Estudio Monte Carlo en Técnicas Poligráficas de Asuntos Múltiples con Dos, Tres y Cuatro Preguntas

Raymond Nelson[1], Mark Handler[2], Stuart Senter[3]

Abstract

Se utilizaron los métodos Monte Carlo y el Análisis de Varianza (ANOVA) para estudiar la precisión del criterio de los exámenes PDD de asuntos múltiples con dos, tres y cuatro preguntas relevantes (RQs) - como los que se realizan utilizando el USAF MGQT - cuando son calificados con los métodos de siete posiciones, tres posiciones y el Sistema Empírico de Puntuación (ESS). La sensibilidad al engaño de la prueba superó el azar (.5) con todos los métodos de puntuación para dos, tres y cuatro RQs. Se observaron algunas diferencias en los distintos tratamientos, con tasas de inconclusos que disminuían dependiendo del número de RQs en los casos con criterio de engaño y aumentaban dependiendo del número de RQs para los casos con criterio de veracidad. La especificidad hacia la veracidad de la prueba fue significativamente mayor que el azar sólo para el modelo de 2 RQ con puntuaciones ESS. No se encontraron diferencias significativas en las tasas de falsos positivos o negativos con las puntuaciones de siete posiciones, tres posiciones o ESS con dos, tres o cuatro RQs. Sin embargo, la probabilidad de error de la prueba aumentó dependiendo del número de RQs para los casos con criterio de veracidad, mientras que disminuyó para los casos con criterio de engaño. Excluyendo los resultados inconclusos, la precisión media no ponderada de las decisiones para los casos con criterio de engaño y veracidad superó el azar, y no se observaron diferencias significativas en la precisión no ponderada para los tres métodos de puntuación con dos, tres y cuatro RQs. En este estudio no fue posible determinar si esta diferencia se debió al método de puntuación o al uso de una puntuación de corte con referencia a la norma y a la corrección de multiplicidad para las puntuaciones de corte ESS en comparación con las puntuaciones de corte tradicionales.

[1] Raymond Nelson es especialista en investigación de la empresa Lafayette Instrument Company (LIC) y miembro electo de la Junta Directiva de la APA. Las opiniones expresadas en este trabajo son las del autor y no las de la LIC o la APA. El Sr. Nelson es psicoterapeuta, examinador de campo del polígrafo, desarrollador del algoritmo de calificación OSS-3 y autor de publicaciones sobre diversos temas relacionados con el polígrafo. Para más información, póngase en contacto con raymond.nelson@gmail.com.

[2] Mark Handler es un experimentado examinador policial e investigador de polígrafo que ayudó a desarrollar el Objective Scoring System, versión 3 y el Empirical Scoring System. Su dirección de correo electrónico es polygraphmark@gmail.com.

[3] Stuart Senter trabaja en el Centro Nacional de Evaluación de la Credibilidad (NCCA). Las opiniones expresadas en este trabajo no reflejan las del NCCA.

[4] Los autores agradecen a Joseph Stainbeck IV, que revisó una versión previa de este manuscrito.

Introducción[4]

Comúnmente, los polígrafos de asunto múltiple se utilizan en polígrafos exploratorios - en ausencia de una acusación o incidente conocido, usando dos, tres y cuatro preguntas relevantes (RQs). El United States Air Force Modified General Question Test (USAF MGQT) (Departament of Defense, 2006; Nelson, Blalock & Handler, 2011; Nelson, Handler, Morgan & O'Burke, 2012; Senter, Waller & Krapohl, 2008) – del que existen dos versiones en la práctica de campo – es un ejemplo de una prueba poligráfica que puede utilizarse con dos, tres y cuatro RQs. Existen también otros formatos poligráficos para asuntos múltiples. El polígrafo de asuntos múltiples puede ser considerado como una variación contemporánea de la técnica de preguntas de comparación descrita por Reid (1947) y Summers (1939). La característica que describe a los polígrafos de asuntos múltiples - que incluye al USAF MGQT y a otros formatos - es que se asume que las preguntas relevantes (RQ) son independientes[5].

Los autores no encontraron estudio alguno publicado que describiera la precisión del criterio de esta técnica al variar o comparar el número de RQs. El presente estudio es un esfuerzo exploratorio para ampliar nuestra base de conocimientos sobre las diferencias que pueden observarse en la precisión del criterio, dependiendo del número de RQs. La hipótesis fue que los polígrafos de asuntos múltiples, con dos, tres y cuatro RQs pueden alcanzar tasas de precisión en su clasificación que son mayores que el azar (50%) cuando se evalúan con los métodos de 7 posiciones, 3 posiciones y ESS. Lo mismo puede afirmarse en términos de los errores de prueba: la hipótesis es que los polígrafos de asunto múltiple con dos, tres y cuatro RQs pueden lograr tasas de error falso-positivo y falso-negativo que son significativamente menores que el azar.

Método

Se utilizaron métodos Monte Carlo para calcular los intervalos de confianza para la exactitud de criterio de los exámenes poligráficos de asunto múltiple con dos, tres y cuatro RQs, incluyendo el de la sensibilidad, la especificidad, las tasas de error de falsos positivos y falsos negativos, junto con la exactitud de la decisión no ponderada y las tasas de inconclusos de la prueba. Los datos se puntuaron e interpretaron utilizando los métodos de análisis de datos de prueba de siete y tres posiciones (Departament of Defense, 2006; Harwell, 2000; Krapohl, 1998; Van Herk, 1990) y el Sistema de Puntuación Empírica (ESS; Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson, & Hicks, 2011; Krapohl, 2010; Nelson, Blalock &

5 La independencia en las pruebas científicas, se refiere a la suposición de que la varianza del criterio o del estado externo de cada estímulo individual de prueba no se ve afectado ni afecta la varianza de criterio de otros estímulos de la prueba. La varianza de criterio está relacionada con la varianza de respuesta, pero es distinta de ella. En la práctica, se asume que tanto los exámenes multifacéticos como los de asunto múltiple se componen de estímulos independientes, por lo que ambos tipos se puntúan e interpretan utilizando las puntuaciones subtotales de las preguntas, aun cuando la independencia de las puntuaciones subtotales de los exámenes multifaceta no ha sido respaldada por estudios previos.

Handler, 2011; Nelson, Blalock, Oelrich & Cushman. 2011; Nelson & Handler, 2010; Nelson et al., 2011; Nelson & Krapohl, 2011; Nelson, Krapohl, & Handler, 2008). Se construyeron modelos Monte Carlo para los tres métodos de puntuación y cada uno de ellos se evaluó utilizando dos, tres y cuatro RQ. Además de estos nueve modelos, se definieron tres modelos Monte Carlo adicionales para evaluar la eficacia de los métodos de puntuación de siete posiciones, tres posiciones y ESS, variando aleatoriamente el número de RQs.

El espacio Montecarlo constó de N = 100 exámenes simulados de asuntos múltiples, para los que se estableció un estado de criterio de forma independiente para cada RQ, comparando un número aleatorio contra una tasa base fija. Se crearon modelos Montecarlo independientes para exámenes con dos, tres y cuatro RQ, y el número de RQ fue uniforme dentro de cada espacio Montecarlo. Se simuló 10.000 veces cada espacio Montecarlo para crear tres distribuciones de resultados Montecarlo - para dos, tres y cuatro RQ – para que pudieran estudiarse con respecto a su precisión en la decisión, errores y resultados inconclusos. Cada distribución Monte Carlo se evaluaría con los métodos de puntuación de siete posiciones, tres posiciones y ESS.

Las puntuaciones subtotales se simularon mediante la estandarización de números aleatorios para cada uno de los parámetros, que fueron las medias y las desviaciones estándar de las puntuaciones subtotales obtenidas por los participantes del estudio de Krapohl y Cushman (2006), después de transformar estas puntuaciones subtotales de los casos de culpables e inocentes desde siete posiciones hacia puntuaciones de tres posiciones y luego a puntuaciones ESS[6]. Krapohl (2010) y Robertson (2012) demostraron que las puntuaciones transformadas a ESS, pueden extraer datos fisiológicos similares a las puntuaciones manuales de 7 y 3 posiciones.

La tabla 1 muestra los parámetros de entrada, las medias y desviaciones estándar del subtotal para las puntuaciones de la muestra Montecarlo. El diseño de este espacio Monte Carlo significaba que el estado del criterio era aleatorio, independiente y conocido para cada RQ en el espacio Monte Carlo, y el número de RQs podía ser manipulado para evaluar los tamaños del efecto.

6 Los casos del ZCT Federal en Krapohl y Cushman (2006) tenían tres preguntas relevantes que se referían a la participación del examinado en una acusación única o incidente conocido. El uso tradicional del ZCT Federal incluía dos preguntas relevantes que describen el comportamiento del examinado, mientras que la tercera pregunta relevante se utiliza para describir el conocimiento del examinado de los detalles incriminatorios del incidente o la acusación. Sin embargo, todas las preguntas relevantes se interpretan de manera uniforme o no independiente cuando se utiliza el ZCT Federal, y no hay publicación existente que describiera los tamaños del efecto para el tratamiento o la interpretación independiente de las preguntas del ZCT Federal. Uno de los casos de la muestra de Marín incluía solamente dos preguntas relevantes. Se utilizaron un total de 299 puntuaciones subtotales, consideradas como uniformes entre inocentes o culpables para las semillas Monte Carlo para los casos de asuntos múltiples en el modelo Monte Carlo. Mientras que el uso tradicional del ZCT federal involucra tanto las puntuaciones de gran total como subtotales, en el presente estudio Montecarlo solamente se utilizó la información subtotal para los parámetros semilla.

Tabla 1. Medias de subtotales y desviaciones estándar.

	Media Engaño	DS Engaño	Media Veracidad	DS Veracidad
7-posiciones	-2.827	4.504	3.556	3.766
3-posiciones	-1.886	3.161	2.427	2.557
ESS	-3.031	4.535	3.265	3.661

Para cada espacio Montecarlo se calculó individualmente la tasa base para el engaño y la veracidad para cada RQ de forma individual, utilizando la inversa de la corrección de Šidák (Abdi, 2007; Šidák, 1967) para comparaciones estadísticas múltiples bajo una condición de varianza independiente (Abdi, 2007). Las tasas base para las preguntas individuales fueron las siguientes: dos RQs = .293, tres RQs = .206, y cuatro RQs = .159. Para cada RQ de cada caso se hizo la comparación de un número aleatorio uniforme contra la tasa base, y el estado del criterio se estableció como veraz si la tasa base era menor que el número aleatorio. Esto garantizó una tasa base para cada distribución Montecarlo que tenía convergencia a .5, al tiempo que se fijaba aleatoriamente el estado del criterio para cada RQ y se permitía la variación de la tasa de incidencia observada de engaño y veracidad para cada iteración de los casos en el espacio Montecarlo. Para cada examen en cada uno de los espacios Montecarlo, el estado del criterio de cada caso se estableció como de engaño si el estado del criterio de una o más de las RQs era de engaño. Los estados de criterio de los casos se establecieron como veracidad si el estado de criterio de todas las RQs era de veracidad.

