores FN y FP, e implementar estrategias

programáticas para lograr objetivos y reducir costos.

Otra limitación de este proyecto es que no se hizo ningún esfuerzo para investigar los efectos de la precisión en exámenes de asuntos múltiples con antecedentes mixtos. Es decir, cuando diferentes objetivos de prueba tienen diferentes tasas de prevalencia o incidencia. Es improbable que todos los asuntos-objetivo de investigación de los programas de exploración poligráfica tengan tasas base de ocurrencia similares. Relacionado con este tema está el hecho de que las actividades de clasificación estadística e inferencia probabilística serán propensas a tasas más altas de índice de falsos positivos (IFP, la proporción de resultados falsos positivos y todos los resultados positivos) cuando la probabilidad previa es baja, mientras que el índice falso negativo (IFN, la proporción de resultados falsos negativos y todos los resultados negativos) será mayor cuando la probabilidad previa sea alta. En muchas circunstancias de prueba - como cuando se intenta identificar o diagnosticar problemas - lo que resulta de mayor interés son la sensibilidad de la prueba y las métricas de falsos positivos. Sin embargo, en otros contextos de pruebas - como aquellos interesados en descartar un diagnóstico o problema en particular - pueden estar más interesados en la especificidad de las pruebas y las métricas de falsos negativos. Aquí es importante que los efectos de multiplicidad pueden afectar resultados de prueba tanto positivos como negativos. Lo que es igualmente importante es que los efectos de multiplicidad pueden ocurrir independientemente de lo anterior.

Una última limitación de este proyecto es que no investigó los beneficios, en términos de precisión o economía, relacionados con las prácticas de prueba post-hoc, a las que los examinadores de polígrafo de campo se refieren como pruebas de ruptura. De alguna manera similar a la investigación científica que involucra métodos analíticos omnibus y pruebas post-hoc, se ha sugerido una variedad de enfoques exploratorios con polígrafo post-hoc. Es posible que cada una de las diferentes sugerencias tenga algún valor y que la selección de un enfoque post-hoc óptimo sea una cuestión de factores económicos y contextuales.

Se deben desarrollar futuras actividades de investigación para avanzar aún más en nuestro conocimiento del papel de la multiplicidad estadística y el uso potencial de correcciones estadísticas simples en los programas poligráficos exploratorios. En este momento, los datos indican que evaluar múltiples asuntos-objetivo independientes producirá más errores de prueba y resultados inconclusos en comparación con evaluar un solo asunto-objetivo, independientemente de si se utilizan técnicas poligráficas tradicionales de asuntos múltiples o el enfoque SSI. Se necesita más información acerca de los diferentes factores económicos que se toman en consideración al desarrollar programas y políticas de práctica de campo del polígrafo. También se



necesita más investigación sobre los efectos de los exámenes de asuntos múltiples con antecedentes mixtos. Es posible que se necesite más capacitación entre los profesionales del campo del polígrafo en torno a la gestión y corrección de los efectos de multiplicidad y el uso estratégico de estrategias analíticas ómnibus y específicas. También es posible que se necesite capacitación adicional para trabajar con información previa y el uso en serie de los resultados de las pruebas en la clasificación bayesiana.

El presente análisis no es el primer intento de desarrollar o lograr resultados analíticos para acrecentar nuestro conocimiento sobre el SSI. Investigaciones anteriores (Barland et al. 1989) tampoco lograron respaldar la noción de que el enfoque SSI para la exploración de asuntos múltiples sea inherentemente superior o ventajoso.

Algunos efectos del enfoque SSI pueden ser deseables. La disminución de los errores FN y el aumento de la sensibilidad de la prueba pueden ser una ventaja útil para situaciones de exploración poligráfica de alto valor con un gran grupo de solicitantes. Sin embargo¹⁷, los factores económicos y el aumento de los errores de FP, el aumento de los resultados inconclusos y la reducción de la precisión general pueden ser problemáticos para otros programas poligráficos exploratorios¹⁸. Los datos en este momento no respaldan el abandono de las técnicas poligráficas de asuntos múltiples. Al igual que en el contexto científico más amplio, la evidencia disponible indica que sigue existiendo un valor importante en el uso estratégico de soluciones ómnibus. Este análisis sugiere que el enfoque del SSI para la exploración poligráfica de asuntos múltiples es efectivamente un examen asuntos múltiples más prolongado - no mitiga los efectos de la multiplicidad estadística y sigue estando sujeto a problemas conocidos asociados con el uso de comparaciones estadísticas múltiples.

