Revista Mexicana de Análisis de la Conducta,
1984, Vol. 10, Núm. 2, págs. 137-160

Algunas consideraciones sobre la utilización del Coeficiente r de Pearson como índice de acuerdo entre observadores[1]

Some comments on the use of Pearson's coefficient as an index of interobserver agreement.

Arturo Silva Rodríguez

Maestría en Modificación de Conducta, Escuela Nacional de Estudios Profesionales Iztacala.

RESUMEN

Uno de los temas poco abordados en la modificación de conducta, es el de los principios psicométricos de las técnicas de evaluación conductual. Sin embargo, recientemente se empezó a cuestionar si las técnicas desarrolladas bajo la tutela de la psicometría clásica eran adecuadas para la evaluación conductural, desarrollada ésta bajo marcos teóricos totalmente diferentes.

Este trabajo tiene como objetivo el hacer un análisis del coeficiente r de Pearson, como una técnica utilizada para estimar las propiedades psicométricas de algunos instrumentos de evaluación en situaciones de observación conductual directa. Así, en primer lugar, se hace un breve bosquejo de las técnicas de observación directa y posteriormente, se analiza al coeficiente r de Pearson como un estadístico de comparación de puntuaciones estándares y como modelo lineal determinado por la función $Y_t = \alpha + \beta X_t$

[1] Una versión preliminar de este trabajo se presentó en el XXIII Congreso Internacional de Psicología, Acapulco, 1984. Se pueden solicitar sobretiros a la Coordinación General de Estudios de Posgrado, ENEP Iztacala, Apdo. Postal 314,54090 Tlalnepantla, Edo. de México.

Se concluye que el índice de acuerdo entre observadores es sólo una caso especial de correlación contemplado en el modelo lineal estadístico y, que por consiguiente, no se justifica la utilización de esta técnica psicométrica para la estimación del acuerdo entre observadores.

De aquí que se plantee la necesidad de que la evaluación conductual debe crear sus propias técnicas psocométricas que tomen en cuenta los principios teóricos en los que se sustenta, ya que la extrapolación de técnicas psicométricas clásicas a esta área de la modificación de conducta, presenta tanto problemas teóricos como prácticos.
DESCRIPTORES: Confiabilidad, evaluación conductual, r de Pearson, acuerdo entre observadores, regresión lineal simple, evaluación tradicional.

ABSTRACT

The aim of this paper was to discuss the use of Pearson's correlation coefficient to evaluate some psychometric properties of assessment techniques used in behavioral studies. The characteristic of these techniques are described first, and then the use of Pearson's coefficient as a comparison statistic of standard scores, and as a linear model of the form $Y_t = \alpha + \beta X_t$ *is discussed.*

It is suggested that the measurement of interobserver agreement is only a special case of the linear regression model, and that the use of this technique in assessing interobserver agreement is not justified. The need for developing new psychometric techniques for behavioral studies is stressed.
DESCRIPTORS: Reliability, behavioral assessment, Pearson's r, interobserver agreement, straight-line regression, traditional assessment.

Uno de los temas recientemente abordados dentro de la modificación de conducta es el de los principios psicométricos de las técnicas de evaluación conductual. Aunque si bien es cierto que existía ya una teoría psicométrica muy desarrollada a partir de modelos estructural-diferenciales, se empezó a cuestionar si todas estas técnicas desarrolladas bajo la tutela teórica de la psicometría clásica eran adecuadas para la evaluación conductual, desarrollada ésta bajo marcos teóricos totalmente diferentes.

Aunque algunas técnicas de evaluación conductual estructuralmente parecen similares a las técnicas tradicionales, difieren significativamente en sus supuestos, objetivos, niveles de inferencia y el uso que se le da a los datos, debido al hecho de que la evaluación conductual es llevada a cabo, la mayoría de las veces, de una manera idiográfica. Además, como señalan Nelson y Hayes (1981), los métodos de la evaluación conductual consisten en la identificación de unidades de respuesta significativas y las variables que las controlan (tanto organísmicas como ambientales), con el propósito de entender y modificar la conducta, mientras que la evaluación tradicional generalmente sostiene que la conducta es el resultado de variables intraorganísmicas; más aún, supone que estas variables son mentales en lugar de físicas.

Entre las técnicas de evaluación conductual más comunes están la entrevista, la observación directa, el auto-reporte, el auto-monitoreo y los aparatos mecánicos y electrónicos. Este conjunto de técnicas surgió a partir de la necesidad de cuantificar y medir los tres tipos de respuesta consideradas dentro de la modificación de conducta como las unidades de análisis fundamen-

tales, por lo cual algunas técnicas se utilizan para evaluar respuestas cognitivas; otras para evaluar respuestas fisiológicas y otras para respuestas motoras.

Como se mencionó anteriormente, si bien es cierto que los objetivos de la evaluación conductual son muy diferentes a los de la evaluación tradicional, la confiabilidad de sus instrumentos sigue siendo una condición necesaria para la validación de los mismos. Por otro lado, la confiabilidad dentro de la evaluación conductual no sólo es importante para determinar la validez del instrumento, sino también para informar sobre la validez del tratamiento, puesto que "... la probabilidad de detectar una diferencia en la ejecución entre condiciones de tratamiento está en función directa de la confiabilidad de la medida usada" (Hartmann, 1977, p. 103).

De esta manera, la confiabilidad es una propiedad necesaria y deseable de los instrumentos de evaluación conductual y para su determinación existen una gran cantidad de herramientas estadísticas disponibles, sin embargo, es importante tener siempre presente que debido a que la estimación de la confiabilidad se deriva de procedimientos estadísticos, ésta variará como una función de los métodos estadísticos usados para su cálculo.

La técnica estadística por excelencia para estimar la confiabilidad en la evaluación tradicional es el coeficiente de correlación r de Pearson y sólo hasta recientemente se ha propuesto su uso dentro de la evaluación conductual (Anguera, 1983; Hartmann, 1977; Hersen y Bellack, 1978; Hollenbeck, 1978; Kent y Foster, 1977). Sin embargo, dichos autores en sus proposiciones no han tomado en cuenta que el concepto de confiabilidad en ambos modelos de evaluación difieren marcadamente.

Conforme a todo lo anterior, este trabajo tiene como objetivo el hacer un análisis del coeficiente de correlación r de Pearson como una técnica utilizada para estimar las propiedades psicométricas de algunos instrumentos de evaluación en situaciones de observación conductual directa, para lo cual, se hará un breve bosquejo de las técnicas de observación directa para posteriormente hacer algunas consideraciones sobre el coeficiente de correlación r de Pearson como índice de acuerdo entre observadores.

LA OBSERVACION CONDUCTUAL

Las técnicas de observación directa son las que mayormente se utilizan dentro de la evaluación conductual, debido a que son congruentes con algunos supuestos teóricos subyacentes a la modificación de conducta tales como: a) que sea conductual (Baer, Wolf y Risley, 1976; Fernández, 1981; Haynes, 1978; Kazdin, 1979;Kratochwill y Weltzel, 1977; Nelson y Hayes, 1979; 1981; Yelton, Wildman y Erikson, 1977) y b) que sea empírica (Haynes, 1978; Kazdin, 1979); pero la preferencia de esta técnica sobre otros métodos es también debida en parte a la cuestionable confiabilidad y validez de las otras técnicas de evaluación conductual (auto-reporte, entrevista, etc).

Generalmente la observación conductual es hecha en ambientes natura-

les, en situaciones análogas o en ambientes estructurados, con la finalidad de identificar y cuantificar el segmento conductual de interés, así como también en la evaluación de los resultados de la intervención. Existen una gran cantidad de factores que influyen en la selección del segmento conductual a ser observado, algunos de ellos son el ambiente en donde se observará, quiénes observarán, el tiempo de observación (Pinkerton, Hughes y Wenrich, 1981). Pero independientemente de estos factores, en la evaluación observacional no sólo interesa identificar las características relevantes del segmento conductual que permita hacer una descripción de dicho segmento, sino, entre otras cosas, cómo medir esas características.