Se utilizaron puntuaciones de corte tradicionales para los métodos TDA de siete y de tres posiciones: los resultados de las pruebas se clasificaron como engaño cuando cualquier puntuación subtotal era de -3 o inferior, y los resultados de las pruebas se clasificaron como veraces cuando todas las puntuaciones subtotales eran mayores o iguales a +3. Cabe señalar que estas puntuaciones de corte tradicionales no se basan en datos normativos, sino que se derivaron a través de la experiencia y el estudio heurístico y son similares a las puntuaciones de corte que se derivan de procedimientos estadísticos (Nelson, et al., 2011; Nelson, 2017; Nelson & Rider, 2018).

Las puntuaciones de corte para los exámenes ESS de los exámenes USAF MGQT se basan en distribuciones estadísticas de referencia para las puntuaciones subtotales individuales de personas culpables e inocentes (Nelson et al., 2011, Nelson, 2017, Nelson & Rider, 2018). La principal diferencia entre las puntuaciones de corte del ESS y las puntuaciones de corte tradicionales es que las puntuaciones de corte del ESS se logran al utilizar una corrección de Šidák para tomar en cuenta los efectos esperados de multiplicidad, que son el resultado del procedimiento requerido de que todas las puntuaciones subtotales deben ser estadísticamente significativas de veracidad para poder clasificar un resultado de prueba como veraz. Las puntuaciones de corte del ESS fueron de -3 y +1, lo que significa que los resultados de las pruebas se clasificarían como engaño cuando cualquier puntuación subtotal fuera de -3 o inferior y se clasificarían como veraces cuando todas las puntuaciones subtotales fueran de +1 o superiores.

Todos los casos en el espacio Monte Carlo se evaluaron utilizando la regla de puntuación subtotal (SSR; Departament of Defense, 2006a, 2006b; Capps & Ansley 1992; Senter Waller & Krapohl; 2008) en la cual el resultado global de la prueba se hereda de la puntuación subtotal/pregunta más baja - mientras que los resultados para cada pregunta en los exámenes de diagnóstico de eventos específicos heredan del resultado global de la prueba [Véase Nelson, Blalock & Handler, 2019 para más información]. Los resultados de las pruebas PDD se clasifican para toda la prueba en su conjunto, independientemente de si la decisión se toma utilizando las puntuaciones totales o subtotales. En términos prácticos, la rúbrica de procedimiento para el SSR, es que los resultados del examen se clasifican como indicativos de engaño - utilizando normalmente el término *reacciones significativas* - siempre que cualquier puntuación subtotal iguale o supere la puntuación de corte para las clasificaciones de engaño, y se clasifican como indicativos veracidad - utilizando el término *sin reacciones significativas*- cuando todas las puntuaciones subtotales igualan o superan la puntuación de corte para las clasificaciones veraces. Los resultados del examen se clasifican como inconclusas o sin opinión (es decir, no son estadísticamente significativos para el engaño o la veracidad) cuando ninguna de las puntuaciones subtotales iguala o supera la puntuación de corte para la clasificación de engaño, y no todas las puntuaciones subtotales igualan o superan la puntuación de corte para las clasificaciones veraces.

Investigaciones previas (Barland, Honts y Barger, 1989; Podlesney y Truslow, 1993; Departament of Defense, 1995a; 1995b) no dieron soporte a la hipótesis de una sensibilidad o especificidad de prueba a un nivel individual en las RQ, y las prácticas de campo dictan que los examinadores no están autorizados a tomar decisiones tanto de engaño como de veracidad dentro de un examen individual. Por este motivo, no se ha intentado reportar engaño en algunas RQs y veracidad en otras dentro de mismos casos en el espacio Monte Carlo.

Resultados

Se calculó la precisión del criterio para cada una de las tres condiciones del USAF MGQT (es decir, dos, tres y cuatro RQ) para los tres métodos de análisis de datos de la prueba (es decir, siete, tres posiciones y ESS). Los índices de exactitud de interés incluyeron: sensibilidad de la prueba al engaño, especificidad de la prueba a la veracidad, tasas de error falso-negativo y falso-positivo, y tasas de inconclusos para los casos de engaño y veracidad. Valor predictivo positivo (VPP; calculado como verdaderos positivos divididos entre todos los resultados positivos), valor predictivo negativo (VPN; calculado como verdaderos negativos divididos entre todos los resultados negativos), las proporciones de decisiones correctas sin resultados inconclusos para los casos con engaño y veracidad, junto con la media no ponderada de las proporciones de decisiones correctas y resultados inconclusos para los casos de engaño y veracidad. Todos los análisis estadísticos se realizaron con un nivel de significancia fijado con un alfa = 0.05. Estos se encuentran en los Apéndices A a D.

Precisión de la decisión para los exámenes USAF MGQT con dos, tres y cuatro RQs.

Los efectos de la precisión de la prueba se evaluaron mediante una prueba de hipótesis de Monte Carlo. Este método involucra el uso de métodos Monte Carlo para calcular el intervalo de confianza estadístico (Efron y Hastie, 2016; Efron y Tibshirani, 1986; 1993), que luego se compara contra el valor de la hipótesis nula o del azar (es decir, .5). Los resultados se interpretan como estadísticamente no significativos cuando el valor del azar no está contenido dentro del intervalo de confianza, o cuando los límites del intervalo de confianza 1 - alfa superan el valor de azar.

Los intervalos de confianza Monte Carlo se calcularon como el percentil alfa/2 = 0.025 y 1-alfa/2 = 0.975 de 10.000 iteraciones de un espacio Montecarlo compuesto por n = 100 exámenes simulados de asuntos múltiples. Se realizaron simulaciones Montecarlo independientes para exámenes de asuntos múltiples con dos, tres y cuatro RQ. Se completaron nueve simulaciones Montecarlo diferentes. Además, se calculó una simulación 10^{th} Montecarlo con un número aleatorio de dos a cuatro RQs.

Para cada simulación Montecarlo, se calculó la precisión de criterio para cada iteración del espacio Montecarlo, que incluyó la sensibilidad, la especificidad, las tasas de error de los falsos positivos y de los falsos negativos de la prueba, junto con el valor predictivo positivo, el valor predictivo negativo, la exactitud de la decisión no ponderada y las tasas de inconclusos para los casos con engaño y veraces de los datos observados. También se calculó la desviación estándar media para la precisión del criterio de cada dimensión, de modo que también se pudieron calcular los ANOVAs factoriales dependiendo del número de RQs x método de puntuación x estado del criterio.

Los resultados se muestran en los Apéndices A, B y C para polígrafos de asuntos múltiples de dos, tres y cuatro RQs. El Apéndice D muestra los resultados para los casos dentro de cada iteración del espacio Monte Carlo, al tiempo que varía el número de RQs.

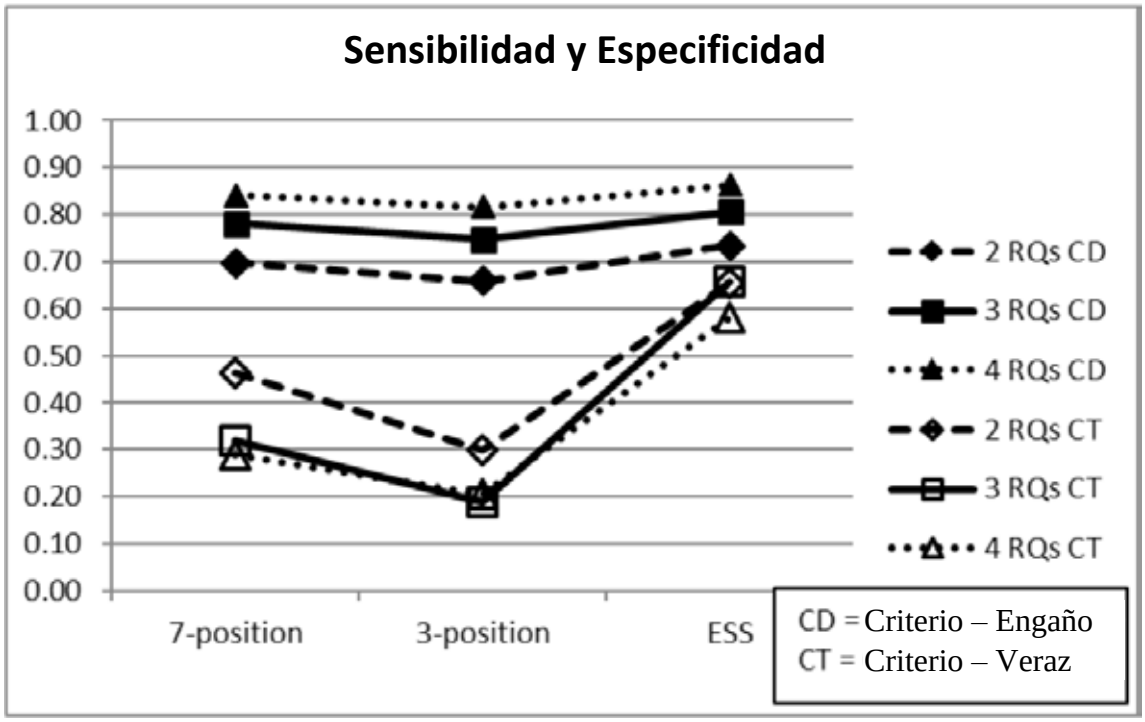
Sensibilidad y especificidad para los exámenes USAF MGQT con dos, tres y cuatro RQs.

Se utilizó el método descrito por Cohen (2002) - junto con los tamaños de las medias de las muestras en el espacio Montecarlo (una media n=50 para los casos de engaño y n=50 para los casos veraces), y las medias y desviaciones estándar Montecarlo - para calcular un ANOVA de tres vías (estado del criterio x método TDA x número de RQs) para la precisión de la decisión, incluyendo los resultados inconclusos (es decir, sensibilidad y especificidad de la prueba). La tabla 2 muestra el resumen del ANOVA de tres vías, y la figura 1 muestra el gráfico de medias para la sensibilidad y la especificidad de la prueba. La interacción de tres vías fue significativa $F(4,882) = 5.705$, $p < .001$). Este resultado indica que pueden existir diferencias en la eficacia de los métodos de puntuación de tres posiciones, siete posiciones y ESS en exámenes con criterio de engaño y veracidad con dos, tres o cuatro preguntas relevantes.