¹⁷ Por ejemplo: seleccionar a los solicitantes de la División Uniformada del Servicio Secreto de los EE. UU., cuyos oficiales están autorizados (obligados) a patrullar áreas fuera de la Casa Blanca mientras están armados con armas automáticas o semiautomáticas. En esta situación, es posible que los directores de programas deseen utilizar el enfoque SSI para reducir los errores FN. Con un grupo suficientemente grande de solicitantes experimentados y de alta calidad, el exceso de errores de FP puede ser menos preocupante. La reducción del alfa de 0,05 a 0,01 también reduciría los errores FN, pero sin el correspondiente aumento en los errores FP.

¹⁸ Por ejemplo: las grandes agencias policiales metropolitanas o estatales, que pueden servir como puntos de entrada a la carrera para muchos profesionales encargados de hacer cumplir la ley, pueden tener un grupo de solicitantes menor que el óptimo y estar más preocupadas por los costos económicos a largo plazo del exceso en errores de FP y resultados inconclusos, además de los costos obvios asociados con los errores FN. Estas agencias pueden preferir soluciones de exploración con mayor precisión general y tasas de sensibilidad y especificidad de las pruebas más equilibradas. Las agencias de aplicación de la ley más pequeñas también pueden tener proporciones más bajas de solicitantes por contratación y pueden experimentar efectos económicos negativos debido a la mayor FP y las tasas de inconclusos del enfoque SSI. En cambio, es posible que prefieran el enfoque tradicional del polígrafo de asuntos múltiples a la selección de candidatos, en el que los errores de prueba se pueden gestionar mediante la selección de umbrales alfa y correcciones estadísticas.



Referencias

- Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. In N.J. Salkind (Ed.) *Encyclopedia of Measurement and Statistics*. Sage.
- American Polygraph Association (2011). *Meta-analytic survey of criterion accuracy of validated polygraph techniques*. *Polygraph*, 2011, 40(4) 195-305. [Electronic version] Retrieved January 15, 2023, from <http://www.polygraph.org>.
- Barland, G. H., Honts, C. R., & Barger, S. D. (1989). *Studies of the Accuracy of Security Screening Polygraph Examinations*. Department of Defense Polygraph Institute.
- Dinnes J, Sharma P, Berhane S, van Wyk SS, Nyaaba N, Domen J, Taylor M, Cunningham J, Davenport C, Dittrich S, Emperador D, Hooft L, Leeftang MMG, McInnes MDF, Spijker R, Verbakel JY, Takwoingi Y, Taylor-Phillips S, Van den Bruel A, & Deeks JJ. (2022). Cochrane COVID-19 diagnostic test accuracy group. Rapid, point-of-care antigen tests for diagnosis of SARS-CoV-2 infection. *Cochrane Database of Systematic Reviews* 2022, Issue 7. Art. No.: CD013705. DOI: 10.1002/14651858.CD013705.pub3.
- Doernberg, S. B., Holubar, M., Jain, V., Weng, Y., Lu, D., Bollyky, J. B., Sample, H., Huang, B., Craik, C. S., Desai, M., Rutherford, G. W., & Maldonado, Y. (2022). CHART Study Consortium. Incidence and Prevalence of Coronavirus Disease 2019 Within a Healthcare Worker Cohort During the First Year of the Severe Acute Respiratory Syndrome Coronavirus 2 Pandemic. *Clinical Infectious Diseases*, 75(9): 1573-1584. Doi: 10.1093/cid/ciac210. PMID: 35279023; PMCID: PMC8992269.
- Food and Drug Administration (November 1, 2022a). [Letter re: Revisions Related to Serial (Repeat) Testing for the EUAs of Antigen IVDs]. Retrieved from <https://www.fda.gov/media/162799/download>.
- Food and Drug Administration (November 17, 2022b). *At-Home COVID-19 Antigen Tests-Take Steps to Reduce Your Risk of False Negative Results: FDA Safety Communication*. <https://www.fda.gov/medical-devices/safety-communications/home-covid-19-antigen-tests-take-steps-reduce-your-risk-false-negative-results-fda-safety>.
- Handler, M. (2016). Low base rate screening survival analysis & successive hurdles. *Police Polygraphist*, 2016(March), 31-38.
- Honts, C. R., Handler, M., Shaw, P., & Gougler, M. (2015). The vasomotor response in the comparison question test. *Polygraph*, 44 (1), 62-78.
- Kalish, H., Klumpp-Thomas, C., Hunsberger, S., Baus, H. A., Fay, M. P., Siripong, N., Wang, J., Hicks, J., Mehalko, J., Travers, J., Drew, M., Pauly, K., Spathies, J., Ngo, T., Adusei, K. M., Croker, J. A., Li, Y., Graubard, B. I., ... Czajkowski, L. (2021). Undiagnosed SARS-CoV-2 seropositivity during the first 6 months of the COVID-19 pandemic in the United States. *Science Translational Medicine*, 13 (601), eabh3826. DOI: 10.1126/scitranslmed.abh3826.
- Kircher, J. C. & Raskin, D. C. (2016). Laboratory and field research on the ocular motor test. *European Polygraph*, 10(4), 159-172.
- Krapohl, D. J. & Stern, B. A. (2003). Principles of multiple-issue polygraph screening: A model for applicant, post-conviction offender, and counterintelligence testing. *Polygraph*, 32, 201-210.