Una vez que el segmento conductual ha sido identificado por medio de la determinación del límite espacio-temporal en donde se va a llevar a cabo la observación y se ha especificado la topografía, es necesario seleccionar una estrategia para medir el parámetro o los parámetros de interés. La cuantificación debe ser hecha sobre el parámetro más relevante para la conducta particular, que puede ser: la frecuencia. la duración, distancia de recorrido, ocurrencia por oportunidades, productos permanentes de la conducta del paciente, la frecuencia de aparición de la respuesta sobre unidad de tiempo o porcentaje de componentes realizados de una tarea.

Otra cuestión importante, dentro de la observación conductual directa, es la selección que el investigador debe hacer para registrar dicho segmento; esto es, cómo registrar el parámetro de interés del segmento elegido. Existen tres procedimientos de registro básicos que pueden ser modificados de acuerdo a las situaciones de observación, éstos son:

a) Registrar el segmento conductual cada vez que aparezca, con lo que se estará midiendo la frecuencia de aparición del evento. Bajo ciertas circunstancias algunas veces la frecuencia es reportada en una unidad de tiempo fija, lo cual dará información sobre la tasa de aparición del segmento conductual.

b) Registrar el tiempo de aparición de la respuesta, que provee medidas de duración de la ocurrencia del evento. Al igual que los registros de frecuencia, la duración del segmento conductual frecuentemente también es presentada por unidades de tiempo (por ejemplo, horas de dormir por la noche, horas de estudio por día, etc.).

c) En ambientes naturales generalmente no es posible registrar el segmento conductual continuamente durante todo el periodo de tiempo, por lo cual la conducta es muestreada seleccionando periodos de observación específicos, en los cuales se anota si la respuesta ocurre o no ocurre durante todos los intervalos de tiempo muestreados. Una variante de esta forma de registro, es cuando el investigador tiene control sobre la oportunidad de la emisión de la respuesta del paciente, llamado registro por ensayo, presentando la señal o estímulo específico al paciente para que éste presente el segmento conductual de interés dentro de un tiempo determinado.

En conclusión, existe una propiedad común en todos estos procedimientos de observación conductual directa, consistente en que todos ellos

requieren la utilización de instrumentos de registro y/o de observadores entrenados, por lo cual, el propósito del estudio, la naturaleza de la variable independiente y dependiente, y otros aspectos de la situación de observación determinarán cuál de estos procedimientos es más adecuado en un momento determinado.

EL CONCEPTO DE CONFIABILIDAD

La confiabilidad por sí misma es un término complejo que abarca tanto medidas de exactitud y estabilidad, así como también las condiciones reales bajo las cuales la medición es hecha, ya que, de acuerdo con Kerlinger (1975), el constructo confiabilidad debe definirse en términos de estabilidad, exactitud y predictibilidad. De aquí, que la confiabilidad debe dar respuesta a preguntas tales como: ¿los datos obtenidos por un instrumento de medición son estables y relativamente predecibles, o inestables y relativamente impredecibles? ¿las medidas obtenidas por medio de un instrumento corresponden a las verdaderas medidas de la propiedad en cuestión?

La Confiabilidad en la Psicometría Clásica

Dentro de la psicometría clásica el acuerdo entre medidas repetidas bajo condiciones similares constituyen el valor numérico de la confiabilidad que es estimado a partir del coeficiente de correlación, "... este coeficiente de correlación es llamado coeficiente de confiabilidad, y puede tomar valores entre cero y uno, pero no puede ser negativo" (Magnusson, 1975, p. 179).

Desde esta aproximación, se han propuesto diferentes diseños para evaluar la confiabilidad de los instrumentos de medida, cuatro son los más representativos:

a) Aplicación de formas paralelas. Con este método se administraran bajo condiciones específicas, durante la misma sesión, dos formas equivalentes del instrumento, con la finalidad de obtener la equivalencia de las medidas.

b) Aplicación del instrumento en dos ocasiones distintas; test-retest. Este método requiere que se administre el mismo instrumento al mismo grupo después de cierto tiempo. Este tipo de diseño pone el acento en la estabilidad temporal de las puntuaciones.

c) Aplicación de formas paralelas en dos ocasiones distintas. En este diseño se aplican las dos formas paralelas mediante cierto espacio temporal, con la finalidad de encontrar además de la equivalencia de medida, la estabilidad temporal.

d) Aplicación de una forma del test una sola ocasión. Con este tipo de diseño se intenta poner de relieve la homogeneidad de los elementos, es decir, el grado de consistencia interna del instrumento.

Cada uno de estos distintos diseños de obtener la confiabilidadd proporciona un tipo diferente de información, sin embargo, presentan un aspecto

en común, debido a que el grado de acuerdo entre dos conjuntos de puntuaciones siempre es determinado mediante la utilización de un coeficiente de correlación, que generalmente es el r de Pearson.

La Confiabilidad en la Evaluación Conductual

Por otro lado, dentro de la evaluación conductual, la suposición común en la mayoría de los estudios de observación conductual es que la confiabilidad puede ser evaluada a través de la estimación del acuerdo entre observadores. Sin embargo, algunos autores (Anguera, 1983; Hollenbeck, 1978; Kazdin, 1977; 1978; Martínez, 1981) han señalado que el acuerdo entre observadores no es una medida de confiabilidad propiamente, puesto que no puede evaluar por sí mismo la exactitud de los observadores, a menos que se compare con un estándar previamente establecido; similarmente, no evalúa la estabilidad de las mediciones, a menos que sea medido en ensayos sucesivos. De tal manera, el acuerdo entre observadores sólo refleja la medida en que ambos observadores acuerdan registrar la ocurrencia del segmento conductual, pero no con la exactitud con que lo registran. Es por esto, que existe la posibilidad de obtener un alto acuerdo entre observadores con una confiabilidad cercana a cero en términos de exactitud.

En este trabajo, al hablar del acuerdo entre dos observadores que registran simultáneamente un segmento conductual como índice de confiabilidad, se hará con la clara convicción de que es un concepto distinto al de exactitud y estabilidad, y que tanto estos dos conceptos como aquél forman parte de una medida llamada confiabilidad; además de que cada uno por si solo no es suficiente como índice indicativo de dicha medida compleja y que, por consiguiente, todos son necesarios para informar sobre la confiabilidad de un instrumento de evaluación observacional.

Algunas veces la evaluación del acuerdo entre observadores ha sido vista solo como un problema de medida, sin embargo, como han mostrado Birkimer y Brown (1979), dicho acuerdo es también importante para evaluar la veracidad de las afirmaciones del investigador acerca de los efectos del tratamiento.

En lo que respecta a la evaluación del acuerdo entre observadores como un elemento de confiabilidad, Hartmann (1977) menciona que se requiere la especificación del espacio de tiempo sobre el cual los datos serán resumidos para propósitos de su evaluación. De esta manera, la confiabilidad puede ser calculada sobre el puntaje de cada intervalo de registro o ensayo por sesión, en la cual dos o más observadores independientes registran el segmento conductual. Este nivel de confiabilidad es llamado confiabilidad por ensayo, puesto que da un porcentaje de acuerdo por intervalo o ensayo, razón por la cual también se le conoce como confiabilidad punto por punto. La confiabilidad también puede ser determinada por unidades de tiempo más grandes, tales como medidas por condición o más comúnmente medidas por sesión. La evaluación sobre medidas por sesión (por ejemplo, la suma de medidas

para múltiples ensayos o intervalos de una sesión) es llamada confiabilidad por sesión.

Recientemente, Gottman (1980) ha señalado que en algunas ocasiones el criterio para evaluar confiabilidad entre observadores no está basada sobre el acuerdo punto por punto o sobre medidas por sesión, sino en lugar de esto, en la medida en la cual ambos observadores independientes registran datos que rinden estructuras conductuales secuenciales similares. Es decir, la evaluación de la confiabilidad depende de la detección por ambos observadores de la misma estructura secuencial de conductas, aún cuando el acuerdo punto por punto sea bajo.

