Efectos de Precisión para Puntuaciones ESS y de Tres Posiciones de Exámenes Federales ZCT Utilizando la Regla de Gran Total con Puntuaciones de Corte Tradicionales/Federales y Multinomiales

Raymond Nelson

Abstract

Este proyecto es una comparación de los efectos de precisión para las puntuaciones ESS y de tres posiciones utilizando puntuaciones de corte numéricas tradicionales y puntuaciones de corte multinomiales. Los efectos estudiados incluyen sensibilidad, especificidad, tasas de error de falsos positivos y falsos negativos de prueba; además del valor predictivo positivo, el valor predictivo negativo, las proporciones de clasificaciones correctas para los casos culpables e inocentes y la media no ponderada de los casos correctos e inconclusos. Se utilizaron muestras de archivo de n=100 casos confirmados de campo utilizando el formato Federal ZCT, lo que permitió una comparación intuitiva con los efectos previamente publicados. También se incluyó en el análisis una segunda muestra de n=60 casos confirmados de campo utilizando el formato Federal ZCT. Las respuestas se extrajeron de los datos registrados y los puntajes se asignaron a través de un algoritmo ESS automatizado que fue diseñado para parecerse mucho al proceso de extracción de características utilizado por expertos humanos al calificar manualmente los datos poligráficos. Luego, las puntuaciones del ESS se convirtieron en puntuaciones de tres posiciones. Se utilizó un bootstrap paramétrico para calcular los intervalos de confianza estadísticos en percentiles .025 y .975, y para estimar la varianza de los efectos observados. Se utilizó un procedimiento ANOVA de efectos mixtos para evaluar el tratamiento de los cuatro casos de muestra: puntuaciones ESS y de tres posiciones con puntuaciones de corte tradicionales y multinomiales. La precisión de los cuatro tratamientos al excluir los casos inconclusos fue similar para el valor predictivo positivo, el valor predictivo negativo y para las proporciones de clasificaciones correctas excluyendo los resultados inconclusos. El uso de puntuaciones de corte multinomiales contribuyó a una reducción estadísticamente significativa de resultados inconclusos y a incrementos estadísticamente significativos en la sensibilidad de la prueba para el engaño y en la especificidad de la prueba para la veracidad tanto para el ESS como para las puntuaciones de tres posiciones. Se observó que en ninguna de las clasificaciones de los casos individuales cambiaron de engaño a veracidad o de veracidad a engaño para ninguno de los cuatro tratamientos con cualquiera de las dos muestras de archivo.

Traductor. rodolfo@poligrafia.com.mx *Polygraph & Forensic Credibility Assessment* , 2020, 49(2)

Efectos de Precisión para Puntuaciones ESS y de Tres Posiciones de Exámenes Federales ZCT Utilizando la Regla de Gran Total con Puntuaciones de Corte Tradicionales/Federales y Multinomiales

Raymond Nelson

Introducción

El Sistema de Puntuación Empírico (ESS; Nelson et al., 2011) es un método para el análisis de datos de pruebas de detección psicofisiológica del engaño (PDD), basado en evidencia, estandarizado y con referencias estadísticas. El ESS puede considerarse como una modificación del método de puntuación Federal de tres posiciones (National Center for Credibility Assessment, 2017), que puede considerarse a sí mismo como una modificación del método de puntuación Federal de Siete Posiciones. Estudios previos sobre el ESS indican que proporciona efectos de precisión similares al método Federal de siete posiciones, con la confiabilidad práctica y la intuición del método de puntuación de tres posiciones. Los resultados del ESS se describieron por primera vez en una comparación de los efectos de precisión (Nelson, Krapohl & Handler, 2008) entre examinadores poligráficos en formación y de examinadores experimentados. El ESS se actualizó (Nelson, 2017a; 2017b) para hacer uso de un clasificador Bayesiano y de puntuaciones de corte que se obtuvieron de una distribución de referencia multinomial que se calculó bajo la teoría analítica de las pruebas PDD (Nelson, 2106). Posteriormente se calcularon las puntuaciones de corte multinomiales y las distribuciones de referencia para el método de puntuación de tres posiciones (Nelson, 2018a). El ESS es ampliamente utilizado por examinadores poligráficos en los EE. UU. y de todo el mundo, incluyendo a los profesionales en la práctica privada y de agencias de investigación/ procuración de la ley municipales, estatales y federales.

El ESS se diferencia del método de puntuación de tres posiciones de tres maneras principales. Primero, las puntuaciones del ESS se asignan duplicando el valor de todas las puntuaciones del EDA. Esto es para que los puntajes electrodérmicos se ponderen de manera que se aproximen a las funciones estructurales y estadísticas descritas en la literatura científica acerca del análisis de datos de pruebas PDD y del desarrollo de algoritmos informáticos (Nelson 2019). Una segunda diferencia importante es que el ESS hace uso de puntuaciones de corte numéricas referenciadas estadísticamente en lugar de puntuaciones de corte tradicionales que se derivaron heurísticamente del método de puntuación de siete posiciones. Una tercera diferencia es que varias agencias han implementado el ESS con reglas de decisión distintas, de acuerdo con sus objetivos operativos y de su misión.

Las reglas de decisión definen los procedimientos estructurados utilizados para interpretar y analizar el resultado categórico de la prueba a partir de la información numérica y estadística. [Consulte Nelson (2018b) para obtener un resumen y una descripción de la literatura sobre las reglas de decisión poligráficas.] Las reglas de decisión de uso común en la práctica de campo

Traductor. rodolfo@poligrafia.com.mx *Polygraph & Forensic Credibility Assessment* , 2020, 49(2)

PDD incluyen la regla de gran total (GTR), la regla de dos etapas (TSR), la Regla de Zona Federal (FZR) y la regla de puntuación subtotal (SSR). Dentro de estos, el GTR ha demostrado consistentemente que proporciona el más alto nivel de precisión de clasificación para exámenes de asunto único, mientras que el SSR ha sido considerado por muchas agencias y examinadores poligráficos de campo como la regla de decisión óptima para exámenes exploratorios de asuntos múltiples.

Este proyecto es una comparación de los efectos de precisión del ESS utilizando el GTR con los cortes tradicionales y los cortes multinomiales. Se utilizaron dos muestras de archivo en este proyecto, lo que permite una comparación intuitiva con los efectos de precisión publicados anteriormente. También se estudiaron los efectos de precisión de clasificación con puntuaciones de tres posiciones.

Métodos y Materiales

Datos

Los datos para este proyecto fueron una muestra de campo confirmada de n=100 exámenes, aplicados con el formato de Prueba de Zonas de Comparación Federal (FZCT) (Departamen of Defense, 2006). Esta muestra fue utilizada previamente por Krapohl y Cushman (2006) con puntajes federales de siete posiciones, y luego por Nelson, Krapohl y Handler (2008) en un estudio inicial sobre el ESS. Los casos de la muestra fueron aplicados por diferentes agencias de procuración de la ley federales, estatales y municipales y posteriormente se incluyeron en el archivo de casos confirmados por el Department of Defense Polygraph Institute (ahora el National Center for Credibility Assessment). El FZCT es un formato de prueba de tres preguntas para evento específico, que se reconoce como uno de los formatos de prueba más útiles para la investigación de incidentes criminales. Todos los casos contaron con tres iteraciones en una secuencia de preguntas que incluía tres preguntas relevantes (RQ) y tres preguntas de comparación (CQ) además de otras preguntas de procedimiento que no están sujetas a un análisis numérico o estadístico. Los examinadores de polígrafo de campo se refieren a las repeticiones o iteraciones de la secuencia de preguntas como "gráficos", haciendo referencia a los instrumentos de polígrafo antiguos que trazaban datos fisiológicos a través de agujas de tinta capilar sobre un rollo de papel cuadriculado. Cuando expertos humanos calificaron manualmente los casos de muestra, describieron algunos de los casos como desafiantes. Aunque quizás no sea lo ideal, el uso de estos mismos datos de muestra puede proporcionar una comprensión práctica e intuitiva de las diferencias en los efectos de precisión para los diferentes métodos de análisis de datos de prueba. [Consulte Nelson (2015) y el Departamen of Defense (2006) para obtener información general sobre esta prueba de preguntas de comparación y cómo se condujeron los casos de muestra.]

Todos los casos incluyeron la recolección estandarizada de sensores PDD, de los cuales podrían extraerse las respuestas fisiológicas y las puntuaciones numéricas que incluyen: sensores de respiración superior e inferior, un sensor de actividad electrodérmica y un sensor de actividad cardiovascular. El conocimiento adquirido acerca del formato FZCT y de sus principios y

procedimientos básicos, se han generalizado para otros formatos PDD, que incluyen casos de asunto único y de asuntos múltiples con dos, tres y cuatro RQs. Esta muestra se utilizó en el estudio inicial y en el desarrollo de distribuciones de referencia empíricas para el ESS, y posteriormente se utilizó en una demostración de precisión de la actualización multinomial del ESS-M (Nelson, 2017b).

Se obtuvo una segunda muestra de archivo, que consta de n=60 exámenes confirmados de campo utilizando el formato FZCT. Estos exámenes también estaban incluidos en el archivo de casos confirmados del DoDPI (ahora NCCA). Esta muestra se utilizó previamente como muestra de soporte para la validación del desarrollo de los algoritmos de puntuación del OSS (Krapohl & McManus, 1999, Krapohl, 2002, Nelson Krapohl & Handler, 2008), y también se utilizó en un estudio de los métodos de puntuación manual Federal de siete posiciones y ESS. Todos los exámenes del segundo conjunto de datos constaban de tres iteraciones de una secuencia de preguntas que incluía tres RQ y tres CQ, además de otras preguntas de procedimiento. Al igual que en la primera muestra de archivo, estos exámenes fueron realizados por un grupo de agencias de procuración de la ley municipales, estatales y federales.