- Meehl, P. & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194–216.
- Nelson, R. (2015). Bonferroni and Sidak corrections for multiplicity effects with subtotal scores of comparison question polygraph tests. *Polygraph*, 44(2), 162-167.
- Nelson, R. (2016). Five-minute science lesson: a second look at successive hurdles screening. *APA Magazine*, 49(1): 115-122.
- Nelson, R. & Turner, F. (2017). Bayesian probabilities of deception and truth-telling for single and repeated polygraph examinations. *Polygraph & Forensic Credibility Assessment*, 46 (1): 53-80.
- Nelson, R., Kircher, J. & Handler, M. (2018). How to calculate the expected agreement and the combined accuracy of two test results. *Polygraph & Forensic Credibility Assessment*, 47(1): 18-25.
- Nelson, R. (2018). Practical polygraph: a survey and description of decision rules. *APA Magazine*, 51(2), 127-133.
- Nelson, R. (2020) Five-minute science lesson: test accuracy metrics. *APA Magazine*, 53(3), 53-57.
- O’Burke, J. P. (2022). Using single-issues in screening examinations. *APA Magazine*, 55(6), 51-59.
- Podlesny, J. A. & Truslow, C. M. (1993). Validity of an expanded-issue (modified general question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology*, 78, 788-797.
- Prado, R. Grajales, C. & Nelson, R. (2015a). Laboratory study of directed lie polygraphs with Spanish speaking examinees. *Polygraph*, 44(1), 79-90.
- Prado, R. Grajales, C. & Nelson, R. (2015b). Laboratory study of a diagnostic polygraph in a single sequence: a replication. *Polygraph*, 44(2), 1-12.
- Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (1988). *A Study of Polygraph Examination in Criminal Investigation*. Final Report to the National Institute of Justice Grant No. 85-IJ-CX-0040. University of Utah, Department of Psychology. Reprinted in *Polygraph & Forensic Credibility Assessment*, 48 (1): 10-39.
- R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Research Division Staff (1995a). *A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope Polygraph and the test for espionage and sabotage question formats*. Report number DoDPI94-R-0008. DTIC AD Number A319333. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in *Polygraph*, 26 (2), 79-106.
- Research Division Staff (1995b). *Psychophysiological detection of deception accuracy rates obtained using the test for espionage and sabotage*. DoDPI94-R-0009. DTIC AD Number A330774. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in *Polygraph*, 27, (3), 171-180.
- Sidak, Z. (1967). Rectangular confidence region for the means of multivariate normal distribution. *Journal of the American Statistical Association*, 62, 626-633.



Apéndice A.

Resumen reportado previamente acerca de métricas de precisión del polígrafo

Tabla 1. Métricas de precisión informadas anteriormente, incluida la media (SE) y [IC del 95 %]. Tomado de la encuesta meta analítica de técnicas poligráficas validadas de la APA (2011). Tabla 7: precisión de los criterios para técnicas de polígrafo de asuntos múltiples (independientes del criterio) y de asunto único (no independientes).

	Criterio independiente Técnicas PDD	No-independiente Técnicas PDD
Porcentaje Correcto	.850 (.039) [.773 to .926]	.896 (.030) [.837 to .955]
Inconclusos	.125 (.029) [.068 to .183]	.106 (.031) [.044 to .167]
Sensibilidad	.771 (.072) [.630 to .911]	.840 (.050) [.743 to .938]
Especificidad	.719 (.047) [.626 to .811]	.775 (.059) [.658 to .891]
Errores FN	.113 (.058) [.001 to .226]	.074 (.032) [.011 to .138]
Errores FP	.144 (.039) [.066 to .221]	.109 (.041) [.029 to .189]
Inc C	.112 (.051) [.013 to .212]	.089 (.039) [.011 to .166]
Inc V	.136 (.031) [.076 to .196]	.122 (.049) [.027 to .218]
VPP	.828 (.059) [.712 to .943]	.893 (.039) [.816 to .969]
VNP	.878 (.049) [.782 to .973]	.910 (.043) [.826 to .995]
C Correctas	.873 (.066) [.744 to .999]	.919 (.036) [.849 to .989]
V Correctas	.831 (.043) [.746 to .915]	.873 (.047) [.780 to .965]