Existen una gran cantidad de índices de acuerdo que pueden ser utilizados para estimar la confiabilidad por ensayo o sesión, así como también la estructura conductual secuencial, puesto que se supone que la confiabilidad por sí misma es un constructo complejo, que puede ser inferido a través de varias operaciones aritméticas; por consiguiente, como se mencionó anteriormente, ésta variará como una función de los métodos estadísticos usados para el cálculo del coeficiente de acuerdo entre observadores (Hollenbeck, 1978).

ÍNDICES ESTADÍSTICOS UTILIZADOS PARA EVALUAR EL ACUERDO EN LA OBSERVACIÓN CONDUCTUAL DIRECTA

El Cálculo de la Confiabilidad por Ensayo o Punto por Punto

Dos aproximaciones generales han sido utilizadas para determinar la confiabilidad entre observadores, uno es el porcentaje de acuerdo y otro los coeficientes de correlación.

En los porcentajes de acuerdo por ensayo se divide la sesión por intervalos y se computa en base a los acuerdos y desacuerdos por intervalo. Sin embargo, este porcentaje de acuerdo, llamado también confiabilidad intervalo por intervalo (I x I), presenta grandes inconvenientes, tales como sobrestimar el acuerdo entre observadores cuando la tasa del segmento conductual es baja, por lo que se ha recomendado (Hawkins y Dotson, 1975) el uso del acuerdo de puntaje (índice S-1), en el que se ignoran aquellos intervalos en que ningún observador haya registrado ocurrencia, así como también su contraparte, el acuerdo de los observadores sobre la no ocurrencia (índice U-I), en el que solo se tiene en cuenta los intervalos en blanco para ambos observadores.

Pero como señalan Birkimer y Brown (1979), estos dos últimos índices siguen presentando serias limitaciones, debido a que los observadores pueden registrar en cada intervalo de observación la presencia o ausencia de la conducta meta; por consiguiente, si la tasa del segmento conductual varía, permaneciendo constante la tasa de desacuerdo en los observadores, tanto la confiabilidad S-I como la confiabilidad U-I fluctuarán.

Estos problemas pueden ser resueltos según Birkimer y Brown (1979), por la presentación de una simple gráfica de la tasa de desacuerdo que permite checar tanto el rango de desacuerdo, como evaluar la credibilidad de las afirmaciones del investigador acerca de los efectos experimentales.

Además del uso del porcentaje de acuerdo por ensayo o punto por punto, se han propuesto el uso de algunos coeficientes de correlación para escalas nominales. Por ejemplo, Hollenbeck (1978) considera adecuado el uso de la chi cuadrada, puesto que puede probar la "bondad de ajuste" de los datos de un observador con un conjunto de valores estándar esperados; sin embargo, una de las desventajas de este estimador es que brinda solamente un índice global de asociación para la totalidad de la distribución del código observacional.

Otro índice utilizado es el coeficiente de Scott, el cual puede ser usado con tablas de contingencia de N observadores y K categorías conductuales. Su principal limitación es que supone que la distribución de proporciones en las categorías conductuales son conocidas, e iguales para ambos observadores.

Igualmente para medir el acuerdo entre observadores por ensayo se han tomado como estimadores los estadísticos phi (ϕ) y Kappa (K). El coeficiente phi (ϕ) es el coeficiente producto momento de Pearson que evalúa la relación entre dos conjuntos de datos observacionales medidos a través de escalas nominales dicotómicas. Por otro lado, el coeficiente Kappa ha sido considerado como un estadístico cuasi-correlacional; sin embargo, como menciona Hartmann (1977), tanto los valores de phi como los de Kappa son casi idénticos, cuando la tasa de ocurrencia del segmento conductual es aproximadamente igual para los dos observadores.

El Cálculo de la Confiabilidad por Sesión

Las dos mismas aproximaciones generales utilizadas para evaluar la confiabilidad por ensayo han sido usadas para medir la confiabilidad entre observadores por sesión, lo único que varía son los índices estadísticos utilizados para su cuantificación. Esto último en base a que, como señala Hartmann (1977), las medidas por sesión son obtenidas al sumar ya sea los datos de frecuencia o duración a través de un periodo de observación entero y, puesto que las medidas varían desde cero a un valor positivo, se puede considerar que poseen las propiedades de la escala de razón. El hecho de que se suponga que estos datos tienen propiedades de la escala de razón, amplía más el rango de aplicación de índices estadísticos correlacionales más potentes.

La confiabilidad entre observadores por sesión proporciona un porcentaje de acuerdo logrado a lo largo de la sesión y es obtenido al dividir el dato más pequeño de los dos de una sesión por el dato mayor, multiplicando esta razón por cien. No obstante, Hartmann (1977) menciona que este porcentaje de acuerdo estadístico depende de la tasa específica del segmento conductual de la sesión en donde éste es calculado. Una alta tasa de conducta produce un porcentaje de acuerdo alto.

Varios autores (Anguera, 1983, Hartmann, 1977; Hersen y Bellack, 1978; Hollenbeck, 1978 Kent y Foster, 1977) recomiendan el uso del coeficiente producto-momento de Pearson como estimador de la confiabilidad entre observadores. "La medida más tradicional de confiabilidad por sesión es el coeficiente de correlación producto momento (r), ". . .Este estadístico es apropiado cuando se recolectan medidas de intervalo, frecuencia y duración, y tiene una gran variedad de ventajas asociadas con su status de estadístico paramétrico" (Kent y Foster, 1977, p. 313).

En resumen, la confiabilidad por sesión entre observadores, al igual que la confiabilidad por ensayo, también puede ser estimada a través de coeficientes de correlación y de porcentajes de acuerdo.

En el caso particular de este trabajo, como se mencionó anteriormente, el interés se centra básicamente en hacer algunas consideraciones sobre el coeficiente de correlación r de Pearson como índice de acuerdo entre observadores, para lo cual primero se hará una conceptualización del coeficiente r como una comparación de puntajes estándar y posteriormente como un modelo lineal.

EL COEFICIENTE DE CORRELACION r DE PEARSON COMO PUNTUACIÓN ESTÁNDAR

El coeficiente r de Pearson es el estimador por excelencia de la confiabilidad de los instrumentos en la psicometría clásica, debido a la suposición de que si el instrumento de medida no es afectado por factores producidos al azar, entonces los puntajes de los individuos en diferentes situaciones de medida serán idénticos, y sus posiciones en las distribuciones que se construyan en cada ocasión de medida serán las mismas, por lo que la correlación entre las dos distribuciones será uno (Magnusson, 1975).

Dos son las suposiciones básicas en la confiabilidad desde este punto de vista. La primera se refiere a la obtención de puntajes idénticos; la segunda se refiere a la obtención de idénticas posiciones relativas en las distribuciones de los puntajes obtenidos por los individuos en medidas sucesivas. De esta manera, el concepto de confiabilidad implica tener puntajes absolutos idénticos o posiciones relativas iguales dentro de cada distribución.

La forma de ejemplificar claramente estas dos suposiciones es calculando el coeficiente de correlación r de Pearson a través de la tipificación de los puntajes de los individuos, ya que por medio de este procedimiento es posible determinar "...hasta qué punto los mismos individuos o sucesos ocupan la misma posición relativa respecto a dos variables" (Haber y Runyon, 1973, p. 121).

Supongamos que aplicamos el mismo instrumento al mismo grupo después de cierto tiempo, y encontramos los resultados de la Tabla 1.

Como se puede observar en la Tabla 1, los cinco sujetos en ambas ocasiones de medida obtienen el mismo puntaje absoluto y, por consiguiente, al

TABLA 1

Resultados hipotéticos de la aplicación de un mismo test en dos ocasiones diferentes a un grupo de cinco sujetos

ss	1o. MEDICIÓN (X)	2o. MEDICION (Y)	Z_x	Z_y
1	20	20	−1.29	−1.29
2	25	25	−0.75	−0.75
3	30	30	−0.22	−0.22
4	45	45	1.40	1.40
5	40	40	0.86	0.86

obtener las puntuaciones tipificadas para el cálculo del coeficiente de correlación, cada individuo obtiene exactamente la misma calificación estándar en ambas aplicaciones del test.