Análisis.

Los datos de la muestra se analizaron utilizando una versión automatizada del ESS. Todos los métodos de análisis de datos de pruebas - independientemente de si es una prueba de polígrafo u otra forma de prueba - consta de funciones similares; extracción de características, transformación numérica y reducción de datos, uso de alguna función de probabilidad o clasificador estadístico y procedimientos estructurados para la interpretación y clasificación del resultado. La extracción de características se refiere a la identificación de información útil o de diagnóstico en los datos de prueba registrados y a la extracción o separación de esta información de otra información que no es útil o ruido. Cuando se califican manualmente los datos de la prueba poligráfica, la transformación numérica es la conversión de las respuestas fisiológicas observadas en valores numéricos - utilizando un sistema de enteros [+, 0, -]. La forma más simple de una función de verosimilitud es una puntuación de corte numérica para la cual se conocen los efectos de clasificación, incluyendo los resultados verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN)). Otra forma de función de verosimilitud mapeará u obtendrá puntuaciones de corte numéricas para una distribución de referencia empírica o teórica - ambas están disponibles en publicaciones sobre el ESS en las que se puede calcular una distribución de referencia multinomial bajo la teoría analítica de las pruebas PDD. Independientemente del tipo de función de verosimilitud, analizar un resultado categórico a partir de datos de prueba numéricos y estadísticos requiere el uso de una regla de decisión estructurada. El ESS automatizado fue diseñado para replicar objetivamente los procedimientos utilizados por los expertos humanos al calificar manualmente los datos de las pruebas PDD, incluyendo la extracción de características, la selección de zonas de análisis de RQ y CQ, la asignación y agregación de puntajes enteros, puntuaciones de corte numéricas y reglas de decisión.

Procesamiento de señal.

Los datos de las series de tiempo para todos los casos de muestra se exportaron al formato NCCA ASCII (Editorial Staff, 2019) y se importaron al R Statistical Computing Language and Environment (R Core Team, 2019) para completar el procesamiento de señales, la extracción de características y el análisis de datos. El procesamiento de la señal de los datos digitalizados se completó a una tasa de datos de 30 ciclos por segundo (cps) para todos los sensores registrados. Los datos de respiración se sometieron a un filtro de suavizado, consistente con publicaciones anteriores, que consiste en un filtro de paso bajo tipo Butterworth de primer orden (Butterworth, 1930) con una frecuencia de esquina de .886Hz (equivalente a un filtro de movimiento promedio móvil con una ventana de .5 segundos). Se ha demostrado que los filtros suavizantes de este tipo mejoran los coeficientes de correlación y diagnóstico obtenidos de los datos respiratorios (Nelson & Handler, 2012).

Todos los exámenes se realizaron utilizando sistemas de polígrafo computarizado Axciton que incluían una opción de filtro de paso alto basado en hardware (EDA de auto centrado) además de la opción del EDA centrado manualmente. Una discusión con profesionales de campo reveló una creencia común de que las prácticas de campo favorecían el uso del EDA centrado manualmente al tiempo en que se realizaron los exámenes. Esto puede haber ocurrido como resultado del hecho de que las especificaciones de ingeniería del filtro de paso alto basado en hardware de los instrumentos de polígrafo análogos de antaño no estaban documentadas en gran medida en cuanto a las frecuencias de esquina o las constantes de tiempo del diseño del filtro. Igualmente, la frecuencia de esquina y la constante de tiempo para el sistema de polígrafo computarizado Axciton se han descrito como desconocidas en publicaciones previas (National Research Council, 2003). No se capturó ni registró información sobre la selección del modo EDA para los casos de muestra. Una consecuencia de esto es que es posible que algunos de los casos de muestra se registraron usando el filtro de paso alto basado en hardware y no se intentó determinar el modo EDA mediante inspección visual. Por esta razón, no se utilizó ningún filtro de paso alto para los datos del EDA, y el procesamiento de la señal para los datos del EDA se limitó a la reducción del ruido de alta frecuencia a través de un filtro de paso bajo tipo Butterworth de primer orden con una frecuencia de esquina de .886Hz.

Los datos cardiovasculares incluyen información acerca de la presión arterial de baja frecuencia y de la frecuencia del pulso de alta frecuencia. Debido a la necesidad de evitar alterar o interrumpir los picos cardiovasculares diastólicos y sistólicos, los datos cardiovasculares no estuvieron sujetos a procesamiento o suavizado adicional de señal.

La especificación NCCA ASCII hace uso de unidades adimensionales que no están asociadas con una medición física estandarizada. Por esta razón, la escala de los datos no tiene ningún efecto sobre los resultados analíticos para casos individuales o para este análisis.

Extracción de características.

La extracción de características se logró mediante un procedimiento automatizado destinado a replicar el utilizado por los expertos humanos al calificar manualmente los datos PDD. Las reacciones fisiológicas se evaluaron utilizando una ventana de evaluación de 15 segundos (EW) para todos los sensores de registro. Se cree que esta EW es suficiente para observar la mayoría de las respuestas fisiológicas ante los estímulos de prueba y se considera robusta para las personas con dificultades comunes de atención sostenida. Para los datos de EDA y cardio, se definió una ventana de inicio de respuesta (ROW) desde el inicio del estímulo hasta cinco segundos después del punto de respuesta verbal, o cinco segundos después del fin del estímulo estímulo si no había una respuesta verbal registrada.

Extracción de la función respiratoria. Para los datos respiratorios, en la extracción de características se excluyó la información durante 1.5 segundos antes y 1.5 segundos después del punto registrado de la respuesta verbal. Esto fue para evitar la inclusión de distorsiones de respuesta que ocurren comúnmente en la extracción de características respiratorias. Los datos respiratorios se midieron usando la excursión de la línea respiratoria (RLE), la suma del cambio absoluto para cada par sucesivo de muestras respiratorias - utilizando una ventana corrediza de tres segundos durante la EW de 15 segundos. Para tasas de respiración dentro del rango normal (10 a 22 ciclos por minuto), la ventana corrediza abarcaría de ½ a 2 ciclos respiratorios. La medición de la respiración fue la media de la ventana de tres segundos durante la EW. El valor de la extracción de características para la EW de 15 segundos a una tasa de datos de 30 cps, fue la media de (15 - 3) * 30 = 360 segmentos de tres segundos. El uso de una ventana corrediza de esta forma significa que los valores de extracción de características no dependen de la longitud de la EW y se pueden comparar y optimizar fácilmente para diferentes longitudes de EW - lo que lleva a una optimización potencialmente más sencilla de la EW. La característica de respuesta de interés es la reducción o supresión de la actividad respiratoria, que se espera que ocurra cuando una persona intenta ocultar, evitar revelar o telegrafiar su engaño. Aunque el algoritmo de extracción de características automatizado utiliza una cuantificación adimensional del RLE, los evaluadores humanos observan visualmente las respuestas respiratorias en patrones de formas de onda trazados/mostrados - como una reducción sutil de la amplitud de la respiración y como una disminución sutil de la frecuencia respiratoria.

Extracción de características EDA. Para los datos de EDA, la información se evaluó desde el inicio del estímulo y hasta el final de la EW. Los picos de la respuesta se identificaron por el cambio positivo en la pendiente del EDA (hacia arriba) a negativo (hacia abajo) desde 2.5 segundos después del inicio del estímulo hasta el final de la EW. También se evaluó el pico de respuesta adicional después del final de la EW si la pendiente de la EDA permanecía positiva por 13.5 segundos hasta el primer pico después de la EW. Esto permite la extracción de información hasta el pico de la respuesta en lugar del final de la EW que de alguna manera es arbitraria y también evita la evaluación de un pico de respuesta después de que la pendiente EDA se hubiera vuelto negativa al final de la EW. Los inicios de respuesta se identificaron como

el inicio de un segmento de pendiente positiva (es decir, un cambio de pendiente negativa o cero a pendiente positiva) desde un punto de latencia (LP) de 0.5 segundos hasta el final del ROW. El uso del LP elimina la necesidad de evaluar la pendiente del EDA antes del inicio del estímulo y asegura que no se evalúen las respuestas que comienzan inmediatamente al inicio del estímulo. Cuando la pendiente del EDA ya es positiva antes del inicio del estímulo y permanece positiva durante todo el ROW, el inicio de la respuesta se imputó estadísticamente en función de un cambio estadísticamente significativo en la varianza de la pendiente positiva (con alfa = .001) para dos ventanas móviles adyacentes de un segundo desde el LP y hasta el final del ROW. Esto puede ser visualizado por los expertos como un cambio sustancial en el ángulo ascendente dentro de un segmento de pendiente positiva durante el ROW. El valor extraído fue la diferencia máxima entre un pico de respuesta y un inicio de respuesta anterior. En términos simples, la característica de respuesta del EDA se puede visualizar como la distancia máxima desde el inicio de una respuesta durante el ROW, después del LP, y hasta un punto máximo durante la EW.

