- Trabajo realizado por la Biblioteca Digital de la Universidad CEU-San Pablo

- Me comprometo a utilizar esta copia privada sin finalidad lucrativa, para fines de investigación y docencia, de acuerdo con el art. 37 de la M.T.R.L.P.I. (Modificación del Texto Refundido de la Ley de Propiedad Intelectual del 7 julio del 2006)

# EL EMPLEO DE OBSERVADORES EN INVESTIGACION APLICADA (EDUCATIVA Y CLINICA): EL CALCULO DE LA CONFIABILIDAD ENTRE OBSERVADORES

* J. GIL ROALES-NIETO [1]
L. VALERO AGUAYO [2]
A. POLAINO-LORENTE [3]

El empleo de observadores que registren los datos básicos de un comportamiento humano cualquiera, se remonta a los primeros tiempos en que la psicología inició su andadura como disciplina. En cada etapa histórica, el uso de observadores se ha adecuado a la naturaleza de los problemas estudiados y a los planteamientos teóricos entonces vigentes entre los investigadores.

Las descripciones de la conducta de los niños realizadas por los

---

(*) JESUS GARCIA ROALES-NIETO.—Doctor en Psicología por la Universidad Complutense de Madrid y en la actualidad Profesor Colaborador de Psicopatología en el Departamento de Psicología de la Universidad de Granada. Su línea de investigación abarca el retraso en el desarrollo, especialmente temas concernientes a evalución, diagnóstico y clasificación, así como otros aspectos de la psicopatología y psicología clínica.

LUIS VALERO AGUAYO.—Licenciado en Filosofía y Letras (sección de Psicología) por la Universidad de Granada y Graduado en Psicología por la misma Universidad. Trabaja como psicólogo clínico privado, sobre todo en retraso en el desarrollo, y mantiene una línea de investigación clínica vinculado a las actividades del Departamento de Psicología de la Universidad de Granada.

AQUILINO POLAINO-LORENTE.—Catedrático de Psicopatología en la Universidad Complutense de Madrid. Ha publicado más de un centenar de artículos en revistas de su especialidad y alrededor de quince libros, entre los que cabe resaltar *Psicología Patológica, Autismo Infantil, Depresión,* etc. Entre los temas a que viene dedicando una particular atención entre sus investigaciones recientes constituyen un lugar destacado aquellos en que se concitan los conocimientos psicopatológicos y la práctica de los procesos de enseñanza-aprendizaje.

adultos que convivían con ellos, se cuentan entre los primeros ejemplos de registros en los que intervienen observadores humanos con el propósito de recabar datos en ambientes naturales y durante largos períodos de tiempo (WILDMAN y ERICKSON, 1977). De este modo se conseguían registros de ciertos datos, que de otra forma el propio investigador no podría haber cubierto mediante su presencia personal.

Con el auge que experimentó el estudio del desarrollo infantil durante las décadas de los años veinte y treinta, se estimuló el empleo de observadores para la obtención y el registro de datos de una amplia gama de comportamientos. Con el masivo empleo de observadores, comenzó a cobrar interés el estudio de las circunstancias metodológicas concurrentes en su uso. De esta forma, cuestiones como las técnicas de entrenamiento, los códigos de registro, la aparición de sesgos y otras se convirtieron en obligado objeto de estudio (ARRINGTON, 1932; THOMAS, LOOMIS y ARRINGTON, 1933).

Hasta los años sesenta y setenta, el empleo de observadores no vuelve a tener un papel especialmente significativo en la recogida de los datos conductuales. Pero en esta ocasión, el papel desempeñado es incluso de mayor importancia, ya que los observadores comienzan a emplearse para registrar y obtener datos, tanto en estudios básicos como aplicados, y tanto en el ámbito educativo como en el clínico.

Al incrementarse su utilización, se presta mayor atención a sus fundamentos metodológicos y a los problemas surgidos durante el mismo proceso de observación. En la actualidad, nos encontramos en un momento en el que la necesidad de objetividad y replicabilidad del trabajo experimental clínico y educativo nos obliga a prestar una mayor atención al hecho de la fiabilidad. Cuando el registro automático del comportamiento del sujeto no es posible o no es suficiente, los observadores humanos —empleados de forma metodológicamente correcta— son la mejor alternativa.