Como en este caso hipotético se encontró un coeficiente r = 1.00, se puede afirmar que existen evidencias de que dicha prueba es confiable, puesto que presenta estabilidad temporal a través de aplicaciones sucesivas. Sin embargo, como se mencionó anteriormente, dentro de este concepto de confiabilidad el interés se centra principalmente en la posición relativa de los puntajes de los individuos en las dos distribuciones, independientemente de las diferencias en los puntajes absolutos obtenidos en las distintas situaciones de medida.

Para ejemplificar esto, supongamos que administramos durante la misma sesión dos formas equivalentes de un test y obtenemos los siguientes resultados.

En la Tabla 2 se observa que los cinco sujetos en las dos formas del test obtienen puntajes absolutos diferentes, sin embargo, las puntuaciones tipificadas en ambas formas son las mismas, por lo tanto al igual que en el ejemplo

TABLA 2

Resultados hipotéticos de la aplicación de dos formas paralelas de un test a un grupo de cinco sujetos.

1 Ss	1o. FORMA (X)	2o. FORMA (Y)	Z_x	Z_y
1	20	24	−1.29	−1.29
2	25	29	−0.75	−0.75
3	30	34	−0.22	−0.22
4	45	49	1.40	1.40
5	40	44	0.86	0.86

anterior, la confiabilidad estimada a través del coeficiente de Pearson es igual a 1.00, pudiéndose afirmar que existen evidencias que guían a suponer que ambas formas del test son equivalentes.

En conclusión, de los dos ejemplos anteriores se desprende que el supuesto fundamental de la confiabilidad en la psicometría clásica es el comparar en qué medida un grupo de sujetos mantienen la misma posición relativa en las dos distribuciones cuando es aplicado un instrumento de evaluación en diferentes situaciones de medida.

De esta forma, en los diferentes diseños para evaluar la confiabilidad en la psicometría clásica, lo que interesa es comparar la posición relativa de cada individuo con respecto al grupo, debido a que siempre el análisis del puntaje absoluto obtenido por un individuo dependerá del status relativo que guarde este puntaje con respecto al grupo. ". . . todos los procedimientos psicométricos clásicos dependen sobre la consistencia en las diferencias entre individuos para demostrar lo adecuado de la medida. . ." (Cone, 1981, p. 42).

Por otro lado, con respecto a la confiabilidad entre observadores en la evaluación conductual, Hartmann (1977) señala que cuando se usa el coeficiente r de Pearson para estimar el acuerdo entre observadores, lo que se obtiene simplemente es la correlación producto momento, basada sobre la serie de pares de medida obtenidos de las sesiones observadas conjuntamente. El rango del coeficiente normalmente va de 0.00 a+1.00 (aunque el rango posible de r se extiende desde −1.00 a + 1.00, coeficientes de confiabilidad de acuerdo entre observadores negativos son raros). Un r = 0.00 indica una carencia total de relación entre las medidas de los observadores, mientras que un r = 1.00 indica un perfecto acuerdo (en el sentido de puntajes estándares idénticos).

En la evaluación conductual, sin embargo, el acuerdo entre observadores como indicador de confiabilidad, solo supone la existencia de medidas absolutas idénticas entre los observadores en los diferentes momentos m_t de observación, independientemente de la posición relativa que tengan las medidas obtenidas en los diferentes m_t de cada observador.

Supongamos que dos observadores han registrado la ocurrencia de la conducta de lavarse las manos de un paciente obsesivo compulsivo, a lo largo de cinco días. Con el propósito de evaluar la confiabilidad se sumó el total de veces que el observador uno (X) reportó la aparición del segmento conductual durante el día, lo mismo se hizo con el registro del observador dos (Y), obteniéndose los resultados de la Tabla 3. De esta forma, conforme a la clasificación de Hartmann (1977) se estaría estimando la confiabilidad por sesión.

Como se puede ver en la Tabla 3, en todos los días los observadores acuerdan en el número de veces que el segmento conductual aparece; por consiguiente, la posición ordinal de aparición de la conducta evaluada a través del rango, durante los cinco días para cada una de las distribuciones de los observadores es la misma, así como también la posición relativa de los puntajes dentro de cada distribución (puntaje Z). De tal manera, que al calcular la

TABLA 3

Datos hipotéticos de la ocurrencia de lavarse las manos de un paciente obsesivo-compulsivo durante cinco días

Día	OBSERVADOR 1 (X)	OBSERVADOR 2 (Y)	RANGO (x)	RANGO (y)	Z_x	Z_y
1	30	30	4	4	0.71	0.71
2	15	15	1	1	−1.41	−1.41
3	25	25	3	3	0.00	0.00
4	35	35	5	5	1.41	1.41
5	20	20	2	2	−0.71	−0.71

confiabilidad de acuerdo entre observadores por medio del r de Pearson el valor será igual a 1.00, el cual indicará un acuerdo perfecto entre los observadores.

Sin embargo, cuando existen cambios en las medidas absolutas entre los observadores, manteniéndose constante la posición ordinal y la relativa, el coeficiente r de Pearson produce índices de acuerdo falsos. Por ejemplo, supongamos que en lugar de haberse obtenido los datos de la Tabla 3, se hubieran encontrado los resultados de la Tabla 4.

La Tabla 4 muestra que no hubo acuerdo entre los dos observadores en ningún día sobre el número de veces que apareció el segmento conductual, puesto que el observador 2 estuvo sobrestimando consistentemente la ocurrencia de la conducta. No obstante, tanto la posición ordinal como la posición relativa de las medidas dentro de cada distribución son las mismas, por lo que si se calcula el r de Pearson indicará un acuerdo perfecto entre los observadores, con un valor estimado igual a 1.00. Esto es debido a que como se mencionó anteriormente, el coeficiente de correlación solo compara las posiciones relativas de los puntajes con respecto a las dos distribuciones, independientemente de los valores absolutos y, como en este caso

TABLA 4

Datos hipotéticos de la ocurrencia de lavarse las manos de un paciente obsesivo-compulsivo durante cindo días

Día	OBSERVADOR 1 (X)	OBSERVADOR 2 (Y)	RANGO (x)	RANGO (y)	Z_x	Z_y
1	30	35	4	4	0.71	0.71
2	15	20	1	1	−1.41	−1.41
3	25	30	3	3	0.00	0.00
4	35	40	5	5	1.41	1.41
5	20	25	2	2	−0.71	−0.71

las posiciones relativas son iguales, el r de Pearson sobrestima el acuerdo entre observadores.

De esto se deriva, que si bien es cierto lo que menciona Hartmann (1977), de que una r de Pearson igual a 1.00 indica perfecto acuerdo entre observadores en el sentido de puntajes estándares idénticos, esta afirmación no tiene ninguna validez, puesto que en la evaluación conductual el concepto de confiabilidad no supone igualdad en los puntajes estándares, sino en los puntajes absolutos. Esto es, que ambos observadores acuerden en el número de veces que la conducta apareció en las diferentes sesiones de observación ($Y_t = X_t$).

Por otro lado, Hartmann (1977) señala que es posible obtener coeficientes de correlación altos cuando un observador sobrestima consistentemente en la misma dirección la tasa de ocurrencia del segmento conductual; para saber si ésto está ocurriendo, propone que se utilize la prueba "t" de student de diferencias entre medidas correlacionadas. No obstante que la prueba "t" de student sea una alternativa para detectar diferencias entre los observadores, ésta no es sensible a diferencias pequeñas entre ambos conjuntos de medidas, además de que no elimina el problema de las comparaciones de las posiciones relativas. De aquí que se vea la necesidad de crear índices estadísticos para evaluar el acuerdo entre observadores, que midan hasta qué punto los valores absolutos de las dos distribuciones están relacionados, sin tomar en cuenta las posiciones relativas de estos puntajes. La piedra angular de la confiabilidad en la evaluación conductual es la existencia de medidas absolutas idénticas entre los observadores, en los diferentes momentos de observación, puesto que estos datos no necesitan ser sometidos a transformaciones, porque tienen significado en términos absolutos (Martínez, 1981).