Extracción de características cardiovasculares. La extracción de características para los datos del cardio fue similar a la del EDA, pero con dos diferencias. Los datos cardiovasculares se extrajeron de un promedio móvil de todos los puntos registrados en los datos cardiovasculares. La media móvil se calculó pasando cuatro veces los datos del cardio a través de un filtro de media móvil de 0.5 segundos. El resultado del filtro de media móvil puede verse como la línea media entre los picos sistólico y diastólico. La inclusión de un pico de respuesta adicional, después de la EW de 15 segundos, se mantuvo si la pendiente del cardio permanecía positiva por 14.5 segundos hasta el pico de respuesta inicial después de la EW. Este cambio fue necesario para mejorar la capacidad del procedimiento de extracción de características cuando se requiere tolerar la complejidad potencial de la actividad cardiovascular, que a veces puede verse influida por la actividad respiratoria además de la influencia de los factores cognitivos, emocionales y conductuales. De manera similar al EDA, la característica de respuesta cardiovascular se puede visualizar como la distancia máxima desde el inicio de una respuesta durante el ROW hasta un punto máximo durante el EW. Una diferencia entre la extracción automática de características y la extracción manual de características es que con mayor frecuencia el examinador humano evalúa los datos cardiovasculares en la línea base diastólica. Se piensa que este procedimiento mejora la confiabilidad de la extracción visual/manual de características y se basa en una fuerte correlación entre la información contenida en la línea diastólica y la línea media.

Transformación numérica y reducción de datos. Todas las respuestas fisiológicas se midieron en unidades adimensionales – no intentan representar una cantidad física. Esto permite escalar los datos para su visualización y trazarlos sin que tenga efecto sobre las transformaciones numéricas en la comparación de valores de las RQ y CQ. Se asumió que los datos eran ordinales e intervalares. Para los datos de EDA y cardio, los valores extraídos más grandes se asociaron con mayores cambios en la actividad fisiológica. Para los datos de

respiración, la característica de interés en las pruebas PDD es una reducción de la actividad respiratoria en respuesta a las RQ y CQ. Esto significó que los valores de respiración más pequeños se asociaron con un mayor cambio en la actividad fisiológica para las puntuaciones de la respiración.

Suposiciones y Restricciones. No se asumió que los datos fisiológicos registrados ni los valores extraídos fueran lineales, y no se emplearon suposiciones de proporciones en la transformación de los valores de respuesta extraídos a valores enteros ESS. Sin embargo, se emplearon algunas restricciones lineales para evitar la extracción y puntuación de valores extremos. Los valores extremos se definieron como menores al 2% del valor máximo de escala para la visualización y el trazo. Los valores inferiores al umbral del 2% se consideraron ruido potencial y por lo tanto, menos probable que fueran una respuesta auténtica ante los estímulos de la prueba, y no se utilizaron para la transformación de los valores extraídos en puntuaciones ESS. Se calcularon las puntuaciones z de dejar-uno-fuera para todos los valores extraídos de cada sensor de registro dentro de cada gráfico de prueba registrado. Para los datos de EDA y cardio, las puntuaciones z superiores a 10 (es decir, 10 desviaciones estándar) se consideraron artefactos de datos, posiblemente como resultado de un movimiento físico, y se excluyeron del análisis. Ninguna de las respuestas superó este valor ($z > 10$) para estos datos de muestreo, aunque otras muestras han incluido artefactos de respuesta que superan las 10 desviaciones estándar. Se asumió que los valores de respuesta extraídos fueron monotónicos en los cambios de fisiología. Es decir, se asumió que los mayores cambios en la actividad fisiológica estaban asociados con diferencias en los valores numéricos extraídos.

Puntuaciones enteras de ESS. Se asignaron los puntajes enteros del ESS utilizando una escala de tres posiciones con valores con signo [+, 0, -], similar al procedimiento del método de puntuación de tres posiciones. Una diferencia entre las puntuaciones del ESS y las de tres posiciones es que las puntuaciones del ESS se obtienen utilizando solamente la característica de respuesta primaria, mientras que los métodos de tres y siete posiciones permiten el uso combinado de respuestas primarias y secundarias (National Center for Credibility Assessment, 2017).

Se le asignaron puntuaciones enteras ESS a cada RQ después de compararla contra la CQ apareada. Los pares RQ y CQ se seleccionaron mediante un algoritmo automatizado utilizando el procedimiento descrito por Nelson (2017c). Los casos FZCT de los datos de muestreo constaron de tres RQ, denominadas R5, R7 y R10. La R5 se compara contra la CQ adyacente (C4 que precede la RQ o C6 posterior a la RQ) dependiendo de qué CQ tenga el mayor cambio en la actividad fisiológica para cada sensor de registro. Se cree que esta forma de uso de dos CQ para R5 beneficia a las personas inocentes/veraces, ya que un cambio en la actividad fisiológica en la RQ tendrá que exceder el de las dos CQ antes de asignársele una puntuación de engaño; también brindará dos oportunidades para producir un cambio en la actividad fisiológica en una de las CQ que exceda el de la RQ. R7 se compara solo con la CQ previa (C6), y R10 también se compara solo con la CQ anterior. Las prácticas de campo permiten la rotación

de algunas preguntas durante las diversas iteraciones de la secuencia de preguntas del examen; esto tiene como objetivo distribuir o equilibrar los efectos relacionados con la posición de cada pregunta en la secuencia y también disuadir a los examinados de memorizar o habituarse a la secuencia de preguntas. Independientemente de la rotación de preguntas, la primera RQ de la secuencia se compara con las dos primeras CQ, mientras que la segunda y la tercera RQ se comparan solo contra las CQ previas. Los examinadores poligráficos de campo a veces se refieren a cada RQ y CQ apareadas como un *punto de análisis (análisis de spot)*.

Restricciones respiratorias y puntuaciones numéricas. Para los datos respiratorios se asignaron puntajes enteros ESS con valor de signo + indicativo de veracidad cuando se observó un mayor cambio fisiológico en respuesta a la CQ, mientras que se asignaron puntuaciones con valor de signo -, como indicativo de engaño, cuando se observó un mayor un cambio fisiológico en respuesta a la RQ. En contraste con los datos de EDA y cardio, en los datos respiratorios se observan los mayores cambios fisiológicos cuando los valores extraídos son más pequeños - lo que indica una mayor reducción o supresión de la actividad respiratoria. Para evitar el análisis de cambios o valores extremos que pueden ser el resultado de una actividad voluntaria o deliberada - como la que a veces se observa en personas que intentan alterar o falsificar sus datos y resultados de prueba - se empleó una proporción de restricción de respiración máxima de 1.5:1 en el análisis de los puntos (*spot*) RQ y CQ. Esta restricción está destinada a evitar la asignación de una puntuación con signo entero cuando se toma una respiración profunda o se contiene la respiración en respuesta a una RQ o CQ. Además, se utilizó una proporción de restricción respiratoria mínima de 1.25:1 para evitar la asignación de una puntuación numérica a las diferencias en la respuesta que no se deben a los estímulos de la prueba - y que pueden considerarse ruido resultante de una variación normal/no controlada en la actividad respiratoria, o debido a la inestabilidad observada que algunas personas presentan en su frecuencia y amplitud respiratoria.

Las puntuaciones del ESS para los sensores de respiración abdominal y torácica se combinan en una puntuación ESS única, mediante el procedimiento descrito por Nelson y Krapohl (2019). Este procedimiento es común en otros métodos de puntuación manual y resultará familiar para muchos profesionales de campo. Las puntuaciones se combinan a un valor de cero (0) cuando los valores de los signos son opuestos para los sensores torácico y abdominal, y se reducen a una única puntuación cuando no son opuestos.

Restricciones para EDA y cardio y puntuaciones numéricas. Para los datos del EDA y cardio, se asignaron puntuaciones enteras ESS de valor con signo +, indicativo de veracidad, cuando se observó un mayor cambio en la actividad fisiológica en la CQ. Se observan mayores cambios en la fisiología de los datos del EDA y cardio cuando los valores extraídos son más grandes. Se asignaron puntuaciones con valor de signo -, indicativo de engaño, cuando se observó un mayor cambio en la actividad fisiológica en respuesta a la RQ. Se asignaron puntuaciones de 0 (valor de signo 0) cuando no había diferencia observable o apreciable entre las respuestas en un punto de análisis RQ y CQ. Se utilizó una restricción mínima para evitar

la asignación de puntuación a EDA y cardio para los puntos de análisis RQ y CQ para los que la diferencia observada en la magnitud de la respuesta era pequeña. La restricción seleccionada para este proyecto fue una proporción de 1.05: 1 para los puntos de análisis RQ y CQ. Esta restricción fue el resultado de la optimización escalonada de los coeficientes de correlación y característica operativa del receptor (ROC) contra otros datos. Es más probable que las diferencias menores al 5% sean el resultado del ruido fisiológico, y también pueden ser el resultado de una influencia desconocida en los datos del EDA cuando se utiliza una solución de EDA con auto centrado cuyas características de diseño son desconocidas o no están documentadas.

Puntuaciones ponderadas del EDA. Las puntuaciones enteras ESS para los datos del EDA se ponderan más que las de los otros sensores de grabación. Esto se logra duplicando todos los valores enteros + y - a +2 y -2. El efecto de esto es aproximar los coeficientes estructurales y estadísticos que se han reportado en numerosos estudios acerca del análisis de datos PDD y del desarrollo de algoritmos informáticos durante un período de casi cinco décadas. [Refiérase a Nelson (2019) para un estudio de literatura sobre coeficientes estructurales y estadísticos para respiración, EDA, datos cardiovasculares y vasomotores.]

Reducción de datos. Debido a que las puntuaciones del ESS son valores con signos enteros, la reducción de datos es una sumatoria simple. Los puntajes subtotales se suman para cada RQ, incluyendo todos los sensores y todas las iteraciones de la secuencia de preguntas. Posteriormente se suman las puntuaciones subtotales para obtener una puntuación de gran total. Aunque las prácticas de campo y otros análisis a menudo pueden hacer uso de puntajes subtotales, este proyecto solo involucró el análisis del puntaje por gran total.