Tabla 2. Resumen de precisión para ANOVA de 3 vías (número de RQs x método TDA x estado del criterio).

Source	SS	df	MS	F	p	F crit .05
# RQs	0.048	2	0.024	2.684	.069	3.006
Status	4.368	1	4.368	486.435	<.001	3.852
Model	8.240	2	4.120	458.787	<.001	3.006
# RQs x Status	28.203	2	14.102	1570.261	<.001	3.006
Status x Model	4.629	2	2.314	257.719	<.001	3.006
# RQs x Model	0.170	4	0.042	4.731	.001	2.382
# RQs x Status x Model	0.204	4	0.051	5.669	<.001	2.382
Error	7.921	882	0.009			
Total	53.784	899				

Figura 1. Gráfico de medias para sensibilidad y especificidad de prueba para los métodos de calificación de 3, 7 posiciones y ESS.



La figura 1 muestra que la sensibilidad media de la prueba hacia el engaño superó el azar (.5) con los tres métodos de puntuación, mientras que la especificidad media de la prueba hacia la veracidad no superó el azar para los métodos de puntuación de siete y tres posiciones.

Traductor. rodolfo@poligrafia.com.mx

Debido a que el ANOVA de 3 vías fue significativo, se completaron por separado ANOVAs post-hoc de 2x2 (método TDA x número de RQs) para los casos de engaño y veracidad en el modelo Monte Carlo. El ANOVA de 2 vías, mostrado en la Tabla 3, fue estadísticamente significativo para los casos de engaño $F(1,441) = 4.848$, $p = .028$), indicando una interacción entre el modelo TDA y el número de RQs. Los ANOVAs de una vía no fueron significativos para el número de RQs ($p = .071$) o para el método de puntuación ($p = .625$) con los casos de engaño.

Tabla 3. Resumen de precisión con ANOVA de dos vías con casos de engaño (modelo TDA x número de RQs).

Source	SS	df	MS	F	p	F crit .05
Model	0.277	2	0.002	0.942	.391	3.016
# RQs	1.564	2	0.010	5.327	.005	3.016
Interaction	0.009	1	0.009	4.848	.028	3.863
Error	0.863	441	0.002			
Total	1.850	446				

Los resultados de un ANOVA de dos vías para los casos veraces se muestran en la Tabla 4. La interacción del método TDA x el número de RQs fue significativa para los casos veraces $F(1,441) = 5.669$, $p = <.001$). Los ANOVAs de una vía mostraron que los efectos principales para los casos veraces no fueron significativos con respecto al número de RQs ($p = 0,799$) o del método de puntuación ($p = 0,056$).

Tabla 4. Resumen de precisión para ANOVA de dos vías con casos veraces (modelo TDA x número de RQs).

Source	SS	df	MS	F	p	F crit .05
Model	12.593	2	0.084	5.428	.005	3.016
# RQs	1.044	2	0.007	0.450	.638	3.016
Interaction	0.364	1	0.364	23.541	<.001	3.863
Error	6.821	441	0.015			
Total	14.001	446				

Estos resultados sugieren que la principal fuente de varianza en la interacción de tres vías puede atribuirse a las diferencias en la capacidad de los tres métodos de puntuación para detectar engaño y veracidad. Para comprender mejor la influencia del método de puntuación en la precisión de la decisión, se calculó un último contraste de tres vías para los resultados de siete y tres posiciones, excluyendo los resultados del EES. La interacción de tres vías para el número de RQs x método de puntuación x estado del criterio no fue significativa [$F(4,588) =$

0.916, p = 0.454] cuando se excluyeron los resultados del ESS. Esto sugiere que la interacción inicial de tres vías puede atribuirse a las diferencias en la precisión de la decisión para los resultados de ESS con casos veraces.

Tasas de inconclusos para los exámenes USAF MGQT con dos, tres y cuatro RQs.

Se realizó un ANOVA de tres vías (estado del criterio x modelo TDA x número de RQs) para los resultados inconclusos. El resumen del ANOVA de tres vías para los resultados inconclusos se muestra en la Tabla 5. La interacción de tres vías para los resultados inconclusos fue significativa $F(4,882) = 2.580$, $p = .036$ para el método TDA x número de RQs x estado de criterio.

Tabla 5. Resumen de resultados inconclusos para ANOVA de tres vías (RQs x método TDA x estado de criterio)

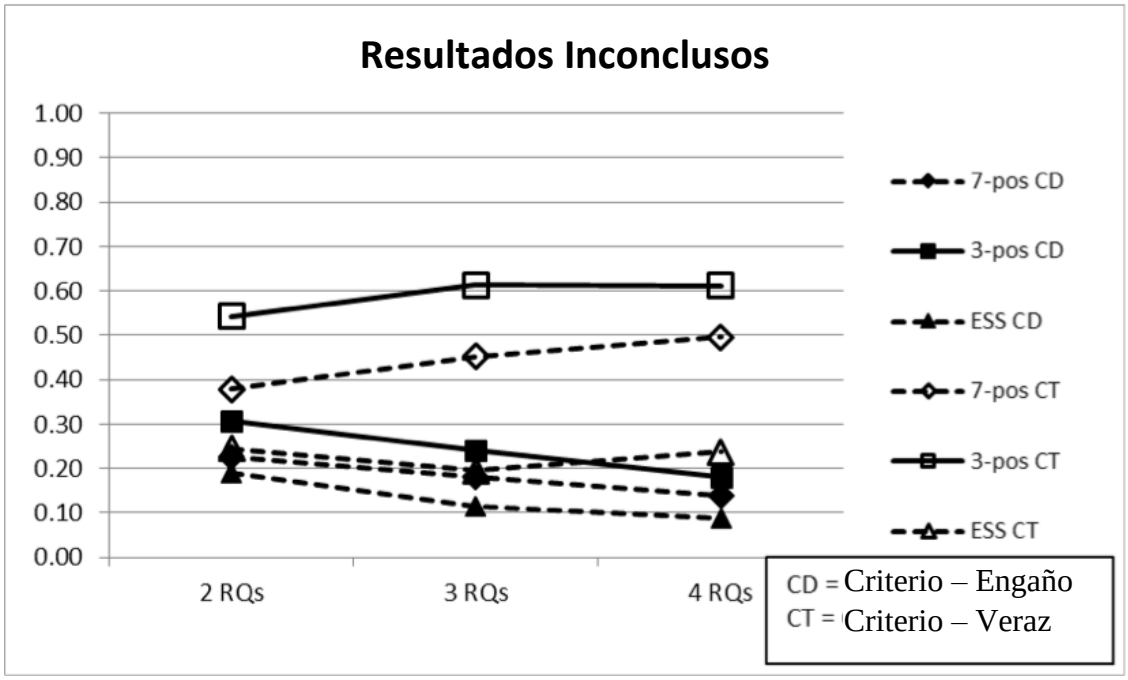
Source	SS	df	MS	F	p	F crit .05
# RQs	0.076	2	0.038	3.103	.045	3.006
Status	1.786	1	1.786	145.371	<.001	3.852
Model	8.482	2	4.241	345.097	<.001	3.006
# RQs x Status	11.506	2	5.753	468.140	<.001	3.006
Status x Model	2.468	2	1.234	100.431	<.001	3.006
# RQs x Model	0.228	4	0.057	4.638	.001	2.382
# RQs x Status x Model	0.127	4	0.032	2.580	.036	2.382
Error	10.839	882	0.012			
Total	35.512	899				

La figura 2 muestra el gráfico de medias de los resultados inconclusos para los casos de engaño y veracidad, con los métodos de siete posiciones, tres posiciones y ESS con dos, tres y cuatro RQ. En general, las tasas medias de los resultados inconclusos fueron más altas para los casos veraces que para los de engaño, y esta diferencia fue más pronunciada con los métodos de tres y de siete posiciones. Los efectos de la media simple no fueron significativos para las diferencias en los resultados inconclusos con el método de siete posiciones ($p = 0.156$) o para el ESS ($p = 0.415$). El efecto de la media simple fue significativo para los resultados inconclusos del método de puntuación de tres posiciones con casos con criterio de engaño y veracidad [$F(1,98) = 4.382$, ($p = 0.039$)].

Los ANOVAs de dos vías para cada método de puntuación mostraron una interacción significativa con respecto al número de RQs x el estado del criterio, incluyendo el de siete posiciones [$F(1,294) = 31.435$, ($p < .001$)], el de tres posiciones [$F(1,294) = 37.143$, ($p < .001$)] y el ESS [$F(1,294) = 17.702$, ($p < .001$)]. Los efectos principales simples para los resultados

inconclusos, en función de las RQ con los resultados de siete posiciones, no fueron significativos para los casos con criterio de engaño (p = 0.316) ni de veracidad (p = 0894).

Figura 2. Gráfica de medias para resultados inconclusos.



Para los resultados con tres posiciones, los efectos principales simples tampoco fueron significativos para los casos con criterio de engaño (p = 0.157) ni para los veraces (p = 0.936). Los efectos principales simples para las puntuaciones del EES tampoco mostraron diferencias significativas para los resultados inconclusos en función del número de RQ con los casos con criterio de engaño (p = 0.161) o de veracidad (p = 0.940).

Un ANOVA de dos vías con respecto al método TDA x número de RQs para los casos con criterio de veracidad fue estadísticamente significativo $F(1,441) = 14.183$, ($p < .001$). Los efectos principales simples para las diferencias con el método de puntuación no fueron significativos para dos RQ (p = 0.083), tres RQ (p = 0.085) o cuatro RQ (p = 0.428). Después de combinar las celdas para los diferentes métodos de puntuación, el efecto principal para las tasas de inconclusos, dependiendo del número de casos RQs, no fue significativo (p = .962) con el criterio de casos veraces. Se realizó un análisis de potencia post-hoc utilizando la función power.anova.test() en Language R y con Environment for Statistical Computing (R Core Team, 2019), indicando una potencia >.99 para detectar una diferencia significativa si es que existiera.

Los efectos principales simples con respecto al número de RQs no fueron significativos para los resultados inconclusos con los casos con criterio de veracidad para el método de puntuación de siete posiciones (p = .866), el método de tres posiciones (p = .936) o el ESS (p =

.940). Después de combinar las celdas para dos, tres y cuatro RQ, el efecto principal para las diferencias entre los resultados inconclusos dependiendo del método de puntuación fue estadísticamente significativo $F(2,447) = 1250.483$, ($p < .001$) para los casos con criterio de veracidad. Esto indica que los efectos de interacción observados en los resultados inconclusos en función de las RQs x método de puntuación pueden atribuirse a las diferencias entre los métodos de puntuación con los casos con criterio de veracidad.