EL COEFICIENTE DE CORRELACION r DE PEARSON COMO UN MODELO LINEAL

En la psicometría clásica el concepto de confiabilidad, además de la igualdad entre las mediciones absolutas y/o la igualdad entre posiciones relativas en las dos distribuciones, supone que existe una relación lineal entre ambos conjuntos de medidas. "El coeficiente de correlación indica entonces el grado en que los puntajes en una de las variables mantiene una relación lineal sistemática con los puntajes en la otra" (Magnusson, 1975, p. 47).

De esta manera, el supuesto de linealidad entre los puntajes de medidas sucesivas estará representado por el modelo matemático de:

$$Y_t = \alpha + \beta X_t$$

Conforme a la suposición de linealidad, la forma de evaluar la confiabilidad es analizando la dispersión de los puntos de ambas mediciones con respecto a un modelo lineal estimado a partir de los datos obtenidos. Cuando

los puntos están ampliamente dispersos alrededor del modelo lineal estimado, el coeficiente de correlación r de Pearson es pequeño, mientras que cuando todos los puntos del diagrama de dispersión caen sobre el modelo estimado, el coeficiente r es igual a 1.00.

Hagamos ahora una interpretación del coeficiente de correlación de Pearson desde este punto de vista, calculando para los datos hipotéticos de las Tablas 1 y 2, el modelo matemático lineal. En este caso particular, los valores obtenidos en la primera ocasión de medida son asignados al eje horizontal (X) y los valores de la segunda ocasión de medida al eje vertical (Y). El diagrama de los resultados hipotéticos de la aplicación de un mismo test en dos ocasiones diferentes a un mismo grupo (Tabla 1) se muestra en la Figura 1, gráfica (a).

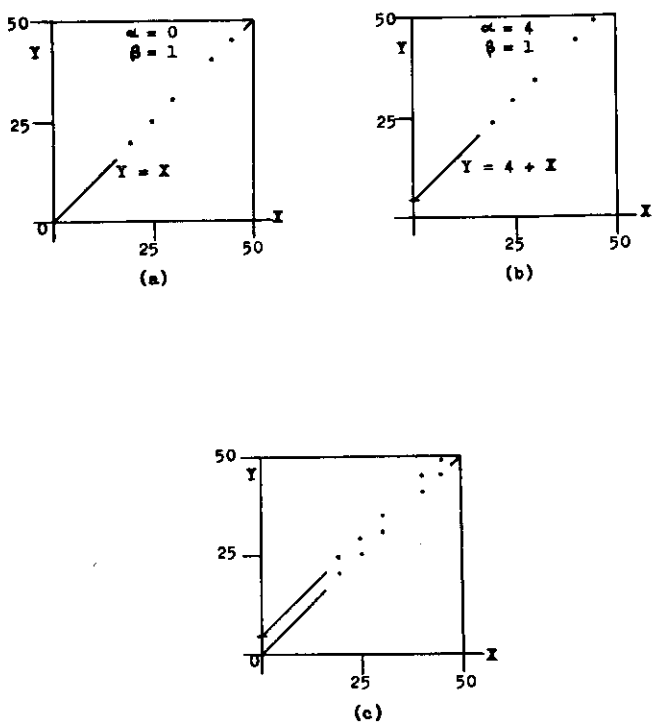

Figura 1. Muestra los diagramas de dispersión y el modelo lineal estimado para la aplicación de un test en dos ocasiones diferentes (a), la aplicación de dos formas paralelas de un test (b) y la representación conjunta de ambas relaciones lineales (c).

El diagrama de dispersión muestra que los individuos que obtuvieron puntajes altos en la primera aplicación del test, también tuvieron puntajes altos en la segunda aplicación. Esta misma gráfica muestra también que estos pares de valores dan como resultado que todos los puntos estén sobre una línea recta, y que por consiguiente el coeficiente de correlación sea igual a 1.00. Al evaluar la relación lineal de las dos aplicaciones del test, a través de los mínimos cuadrados, los estimadores de los parámetros α y β son cero y uno respectivamente. El modelo lineal estimado quedaría:

$$Y_t = O + 1\ (X_t), \text{ decir, } Y_t = X_t$$

El diagrama de dispersión que representa los resultados hipotéticos de la aplicación de dos formas paralelas de un test (Tabla 2), se muestra en la Figura 1, gráfica (b). En esta gráfica se ve que al igual que los puntos del diagrama de dispersión anterior, éstos también caen sobre una línea recta, solo que aquí los puntos se han desplazado un poco más hacia arriba en el eje de la ordenada. Este es el efecto característico que sucede en el modelo lineal, cuando los valores absolutos en ambas distribuciones no son los mismos.

Sin embargo, puesto que el coeficiente r de Pearson visto como modelo lineal evalúa el grado en que los pares de valores dan como resultado puntos que caen sobre una recta, y en el diagrama de dispersión de la gráfica (b) todos los puntos caen sobre la recta, el coeficiente de correlación también es igual a 1.00.

Si estimamos la relación lineal para los datos de esta gráfica, los parámetros serán 4 y 1, por lo que la recta cruza el eje Y en el 4, y a incrementos unitarios en X corresponden incrementos unitarios en Y. El modelo lineal estimado quedaría:

$$Y_t = 4 + X_t$$

En la gráfica (c) de la Figura 1, se representan conjuntamente ambas relaciones lineales. En esta gráfica se observa que la única diferencia en estas dos relaciones es el punto en donde cruzan el eje de las ordenadas, puesto que ambas tienen el mismo valor estimado de β, que es 1.00. En la primera relación lineal la ordenada al origen es igual a 0 y en la segunda igual a 4.

De esta manera, se puede concluir que si bien es cierto que en la psicometría clásica el concepto de confiabilidad también presupone una relación lineal entre ambos conjuntos de mediciones, ésta puede ser cualquier tipo de relación lineal positiva, puesto que como señala Magnusson (1975), la confiabilidad puede tomar valores entre cero y uno, pero jamás puede tener valores negativos. Es decir, las relaciones lineales negativas, que comprenden las relaciones en donde a aumentos en X corresponden disminuciones en Y, no son parte constituyente de la confiabilidad. De esta manera, el supuesto de linealidad en la confiabilidad, presupone cualquier tipo de recta con pendiente positiva.

Dentro del campo de la evaluación conductual, algunos autores (Anguera, 1983; Hartmann, 1977; Kent y Foster, 1977) han señalado que es necesario que exista una linealidad en los puntos correspondientes a los valores obtenidos en las sesiones de observación. De esta forma, al igual que en la psico-

metría clásica, el concepto de confiabilidad dentro de la evaluación conductual supone también la existencia de una linealidad entre el conjunto de datos de los observadores.

Igualmente, la relación lineal específica que se presupone es de tipo positiva, puesto que como Hartmann (1977) menciona, coeficientes de confiabilidad de acuerdo entre observadores negativos no son posibles; esta afirmación se debe al hecho de que para asegurar que existe acuerdo entre observadores es necesario que ambos registradores obtengan los mismos datos, esto es, que los puntajes del observador 1 (X_t) sean iguales a los datos del observador 2 (Y_t). En otras palabras, se espera que al bajar la frecuencia de aparición real del segmento conductual, ambos registren dicho decremento; de la misma manera, si la frecuencia real de aparición de la conducta aumenta, ambos observadores deberán detectar dicho incremento. Es así, que a una variación en los datos de un observador debe corresponder una variación en el mismo sentido en los datos del otro observador, puesto que en el caso de que se encontrara consistentemente que a una variación en los datos de un observador le correspondiera una variación de sentido contrario en los datos del otro observador, automáticamente se podría afirmar que el acuerdo entre observadores es nulo, debido a que mientras un observador registra un aumento en la frecuencia, el otro registra una disminución.