La necesidad de estimar la consistencia de las medidas de un determinado evento observado deviene en la razón principal para el uso de observadores independientes en el marco de la evaluación clínico-educativa. Una evaluación sólo es útil en la medida que sea obtenida con algún grado de consistencia (KAZDIN, 1981 *a*).

Las ventajas de contar con observadores fiables y precisos son múltiples:

1.ª Las variaciones que se obtengan en los datos no podrán entonces ser interpretados como una mera función de las inconsistencias en la medición.
2.ª Podemos utilizarlos como termómetro de la objetividad y fiabilidad del instrumento de medición que estamos utilizando, ya que, como KENT y FOSTER (1977) precisan, los observadores fiables proveen las bases para asegurar que los registros conductuales obtenidos son un producto replicable al haberse empleado para su obtención procedimientos de registro bien especificados y no juicios idiosincrásicos.
3.ª El grado de dificultad por conseguir niveles de acuerdo deseables, cuando los observadores son fiables y precisos puede ofrecernos una medida rigurosa acerca de la aplicabilidad o no, en la práctica, de un determinado instrumento concreto de evaluación.
4.ª Nos permite asegurar, además, que los datos conseguidos por los observadores pueden generalizarse a otros observadores.
5.ª Y, por último, la concordancia entre los observadores refleja si la conducta problema está bien o mal definida.

En definitiva, no hemos de olvidar que los hallazgos de una investigación no pueden ser más confiables y válidos que los procedimientos de medición a través de los cuales aquéllos se obtienen (KENT y FOSTER, 1977).

La concordancia entre observadores es, por tanto, imprescindible si optamos por una mayor objetividad en los datos obtenidos. Si existieran diferencias en los registros, no podríamos saber cuál es la verdadera ejecución del sujeto, y de esa forma, a las naturales fluctuaciones de la conducta, uniríamos una fluctuación más producida por los observadores.

*Necesidad de varios observadores*

De todo lo anterior se deduce que si utilizamos un solo observador (o si el propio investigador, terapeuta o educador es el observador único) para registrar la ocurrencia de una conducta determinada, cualquier cambio anotado (o una ocurrencia de dicha conducta) puede haberse producido realmente en la conducta, pero puede también haber aparecido como consecuencia del relajamiento —o la rigidez— observacional del

observador con respecto a la definición de la conducta-objetivo. Distracciones, interpretaciones erróneas u otros sesgos podrían influir también en los registros del observador, sin posibilidad de que sean detectadas las ocurrencias de esas alteraciones de la fiabilidad.

Sólo el empleo de diferentes observadores independientes, que registran la misma conducta simultáneamente (y en el mismo sujeto o grupo de sujetos), nos permitirá detectar la posible aparición de las amenazas y sesgos anteriores en torno a la fiabilidad terminal de los datos obtenidos. Todas y cada una de las ventajas señaladas más arriba, provinientes de una observación fiable y precisa, pueden lograrse con el empleo de observadores simultáneos e independientes.

*Investigación aplicada y empleo de observadores*

En todas aquellas investigaciones aplicadas (educativas, clínicas, sociales...) en las que estemos interesados en obtener registros fiables de la conducta o conductas-objetivo, motivo de nuestro estudio, y no podamos utilizar medios automáticos de registro, resula una necesidad imprescindible el empleo de observadores humanos.

En investigaciones educativas individuales o de grupo podemos necesitar observadores que registren, por ejemplo, el número de niños que emitan tal o cual comportamiento o el número de ocasiones que un sujeto determinado se levanta o alborota en clase, interacciona bien con sus compañeros o realiza cualquier otro tipo de comportamiento. En la práctica clínico-educativa en educación especial resultan también muy eficaces los observadores, ya que el terapeuta directamente implicado en la interacción educativa con el sujeto puede encontrarse en serias dificultades para obtener datos sobre la ejecución del niño, así como sobre su propia ejecución. Datos que, por otra parte, van a ser vitales para el continuo control de su diseño y la efectividad de las técnicas que está empleando.

Igual sucede en la práctica clínica psiquiátrica y psicológica. Registrar, por ejemplo, la frecuencia de verbalizaciones de autodesprecio, las interacciones y quejas de un paciente depresivo; registrar las interacciones de un esquizofrénico crónico con los demás en las sesiones de terapia de grupo; obtener una línea de base de las conductas autoestimuladas de un niño autista, podrían ser algunas de las muchas situaciones que demandarían la utilización de observadores cuando el objetivo es obtener datos confiables.