Otro ANOVA de dos vías con respecto al método TDA x número de RQs mostró una interacción estadísticamente significativa para los casos con criterio con engaño $F(1,441) = 17.789$, $p = <.001$). Los efectos principales simples no fueron significativos para las diferencias en los resultados inconclusos entre los casos con criterio de engaño, en función de los diferentes métodos de puntuación con dos RQs ($p = .218$), tres RQs ($p = .080$) o cuatro RQs ($p = .218$). Después de combinar las celdas para los diferentes métodos de puntuación, el efecto principal de las RQs sobre los resultados inconclusos no fue estadísticamente significativo para los casos con engaño ($p = 0,209$). Un análisis de potencia post-hoc indicó una potencia > 0.99 para detectar un efecto significativo para el número de RQs si es que existiera.

Los efectos principales simples para el número de RQs no fueron significativos para los métodos de siete posiciones ($p = .316$), tres posiciones ($p = .157$) o ESS ($p = .161$). Después de combinar las celdas para dos, tres y cuatro RQs, el efecto principal con respecto a las diferencias en los resultados inconclusos, dependiendo del método de puntuación, fueron estadísticamente significativos $F(2,447) = 3.424$, ($p = .033$) para los casos con criterio de engaño. Esto sugiere que las tasas de resultados inconclusos para los casos con criterio de engaño variaron más en función del método de puntuación que del número de RQs.

La inspección del gráfico de la Figura 2 muestra que las tasas medias de inconclusos para los casos con criterio de veracidad con el ESS pueden tener una pendiente diferente en comparación con los otros resultados. Para una mejor comprensión de la influencia del método de puntuación en las tasas observadas de inconclusos, se calculó un contraste ANOVA de tres vías para las puntuaciones de siete y tres posiciones, excluyendo las puntuaciones del ESS. La interacción de tres vías para los resultados inconclusos no fue significativa [$F(4,588) = 0.051$, ($p = 0.995$)] para los métodos de puntuación de siete y de tres posiciones cuando se excluyeron los resultados del ESS. Estos resultados sugieren que la interacción de tres vías para los resultados inconclusos puede atribuirse a las diferencias en los resultados de los casos con criterio de veracidad con el ESS. Las interacciones de dos vías para cada método de puntuación indican que se puede esperar que los índices de inconclusos aumentan con el número de RQs para los casos con criterio de veracidad y disminuyen con el número de RQs para los casos con criterio de engaño.

Errores falsos negativos y falsos positivos para los exámenes USAF MGQT con dos, tres y cuatro RQs.

La figura 3 muestra el gráfico de medias de los errores falsos positivos y falsos negativos. Se completó un ANOVA de tres vías (estado del criterio x método TDA x número de RQs) para los

errores de decisión. El resumen del ANOVA para los errores de decisión se muestra en la Tabla 6. La interacción de tres vías no fue estadísticamente significativa $F(4,882) = 0.943$, $p = 0.438$.

Figura 3. Gráfica de medias de errores falso-positivo y falso-negativo.

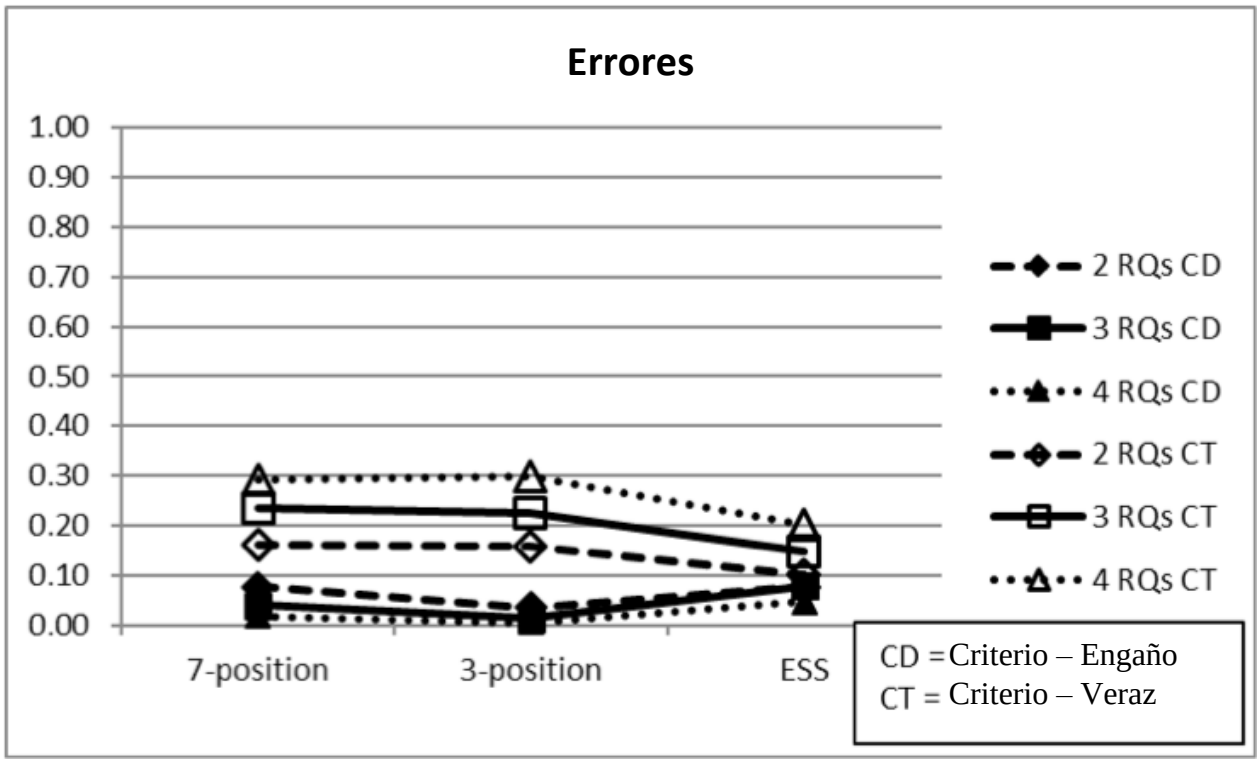


Tabla 6. Resumen para errores con ANOVA de tres vías (RQs x método TDA x criterio de estado)

Source	SS	df	MS	F	p	F crit .05
# RQs	0.273	2	0.137	14.086	<.001	3.006
Status	0.827	1	0.827	85.294	<.001	3.852
Model	0.123	2	0.062	6.358	.002	3.006
# RQs x Status	5.862	2	2.931	302.373	<.001	3.006
Status x Model	0.684	2	0.342	35.283	<.001	3.006
# RQs x Model	0.015	4	0.004	0.394	.813	2.382
# RQs x Status x Model	0.037	4	0.009	0.943	.438	2.382
Error	8.550	882	0.010			
Total	16.371	899				

Debido a que la interacción de tres vías no fue significativa, se calculó un ANOVA de dos vías para las RQs x estado del criterio después de combinar las celdas de los tres métodos TDA. La

figura 4 muestra el gráfico de medias. El resumen del ANOVA de dos vías que se muestra en la Tabla 7 indica una interacción significativa [$F(1,894) = 104.051$, ($p < .001$)] para los errores de decisión en función del número de RQs y del estado del criterio.

Aunque los errores parecen aumentar con el número de RQs para los casos con criterio de veracidad y disminuir con el número de RQs para los de engaño, los efectos principales simples para el número de RQs no fueron estadísticamente significativos para los casos con criterio de engaño ($p = .459$) o de veracidad ($p = .814$)

Figura 4. Gráfica de medias para errores de decisión con métodos de calificación combinados.

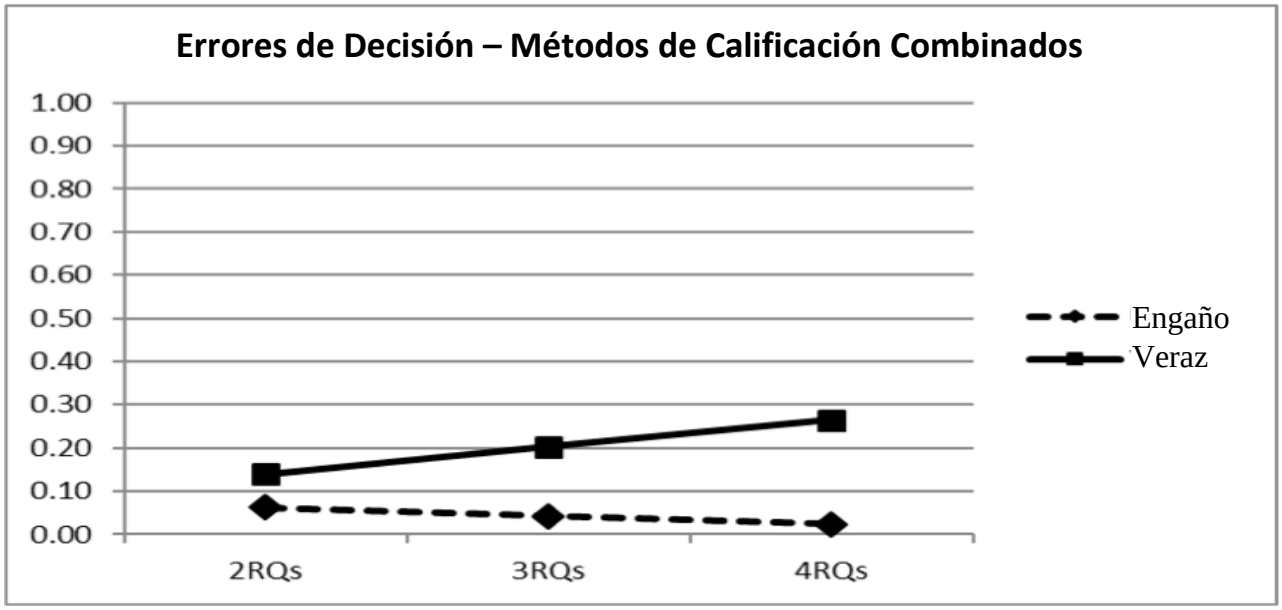


Tabla 7. Resumen de errores de decisión con ANOVA de dos vías con puntuaciones de 7 posiciones (RQs x estado de criterio).

Source	SS	df	MS	F	p	F crit .05
# RQs	0.273	1	0.001	0.094	.759	3.852
Status	5.680	1	0.013	1.302	.254	3.852
Interaction	1.009	1	1.009	104.051	<.001	3.852
Error	8.666	894	0.010			
Total	6.962	897				

Un cálculo de potencia post-hoc para los efectos principales simples de una vía, con n = 50 para cada celda, tuvo una potencia >.99 para detectar un efecto significativo si realmente pudiera existir. Esto sugiere que la interacción observada puede atribuirse al hecho de que, aunque la diferencia para dos, tres o cuatro RQs no es significativa dentro de los casos veraces o con engaño, la probabilidad de error de la prueba para polígrafos de asuntos múltiples aumenta con el número de RQs para los casos con criterio de veracidad al tiempo que disminuye para los casos con criterio de engaño.