Función de verosimilitud multinomial y puntuaciones de corte numéricas.

Función de verosimilitud. La forma más simple de una función de verosimilitud es una puntuación de corte numérica que corresponde a una probabilidad empírica conocida de una clasificación correcta o incorrecta. Formalmente, una función de verosimilitud es una herramienta - que probablemente incluye una fórmula matemática o estadística, una función de computadora o una tabla de referencia publicada - que se puede utilizar para calcular u obtener un valor estadístico de los datos de prueba observados. Las puntuaciones de corte para las puntuaciones del ESS se pueden obtener a partir de tablas de referencia multinomiales y análisis Bayesianos.

Tanto las puntuaciones de gran total como las de subtotales se pueden utilizar para clasificar los datos de la prueba como indicativos de engaño o de veracidad – que es la alegoría contextual de los términos generales de *positivo y negativo*. Los estudios publicados han demostrado sistemáticamente que las puntuaciones de gran total proporcionan las tasas más altas de

precisión al realizar una clasificación. [Consulte APA (2011) para obtener un resumen de los tamaños del efecto para las técnicas de polígrafo validadas.] Esto no es sorprendente si se considera que los puntajes de gran total utilizan más información que los puntajes subtotales y por lo tanto, proporcionarán una variación más reducida y una mayor oportunidad de convergencia de datos. Esto a veces se conoce como la *débil ley de los grandes números* (Dekking, 2005) - relacionada con el teorema del límite central que establece que las medias de las muestras seleccionadas al azar se distribuirán normalmente y convergerán hacia la media poblacional desconocida. Es la razón principal por la que es más ventajoso tener muchas muestras (para lo cual también se puede utilizar un meta análisis) y la razón por la que se prefieren las muestras más grandes a las más pequeñas.

La pregunta de mayor importancia es la siguiente: ¿qué puntuaciones de corte numéricas son más efectivas o eficientes para clasificar los resultados de las pruebas como indicativos de engaño o de veracidad? O de manera más precisa, ¿qué inferencias probabilísticas sobre el engaño y la veracidad se pueden hacer acerca de las puntuaciones numéricas y las clasificaciones resultantes? Debido a que las pruebas científicas y el análisis de datos de pruebas científicas son intrínsecamente probabilísticos (dado que el propósito de cualquier prueba científica es cuantificar un fenómeno de interés que no puede estar sujeto a una medición física), los examinadores de campo y los administradores de programas estarán principalmente interesados en esta versión práctica de la misma pregunta: ¿qué puntuaciones numéricas proporcionarán una experiencia óptima de resultados correctos vs incorrectos? En términos científicos, esta es la cuestión de seleccionar puntajes de corte numéricos que optimizarán la observación deseada de los resultados de TN, TP, FN y FP. En el contexto poligráfico, la respuesta a esta pregunta se toma en cuenta con respecto al potencial del resultado adicional inconcluso.

Teoría analítica de las pruebas poligráficas. La teoría analítica de las pruebas poligráficas - bajo la cual se calculan las distribuciones multinomiales de las puntuaciones ESS y de tres posiciones - sostiene que los mayores cambios fisiológicos se cargarán ante diferentes tipos de estímulos de prueba (es decir, preguntas relevantes y de comparación) en función del engaño o veracidad, en respuesta a los estímulos relevantes u objetivo de prueba. [Consulte Nelson (2016) para una discusión de la teoría analítica de la prueba del polígrafo.] En la prueba del polígrafo, se espera cierta variación no controlada al nivel de cada sensor y de cada presentación de cada RQ (por ejemplo, no se espera que los puntajes tendrán un valor uniforme de signo). En la medida en que la teoría de las pruebas de PDD es válida (respaldada por evidencia), y los sensores de PDD registran datos válidos (datos y puntajes que se cargan en función del engaño o veracidad y no por pura aleatoriedad), la convergencia de los puntajes de subtotal y gran total se pueden utilizar para hacer inferencias estadísticas sobre la realidad - el grado en que una persona es *probablemente engañosa o veraz*. En otras palabras, es la suma de las puntuaciones subtotales y de gran total lo que se utilizará para clasificar los datos de la prueba como indicativos de engaño o de veracidad.

La suma de puntajes de múltiples RQ, múltiples CQ, múltiples sensores de grabación y múltiples iteraciones de la pregunta se basa en la ley de los grandes números (LLN) - que para la sumatoria de los puntajes PDD convergerán para cargarse a un valor de signo ya sea + o - en función del engaño o veracidad en respuesta a las RQ. El LLN también explica el por qué se espera que en general la precisión de clasificación con puntajes de gran total siga superando la precisión de clasificación con los puntajes subtotales.

La distribución multinomial se calcula bajo la teoría analítica. La distribución matemática/estadística del valor de los datos (es decir, todos los puntajes posibles del ESS y las probabilidades asociadas con cada uno de ellos) se puede caracterizar empíricamente cuando se obtienen datos de la realidad. También se puede calcular una distribución de las puntuaciones del ESS al solo utilizar la información sujeta a pruebas matemáticas y lógicas según una teoría propuesta. A menudo, la caracterización matemática de una distribución de puntajes se logra bajo la hipótesis nula de una teoría. Esto se debe a que a menudo es difícil (léase: imposible) caracterizar matemáticamente la teoría propuesta, mientras que la (hipótesis nula) a menudo puede caracterizarse fácilmente como una distribución de valores aleatorios. Una distribución bien conocida es la distribución Gaussiana o normal (curva de campana). Usamos nuestro conocimiento matemático de la distribución estadística para hacer inferencias sobre casos individuales con respecto a la población de todas las posibilidades que están representada por la distribución estadística. En el contexto poligráfico, debido a que existe un número finito, aunque grande, de todas las combinaciones posibles de puntajes ESS para todas las iteraciones de todas las preguntas y todos los sensores, la distribución estadística de los puntajes ESS no es Gaussiana, sino multinomial. La distribución de las puntuaciones ESS es multinomial porque hay tres valores posibles para cada puntuación.

La distribución multinomial de las puntuaciones del ESS exhibirá una forma de campana, de alguna manera similar a la distribución normal, aunque con valores discretos para cada puntuación posible de prueba. Bajo la hipótesis nula - que las puntuaciones no se cargan de manera sistemática y, por lo tanto, pueden caracterizarse como aleatorias - la mayoría de las puntuaciones multinomiales ocurrirán cerca de la mitad de la distribución (cerca de cero) con solo una forma posible de lograr la puntuación máxima o mínima. (puntuaciones + o - uniformes para cada iteración de cada pregunta y cada sensor). Existe un número finito, aunque grande, de combinaciones posibles de las puntuaciones [+, 0, -] para cada examen. También hay un número finito de formas de organizar las puntuaciones [+, 0, -] para lograr cada puntuación posible.

La distribución multinomial de puntuaciones es una lista de todas las puntuaciones posibles y la probabilidad asociada con cada una de ellas; se puede calcular mediante una fórmula combinatoria. También se puede calcular (a veces de manera más fácil y rápida) mediante la simulación Monte-Carlo. (Los cálculos multinomiales durante el Proyecto Manhattan fueron un impulso para el desarrollo de métodos estadísticos Monte Carlo.) Las distribuciones multinomiales para puntajes ESS y de tres posiciones (Nelson, 2017a; 2018a) son un cálculo exacto. Lo más importante es que nuestro conocimiento e información sobre la distribución

multinomial de las puntuaciones del ESS se puede utilizar para hacer inferencias estadísticas sobre la realidad (es decir, clasificaciones bajo incertidumbre). Todo lo que se necesita hacer es calcular primero la estadística de probabilidad para una puntuación observada, si se carga hacia el engaño o veracidad, y luego usar el valor estadístico de la distribución multinomial como una función de verosimilitud en el análisis Bayesiano, de la probabilidad de engaño o verdad.

Además de las puntuaciones del ESS, también se han publicado distribuciones multinomiales para las puntuaciones de tres posiciones (Nelson, 2018a). Esto es posible porque el método de tres posiciones se basa en la regla de *más grande es mejor* en la cual las reacciones que se registran y miden, independientemente de si se utilizan unidades de medida estandarizadas o adimensionales/arbitrarias, son objetivamente más grandes, más pequeñas o equivalentes. Estas diferencias son mayores porque existe una prueba matemática de que los números sucesivos, ya sean positivos o negativos, pueden en realidad representar cantidades más grandes o pequeñas - incluso cuando a esas cantidades no se les asigna una unidad de medida estandarizada. Por esta razón, los resultados de este análisis también se calcularon para las puntuaciones de gran total de tres posiciones federales. Desafortunadamente, no existe una distribución multinomial para los puntajes federales de siete posiciones, debido a decisiones arbitrarias (es decir, sin prueba matemática) de las diferencias en los valores de la escala de siete posiciones que correspondan a la actividad fisiológica. La automatización de los puntajes de siete posiciones no se puede lograr utilizando solo hechos e información sujeta a pruebas matemáticas y lógicas, y por esta razón las preguntas sobre la precisión de la clasificación del puntaje de gran total de siete posiciones no se abordaron en este proyecto.