No obstante, a pesar de las amplias posibilidades de aplicación y del trasfondo de objetividad que caracteriza el trabajo aplicado que emplea observadores, el entrenamiento en observación es básicamente inexistente todavía en la mayoría de los diseños curriculares de las profesiones en que su uso potencial parece más obligado.

Aceptar, sin embargo, la necesidad de usar observadores no resuelve los problemas que su uso conlleva. Tres serían los aspectos capitales que deberíamos considerar en el empleo de observadores para la obtención de datos comportamentales fiables. Uno de ellos se refiere al entrenamiento que dichos observadores deben recibir; un segundo aspecto consiste en las posibles fuentes de error que amenazan la fiabilidad y validez de la observación, y en tercer lugar, la forma en que debemos calcular (cuantificar) el grado de confiabilidad y precisión con que los observadores han trabajado. El propósito de este artículo reside en este último.

Por *precisión* de los observadores se entiende la adecuación de su ejecución a un criterio previamente establecido (HAYNES, 1978). Por *confiabilidad* de los observadores se entiende el grado de acuerdo logrado entre ellos durante la observación (HAYNES, 1978). La confiabilidad fundamenta el grado máximo de confianza con que pueden ser valorados los datos de observación. Un bajo acuerdo entre los observadores sugiere posibles deficiencias en su entrenamiento, excesiva complejidad en el sistema de registro, presencia de sesgos en los observadores o deficiencias en la definición de las conductas a observar. Pero antes que confiables es necesario que los observadores sean precisos en su observación, es decir, que adecúen su ejecución a los criterios concretos establecidos previamente para esa observación. La confiabilidad de dos observadores es útil *sólo* si previamente hemos determinado su precisión, su adecuación al criterio, hasta el punto de que si los dos observadores son precisos, necesariamente sus resultados serán también confiables. Por contra, dos observadores pueden ser confiablemente imprecisos, esto es, puede haber confiabilidad entre los resultados obtenidos independientemente por cada uno de ellos, por el hecho de que estén cometiendo sistemáticamente los mismos errores.

La precisión de los observadores es una cuestión que puede optimizarse a través fundamentalmente del entrenamiento de ellos. Con unos observadores precisos, la cuestión de la confiabilidad de sus registros pasa a primer plano.

Obtener el grado de confiabilidad de una observación es útil por varias razones. En primer lugar, porque la evaluación (de una conducta, problema u objetivo de intervención clínico-educativo) sólo será de utilidad si puede obtenerse con cierta consistencia. La medición inconsistente introduce variaciones en los datos, las cuales se suman a las fluctuaciones naturales de la ejecución del sujeto (o grupo), resultando imposibles de discernir y pasando inadvertidas al investigador (KAZDIN, 1975). En segundo lugar, porque el acuerdo entre observadores es un buen indicador del grado de replicabilidad de los datos observacionales (WILDMAN y ERICKSON, 1977). En tercer lugar, porque la confiabilidad entre observadores es un buen predictor de si la conducta a observar está bien definida o si aparecen sesgos en la ejecución de alguno de los observadores (KAZDIN, 1975). Y, en cuarto lugar, porque un registro confiable resulta irrenunciable si queremos que el cambio conductual sea reflejado en función de las variaciones operadas en esa conducta, ya que hemos de comparar, a través del tiempo, patrones estables de conducta (KAZDIN, 1975).

En definitiva, como establecen HAWKINS y FABRY (1979), la principal función de la valoración de la confiabilidad de los observadores es ofrecernos garantías acerca de la verosimilitud de que los efectos experimentales conseguidos (vale decir terapéuticos o educativos) son una consecuencia real de nuestra intervención y no un mero artefacto metodológico.

*El cálculo de la precisión y la confiabilidad de los observadores*

La *precisión* de los observadores se calcula comparando su registro con algún registro criterio, que puede ser otro observador con amplia experiencia, un protocolo desarrollado por «expertos», un registro permanente de la conducta observada *(videotape* o grabación magnetofónica) o la observación realizada por el propio investigador. La *confiabilidad* de los observadores, en cambio, se calcula comparando entre sí sus registros.