Similarmente, la forma de evaluar la linealidad de los datos de los dos observadores es por medio del diagrama de dispersión, obtenido a partir de la representación de los pares de datos, en donde los valores obtenidos por el primer observador son asignados al eje horizontal y los valores del segundo al eje vertical. Los datos hipotéticos de la Tabla 3, de la ocurrencia de lavarse las manos de un paciente obsesivo-compulsivo, se muestran en la Figura 2 gráfica (a).

El diagrama de dispersión de la gráfica (a) muestra que, debido a que ambos observadores acordaron perfectamente sobre la frecuencia de ocurrencia de la conducta durante los cinco días, es decir $Y_t = X_t$, estos pares de valores dan como resultado que todos los puntos estén sobre una línea recta; además, que la correlación entre los pares de valores sea igual a 1.00.

Los parámetros estimados de esta relación lineal dan como resultado un $\alpha = 0$ y $\beta = 1$, de los cuales se desprende que para afirmar la existencia de un perfecto acuerdo entre observadores, es necesario que los parámetros del modelo lineal tengan una ordenada al origen y una pendiente igual a uno.

Con respecto al diagrama de dispersión de los datos hipotéticos de la Tabla 4, en donde a simple vista se observa que no hubo absoluto acuerdo entre los observadores, es decir $Y_t \neq X_t$, sin embargo los pares de valores dan también como resultado una nube de puntos que caen sobre una línea recta, como se puede observar en la Figura 2 gráfica (b).

De esta forma se puede explicar por qué al calcular el coeficiente de correlación r de Pearson, éste es igual a 1.00, puesto que como se recordará, este coeficiente de correlación evalúa la dispersión de los puntos en relación a un modelo lineal y, como en este caso todos los puntos caen sobre la línea recta, la dispersión es nula.

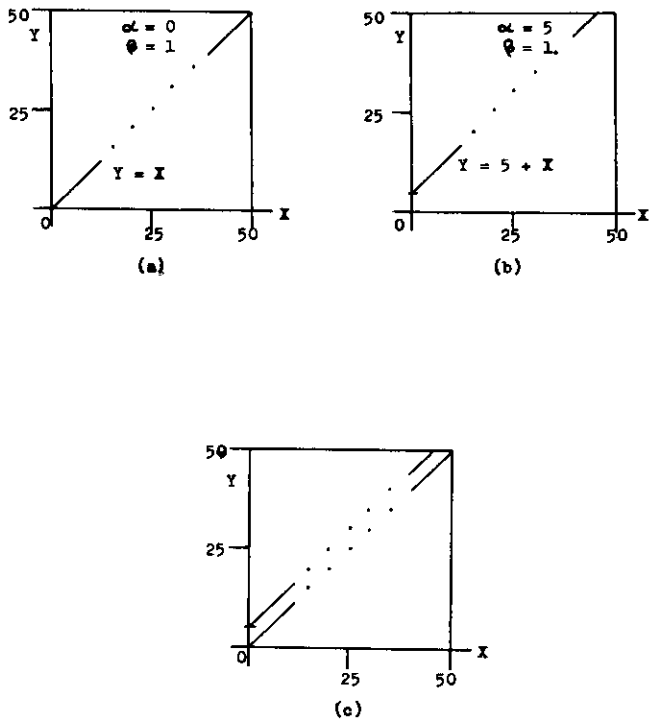

Figura 2. Muestra los diagramas de dispersión y el modelo lineal estimado para la ocurrencia de lavarse las manos de un paciente obsesivo-compulsivo durante cinco días (a) y (b), y la representación conjunta de las funciones lineales obtenidas en ambas situaciones hipotéticas de observación (c).

Sin embargo, al comparar las funciones lineales obtenidas en ambas situaciones hipotéticas de observación, como lo muestra la gráfica (c) de la Figura 2, se observa que si bien ambas funciones son lineales, difieren en el punto de intersección con el eje de las ordenadas, puesto que los datos de la Tabla 3 arrojaron un $\alpha = 0$ y $\beta = 1$, mientras que los datos de la Tabla 4 dieron un $\alpha = 5$ y $\beta = 1$.

De lo anterior se desprende en conclusión, que si bien el concepto de confiabilidad en la evaluación conductual, al igual que en la psicometría clásica, supone la existencia de una linealidad entre el conjunto de datos de los dos observadores, esta suposición de linealidad no presupone cualquier tipo de relación lineal, sino una relación especial que parte del origen y su inclina-

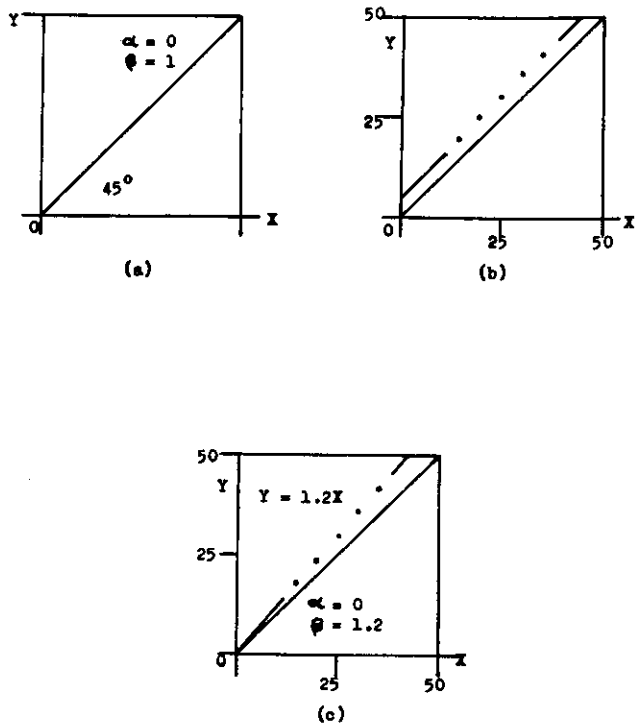

Figura 3. Muestra: (a) la relación lineal especial que presupone el concepto de confiabilidad en la evaluación conductual, (b) la comparación de la función lineal obtenida a partir de los datos de la Tabla 4 con la función lineal óptima de confiabilidad y (c) la comparación de la función lineal estimada a partir de los datos de la Tabla 5 con la función lineal óptima.

ción con respecto al eje de las X'x es de 45°, cuyos parámetros óptimos son $\alpha = 0$ y $\beta = 1$; como se muestra en la gráfica (a) de la Figura 3.

De tal modo, debido a que el coeficiente de correlación r de Pearson evalúa cualquier tipo de relación lineal, su uso con cualquier otra recta que se aleje ya sea tanto del origen como de los 45°, pondrá en tela de juicio la validez de su uso como estimador de confiabilidad entre observadores.

Sin embargo, es posible determinar si el coeficiente de correlación r de Pearson es un índice adecuado de acuerdo entre observadores, a través de considerar la relación entre los dos conjuntos de datos como un modelo de regresión lineal simple, en el cual ambas variables (los puntajes de los observa-

dores 1 y 2) son consideradas variables aleatorias. Conforme a esta aproximación, es factible probar la hipótesis nula de que $\alpha = 0$ y $\beta = 1$; en caso de comprobarse esta hipótesis, existirían evidencias que permitirían afirmar que el coeficiente r de Pearson es un índice adecuado para evaluar la confiabilidad entre observadores.

Esto se deriva del hecho de que como se mencionó anteriormente, la linealidad dentro del concepto de confiabilidad entre observadores presupone medidas absolutas idénticas en ambos conjuntos de datos, es decir, $Y_t = X_t$. De tal manera que un error aditivo en uno de los observadores, arrojaría medidas absolutas diferentes ($Y_t \neq X_t$), cuyo efecto estaría reflejado en cambios en el parámetro α de la función lineal, como se muestra en la gráfica (b) de la Figura 3.