Puntajes de corte multinomiales Bayesianos. Sería de gran conveniencia determinar los puntajes de corte numéricos que proporcionen tanto un clasificador estadístico como información sobre el significado práctico de la fuerza probabilística de la clasificación. Eso hacen precisamente los puntajes de corte multinomiales para calificaciones ESS (y de tres posiciones) junto con el análisis Bayesiano. Mientras que los primeros trabajos sobre algoritmos poligráficos se basaron en clasificadores estadísticos que no tenían la intención de ofrecer una intuición o inferencia práctica, las puntuaciones de corte multinomiales, calculadas mediante métodos Bayesianos, cuantifican las probabilidades prácticas o sistemáticas asociadas con el engaño o veracidad, además de la estimación del error aleatorio asociado con los diferentes resultados del análisis y otros datos no disponibles en el presente análisis. El análisis Bayesiano se basa en la suposición de que los datos disponibles de muestra/prueba son toda la información disponible para llegar a una conclusión (Stone, 2013; Winkler, 1972). Por el contrario, la inferencia frecuentista se basa en la suposición de que los datos disponibles de muestra/prueba son informativos de otros datos e información que potencialmente podrían obtenerse del universo y la realidad en lo que respecta al individuo y al objetivo conductual de una investigación en PDD. [Véase Nelson (2017d) para una breve descripción del análisis Bayesiano y de las pruebas de significancia de hipótesis nulas.] A menudo, los científicos y los métodos científicos pueden utilizar una combinación de suposiciones frecuentistas y Bayesianas. [Consulte Nelson (2018c) para obtener una descripción del análisis Bayesiano y del ESS-M.]

Puntajes de corte de ESS Multinomial para calificaciones de gran total. Para las puntuaciones de gran total de los casos de muestra FZCT, las puntuaciones de corte multinomiales son de -3 o menos para clasificaciones de engaño y +3 o más para clasificaciones de veraces. Se seleccionaron puntuaciones de corte de la distribución multinomial de todas las puntuaciones posibles de ESS (también la distribución de todos los puntajes de corte posibles de ESS) hasta el punto en el cual la estimación del error aleatorio - indicada por el límite inferior del intervalo creíble Bayesiano - proporciona una probabilidad estadísticamente significativa (con un alfa =.05) que mantuviera el mismo resultado analítico, a pesar de la variación esperada, si fuera posible repetir el examen o análisis varias veces. [Consulte Nelson (2018d) para obtener una ilustración gráfica sobre el cálculo de los puntajes de corte Bayesianas de ESS-M.]

Puntajes de corte multinomiales para puntajes totales de tres posiciones. Para las calificaciones con tres posiciones, los puntajes de corte multinomiales se pueden calcular utilizando los mismos métodos analíticos Bayesianos que para el ESS. Los puntajes de corte multinomiales para las puntuaciones de gran total de tres posiciones son -2 o menos para las clasificaciones de engaño y +2 o más para las clasificaciones veraces. [Consulte también Nelson (2020) para ver una demostración tabular del ESS-M y puntajes de corte de tres posiciones para un rango de probabilidades previas y diferentes niveles alfa para clasificaciones de engaño y veracidad.] La Tabla 1 muestra las puntuaciones de corte multinomiales para puntuaciones de gran total con el método de puntuación de tres posiciones, junto con las puntuaciones de corte tradicionales para puntuaciones de gran total.

Tabla 1. Puntajes de corte tradicionales y multinomiales para puntuación de gran total con ESS y Federal de 3 posiciones

	Tradicional		Multinomial	
	Engaño Indicado	Engaño No Indicado	Engaño Indicado	Engaño No Indicado
ESS	-6	+6	-3	+3
Tres posiciones	-6	+6	-2	+2

Puntajes de corte tradicionales. Las puntuaciones de corte numéricas tradicionales se seleccionaron inicialmente para los métodos de puntuación previos y complejos de siete posiciones; ellos también se derivaron inicialmente de manera empírica y heurística y posteriormente se sometieron a análisis para determinar su eficiencia de clasificación. Las puntuaciones de corte tradicionales para los puntajes de gran total de los exámenes FZCT son -6 o menos para las clasificaciones de engaño y +6 o más para las clasificaciones de veracidad. Una consideración importante es que los estándares de práctica de campo para los examinadores federales que usan el FZCT, no involucran solamente el uso de puntajes de gran total, sino que involucran una combinación de puntajes totales y subtotales. Aunque es tentador ahondar aquí y ahora en una investigación empírica de esos procedimientos, y aunque se ha publicado poco trabajo sobre el tema de las reglas de decisión desde Senter y Dollins (2003), el propósito de este proyecto era solo avanzar en el conocimiento disponible

sobre el tamaño del efecto para las puntuaciones de corte numéricas para las puntuaciones de gran total.

Otra consideración de importancia es que las puntuaciones de corte tradicionales de gran total también se utilizan con el método de puntuación de tres posiciones, lo que con este método se generan tasas más altas de resultados inconclusos (APA, 2011) y la necesidad de dedicar recursos adicionales a la resolución de estos resultados. Las tasas más altas de inconclusos también crean condiciones para el surgimiento o dependencia de soluciones encubiertas para reducir su ocurrencia. Lo más importante es que las puntuaciones de corte tradicionales se sugirieron por primera vez hace décadas para los métodos de puntuación de siete posiciones previos y complejos, y se han mantenido sin cambios a pesar de las innovaciones científicas en el análisis de datos de PDD y a pesar de las diferencias conocidas y esperadas en la distribución de las puntuaciones posibles. Mantenerse en el uso de estas puntuaciones de corte tradicionales es un reflejo del hecho de que, aunque tal vez sean subóptimos, los efectos de los resultados se conocen razonablemente y aún no se ha decidido una solución óptima, idealmente respaldada tanto por la teoría como por la evidencia científica.

Interpretación y clasificación de los resultados analíticos.

Para su uso en este documento, interpretación se refiere a la traducción de los resultados numéricos y estadísticos de las pruebas a resultados categóricos de prueba con los cuales se pueden tomar decisiones consistentes y racionales. La interpretación y clasificación del resultado de la prueba se logra de manera procedimental mediante el uso de reglas de decisión estructuradas. Debido a que este proyecto implica el estudio de las puntuaciones de corte de gran total, la regla de decisión de interés es la GTR.

Regla de gran total. La ejecución de la GTR es una suma de las puntuaciones subtotales para después obtener una puntuación de gran total. Luego, la puntuación de gran total se compara contra las puntuaciones de corte numéricas de las del gran total. Las puntuaciones de corte multinomiales para los métodos ESS y de tres posiciones se muestran en la Tabla 3. Para ESS, las puntuaciones de corte multinomiales son -3 o menos para las clasificaciones de engaño y +3 o más para las clasificaciones de veracidad. Para puntuaciones de tres posiciones, las puntuaciones de corte multinomiales son -2 o menos para clasificaciones de engaño y +2 o más para clasificaciones de veracidad. Estas puntuaciones de corte suponen una probabilidad previa de 0.5. Las puntuaciones de corte numéricas tradicionales para el gran total son -6 o más bajas para las clasificaciones de engaño y +6 o más para las clasificaciones de veracidad. Las puntuaciones de corte tradicionales se obtuvieron para los primeros métodos de puntuación de siete posiciones. La intuición sugiere que pueden ser ineficaces para las puntuaciones del ESS y de tres posiciones - lo que genera tasas más altas de resultados inconclusos y una dependencia excesiva de las puntuaciones subtotales. Aunque el uso de puntajes subtotales junto con los puntajes de gran total puede mejorar la clasificación con casos de engaño, va a introducir efectos de multiplicidad estadística y puede sesgar la precisión

general de maneras desafortunadas o no intencionales. Por esta razón, la comprensión y la selección de puntuaciones de corte óptimas de gran total pueden aumentar los tamaños del efecto de precisión para los casos FZCT.

Clasificación analítica Bayesiana del engaño o veracidad. Las puntuaciones de corte de gran total multinomiales, tanto para puntuaciones ESS como para tres posiciones, proporcionan una estimación de posibilidades posteriores Bayesianas (error sistemático) de aproximadamente 2:1 de engaño y veracidad, lo que permite una probabilidad de 1-alfa x 100% = 95% de observar otro resultado analítico de al menos este valor. En términos prácticos, los puntajes de las pruebas a este nivel, son suficientes para aceptar la noción de que la actividad fisiológica registrada se cargó sistemáticamente, y para rechazar la noción de que los puntajes se cargaron de una forma aleatoria o sin sentido, no interpretable/no clasificable. Aunque las posibilidades 2:1 pueden no ser espectaculares, es importante reconocer que con clasificaciones realizadas con una puntuación de +/-2 o +/-3 no puede esperarse razonablemente que proporcionen precisiones espectaculares, cuando se considera el rango de distribución de las puntuaciones posibles. De igual importancia, las posibilidades posteriores de 2:1 pueden proporcionar conocimiento procesable bajo ciertas circunstancias. Por ejemplo: considere la información donde las probabilidades de que un puente en particular colapse se estiman en 2:1. Muchas personas razonables podrían dudar realmente en hacer uso de ese puente. Por supuesto, también existirán circunstancias que requieran una base de información probabilística procesable más sólida que las posibilidades posteriores de 2:1. Para la mayoría de las estimaciones de las probabilidades previas, estas necesidades se pueden satisfacer mediante los datos de referencia multinomiales y la selección de puntuaciones de corte numéricas que limitarán las tasas de error sistemático y aleatorio a los niveles requeridos.

Resultados

Se calcularon clasificaciones para el engaño y veracidad para cada una de las dos muestras del FZCT. Debido a que los practicantes de campo de polígrafo comúnmente discuten los efectos de precisión de las pruebas en términos de las proporciones de conclusiones correctas, incorrectas e inconclusos, se presentan los efectos de precisión de esta manera, en lugar de usar tamaños de efectos que comparan clasificaciones con respecto a los niveles de azar.