En ambos casos, el índice obtenido ha sido tradicionalmente expresado como un *porcentaje de acuerdo* entre ambos registros (observador-criterio y observador-observador). Este porcentaje de acuerdo se ha definido, en unos casos, como la proporción de acuerdos (de ocurrencia y no ocurrencia de una conducta) dividido entre el número total de conductas

registradas (acuerdos más desacuerdos) (HARTMANN, 1977), y en otros, como la proporción de acuerdos (sólo de ocurrencia de la conducta) dividido entre el número total de conductas (acuerdos más desacuerdos) (HAWKINS y DOTSON, 1972). Este índice, por otra parte, bastante utilizado (KELLY, 1977), se revela, sin embargo, insensible a los acuerdos o desacuerdos que podrían ocurrir por azar, además de indicar una fiabilidad muy diferente, según ocurran mayor o menor número de conductas cuando se realiza un registro de intervalos; es decir, se ve muy influido por la tasa de respuestas.

Otro índice utilizado con frecuencia ha sido el *coeficiente de correlación de Pearson,* en especial cuando se pretende hallar la fiabilidad de las observaciones continuadas en varias sesiones (clínicas o experimentales). Cuenta con la ventaja de indicar no sólo la fiabilidad de las medidas entre sesiones, sino también su significación estadística al contrastar ese coeficiente de correlación con la probabilidad normal de obtenerlo al azar. Sin embargo, tiene el inconveniente de que depende altamente de la varianza de las puntuaciones y, además, si por cualquier causa un observador obtiene medidas sistemáticamente sesgadas, sus datos mostrarán una correlación invariable.

Para tratar de salvar los problemas que plantean los porcentajes de acuerdo, por una parte, y el empleo del coeficiente de correlación, por otra, han aparecido diversos índices que pretenden ofrecer una información más completa sobre la confiabilidad del registro. El índice *kappa,* propuesto por COHEN (1960, 1968), tiene en cuenta la proporción de acuerdos y desacuerdos, así como la probabilidad de que éstos sean obtenidos al azar. Igual sucede con el índice *phi,* propuesto por SCOTT (1955). Ambos índices se basan en la fórmula $Po\text{-}Pc/1\text{-}Pc$ (donde $Po$ es la proporción de acuerdos entre los observadores y $Pc$ traduce la probabilidad de realizar un acuerdo al azar).

Para que estos estadísticos puedan ser utilizados en la determinación de la confiabilidad de una observación, son necesarias tres condiciones:

1. Las categorías utilizadas en la observación han de ser nominales y mutuamente excluyentes, esto es, deben constar de dos o más categorías.
2. Las clases o categorías empleadas han de ser entre sí independientes.
3. Los observadores han de actuar también independientemente.

Estos estadísticos podrían ser utilizados, por ejemplo, con categorías nominales (es decir, ordenadas en orden creciente o decreciente) y con categorías dicotómicas (es decir, «sí-no», «correcto-incorrecto», «presencia-ausencia», etc.), categorías que son coincidentes con las más utilizadas en la investigación clínico-educativa. Sin embargo, según parece, una de las razones que impiden el uso extendido de estos índices es su mayor complejidad y dificultad de cálculo (HARTMANN, 1977). Si esto es así, una forma de solucionar dicha dificultad sería, por ejemplo, la génesis de sencillos programas de cálculo automatizados.

Con todo, no existe un método universal de calcular el acuerdo entre observadores. Según qué métodos, resultarán adecuados para según qué casos y condiciones.

A fin de valorar mejor este último punto, podemos seguir la diferenciación que hizo HARTMANN (1977) entre *confiabilidad de ensayo y confiabilidad de sesión.* El índice que utilicemos para obtener la confiabilidad de nuestra observación va a estar determinado, en primer lugar, por el tipo de datos de que se trate. En la confiabilidad de ensayo se utilizan como unidad de análisis los datos registrados en cada ensayo realizado (o intervalo registrado), dentro de cada sesión. Por el contrario, en la confiabilidad de sesión, las unidades de análisis son más globales y comprenden las puntuaciones totales de cada sesión o fase completa de una investigación o tratamiento. En el cuadro 1 resumimos las características más relevantes de ambos índices.

Como medidas de confiabilidad de sesión, se han recomendado los porcentajes de acuerdo (HARTMANN, 1977; HAYNES, 1978; WALLS y colaboradores, 1977), también llamados porcentajes de confiabilidad total por BIRKIMER y BROWN (1979). Han sido recomendados también el porcentaje de acuerdo completo (HARTMANN, 1977) y los coeficientes de correlación tradicionales (HARTMANN, 1977, y WALLS y cols., 1977).