La función lineal de esta gráfica fue estimada a partir de los datos hipotético de la Tabla 4. En esta función se observa que el único parámetro que difiere con los valores óptimos de linealidad es α , puesto que el valor estimado es igual a 5 y el valor óptimo es igual a cero, mientras que el valor de β estimado es igual a uno, por lo que no difiere con el valor óptimo, de lo cual se obtiene el modelo lineal siguiente:

$$Y_t = 5 + X_t$$

De esta forma, se ve claramente que un error aditivo sistemático en cualquiera de los observadores produce un efecto de desplazamiento de la función lineal a lo largo de la ordenada, mientras que la inclinación seguirá siendo de 45°, por lo que será importante saber si la magnitud del desplazamiento es lo suficientemente significativa como para invalidar el uso del coeficiente r de Pearson como índice de confiabilidad. La manera de saber esto es a través de probar si la diferencia entre el valor α óptimo y α estimado es lo suficientemente grande como para producir una r de Pearson que sobrestime el acuerdo entre observadores, puesto que esto es un efecto característico cuando α estimado se aleja de cero.

La forma de probar si el valor estimado de α se aleja de cero es por medio del estadístico "tα", que permite probar la hipótesis nula de que α es igual a cero. El hecho de aceptar la hipótesis nula, es decir, que sea igual a cero, no garantiza que el valor estimado de acuerdo entre observadores a partir de la r de Pearson sea un valor adecuado de confiabilidad, puesto que sólo se habrá probado que no existe un error aditivo sistemático en los observadores que afecte el valor del coeficiente de correlación, ya que un error sistemático multiplicativo en los observadores, también produce que la r de Pearson sobrestime el acuerdo entre ellos. Pongamos por ejemplo la observación del mismo segmento conductual de lavarse las manos en un paciente obsesivo compulsivo por dos observadores independientes. Los datos hipotéticos obtenidos por ambos registradores durante cinco días se muestran en la Tabla 5, y el diagrama de dispersión que generan estos datos se muestra en la gráfica (c) de la Figura 3.

A simple vista se observa en la Tabla 5, que el supuesto básico de acuer-

TABLA 5

Datos hipotéticos de la ocurrencia de lavarse las manos de un paciente obsesivo-compulsivo durante cinco días.

Día	OBSERVADOR 1 (X)	OBSERVADOR 2 (Y)
1	30	36
2	15	18
3	25	30
4	35	42
5	20	24

do de $Y_t = X_t$ no fue obtenido, puesto que en todos los días de registro se ve que $Y_t \neq X_t$. Sin embargo, al graficar los pares de valores para cada día, estos dan como resultado que todos los puntos estén sobre una línea recta. De tal manera que si se utiliza el coeficiente de correlación r de Pearson para estimar la confiabilidad, éste estará sobrestimando el acuerdo entre observadores. Por consiguiente, el valor que se obtiene eneste caso particular es r = 1.00.

Los parámetros estimados de la relación lineal, generados por estos datos hipotéticos, dan como resultado una $\alpha = 0$ y una $\beta = 1.20$, de lo que se obtiene el modelo lineal siguiente:

$$Y_t = 1.20X_t$$

De esta forma, se ve claramente que un error multiplicativo sistemático en cualquiera de los observadores produce un efecto de desplazamiento de la función lineal, como se observa en la gráfica (c) de la Figura 3, puesto que esta recta aumenta su ángulo de inclinación, mientras que la ordenada al origen sigue teniendo el parámetro óptimo, es decir, $\alpha = 0$.

De aquí que también sea importante saber si la magnitud de aumento del ángulo de inclinación de la recta, es lo suficientemente grande para invalidar el uso de la r de Pearson como índice de confiabilidad entre observadores.

La forma de probar esta suposición es por medio del estadístico "t_β", que permite probar la hipótesis nula de que $\beta = 1$. Si se acepta la hipótesis nula, es posible asegurar que existen evidencias que permiten afirmar la ausencia de un error multiplicativo sistemático en los dos conjuntos de datos, que influirían sobre el valor estimado de confiabilidad a partir de la r de Pearson.

En conclusión, solo hasta que se demuestre que la función lineal generada por los datos de los observadores no se aleja significativamente del modelo lineal $Y_t = X_t$, que presupone el concepto de confiabilidad en la evaluación conductual, el coeficiente de correlación r de Pearson no puede ser tomado como estimador adecuado de confiabilidad entre observadores. Es

decir, será necesario probar que $\alpha = 0$ y $\beta = 1$ para legitimizar el uso de la r de Pearson.

CONCLUSIÓN

En este trabajo, se mostró que el concepto de confiabilidad en ambos modelos difieren marcadamente, puesto que mientras en la evaluación tradicional se supone la existencia de confiabilidad cuando los instrumentos de medición producen posiciones relativas iguales en las distribuciones de los puntajes en cada ocasión de medida, en la evaluación conductual, la suposición en la mayoría de los estudios observacionales es que la confiabilidad puede ser evaluada a partir de la estimación del acuerdo entre observadores, para lo cual es necesario que los puntajes absolutos en ambos observadores en cada momento de registro sean iguales, esto es, $Y_t = X_t$.

Como se habrá notado, todas las anteriores consideraciones sobre el uso de la r de Pearson como índice de acuerdo, se han abocado a través del modelo de regresión lineal simple a proponer un procedimiento cuantitativo que permita tomar una decisión acerca de lo adecuado de dicho coeficiente de correlación para estimar la confiabilidad entre observadores; pero de ninguna manera es factible, a partir de este modelo, obtener información sobre la magnitud óptima de dicho coeficiente para asegurar confiabilidad entre observadores, en caso de que se haya mostrado su adecuación. De tal manera que el problema de cuál es la magnitud mínima necesaria para afirmar acuerdo entre observadores se mantiene.

De aquí que si el concepto de confiabilidad en la evaluación conductual supone una relación lineal positiva específica con parámetros $\alpha = 0$ y $\beta = 1$ y, por otro lado, la r de Pearson evalúa la dispersión de los datos con respecto a cualquier relación lineal, además de que el modelo de regresión simple solo permite saber si dicho coeficiente es adecuado como índice de confiabilidad; ¿no sería mejor desecharlo y buscar otras alternativas de cuantificar la confiabilidad?

Sin embargo, el hecho de que se desprecie el coeficiente correlacional r de Pearson como índice de confiabilidad, no invalida el uso de modelo de regresión lineal simple dentro de la evaluación conductual, puesto que el conceptualizar el acuerdo entre observadores como un modelo de regresión, permitiría detectar errores aditivos y/o multiplicativos en las medidas de los observadores, independientemente de la confiabilidad que existiera entre ellos. De esta forma, la detección de estos tipos de errores en los puntajes de los observadores tiene implicaciones metodológicas y de procedimiento más importantes, que el demostrar únicamente que la r de Pearson es un estimador adecuado de la confiabilidad.

Para finalizar, es importante mencionar que el uso del modelo de regresión lineal simple para detectar errores aditivos y/o multiplicativos entre los datos de los observadores, se podría realizar a través del contraste de hipó-

tesis de los valores estimados de los parámetros, en relación a los valores óptimos de dichos parámetros. El desarrollo de estas consideraciones sería muy extenso para ser abordado en este mismo escrito, por lo cual se deja para un trabajo posterior.

DISCUSIÓN

Tanto el modelo conductual de evaluación como el tradicional, se enfrentan con la imperiosa necesidad de determinar la confiabilidad y la validez de los instrumentos de evaluación, con la finalidad de tener confianza en los datos que proporcionan. En la mayoría de los casos, los principios y técnicas psicométricas en la evaluación conductual han sido tomados de la psicometría clásica. Sin embargo, estas extrapolaciones han pasado por alto las diferencias marcadas que existen en las cuestiones conceptuales subyacentes en ambos modelos, en términos de los determinantes de la conducta, la unidad de análisis, la consistencia de la respuesta y la forma en que una respuesta es interpretada; por lo que, como menciona Cone (1981), la evaluación conductual requiere un paradigma radicalmente diferente al utilizado para evaluar los rasgos y los constructos hipotéticos.

La evaluación tradicional se fundamenta teóricamente dentro del modelo de los rasgos; por consiguiente, la conduce a considerar las características de las personas estables en el tiempo y a intentar apreciarlas por medio de los instrumentos de diagnóstico. Es así que, "Si existieran algunas inconsistencias test-retest, éstas serían inmediatamente imputadas al error de medida y no tomadas como prueba de la inconsistencia de esos rasgos. . . Y en caso de no verificarse la consistencia de tal rasgo, . . .esto sería imputado a la falta de fiabilidad del instrumento" (Fernández, 1981, p. 72).