Resultados de la muestra 1, n=100 exámenes confirmados de campo FZCT.

No se presentaron casos que cambiaron de clasificación positiva a negativa y ningún caso cambió de clasificación negativa a positiva debido al método de puntuación o al tipo de puntuación de corte para esta muestra (n=100) de casos confirmados de campo. La Tabla 2 muestra las métricas de precisión de la prueba para las clasificaciones que utilizan el GTR con puntuaciones de corte tradicionales y multinomiales para puntuaciones ESS y de tres posiciones. En la Tabla 2 se incluyen las tasas de error y de inconclusos, junto con la

proporción de clasificaciones correctas excluyendo los resultados inconclusos. También se incluyen en la Tabla 2 las tasas de sensibilidad (TP) y especificidad (TN), junto con las tasas de error FN y FP. Las otras métricas en la Tabla 2 son el valor predictivo positivo (VPP) calculado como la proporción de los resultados positivos verdaderos y de todos los resultados positivos, y del valor predictivo negativo (VPN) que es la proporción de clasificaciones negativas verdaderas y de todas las clasificaciones negativas. También se incluye la proporción de decisiones correctas para los casos culpables e inocentes excluyendo los resultados inconclusos, junto con la precisión no ponderada y las tasas no ponderadas de inconclusos.

Tabla 2. Clasificaciones de gran total para n=100 FZCT muestras de campo con puntuaciones ESS y tres-posiciones

	puntuaciones ESS		puntuaciones tres-posiciones	
	Puntajes de Corte Tradicionales	Puntajes de Corte Multinomial	Puntajes de Corte Traditionales	Puntajes de Corte Multinomial
Error	[5] .05 (.02) {.01 to .10}	[5] .05 (.02) {.01 to .10}	[3] .03 (.02) {.01 to .07}	[5] .05 (.02) {.01 to .10}
Inconclusos	[35] .35 (.08) {.15 to .46}	[11] .11 (.05) {.01 to .18}	[53] .53 (.09) {.25 to .58}	[14] .14 (.05) {.01 to .21}
Correctos	[60] .92 (.03) {.85 to .98}	[84] .94 (.02) {.89 to .99}	[44] .94 (.04) {.85 to .99}	[81] .94 (.03) {.88 to .99}
Sensibilidad (TP)	[33] .66 (.07) {.53 to .79}	[44] .88 (.05) {.78 to .96}	[28] .56 (.07) {.43 to .70}	[43] .86 (.05) {.75 to .95}
Epecificidad (TN)	[27] .54 (.07) {.40 to .68}	[40] .80 (.06) {.69 to .91}	[16] .32 (.07) {.19 to .46}	[38] .76 (.06) {.64 to .87}
Errores FN	[2] .04 (.03) {.01 to .10}	[2] .04 (.03) {.01 to .10}	[1] .02 (.02) {.01 to .07}	[2] .04 (.03) {.01 to .10}
Errores FP	[3] .06 (.03) {.01 to .13}	[3] .06 (.03) {.01 to .13}	[2] .04 (.03) {.01 to .10}	[3] .06 (.03) {.01 to .13}
Casos Inc culpables	[15] .30 (.06) {.18 to .43}	[4] .08 (.04) {.02 to .16}	[21] .42 (.07) {.29 to .55}	[5] .10 (.04) {.02 to .19}
Casos Inc inocentes	[20] .40 (.07) {.27 to .54}	[7] .14 (.05) {.05 to .24}	[32] .64 (.07) {.50 to .78}	[9] .18 (.05) {.08 to .29}
PPV	.92 (.05) {.82 to .99}	.94 (.04) {.86 to .99}	.93 (.05) {.83 to .99}	.93 (.04) {.86 to .99}
NPV	.93 (.05) {.83 to .99}	.95 (.03) {.88 to .99}	.94 (.06) {.80 to .99}	.95 (.03) {.88 to .99}
Casos culpables correctos	.94 (.04) {.86 to .99}	.96 (.03) {.89 to .99}	.97 (.03) {.89 to .99}	.96 (.03) {.88 to .99}
Casos inocentes correctos	.90 (.06) {.78 to .99}	.93 (.04) {.85 to .99}	.89 (.08) {.72 to .99}	.93 (.04) {.84 to .99}
Inc. no ponderados	.35 (.05) {.26 to .44}	.11 (.03) {.05 to .18}	.53 (.05) {.43 to .62}	.14 (.03) {.07 to .21}
Precisión no ponderada	.92 (.03) {.85 to .98}	.94 (.02) {.89 to .99}	.93 (.04) {.84 to .99}	.94 (.03) {.88 to .99}

* Las celdas muestran la [frecuencia] aunado al boostrap estimado de la media, (desviación estándar) y {95% de intervalo de confianza}.

La inspección de las filas de la Tabla 2 nos indica que los intervalos de confianza se superponen sustancialmente en cuanto a las proporciones de errores producidos por los cuatro tratamientos. Sin embargo, se pueden observar algunas diferencias en la sensibilidad, especificidad y resultados inconclusos. Tanto la sensibilidad al engaño como la especificidad a la veracidad fueron mayores en las puntuaciones ESS y en las multinomiales. Las tasas de inconclusos fueron más bajas en las puntuaciones ESS y en las de corte multinomiales. La frecuencia de los resultados TP y TN fue mayor para las puntuaciones de corte multinomiales y ESS y más baja con las de corte tradicionales y de tres posiciones. Las frecuencias de resultados inconclusos fueron más altas para los cortes tradicionales y más bajas para los multinomiales.

Resultados de la muestra 2, n=60 exámenes confirmados de campo FZCT.

Para la segunda muestra de n=60 casos de FZCT, se presentaron casos en los que la clasificación cambiara de positivo a negativo o de negativo a positivo como resultado del método o del tipo de puntuación.

La Tabla 3 muestra las mismas métricas de precisión de prueba para los puntajes de tres posiciones ESS y Federal para la segunda muestra de archivo.

Tabla 3. Clasificaciones por gran total para la muestra de campo n=60 FZCT con puntajes ESS y de tres posiciones*

	puntajes ESS		puntajes de tres posiciones	
	Puntajes de corte tradicionales	Puntajes de corte Multinomial	Puntajes de corte tradicionales	Puntajes de corte Multinomial
Error	[2] .03 (.02) {.01 to .08}	[2] .03 (.02) {.01 to .08}	[1] .02 (.02) {.01 to .05}	[5] .08 (.04) {.02 to .15}
Inconclusos	[21] .35 (.10) {.10 to .50}	[7] .12 (.04) {.01 to .13}	[32] .53 (.11) {.22 to .63}	[6] .10 (.06) {.01 to .2}
Correctos	[37] .95 (.04) {.87 to .99}	[51] .96 (.03) {.91 to .99}	[27] .96 (.04) {.89 to .99}	[49] .91 (.04) {.82 to .98}
Sensibilidad (TP)	[21] .68 (.09) {.50 to .84}	[29] .94 (.05) {.83 to .99}	[17] .55 (.09) {.37 to .73}	[27] .90 (.05) {.79 to .99}
Especificidad (TN)	[16] .53 (.09) {.34 to .72}	[22] .73 (.08) {.57 to .89}	[10] .33 (.09) {.17 to .50}	[22] .73 (.08) {.56 to .88}
Errores FN	[0] .03 (.03) {.01 to .11}	[0] .03 (.03) {.01 to .11}	[0] .03 (.03) {.01 to .11}	[1] .03 (.03) {.01 to .11}
Errores FP	[2] .07 (.05) {.01 to .17}	[2] .07 (.05) {.01 to .17}	[1] .03 (.03) {.01 to .11}	[4] .13 (.06) {.03 to .26}
Casos culpables Inc	[9] .29 (.08) {.14 to .46}	[1] .03 (.03) {.01 to .11}	[13] .42 (.09) {.24 to .60}	[2] .07 (.05) {.01 to .17}
Casos inocentes Inc	[12] .40 (.09) {.23 to .59}	[6] .20 (.07) {.07 to .35}	[19] .64 (.09) {.45 to .81}	[4] .14 (.06) {.03 to .27}
PPV	.91 (.06) {.77 to .99}	.94 (.04) {.83 to .99}	.95 (.06) {.82 to .99}	.87 (.06) {.74 to .97}
NPV	.94 (.06) {.80 to .99}	.96 (.04) {.85 to .99}	.91 (.09) {.71 to .99}	.96 (.04) {.86 to .99}
Casos culpables correctos	.96 (.05) {.85 to .99}	.97 (.03) {.88 to .99}	.95 (.05) {.81 to .99}	.96 (.03) {.88 to .99}
Casos inocentes correctos	.89 (.08) {.71 to .99}	.92 (.06) {.79 to .99}	.91 (.09) {.70 to .99}	.85 (.07) {.70 to .97}
Inc. no ponderados	.35 (.06) {.23 to .47}	.12 (.04) {.05 to .20}	.53 (.06) {.40 to .65}	.10 (.04) {.03 to .18}
Precisión no ponderada	.92 (.05) {.82 to .99}	.94 (.03) {.87 to .99}	.93 (.05) {.81 to .99}	.91 (.04) {.82 to .98}

* Las celdas muestran la [frecuencia] aunado al boostrap estimado de la media, (desviación estándar) y {95% de intervalo de confianza}.

La inspección de las filas en la Tabla 3 indica que los resultados de la segunda muestra de FZCT coincidieron con los de la primera muestra. Los intervalos de confianza se superponen sustancialmente para las métricas de precisión para las clasificaciones correctas. La sensibilidad al engaño y la especificidad con la veracidad fueron mayores para las puntuaciones del ESS y las multinomiales. Las tasas de inconclusos fueron más bajas para las puntuaciones del ESS y para los cortes multinomiales.