Para la confiabilidad de ensayo, la oferta resulta aún más variada. Los índices varían desde el clásico porcentaje de acuerdo hasta los más sofisticados estadísticos *kappa* y *phi* —que anteriormente comentamos—, pasando por los llamados porcentajes de acuerdo efectivo (de ocurrencias y de no ocurrencias), así como diversas formas basadas en juicios gráficos y reglas (por ejemplo, ver BIRKIMER y BROWN, 1979 *a*, 1979 *b*).

CUADRO 1

*Tipos generales de índices y principales características*

| *Confiabilidad calculada* | *U t i l i d a d* |
|---|---|
| DE SESION | — Indica la generalidad de las puntuaciones a través de los observadores.<br>— Refleja la adecuación de las definiciones de la conducta y el grado de perfección del entrenamiento de los observadores al usar las definiciones. |
| DE ENSAYO | — Indica la adecuación de las definiciones conductuales con mayor precisión.<br>— Indica la precisión conseguida en el entrenamiento de los observadores en el uso de las definiciones y del equipo observacional ensayo a ensayo.<br>— Indica la precisión del equipo observacional (código, instrumentos de registro, etc.). |

Las tablas 1 y 2 recogen los aspectos más importantes de estos índices, así como los de confiabilidad por sesión, enumerando algunas de sus ventajas e inconvenientes más importantes.

Las ventajas de utilizar índices de confiabilidad por ensayo y, en concreto, aquellos corregidos al azar, se menciona brevemente en las anteriores tablas. No obstante, también pueden resultar ciertas desventajas de su utilización. Un ejemplo ilustrará mejor lo que aquí se dice. La figura 1 muestra los datos obtenidos por dos observadores en cuatro

| | 1 (2: 26 \| 1 / 0 \| 1) | 2 (2: 1 \| 0 / 0 \| 62) | 3 (2: 34 \| 1 / 1 \| 1) | 4 (2: 113 \| 1 / 0 \| 2) |
|---|---|---|---|---|
| Pa | 96 | 100 | 94 | 99 |
| k | .65 | 1 | .47 | .79 |
| V.C. | 1.15 | 1.42 | 0.96 | 1.79 |
| Signific. | .05 | .05 | .05 | .05 |

FIGURA 1.—*Matriz de datos resultante de cuatro diferentes evaluaciones.*
Se muestra los acuerdos y desacuerdos, según la distribución que puede consultarse en la figura 2.
V.C. significa valor crítico de kappa.

TABLA 1

*Indices de confiabilidad por sesión*

| *Nombre que recibe el Método* | *Fórmula de cálculo* | *Ventajas y aplicaciones* | *Inconvenientes* |
|---|---|---|---|
| Porcentaje de acuerdo (HARTMANN, 1977)<br>Coeficiente de Acuerdo (*) (HAYNES, 1978)<br>Confiabilidad total (*) (BIRKIMER y BROWN, 1979a) (HAWKINS y DOTSON, 1975) | $PA = \frac{P.\ Máx.}{P.\ Mín.} \times 100$<br>$Ao = 1 - \frac{Pmáx - Pmín}{Pmáx}$<br>(fórmula alternativa) | — Simplicidad de obtención e interpretación.<br>— Util para evaluar si la diferencia entre las puntuaciones de sesión son reales o debidas a errores de observador. | — Ausencia de un límite inferior significativo de acep tabilidad.<br>— Ausencia de valor que indique el desacuerdo.<br>— Su valor es muy dependiente de la tasa específica de R (tasas altas = acuerdos elevados).<br>— No mide el posible acuerdo al azar. |
| Porcentaje de acuerdo completo (HARTMANN, 1977) | $PAC = \frac{N.°\ A.\ c.}{N.°\ total} \times 100$ | — Idénticas al anterior. | — Valor limitado por ser muy restrictivo.<br>— En realidad es un índice de acuerdo por ensayos. |

TABLA 1 (continuación)

| *Nombre que recibe el Método* | *Fórmula de cálculo* | *Ventajas y aplicaciones* | *Inconvenientes* |
|---|---|---|---|
| Coeficiente de fiabilidad ($r_{kk}$) | | — Indica el grado de confianza que puede darse a las puntuaciones de la sesión.<br>— El error estándar de medida —que es una función de $r_{kk}$— puede usarse para generar un *intervalo de confianza* que indica la diferencia más pequeña entre puntuaciones de sesión que puede ser interpretada significativamente.<br>— Proporciona una descripción precisa del grado de dependencia lineal o correlación en las puntuaciones de los observadores.<br>— Los sesgos del observador pueden evaluarse por la *prueba t* de la diferencia entre las puntuaciones correlacionadas. | — Poco indicado para puntuaciones con escasa variabilidad.<br>— Lleva al clínico demasiado lejos de los datos directos (HAWKINS y FABRY, 1979). |