Precisión media no ponderada.

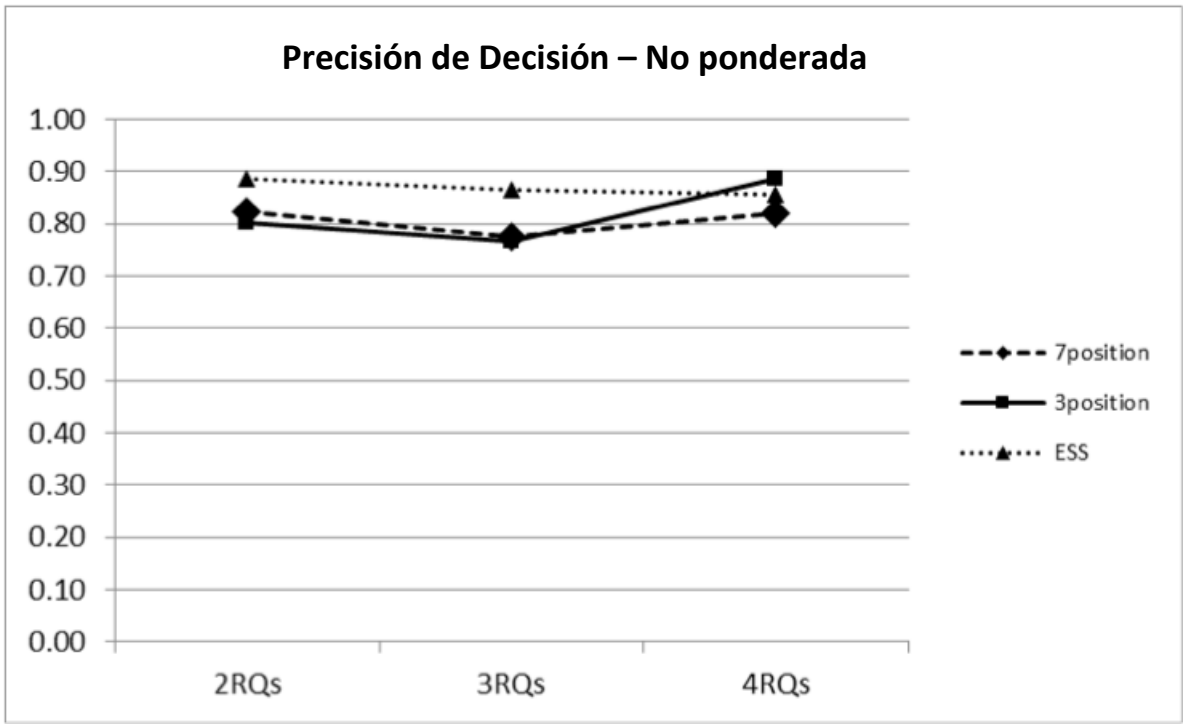
La precisión de la decisión no ponderada, excluyendo los resultados inconclusos, se muestra en la Tabla 2, y fue significativamente mayor que el azar (.5) para los tres métodos TDA con dos, tres y cuatro RQs ($p < .05$). La Tabla 8 también muestra que la variación en la precisión de la prueba aumenta en función del número de RQs para los tres métodos de puntuación a dos vías. Del mismo modo, como se muestra en los apéndices, tanto los errores falsos negativos como positivos se redujeron a un nivel estadísticamente significativo inferior al azar para todas las versiones de TDA con dos, tres y cuatro RQ.

Tabla 8. Precisión no ponderada: media (SD) {95% CI}.

	7-position	3-position	ESS
2RQs	.822 (.061) {.702 to .942}	.802 (.073) {.659 to .945}	.886 (.047) {.795 to .978}
3RQs	.775 (.104) {.571 to .979}	.766 (.128) {.515 to .999}	.866 (.067) {.734 to .998}
4RQs	.820 (.146) {.533 to .999}	.887 (.149) {.595 to .999}	.855 (.101) {.657 to .999}

La figura 5 muestra el gráfico de medias de la precisión promedio no ponderada (es decir, la media no ponderada de la exactitud en la decisión con los casos con criterio de engaño y veracidad). Una interacción de dos vías fue significativa para el número de RQs x método de puntuación [$F(1,891) = 51.009$, ($p < .001$)]. Sin embargo, los efectos principales simples no fueron significativos para los diferentes métodos de puntuación para dos RQs ($p = 0.711$), tres RQs ($p = 0.824$), o 4 RQs ($p = 0959$). Los efectos principales simples tampoco fueron significativos para el método de siete posiciones ($p = 0,975$), tres posiciones ($p = 0,839$) o el ESS ($p = 0,871$). Aunque las líneas de la Figura 1 presentan una pendiente diferente, ninguna de las líneas es en sí misma significativamente diferente de cero.

Figura 5. Gráfica media de precisión promedio no ponderada.



Después de combinar las celdas para los diferentes métodos de puntuación, un ANOVA de una vía mostró que las diferencias en la precisión no ponderada, en función del número de RQ, no fueron estadísticamente significativas [$F(2,897) = 0.046$, ($p = .955$)]. Un análisis de potencia post-hoc indicó que el ANOVA tenía una potencia >.99 para detectar un efecto significativo. Estos resultados indican que no hay una diferencia real en la precisión no ponderada para los resultados de PDD con 2RQ, 3RQ o 4RQ, excluyendo los resultados inconclusos.

Precisión del criterio para dos, tres o cuatro preguntas al azar.

Se utilizaron otros tres modelos Monte Carlo para una mejor comprensión de las diferencias entre los métodos de puntuación de siete, tres posiciones y ESS, cuando se aleatoriza el número de RQ para cada caso en el espacio Monte Carlo. Para cada caso, el número de dos, tres o cuatro RQs varió aleatoriamente comparando un número aleatorio contra los valores .3333333 y .666666. Las proporciones de casos con dos, tres y cuatro RQ podrían variar para cada iteración dentro del espacio Monte Carlo, y podrían converger en iguales proporciones dentro de la distribución de los resultados Monte Carlo compuesta de 10,000 iteraciones en el espacio Monte Carlo.

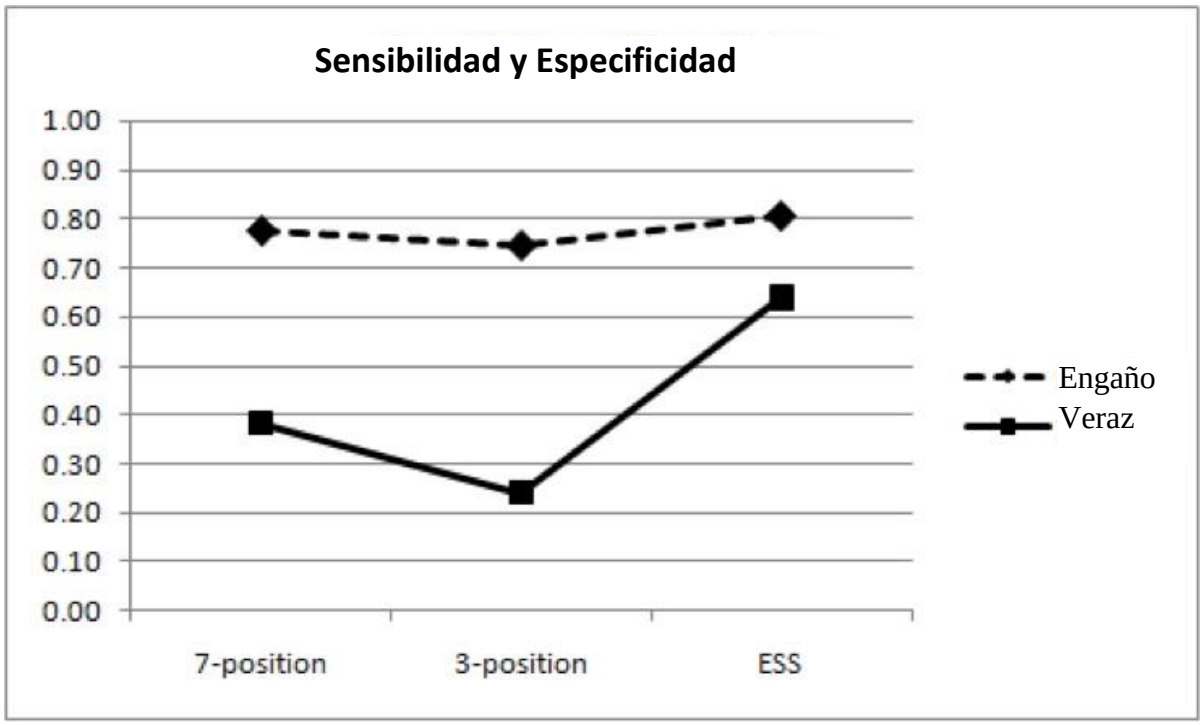
Las tasas base para el estado de criterio de las preguntas individuales fueron las siguientes; para los casos con dos RQ, la tasa base fue = .293, para los casos con tres RQ = .206 y con cuatro RQ = .159. En cada caso para cada RQ, se comparó un número uniforme aleatorio contra la tasa base, y el estado del criterio se estableció como verdadero si la tasa base era menor que el número aleatorio. Esto aseguró que aunque la proporción de casos con criterio con engaño y de veracidad variara para cada iteración dentro del espacio Monte Carlo, la tasa base para el engaño convergería a .5 para la distribución de los resultados Monte Carlo, cuando se establecía aleatoriamente el número de RQ para cada examen y se establecía aleatoriamente el estado del criterio para cada RQ. Cada caso se evaluó

con los métodos de puntuación de siete, tres posiciones y ESS utilizando la SSR anteriormente descrita. El Apéndice D muestra las medias, las desviaciones estándar y los intervalos de confianza del 95% para la distribución de resultados de Monte Carlo, mientras se varía el número de dos, tres o cuatro RQ.

Sensibilidad y especificidad de los exámenes USAF MGQT con dos, tres o cuatro RQ al azar.

Un ANOVA bidireccional para la precisión de la decisión mostró una interacción significativa entre el método de puntuación y el estado del criterio $F(1.294) = 177.039$, $p < .001$. La Figura 6 muestra un gráfico de las medias para la sensibilidad y la especificidad de la prueba. Los efectos principales simples no fueron estadísticamente significativos para la sensibilidad de la prueba ante el engaño ($p = .659$) o para la especificidad ante la veracidad ($p = .064$). Un análisis de potencia post-hoc indicó una probabilidad de potencia >.99 para detectar una diferencia significativa si existiera.