En contraste con esta orientación tradicional, la evaluación conductual considera que la conducta debe ser explicada en función de la situación y, por tanto, mantiene que la conducta depende de la situación en que la evaluación es llevada a cabo; más aún, de la interacción entre las variables organísimicas y las variables de la situación (Cone, 1981; Fernández, 1981; Martínez, 1981; Mischel, 1973; Nelson y Hayes, 1979; 1981). Conforme a ésto, los psicómetras orientados conductualmente se enfrentan al problema de deslindar cuándo las diferencias entre las medidas se deben a cambios en el instrumento mismo (inconfiabilidad) y cuándo se deben a cambios en la conducta propiamente. "El razonamiento de que los coeficientes bajos se deben a la falta de confiabilidad en las mediciones, pueden confundirnos en lo que se refiere hasta dónde estos cambios realmente ocurren y hasta dónde pueden servir para explicar la inestabilidad a base de la propia inestabilidad" (Mischel, 1973, p. 50).

De esta forma, no es posible utilizar una misma técnica estadística cuando, por ejemplo, en la psicometría clásica, la presencia de inconsistencia en los datos entre medidas repetidas bajo condiciones similares es inmediatamen-

te imputada al error de medida, es decir, a la falta de confiabilidad del instrumento; mientras que en la evaluación conductual esta inconsistencia en los datos, no descarta la posibilidad de que estén representando cambios reales de aparición del segmento conductual medido, debido que a partir del modelo conductual, se sostiene que la conducta puede ser explicada en función de la situación en que se presenta (Baer y Cols., 1976; Cone, 1981; Fernández, 1981; Haynes, 1978; Kazdin, 1979; Martínez, 1981; Mischel, 1973; Nelson y Hayes, 1981; Skinner, 1953). Si por otra parte, se toma en consideración que los instrumentos de evaluación tradicional han sido construídos para poner de manifiesto diferencias interindividuales, a través de la comparación de la posición relativa que guarda un individuo con respecto al grupo; mientras que los instrumentos de evaluación conductual están interesados en encontrar diferencias intraindividuales (antes y después del tratamiento), se pone de manifiesto más claramente la dificultad de extrapolar técnicas estadísticas de la psicometría clásica a la evaluación conductual.

Otra cuestión importante a considerar para la extrapolación de las técnicas estadísticas, es el hecho de que los datos obtenidos en la evaluación conductual no necesitan ser sometidos a transformaciones, porque tienen significado en términos absolutos, mientras que en la psicometría clásica no tienen significado en sí mismos debido a que necesitan ser transformados en otras medidas referidas al grupo al que pertenece el sujeto. Las transformaciones más frecuentemente hechas son las puntuaciones tipificadas (Z_x y Z_y), como se mostró en este trabajo.

En resumen, uno de los problemas más polémicos dentro de la evaluación conductual, es el hecho de la extrapolación de los índices estadísticos utilizados en la psicometría clásica, derivados del modelo tradicional de evaluación, para estimar la confiabilidad. De aquí que la evaluación conductual, aunque comparte la misma teoría de la medida, no puede extrapolar las técnicas estadísticas utilizadas en la psicometría clásica, ya que esta última no solo desarrolló principios teóricos que la fundamentasen, sino que a la par, también creó técnicas estadísticas que le dieran un soporte formal más estructurado. Por lo tanto, la evaluación conductual debe analizar críticamente dicha extrapolación y, por otro lado, crear sus propias técnicas estadísticas que tomen en cuenta los principios teóricos en los cuales se sustenta.

REFERENCIAS

ANGUERA, A.A.T. (1983). *Manual de prácticas de observación.* México, Trillas.

BAER, D.M., WOLF, M.M. y RISLEY, T.R. (1976). Algunas dimensiones actuales del análisis conductual aplicado. En R. Ulrich, T. Stachnick y J. Mabry (Eds). *Control de la conducta humana.* México: Trillas, Vol. 2.

BIRKIMER, J.C. y BROWN, J.H. (1979). A graphical judgmental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. *Journal of Applied Behavior Analysis, 12, 523-533.*

CONE, J.D. (1981). Psychometric considerations. En M. Hersen y A.S. Bellack (Eds.) *Behavioral assessment: A practical handbook.* New York: Pergamon Press.

FERNANDEZ, B.R. (1981). Comparaciones entre la evaluación tradicional y la evaluación conductual. En R.B. Fernández y J.A. Carrobles (Eds). *Evaluación conductual: Metodología y aplicaciones.* Madrid: Piramide.

GOTTMMAN, J.M. (1980). Analyzing for sequential connection and assessing interobserver reliability for the sequential analysis of observational data. *Behavioral assessment, 2,* 361-368.

HABER, A. y Runyon, R.P. (1973). *Estadística general.* México: Fondo Educativo Interamericano.

HARTMANN, D.P. (1977). Considerations in the choice of interobserver agreement. *Journal of Applied Behavior Analysis, 10,* 103-116.

HAYNES, S.N. (1978). *Principles of behavioral assessment.* New York: Halsted Press.

HAWKINS, R.P. y DOTSON, V.A. (1975). Reliability scores that delude: An Alice in Wonderland trip throught the misleading characteristic of interobserver agreement scores in interval recording. En E. Ramp y G. Semb (Eds). *Behavior analysis: Areas of research and application.* New York: Prentice-Hall.

HERSEN, M. y BELLACK, A.S. (1978). *Behavior therapy in the psychiatric setting.* Baltimore: Williams y Wilkins.

HOLLENBECK, R.A. (1978). Problems of reliability in observational research. En G.P. Sackett (Ed). *Observing behavior: Data collection and analysis methods.* Baltimore: University Park Press.

KAZDIN, A.E. (1977). Artifact bias and complexity of assessment: The ABC's of reliability. *Journal of Applied Behavior Analysis, 10,* 141-150.

KAZDIN, A.E. (1978). Methodological and interpretative problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 4,* 629-642.

KAZDIN, A.E. (1979). Fictions, factions and functions of behavior therapy. *Behavior Therapy, 10.* 629-654.

KENT, R.N. y FOSTER, S.L. (1977). Direct observational procedure: Methodological issues in naturalistic setting. En A.R. Cimimero, K.S. Calhoun y H.E. Adams (Eds). *Handboook of behavioral assessment.* New York: Wiley and Sons.

KERLINGER, N.F. (1975). *Investigación del comportamiento: Técnica y metodología.* México: Interamericana.

KRATOCHWILL, T.R. y WETZEL, R.J. (1977). Observer agreement, credibility and judgment: Some considerations in presenting observer agreement data. *Journal of Applied Behavior Analysis, 10,* 133-139.

MAGNUSSON, D. (1975). *Teoría de los tests.* México: Trillas.

MARTINEZ, A.M.R. (1981). Principios psicométricos de las técnicas en evaluación conductual. En R.B. Fernández, y J. A. Carrobles (Eds). *Evaluación conductual: Metodología y aplicaciones.* Madrid: Piramide.

MISCHEL, W. (1973). *Personalidad y evaluación.* México: Trillas.

NELSON, R.O. y HAYES, S.C. (1979). Some current dimensions of behavioral assessment. *Behavioral Assessment, 1,* 1-16.

NELSON, R.O. y HAYES, S.C. (1981). Nature of behavioral assessment. En M. Hersen y A.S. Bellack (Eds). *Behavioral assessment: A practical handbook.* New York: Pergamon Press.

PINKERTON, S.S., HUGHES, H. y WENRICH, W.W. (1981). *Behavioral medicine: Clinical applications.* New York: Wiley and

SKINNER, B.F. (1953). *Science and human behavior.* New York: The Macmillan Company.

YELTON, A.R., WILDMAN, B.G. y ERICKSON, M.T. (1977). A probability-based formula for calculating interobserver agreement. *Journal of Applied Behavior Analysis, 10,* 127-131.