**TABLA 2**

*Indices de confiabilidad por ensayo*

| | *Indices y fórmula de cálculo* | *Ventajas y aplicaciones* | *Inconvenientes* | *Indices similares* |
|---|---|---|---|---|
| INDICES GLOBALES DE CONFIABILIDAD | Porcentaje de acuerdo: P*a* (p. e., HARTMANN, 1977)<br>$Pa = \frac{B + C}{T} \times 100$<br>también:<br>$Pa = \frac{\text{acuerdos}}{\text{acuerd.} + \text{desac.}} \times 100$ | — Indica la proporción de observaciones totales en las que los dos observadores concordaron.<br>— Resulta muy sencillo de calcular. | — Confunde errores al azar con errores sistemáticos. | — Confiabilidad I × I (HAWKINS y DOTSON, 1975)<br>— Confiabilidad momento a momento. (BIRKIMER y BROWN, 1979a)<br>— Acuerdo por intervalo. (HAYNES, 1978) |
| INDICES PARCIALES DE CONFIABILIDAD | Porcentaje de acuerdo efectivo sobre ocurrencias: P*o* (HARTMANN, 1977)<br>$Po = \frac{B}{(A+B+D)} \times 100$ | — Indican los acuerdos efectivos sobre ocurrencias y no ocurrencias (o correcto o incorrecto). | | — Confiabilidad de intervalo (o ensayo) puntuado (SI). (HAWKINS y DOTSON, 1975)<br>— Fiabilidad de ocurrencia (Fo). (KENT y FOSTER, 1977) |
| | Porcentaje de acuerdo efectivo sobre no ocurrencias: P*no* (HARTMANN, 1977)<br>$Pno = \frac{C}{(A+C+D)} \times 100$ | — Proveen una medida de la fiabilidad más sensible al excluir, al menos parcialmente, las contribuciones de los acuerdos que puedan ser debidos al azar (en función de la tasa de respuesta).<br>— Previenen, por tanto, de la sobreestimación del acuerdo cuando las tasas son extremas. | — Cálculos más largos y complejos al ser necesario computar ambos índices.<br>— No eliminan totalmente los efectos de los acuerdos posibles al azar.<br>— Considera igual error o acierto al azar que error o acierto reales. | — Confiabilidad de intervalo (o ensayo) no puntuado (UI). (HAWKINS y DOTSON, 1975)<br>— Fiabilidad de no ocurrencia (Fno). (KENT y FOSTER, 1977) |

*Abreviaturas empleadas:* A: ver figura 2. B: ídem. C: ídem. D: ídem. T: total de ensayos o intervalos.

TABLA 2 (continuación)

| | *Indices y fórmula de cálculo* | *Ventajas y aplicaciones* | *Inconvenientes* | *Indices similares* |
|---|---|---|---|---|
| INDICES CORREGIDOS | KAPPA: $k$<br>(COHEN, 1960)<br>$$k = \frac{p_o - p_o}{1 - p_o}$$<br>PHI: $\varnothing$<br>(SCOTT, 1955)<br>$$\varnothing = \frac{(A \times D - (B \times C)}{(B+A)\ (D+C)\ (B+D\ (A+C)}$$ | — Excluyen los acuerdos al azar, por lo que ofrecen una confiabilidad «rectificada».<br>— No parecen verse afectados por las tasas de ocurrencia extremas.<br>— Resumen toda la información en una puntuación única, fácil de comparar interestudios.<br>— $k$ puede también calcularse sobre ocurrencias y no ocurrencias.<br>— Ofrecen la posibilidad de calcular su significatividad estadística. | — Confunden errores al azar y errores sistemáticos.<br>— Con tasas extremas altas o bajas, las posibilidades de acuerdo al azar se incrementan en una u otra dirección, por lo que pueden confundir en estos casos la naturaleza de los acuerdos.<br>— Implican un grado de sofisticación estadística elevado, lo que dificulta su cálculo (y, por tanto, su uso) en el campo aplicado. | |

sesiones completas de registro ocurridas durante la evaluación del comportamiento de un niño de cinco años de edad, en cuatro subáreas de evaluación de repertorios diferentes («discriminación de colores», «conceptos temporales», «identificación de objetos» y «habilidades de aseo»).