Figura 6. Estimaciones de medias Monte Carlo para la sensibilidad y especificidad de la prueba

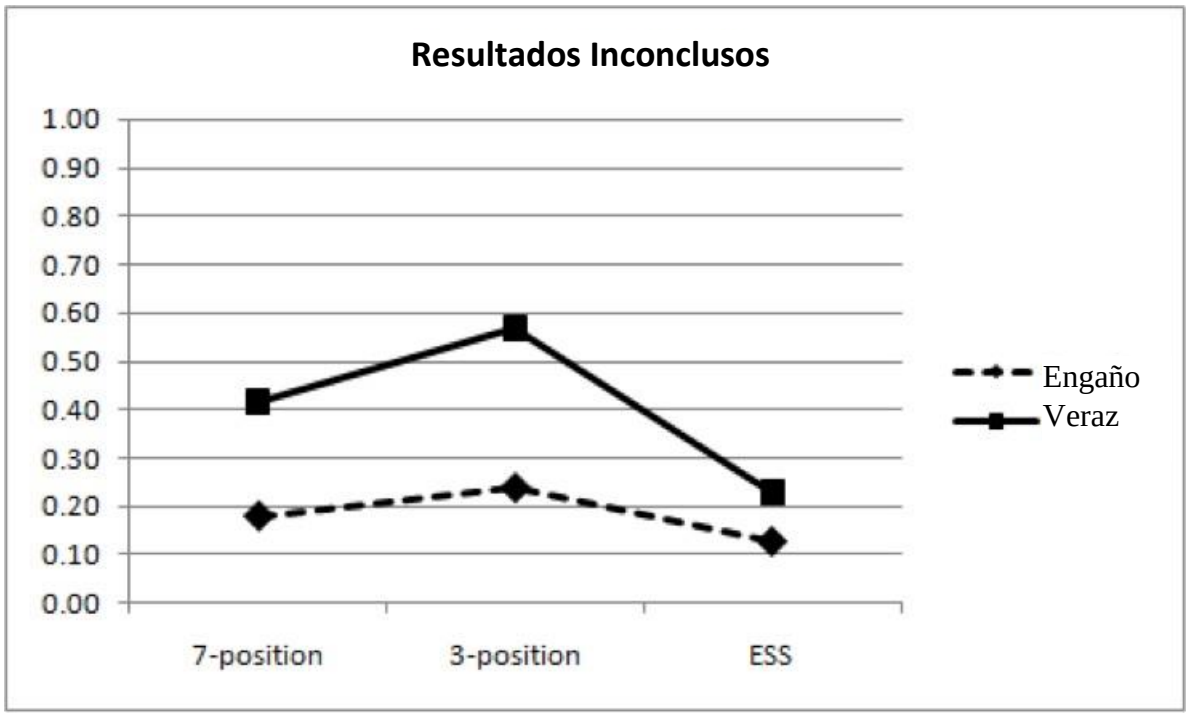


La evaluación de los efectos principales simples dependiendo del método de puntuación, mostró que la diferencia en la detección del engaño difería significativamente de la detección de la veracidad con el método de puntuación de siete posiciones [$F(1,98) = 8.307$, ($p = .005$)] y para el de tres posiciones [$F(1,98) = 19.438$, ($p < .001$)]. El efecto principal simple para los casos con criterio de engaño y de veracidad no fue significativo para el ESS ($p = .222$). Estos resultados indican que la interacción de dos vías se puede atribuir a las diferencias en la sensibilidad y la especificidad de la prueba para el método de puntuación ESS en comparación con la de los métodos de siete y tres posiciones. Como se muestra en el Apéndice D, aunque la sensibilidad de la prueba al engaño fue significativamente mayor que el azar (.5) para los tres métodos de puntuación, la especificidad de la prueba para la veracidad no superó el azar para los métodos de siete o tres posiciones.

Resultados inconclusos de los exámenes USAF MGQT con dos, tres o cuatro RQ al azar.

Un ANOVA de dos vías para los resultados inconclusos (método de puntuación x estado de criterio) mostró diferencias significativas en los resultados inconclusos para los tres métodos de TDA $F(1,294) = 71.927$, $p < 0.001$. La Figura 7 muestra la media Monte Carlo para tasas de inconclusos con los tres métodos de TDA. Los efectos principales simples para los resultados inconclusos no fueron significativos para los casos de engaño ($p = .185$) o los de veracidad ($p = .177$).

Figura 7. Estimaciones medias Monte Carlo para tasas de inconclusos.

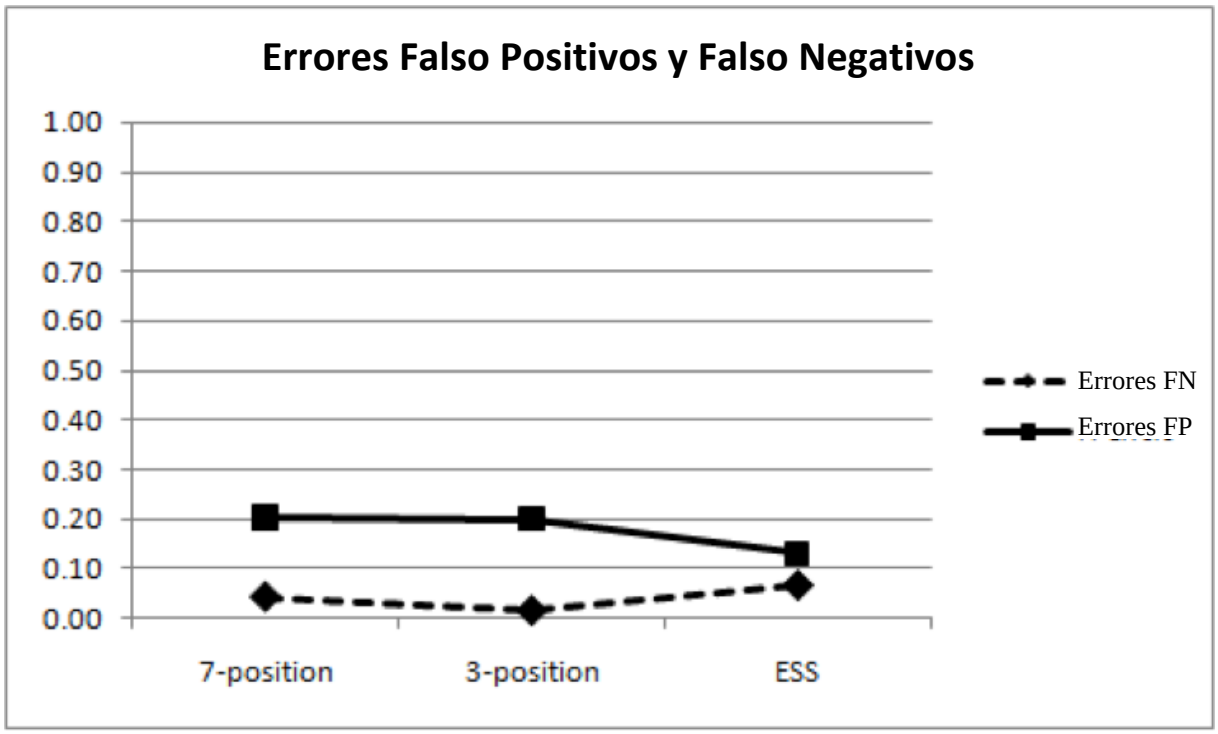


El efecto principal simple, para las diferencias en las tasas de inconclusos con los casos con criterio de engaño y de veracidad, fue significativo con las puntuaciones de tres posiciones [$F(1,98) = 5.147$, ($p = 0.025$)], pero no con las puntuaciones de siete posiciones ($p = 0.084$) o el ESS ($p = 0.413$). Estos resultados indican que la interacción observada en dos vías (método TDA x estado del criterio) para los resultados inconclusos puede atribuirse a la diferencia significativa entre las tasas de inconclusos para los casos con criterio de engaño y veracidad con el método de puntuación de tres posiciones. Las tasas medias de resultados inconclusos fueron elevadas para los resultados con puntuación de tres posiciones en comparación con los resultados con siete posiciones y del ESS, y fueron mayores para los casos con criterio de veracidad.

Errores de decisión en los exámenes USAF MGQT con dos, tres o cuatro preguntas RQs al azar.

Un ANOVA de dos vías para los errores de decisión según el estado del criterio mostró una interacción significativa entre el método TDA y el estado del criterio $F(1,294) = 31.456$, $p < .001$. La figura 8 muestra las medias Monte Carlo para las tasas de error de los tres métodos de TDA.

Figura 8. Estimaciones de medias Monte Carlo para las tasas de inconclusos.



Los efectos principales simples no fueron significativos para los errores falsos negativos (p = 0.229) ni para los errores falsos positivos (p = 0.874). Además, ninguno de los efectos principales simples fue estadísticamente significativos para el método de puntuación de siete posiciones (p = 0.223), el método de puntuación de tres posiciones (p = 0.097) o el ESS (p = 0.510). El análisis de potencia post-hoc mostró que el experimento tenía una potencia > 0.99 para detectar un efecto significativo, si es que existía. La interacción observada en los errores de decisión puede considerarse como un indicador de que las dos líneas de la Figura 8 tienen una pendiente diferente y significativa, aunque ninguna de las líneas es en sí misma significativamente diferente de la pendiente cero, lo que significa que las diferencias observadas están dentro del rango esperado de variación no controlada/no explicada. Estos resultados indican que no existe ninguna diferencia real entre las tasas de falsos negativos y que no existe ninguna diferencia real en las tasas de falsos positivos para los métodos de siete posiciones, tres posiciones y ESS.

Discusión

Este proyecto es un estudio Monte Carlo sobre los efectos de la precisión de criterio de los polígrafos de asunto múltiple con dos, tres y cuatro RQ, como lo es el USAF MGQT. Aunque se esperan diferencias en la precisión del criterio en función del número de RQs, estudios previos no habían investigado estas diferencias. Los polígrafos de asuntos múltiples se utilizan comúnmente en los programas poligráficos exploratorios - en ausencia de cualquier acusación o incidente conocido.

Una característica que define a los polígrafos exploratorios de asuntos múltiples es que las preguntas se interpretan bajo la suposición de criterio de varianza independiente. Los resultados globales de las pruebas poligráficas de asuntos múltiples se heredan de los resultados en las preguntas. En términos prácticos, los resultados de prueba de los exámenes de asuntos múltiples se heredan de la puntuación

de la pregunta más baja. Esto difiere de los polígrafos de asuntos específicos en los que el resultado de la prueba se determina por la prueba en su conjunto, y donde los resultados de las preguntas se heredan del resultado global de la prueba. Existen algunas dificultades conocidas en el estudio de los polígrafos de asuntos múltiples. Una de ellas es la de poder conocer el estado del criterio para cada una de las preguntas de la prueba.