Análisis de los datos de muestra combinados

Un ANOVA de mediciones repetidas en dos vías (método de puntuación con tipo de puntuación de corte x) mostró una interacción significativa para los resultados inconclusos [F (1,636) = 329,671, (p <.001)], lo que indica que las diferencias observadas en las tasas de inconclusos para las puntuaciones de corte multinomiales y tradicionales fueron diferente para los puntajes ESS y de tres posiciones. El ANOVA de una sola vía mostró que la reducción de los resultados inconclusos fue estadísticamente significativa a un nivel =.05 para las puntuaciones ESS [F (1318) = 4, (p = .046)] y también para las de tres posiciones [F (1318) = 12,308, (p <0,001)].

Un ANOVA de mediciones repetidas de tres vías para las clasificaciones correctas positivas y negativas mostró una interacción significativa de tres vías, para el estado de criterio (culpabilidad vs inocencia), el método de puntuación (ESS, tres posiciones) y el tipo de puntaje (tradicional, multinomial) [F (1.632) = 54.282, (p <.001)]. La Tabla 4 muestra el resumen ANOVA. Todos los efectos principales y las interacciones de dos vías también fueron significativos en el ANOVA de tres vías, pero no fueron interpretables debido a los efectos significativos de interacción.

Debido a que el interés principal de este proyecto fueron las diferencias en los efectos de precisión en función del tipo de puntaje de corte, se llevó a cabo una serie de contrastes unidireccionales. Para las puntuaciones ESS, el efecto unidireccional fue estadísticamente significativo para una mayor sensibilidad de la prueba para el engaño *[F (1,158) = 7.533, (p = .007)]* y para una mayor especificidad de la prueba para la veracidad *[F (1,158) = 5.347, (p = .022)].* Para las puntuaciones de tres posiciones, el efecto unidireccional también fue estadísticamente significativo tanto para el aumento de la sensibilidad de la prueba al engaño *[F (1,158) = 11.148, (p = .001)]* como para la especificidad de la prueba para la veracidad *[F (1,158) = 17,663 , (p <.001)].*

Proporciones de riesgo

Para comprender adecuadamente las diferencias en las tasas de sensibilidad, especificidad y de inconclusos que muestran estos datos, se calcularon las proporciones de riesgo después de transformar las proporciones observadas a posibilidades. Las proporciones de riesgo se basan en el supuesto de que las proporciones observadas son una estimación de la probabilidad o la fuerza de la posibilidad de observar un resultado similar con cualquier miembro de la población

seleccionado al azar, y se calculan como las proporciones observadas para dos métodos diferentes. En este proyecto, la comparación de interés es la proporción de riesgo para las diferencias entre los resultados de las puntuaciones de corte tradicionales y multinomiales. La Tabla 4 muestra las proporciones de riesgo para resultados verdaderos positivos, verdaderos negativos e inconclusos.

Tabla 4. Proporciones de riesgo para resultados TP, TN, e Inconclusos para puntajes de corte tradicionales y multinomial

	Puntajes ESS	Puntajes de tres posiciones
Inconcluso	.34 (2.9)	.19 (5.3)
Sensibilidad (TP)	1.38	1.38
Especificidad (TN)	1.64	2.21

Las proporciones de riesgo dan información con respecto a la probabilidad práctica de las diferencias en los resultados observados y cómo estos podrían experimentarse en un individuo o en un grupo de casos. Los índices de riesgo en la Tabla 4 sugieren que el uso de puntuaciones de corte multinomiales con puntajes ESS puede reducir la probabilidad o la ocurrencia de resultados inconclusos al 34% en comparación con las puntuaciones de corte tradicionales. Con las puntuaciones de tres posiciones, el riesgo de resultados inconclusos puede reducirse en aproximadamente el 20% utilizando las puntuaciones de corte tradicionales. Sin embargo, en la práctica de campo la diferencia observada en las tasas reales puede que no logre estas diferencias estimadas porque los profesionales de campo se involucran en una variedad de actividades para resolver o reducir la aparición de resultados inconclusos. La información que se muestra en la Tabla 4 indica que las personas culpables tienen 1.4 veces más probabilidades de ser detectadas si se utilizan puntajes de corte multinomiales, al tiempo que las personas inocentes pueden tener 1.6 veces más probabilidades de ser clasificadas como veraces.

Conclusión

Este proyecto, relacionado con el estudio de las puntuaciones de corte de gran total, involucró el uso de la GTR con puntuaciones ESS y de tres posiciones utilizando tanto las puntuaciones tradicionales como multinomiales. Se ha demostrado que la GTR, aunque quizás sea la más simple de todas las reglas de decisión PDD, proporciona las tasas más altas de precisión de clasificación general con respecto al resto de las reglas de decisión PDD. Se compararon los efectos de precisión para las puntuaciones ESS y de tres posiciones de los exámenes poligráficos de eventos específicos utilizando puntuaciones de corte numéricas tradicionales y multinomiales para las puntuaciones de gran total. Aunque el desarrollo de habilidades en la calificación manual con el ESS - como con las pruebas PDD en general - se adquiere de manera más efectiva en función tanto del conocimiento didáctico como académico de los procedimientos estandarizados, y de la supervisión y orientación práctica de otros profesionales experimentados, los conceptos básicos del ESS son simples y altamente

estructurados, lo que conduce a la posibilidad potencial de un proceso automatizado que se aproxima mucho a las actividades de los expertos humanos. Para asegurar que la varianza observada se pueda atribuir a diferencias en las puntuaciones de corte numéricas, y no debido a la variación esperada dentro de los límites de confiabilidad entre evaluadores con métodos de puntuación manual, durante este proyecto las puntuaciones y resultados del ESS y de tres posiciones se obtuvieron mediante un algoritmo informático incluyendo extracción de características, selección de puntos de análisis RQ y CQ, transformación numérica, reducción de datos del GTR con puntuaciones de corte tradicionales y multinomiales.

Se utilizaron datos de dos muestras diferentes de archivo de exámenes confirmados de campo FZCT. Estas muestras de archivo fueron descritas por los calificadores humanos como desafiantes, aunque los tamaños del efecto reportados previamente para algoritmos humanos y computacionales son consistentes con los que se muestran aquí. Los resultados se muestran por separado en forma de tabla para cada una de las muestras. Después, los datos de las dos muestras se combinaron para el análisis estadístico de las posibles diferencias en los tamaños del efecto de las puntuaciones ESS y de tres posiciones. Para las muestras combinadas, se observaron diferencias significativas en la sensibilidad de la prueba al engaño, la especificidad de la prueba para la veracidad y de la proporción de resultados inconclusos. El uso de puntuaciones de corte multinomiales redujo la aparición de resultados inconclusos y aumentó tanto la sensibilidad al engaño como la especificidad para la veracidad. El uso de datos de archivo permite la comparación directa de los tamaños del efecto observado con el efecto informado previamente, utilizando los mismos datos con otros métodos de puntuación para los cuales los examinadores de campo tienen un conocimiento intuitivo y experiencia de su efectividad.

Es de particular interés que en este proyecto ninguna de las clasificaciones de engaño o veracidad se revirtió como resultado de la selección de puntajes de corte tradicionales o multinomiales para puntajes de corte de gran total con puntajes ESS o de tres posiciones. Otra observación interesante de este proyecto es que los efectos de precisión para las clasificaciones correctas, incluyendo el VPP, el VPN y la proporción de clasificaciones correctas que excluyen los resultados inconclusos dentro de los culpables e inocentes, junto con la precisión no ponderada que excluyen los resultados inconclusos, fueron sustancialmente similares tanto para puntuaciones de corte tradicionales tanto como con multinomiales tanto con puntajes numéricos ESS como de tres posiciones. Sin embargo, el uso de puntuaciones de corte multinomiales aumentó la sensibilidad al engaño por un factor de 1.4 para las puntuaciones de ESS y de tres posiciones, al tiempo que aumentó la especificidad para la veracidad en un factor de 1.6 con las puntuaciones ESS y por un factor de 2.2 para las puntuaciones de tres posiciones. Las puntuaciones de corte multinomiales redujeron la aparición de resultados inconclusos en un factor de 2.9 con las puntuaciones ESS y en un factor de 5.3 con las puntuaciones de tres posiciones. El uso de puntajes de corte multinomiales y de gran total con el formato FZCT, que es un formato de asunto único con tres preguntas, logra el nivel de precisión requerido por los Estándares de Práctica de la American Polygraph Association para los exámenes evidenciarios - exámenes realizados bajo la expectativa de que los resultados y

la información puede ser utilizada en un procedimiento legal (American Polygraph Association, 2018)) – tanto para puntuaciones ESS y de tres posiciones.

Las posibles implicaciones prácticas de estos resultados incluyen la posibilidad de aumentar la efectividad de los polígrafos de campo para poder clasificar correctamente tanto el engaño como la veracidad. Otra implicación práctica, que es una consideración razonable en términos de precisión de clasificación, es que los programas de polígrafo hagan uso de las puntuaciones de corte numéricas tradicionales, pero con puntuaciones de ESS, si la perspectiva de cambiar tanto el tipo de puntuación (puntuaciones ESS o de tres posiciones) como las puntuaciones de corte representa un número incómodo de grados de libertad para quienes crean políticas. De hecho, hay mucha sabiduría práctica y científica al cambiar una variable a la vez mientras se observa y se adquiere experiencia con diferentes métodos. De forma más abierta, estos resultados respaldan que es razonable para los examinadores de campo y/o los creadores de políticas considerar solamente el puntaje de gran total para el formato FZCT - un hallazgo para el cual no hay base de información o justificación teórica que sugiera que no se pueda generalizar para todos los formatos poligráficos de asunto único.