Otra dificultad es el manejo de los efectos de la multiplicidad - la sumatoria del error estadístico cuando se realizan conclusiones basadas en eventos de probabilidad múltiples. Por último, existe la dificultad de poder adquirir una muestra de datos, idealmente una muestra equilibrada, con un número igual de casos en cada una de las diferentes condiciones de prueba que sea de tamaño adecuado para su estudio y análisis.

Una ventaja del enfoque Montecarlo para este proyecto, es la reducción del gasto en términos de actividad humana y de otros recursos para la adquisición de datos en los que se pueda conocer con certeza el estado del criterio de cada RQ. Otra ventaja del enfoque Montecarlo para este proyecto fue la posibilidad de comparar fácilmente la eficacia de los distintos métodos de puntuación - el de siete posiciones, el de tres posiciones y el ESS.

Los resultados de este estudio indican que existen algunas diferencias en la efectividad de los distintos métodos de puntuación para los casos con criterio de engaño y de veracidad con, dos, tres o cuatro RQ. Sin embargo, estas diferencias no se observan con respecto a la precisión de la decisión no ponderada - la precisión de la decisión media no ponderada con casos con criterio de engaño y veracidad, excluyendo los resultados inconclusos. No se encontraron diferencias reales en la precisión no ponderada en función del número de RQs. La precisión de la decisión media no ponderada para polígrafos de asuntos múltiples con dos, tres o cuatro RQs excedió significativamente el azar (.5) para los tres métodos TDA.

A pesar de que la precisión no ponderada no difiere para los polígrafos de asuntos múltiples con dos, tres o cuatro RQs, los resultados del estudio indican que existen algunas diferencias cuando se consideran otras dimensiones de la precisión de la prueba. La sensibilidad media de la prueba al engaño superó el azar (.5) para los tres métodos de puntuación. Sin embargo, la especificidad media de la prueba para la veracidad no superó el azar para los métodos de puntuación de siete y tres posiciones, y la especificidad de la prueba fue significativamente mayor que el azar sólo para el modelo de dos RQ con el ESS.

Se observaron diferencias en las tasas de inconclusos en función del número de RQs y en función del método de puntuación. Se puede esperar que los índices de inconclusos aumenten con el número de RQs para los casos con criterio de veracidad y disminuyan con el número de RQs para los casos con criterio de engaño. Sin embargo, los resultados con el EES pueden producir un patrón diferente de tasas de inconclusos con los casos con criterio de veracidad en comparación con otros métodos de puntuación. Una posible razón para esto, no explorada en este estudio, es el uso de una corrección estadística para efectos de multiplicidad para la puntuación de corte del ESS para clasificaciones de veracidad. Es posible que el uso de puntuaciones ESS con puntuaciones de corte tradicionales pueda dar lugar a índices de inconclusos que se parecerían a la tendencia mostrada por los resultados de siete posiciones y tres posiciones en este estudio.

No se encontraron diferencias significativas en las tasas de error falso-positivo o falso-negativo en función del número de RQs. Los análisis de potencia post-hoc sugieren que este estudio tuvo suficiente potencia para detectar efectos significativos en la prueba, si es que existen. Aunque las diferencias

para dos, tres o cuatro RQs no fueron significativas dentro de los casos con criterio de veracidad o de engaño, la probabilidad de error de prueba aumentó con el número de RQs para los casos con criterio de veracidad mientras que disminuyó para los casos con criterio de engaño.

Además de la investigación de las diferencias en la precisión de criterio que pueden existir en función del número de RQs en polígrafos de asunto múltiple, se utilizaron métodos Monte Carlo para comparar los resultados de los métodos de siete posiciones, tres posiciones y ESS. Los resultados de este análisis mostraron que los tres métodos alcanzaron una precisión de decisión no ponderada que superó significativamente el nivel de azar (.5). La sensibilidad de la prueba al engaño superó el azar con los tres métodos de puntuación. Sin embargo, la especificidad de la prueba para la veracidad no superó el azar con los métodos de siete o de tres posiciones. Las tasas medias de inconclusos fueron más altas para el método de puntuación de tres posiciones, y esto se cargó hacia los casos con criterio de veracidad. A pesar de estas diferencias observadas, los resultados no mostraron diferencias significativas en las tasas de falsos negativos ni en las tasas de falsos positivos para los métodos de siete posiciones, tres posiciones y ESS.

Una limitación de este estudio es que no se hizo ningún esfuerzo para evaluar la diferencia en la precisión del criterio de los tres métodos de puntuación en función de las diferencias en las puntuaciones numéricas. Los resultados de los métodos de puntuación de siete posiciones y de tres posiciones se obtuvieron utilizando puntuaciones numéricas tradicionales (-3 o menos en cualquier subtotal para las clasificaciones de engaño, +3 o más en todos los subtotales para las clasificaciones veraces) sin corrección estadística por los efectos de la multiplicidad. Los resultados del ESS se obtuvieron utilizando puntuaciones de corte referenciadas estadísticamente para las que se utilizó una corrección estadística para manejar los efectos de multiplicidad con los resultados veraces. Las puntuaciones de corte del ESS fueron de -3 o menos en cualquier subtotal para el engaño y de + 1 o más en todos los subtotales para la veracidad. Es posible que algunas interacciones y algunos efectos difieran si todos los resultados se hubieran obtenido utilizando puntuaciones de corte optimizadas a través de puntuaciones de corte optimizadas estadísticamente (o si todos los resultados se obtuvieran utilizando puntuaciones de corte tradicionales). Es posible que con reglas de decisión diferentes, que involucran el uso de la puntuación de gran total, se podría lograr una mejora en la especificidad de la prueba y en los resultados inconclusos sin comprometer de manera no deseada la sensibilidad de la prueba y las tasas de falsos negativos. Esto debería ser objeto de futuras investigaciones.

Otra limitación de este proyecto es el diseño general mediante la simulación Monte Carlo. Los modelos de Monte Carlo, aunque son insuficientes para dar una respuesta final o definitiva a cuestiones hipotéticas, son muy útiles para estudiar problemas de alto costo y riesgo, así como de problemas complejos y difíciles. Los resultados de los estudios Monte Carlo deben ser replicados y evaluados junto con los resultados de otros estudios de laboratorio y de campo. El uso de los parámetros subtotales, que se obtuvieron a partir de los subtotales de los exámenes de asuntos únicos confirmados representa otra limitación. Sin embargo, los parámetros de las puntuaciones subtotales de los exámenes de asunto único, aunque son imperfectos en su capacidad de representar las puntuaciones subtotales de los exámenes de asuntos múltiples, ofrecen la ventaja de tener un estado de criterio razonablemente conocido para utilizarlo como parámetro en la simulación Monte Carlo.

Otra limitación destacable del presente estudio es que no se intentó investigar la sensibilidad o la especificidad de las pruebas al nivel de las preguntas individuales. Aunque las reglas de decisión se ejecutaron al nivel de las puntuaciones subtotales de las preguntas individuales, las clasificaciones de engaño y veracidad se hicieron al nivel de la prueba en su conjunto. Dentro de los casos Montecarlo

no se intentó determinar la veracidad en algunas preguntas y el engaño en otras. Estos procedimientos son consistentes con las prácticas poligráficas de campo.

En resumen, los resultados de este estudio apoyan la validez de la hipótesis de que los exámenes PDD de asuntos múltiples con dos, tres o cuatro RQs, pueden diferenciar el engaño de la veracidad con tasas que son significativamente mayores que el azar cuando se califican con los modelos TDA de siete posiciones, tres posiciones y ESS. Las sugerencias para la investigación futura incluyen el estudio adicional de los efectos de la multiplicidad, la optimización estadística de las puntuaciones de corte para la decisión y las reglas de decisión para los polígrafos de asuntos múltiples. Los formatos de polígrafo de asunto múltiple que pueden usarse con dos, tres o cuatro RQs, como el USAF MGQT, ofrecen el potencial de una gran adaptabilidad y utilidad en una variedad de escenarios de práctica de campo, y se indica un interés continuo en los formatos de PDD de asunto múltiple.

Referencias

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N.J. Salkind (Ed.),
Encyclopedia of Measurement and Statistics. Sage.

Barland, G. H., Honts, C. R. & Barger, S.D. (1989). *Studies of the accuracy of security screening polygraph examinations. Department of Defense Polygraph Institute.*

Blalock, B., Cushman, B. & Nelson, R. (2009). A replication and validation study on an empirically based manual scoring system. *Polygraph, 38, 281-288.*

Capps, M. H. & Ansley, N. (1992). Analysis of federal polygraph charts by spot and chart total.
Polygraph, 21, 110-131.

Cohen, B. (2002). Calculating a factorial ANOVA from means and standard deviations. *UnderstandingStatistics 1(3):191-203.*

Department of Defense (2006). *Federal psychophysiological detection of deception examiner handbook.* Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007. Reprinted in Polygraph, 40(1), 2-66.

Department of Defense (2006). *Psychophysiological Detection of Deception Analysis II - Course #503.* Test data analysis: DoDPI numerical evaluation scoring system. Available from the author. (Retrieved from http://www.antipolygraph.org/documents/federal-polygraph- handbook-02-10-2006.pdf on 3-31-2007).

Efron, B. & Tibshirani R. J. (1986). Bootstrap methods for standard errors, confidence intervals, andother measures of statistical accuracy. *Statistical Science,* 1(1), 54-77.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Handler, M., Nelson, R., Goodson, W. & Hicks, M. (2010)). Empirical Scoring System: A cross-culturalreplication and extension study of manual scoring and decision policies. *Polygraph,* 39(4), 200-215.

Harwell, E.M. (2000). A comparison of 3- and 7-position scoring scales with field examinations.
Polygraph, 29, 195-197.

Krapohl, D. J. (1998). A comparison of 3- and 7- position scoring scales with laboratory data.
Polygraph, 27, 210-218.

Krapohl, D.J., & Cushman, B. (2006). Comparison of evidentiary and investigative decision rules: Areplication. *Polygraph,* 35(1), 55-63.

Krapohl, D.J. (2010). Short Report: A Test of the ESS with Two-Question Field Cases. Polygraph, 39,124-126.

Light, G.D. (1999). Numerical evaluation of the Army zone comparison test. *Polygraph,*

28, 37-45.Marin, J. (2000). He said/She said: Polygraph evidence in court. *Polygraph,*

29, 299-304.

Marin, J. (2001). The ASTM exclusionary standard and the APA 'litigation certificate' program.
Polygraph, 30, 288-293.

Nelson, R. (2017). Multinomial reference distributions for comparison question polygraphs. *Polygraphand Forensic Credibility Assessment, 46(2),* 81-115.

Nelson, R. & Blalock, B. (2016). Extended analysis of Senter, Waller and Krapohl's USAF MGQT examination data with the Empirical Scoring System and the Objective Scoring System, version 3. *Polygraph, 45(1*), 90-94.

Nelson, R., Blalock, B. & Handler, M. (2011). Criterion validity of the Empirical Scoring System andthe Objective Scoring System, version 3 with the USAF Modified General Question Technique.*Polygraph, 40(11),* 172-179.

Nelson, R., Blalock, B. & Handler, M. (2019). Practical Polygraph: How to Parse Categorical Resultsfor Test Questions of Diagnostic and Screening Polygraphs. *APA Magazine, 52(3)*, 60-65.

Nelson, R., Blalock, B., Oelrich, M. & Cushman, B. (2011). Reliability of the Empirical Scoring System with expert examiners. *Polygraph,* 40.

Nelson, R. & Handler, M. (2010). Empirical Scoring System: NPC Quick Reference. Lafayette Instrument Company. Lafayette, IN.

Nelson, R., Handler, M., Morgan, C., & O'Burke, P., (2012). Short Report: Criterion validity of the United States Air Force Modified General Question Technique and Iraqi scorers. *Polygraph,* 41 (1).

Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B., and Oelrich,M.(2011). Using the Empirical Scoring System, *Polygraph, 40,* (In press).

Nelson, R. & Krapohl, D. (2011). Criterion Validity of the Empirical Scoring System with Experienced Examiners: Comparison with the Seven-Position Evidentiary Model Using the Federal Zone Comparison Technique. *Polygraph,* (In press).

Nelson, R., Krapohl, D. & Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. *Polygraph, 37,* 185-215.

Nelson, R. & Rider, J. (2018). Practical polygraph: ESS-M made simple. *APA Magazine, 51(6),* 55-62.

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.

Reid, J. E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law andCriminology,* 37, 542-547.

R Core Team (2019). R: *A language and environment for statistical computing. R Foundation for Statistical Computing,* Vienna, Austria. URL https://www.R-project.org/.

Research Division Staff (1995). *A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope Polygraph and the test for espionage and sabotage question formats.* DTIC AD Number A319333. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in Polygraph, 26(2), 79-106.

Research Division Staff (1995). *Psychophysiological detection of deception accuracy rates obtained using the test for espionage and sabotage.* DTIC AD Number A330774. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in Polygraph, 27, (3), 171-180.

Senter, S M. (2003). Modified general question test decision rule exploration. *Polygraph, 32,* 251-263.

Robertson, B. (2012). The Use of an Enhanced Polygraph Scoring Technique in Homeland Security:The Empirical Scoring System—Making a Difference. . Naval Postgraduate School, Dudley Knox Library: Retrieved from: https://www.hsdl.org/?abstract&did=710340.

Senter, S., Waller, J. & Krapohl, D. (2008). Air Force Modified General Question Test Validation Study. *Polygraph, 37(3),* 174-184.

Šidàk, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions.
Journal of the American Statistical Association, 62, 626–633.

Summers, W. G. (1939). Science can get the confession. *Fordham Law Review, 8,* 334-354.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph, 28,* 10-27.

Van Herk, M. (1990). Numerical evaluation: Seven point scale +/-6 and possible alternatives: A discussion. T*he Newsletter of the Canadian Association of Police Polygraphists, 7,* 28-47. Reprinted in Polygraph, 20(2), 70-79.

Apéndice A.

Precisión de Criterio de Polígrafos de Asunto Múltiple con Dos RQs

	7-position Mean (SE) {95% CI}	3-position Mean (SE) {95% CI}	ESS Mean (SE) {95% CI}
Unweighted Accuracy	.822 (.061) {.702 to .942}	.802 (.073) {.659 to .945}	.886 (.047) {.795 to .978}
Unweighted INC	.302 (.055) {.195 to .409}	.424 (.054) {.319 to .529}	.217 (.050) {.119 to .316}
D INC	.226 (.050) {.128 to .324}	.306 (.054) {.201 to .412}	.190 (.043) {.105 to .275}
T INC	.378 (.097) {.188 to .567}	.542 (.095) {.355 to .729}	.245 (.088) {.072 to .417}
Sensitivity	.697 (.053) {.593 to .800}	.659 (.055) {.550 to .767}	.734 (.049) {.637 to .831}
Specificity	.462 (.101) {.265 to .659}	.300 (.090) {.123 to .476}	.655 (.076) {.506 to .804}
FN	.077 (.032) {.015 to .140}	.035 (.021) {.001 to .076}	.076 (.030) {.018 to .135}
FP	.160 (.077) {.010 to .310}	.158 (.071) {.018 to .298}	.100 (.060) {.001 to .217}
PPV	.929 (.035) {.861 to .998}	.925 (.036) {.854 to .996}	.957 (.027) {.905 to .999}
NPV	.666 (.116) {.439 to .893}	.743 (.142) {.465 to .999}	.737 (.098) {.545 to .929}
D Correct	.900 (.040) {.821 to .979}	.950 (.030) {.891 to 1.009}	.906 (.037) {.834 to .977}
T Correct	.743 (.118) {.513 to .974}	.654 (.145) {.369 to .940}	.867 (.080) {.710 to .999}

Traductor. rodolfo@poligrafia.com.mx

Apéndice B.

Precisión de Criterio de Polígrafos de Asunto Múltiple con Tres RQs

	7-position Mean (SE) {95% CI}	3-position Mean (SE) {95% CI}	ESS Mean (SE) {95% CI}
Unweighted Accuracy	.775 (.104) {.571 to .979}	.766 (.128) {.515 to .999}	.866 (.067) {.734 to .998}
Unweighted INC	.317 (.074) {.171 to .462}	.427 (.071) {.288 to .567}	.156 (.063) {.032 to .279}
D INC	.180 (.038) {.106 to .254}	.242 (.046) {.152 to .331}	.116 (.035) {.048 to .184}
T INC	.453 (.142) {.175 to .732}	.613 (.136) {.346 to .880}	.195 (.121) {.001 to .432}
Sensitivity	.781 (.041) {.701 to .862}	.747 (.046) {.656 to .837}	.806 (.042) {.724 to .889}
Specificity	.320 (.130) {.066 to .574}	.188 (.094) {.004 to .372}	.659 (.141) {.383 to .934}
FN	.039 (.021) {.001 to .08}	.012 (.012) {.001 to .035}	.078 (.028) {.023 to .133}
FP	.235 (.131) {.001 to .493}	.226 (.114) {.002 to .450}	.146 (.108) {.001 to .359}
PPV	.960 (.024) {.914 to .999}	.959 (.022) {.915 to .999}	.975 (.019) {.938 to .999}
NPV	.545 (.190) {.173 to .917}	.728 (.244) {.250 to .999}	.549 (.128) {.298 to .800}
D Correct	.953 (.025) {.903 to .999}	.984 (.016) {.954 to .999}	.912 (.032) {.850 to .974}
T Correct	.589 (.203) {.190 to .987}	.475 (.198) {.086 to .864}	.819 (.131) {.563 to .999}

Apéndice C.

Precisión de Criterio de Polígrafos de Asunto Múltiple con Cuatro RQs

	7-position Mean (SE) {95% CI}	3-position Mean (SE) {95% CI}	ESS Mean (SE) {95% CI}
Unweighted Accuracy	.820 (.146) {.533 to .999}	.887 (.149) {.595 to .999}	.855 (.101) {.657 to .999}
Unweighted INC	.318 (.108) {.107 to .528}	.396 (.112) {.177 to .615}	.163 (.096) {.001 to .351}
D INC	.140 (.035) {.071 to .208}	.180 (.039) {.103 to .257}	.089 (.031) {.028 to .150}
T INC	.496 (.211) {.082 to .91}	.612 (.220) {.181 to .999}	.237 (.191) {.001 to .611}
Sensitivity	.842 (.037) {.771 to .914}	.816 (.039) {.738 to .893}	.864 (.036) {.793 to .934}
Specificity	.289 (.148) {.001 to .580}	.205 (.109) {.001 to .419}	.581 (.200) {.190 to .972}
FN	.018 (.014) {.001 to .046}	.005 (.007) {.001 to .018}	.047 (.022) {.003 to .091}
FP	.292 (.198) {.001 to .680}	.298 (.205) {.001 to .700}	.202 (.180) {.001 to .555}
PPV	.976 (.018) {.940 to .999}	.976 (.017) {.942 to .999}	.985 (.013) {.959 to .999}
NPV	.581 (.262) {.067 to .999}	.815 (.25) {.324 to .999}	.454 (.185) {.092 to .816}
D Correct	.979 (.017) {.946 to .999}	.995 (.008) {.978 to .999}	.948 (.024) {.900 to .996}
T Correct	.546 (.257) {.042 to .999}	.505 (.259) {.001 to .999}	.754 (.201) {.359 to .999}

Traductor. rodolfo@poligrafia.com.mx

Apéndice A.

Precisión de Criterio con (2, 3, o 4) RQs Combinadas/Al Azar

	7-position Mean (SE) {95% CI}	3-position Mean (SE) {95% CI}	ESS Mean (SE) {95% CI}
Unweighted Average Accuracy	.799 (.088) {.627 to .971}	.775 (.107) {.565 to .984}	.878 (.060) {.760 to .996}
Unweighted Inconclusives	.294 (.072) {.154 to .434}	.403 (.071) {.263 to .543}	.178 (.059) {.062 to .294}
D INC	.177 (.044) {.091 to .263}	.238 (.047) {.146 to .331}	.129 (.036) {.059 to .198}
T INC	.411 (.133) {.149 to .672}	.568 (.136) {.300 to .835}	.228 (.114) {.004 to .453}
Sensitivity	.780 (.047) {.689 to .871}	.746 (.048) {.651 to .841}	.805 (.043) {.722 to .889}
Specificity	.382 (.128) {.131 to .633}	.241 (.110) {.025 to .456}	.642 (.130) {.387 to .897}
FN	.043 (.022) {.001 to .085}	.016 (.013) {.001 to .042}	.066 (.027) {.014 to .118}
FP	.208 (.111) {.001 to .427}	.200 (.109) {.001 to .414}	.130 (.092) {.001 to .310}
PPV	.956 (.025) {.907 to .999}	.956 (.025) {.906 to .999}	.974 (.019) {.936 to .999}
NPV	.605 (.165) {.281 to .929}	.733 (.205) {.331 to .9994}	.622 (.127) {.373 to .870}
D Correct	.948 (.026) {.897 to .999}	.979 (.017) {.945 to .999}	.924 (.031) {.864 to .984}
T Correct	.649 (.174) {.309 to .989}	.555 (.203) {.157 to .954}	.832 (.117) {.603 to .999}