Como todos los proyectos, este está sujeto a algunas limitaciones. Una diferencia entre los procedimientos utilizados durante este proyecto y los utilizados por los profesionales de campo cuando se puntúan manualmente, es que este proyecto se limita al análisis de puntuaciones de gran total. Los expertos humanos en la práctica de campo pueden depender de diferentes reglas de decisión según la política de su agencia. Aunque no todas las agencias optarán por utilizar solo el puntaje de gran total, los estudios publicados han mostrado consistentemente que los puntajes de gran total proporcionan las tasas generales más altas de precisión de clasificación. De esta manera se cree que el uso de puntajes de gran total proporciona una mayor comprensión de la influencia de los puntajes de gran total en la precisión general de la prueba, independientemente de las reglas de decisión utilizadas en la práctica de campo.

Este proyecto no intentó estudiar los efectos con reglas de decisión distintas a la GTR. El estudio de los efectos de la interacción tanto de puntajes numéricos como reglas de decisión que pueden hacer uso de puntajes de gran total y subtotales requeriría un análisis multivariante que está más allá del alcance de este análisis – que intentó ser una revisión descriptiva simple e intuitiva de efectos de precisión con puntuaciones de gran total. Un proyecto más completo habría investigado tanto el tipo de puntuación de corte como la regla de decisión. Sin embargo, un proyecto de este tipo ampliaría considerablemente la complejidad del análisis, junto con un aumento correspondiente en la complejidad y la información del análisis. Se pensó que limitar este proyecto a dos dimensiones - tipo de puntuación y corte de puntuación- proporcionaría información de uso potencialmente práctico y, al mismo tiempo, abordaría las preguntas analíticas con cierto grado de profundidad.

Hay motivos para esperar que exista alguna interacción entre la regla de decisión y la selección de puntuaciones numéricas. Una implicación obvia de estos resultados analíticos es que las puntuaciones de corte numéricas tradicionales para las puntuaciones de gran total, aunque no son inexactas, parecen ser ineficientes. Una consecuencia de esto es que los practicantes de campo, además de los instructores, el personal de control de calidad y los directores de

programas de poligrafía, pueden haber llegado a depender más de lo ideal en la puntuación subtotal para poder remediar ineficiencias. Una consecuencia de la dependencia excesiva en los puntajes subtotales para remediar ineficiencias, es que el uso de puntajes subtotales provoca efectos de multiplicidad estadística que complicarán los efectos de precisión de la prueba - con mucha frecuencia de manera que pueden hacer que la prueba parezca sesgada en contra de personas inocentes. La selección de una puntuación de corte de gran total óptima puede aumentar la precisión de la prueba tanto con personas culpables como con personas inocentes, al tiempo que alivia potencialmente la carga de los efectos de multiplicidad. La interacción entre las reglas de decisión y las puntuaciones de corte numéricas debería ser tema de análisis y estudio adicional.

Otra limitación de este estudio tiene que ver con las puntuaciones de tres posiciones. En este proyecto se lograron los puntajes de tres de posiciones mediante el aplanamiento de los puntajes ESS del EDA. Se desconoce hasta qué punto estos puntajes de tres posiciones pueden diferir de los logrados en contextos de práctica de campo donde los examinadores pueden hacer uso de características de respuesta secundarias y otras prácticas semi subjetivas que no están incluidas en el ESS y que no pueden estar sujetas a una automatización. En principio, se pueden extraer las puntuaciones de tres posiciones utilizando exactamente el mismo procedimiento automatizado que para las puntuaciones del ESS. Las puntuaciones de tres posiciones también se pueden lograr utilizando un sistema más complejo con características primarias y secundarias que se desarrolló para las puntuaciones de siete posiciones, y que es menos fácil de automatizar. Independientemente de esta limitación, es la opinión del autor que los valores de tres posiciones son suficientemente representativos para que estos resultados proporcionen información potencialmente útil.

Una limitación final es que las tasas de inconclusos observadas en este estudio, como en todos los estudios científicos, pueden ser mayores que las observadas en la práctica de campo. Esto era de esperarse. Se considera que los examinadores de campo y los programas de polígrafo actúan de manera razonable si se esfuerzan por resolver los resultados inconclusos con cada caso individual. En investigaciones de este tipo, tales esfuerzos equivaldrían a manipular los resultados de la investigación. Por esta razón, en proyectos de este tipo no se hace ningún esfuerzo para resolver o reducir la ocurrencia de resultados inconclusos para cada caso individual. Las diferencias en los resultados inconclusos son un reflejo del método de análisis y serán necesariamente mayores que las tasas de inconclusos en la práctica de campo.

Teniendo en cuenta las limitaciones reconocidas, los efectos de precisión observados en este análisis colocan al FZCT en el rango requerido por los estándares de práctica de la APA para los exámenes evidenciarios (APA, 2018). Los exámenes evidenciarios son aquellos para los que se realiza la prueba con la intención de presentar el resultado de la prueba como base de información en un proceso judicial. Los índices de precisión aquí observados igualan o superan los de otros formatos de PDD evidenciarios. Se recomienda un interés y una investigación continua en las puntuaciones ESS, el GTR y en el uso de puntuaciones multinomiales.

Referencias

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph, 40*, 194-305. [Electronic version] Retrieved August 20, 2012, from http://www.polygraph. org/section/research-standards-apa-publications.

American Polygraph Association (2018). *APA Standards of Practice (Effective September 1, 2018)*. Retrieved (January 20, 2019) from https://www.polygraph.org/apa-bylaws-and-standards.

Butterworth, S. (1930). On the theory of amplifiers. *Experimental Wireless and the Wireless Engineer, 7*, 536-541.

Dekking, Michel (2005). *A Modern Introduction to Probability and Statistics.* Springer.

Department of Defense (2006). *Federal psychophysiological detection of deception examiner handbook.* Available from the author. Reprinted in Polygraph, 40 (1), 2-66.

Editorial Staff. (2019). Introduction to the NCCA ASCII Standard. *Polygraph and Forensic Credibility Assessment, 48(2)*, 125-135.Polygraph & Forensic Credibility Assessment,

Krapohl, D. J. (2002). Short report: Update for the objective scoring system. *Polygraph, 31*, 298-302.

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

Krapohl, D. J. & McManus, B. (1999). An objective method for manually scoring polygraph data. *Polygraph, 28*, 209-222.

National Center for Credibility Assessment (2017). *Test Data Analysis: Numerical Evaluation Scoring System Pamphlet.* Available from the author. (Retrieved from http://www.antipolygraph.org on 4-13-2019).

National Research Council (2003). *The Polygraph and Lie Detection.* Washington, D.C.: National Academy of Sciences.

Nelson, R. (2015). Scientific basis for polygraph testing. *Polygraph 41(1),* 21-61.

Nelson, R. (2016). Scientific (analytic) theory of polygraph testing. APA Magazine, 49(5), 69-82.

Nelson, R. (2017a). Multinomial reference distributions for the Empirical Scoring System. *Polygraph and Forensic Credibility Assessment, 46 (2)*. 81-115.

Nelson, R. (2017b). Updated numerical distributions for the Empirical Scoring System: An accuracy demonstration with archival datasets with and without the vasomotor sensor. *Polygraph and Forensic Credibility Assessment, 46 (2)*, 116-131.

Nelson, R (2017c). Heuristic principles to select comparison and relevant question pairs when scoring any CQT format. *APA Magazine, 50(1)*, 73-83.

Nelson, R. (2017d). Five-minute science lesson: Bayesian and frequentist statistics – what's the deal? *APA Magazine, 50(4)*, 89-95.

Nelson, R. (2018a). Multinomial reference distributions for three-position scores of comparison question polygraph examinations. *Polygraph and Forensic Credibility Assessment, 47(2)*, 158-175.

Nelson, R. (2018b). Practical polygraph: a survey and description of decision rules. *APA Magazine, 51(2)*, 127-133.

Nelson, R. (2018c). Five minute science lesson: Bayes' theorem and Bayesian analysis. *APA Magazine, 51(5)*, 65-78.

Nelson, R. (2018d). Practical polygraph: a tutorial (with graphics) on posterior results and credible intervals using the ESS-M Bayesian classifier. *APA Magazine, 51(4)*, 66-87.

Nelson, R. (2019). Literature survey of structural weighting of polygraph signals: why double the EDA? *Polygraph and Credibility Assessment, 48(2)*, 105-112.

Nelson, R. (2020). Multinomial cutscores for Bayesian analysis with ESS and three-position scores of comparison question polygraph tests. *Polygraph and Forensic Credibility Assessment, 49(1),* 61-72.

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, 40,* 79-85.

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. Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B., & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph, 40,* 67-78.

Nelson, R. & Krapohl, D. K. (2017) Practical polygraph: a recommendation for combining the upper and lower respiration data for a single respiration score. *APA Magazine. 50(6)*, 31-41.

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/.

Senter, S. M. & Dollins, A. B. (2003). *New Decision Rule Development: Exploration of a two-stage approach. Report number DoDPI00-R-0001.* Department of Defense Polygraph Institute Research Division, Fort Jackson, SC. Reprinted in Polygraph, 37(2), 149-164.

Stone, J. (2013). *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis.* Sebtel Press.

Winkler, R. L. (1972). *An Introduction to Bayesian Inference and Decision.* Holt Mc-Dougal.