Las matrices de datos que aparecen en la figura 1 corresponden a los acuerdos y desacuerdos de los dos observadores, y se distribuyen según el modelo que se aprecia en la figura 2. Las casillas B y C representan, respectivamente, los *acuerdos* sobre *ocurrencias* entre el observador 1 ($C_1$) y el observador 2 ($C_2$), y los *acuerdos* sobre *no ocurrencias* (u ocurren-

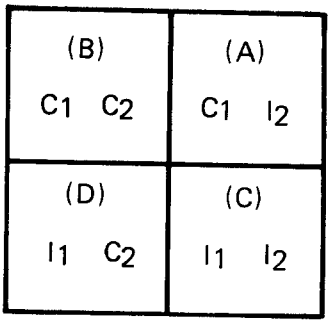


FIGURA 2.—*Matriz típica de representación de los datos para el cálculo de la confiabilidad por ensayo.*

cias incorrectas del comportamiento observado) entre el observador 1 ($I_1$) y el observador 2 ($I_2$). Las casillas A y D representan, respectivamente, los *desacuerdos* entre los observadores; en concreto, la casilla A representa cuándo el observador 1 registra ocurrencia correcta ($C_1$) y el observador 2 ocurrencia incorrecta ($I_2$), mientras que en la casilla D se representan las ocasiones contrarias. La suma total de las cuatro casillas representan el número total de ensayos realizados en esa sesión.

Los índices utilizados en el ejemplo expuesto en la figura 1 son dos para la confiabilidad por ensayo (ver cuadro 2), cuyas características, en nuestra opinión, los hacen complementarios. Por una parte, utilizamos el porcentaje de acuerdo —P*a*— (HARTMANN, 1977), y, por otra, el índice *kappa* —*k*— (COHEN, 1960). El primero de ellos es un índice global, que no rectifica los datos para evitar la posible influencia de los aciertos por azar. El *kappa* es un índice rectificado, en el sentido que tiene en

cuenta la probabilidad de que los índices de acuerdo obtenidos sean debidos a acuerdos por azar.

De la aplicación de ambos índices pueden extraerse, en el ejemplo que citamos, ciertas conclusiones generalizables a la totalidad de su empleo. En las matrices que aparecen en la figura 1 puede observarse cómo en todos los casos el P*a* resultante es muy elevado; el *k* resultante es bajo en tres de los casos, y máximo (igual a 1) en un caso, pero es negativa su significación al .05 en tres de los casos (incluido cuando es igual a 1), y positiva al .05 en el caso restante. Estos datos tal vez resuman bien las características contrapuestas de ambos índices.

Por una parte, es posible que en los casos vistos el azar haya influido en la obtención de una alta tasa de acuerdos sobre ocurrencias y no ocurrencias, pero que «pueda» hacerlo no significa que *necesariamente* lo haya hecho. Y es en este punto donde el índice *kappa* podría resultar excesivamente «severo», en la medida que potenciaría el valor del posible acuerdo al azar en detrimento del acuerdo *verdadero*. Además, si tenemos en cuenta que los ensayos discretos (presentación de un $S^d$ y observación de la ocurrencia o no de una respuesta previamente definida) es la forma utilizada en la obtención de los datos que nos sirven de ejemplo —y una de las más frecuentes formas de trabajar en evaluación y entrenamiento de deficientes y niños normales—, parecería obvio no considerar que las probabilidades de un acuerdo por azar se reducen más que considerablemente en este tipo de ensayos. Acordar la ocurrencia de una respuesta (o su ocurrencia correcta) en 34 ocasiones o en 113, así como su no ocurrencia (o su ocurrencia incorrecta) en 62 ocasiones, como sucede en algunas de las matrices de datos que aparecen en la figura 1, no parece que pueda ser debido al azar, sino más bien al *acuerdo real* entre dos observadores en la aplicación de un código de observación y una definición de respuesta. Si tan sólo utilizáramos el índice *k*, su baja significación podría conducirnos a desdeñar unos datos que, por las razones que apuntamos, no parece que sean desdeñables. El hecho de que otro índice de acuerdo, como el P*a*, ofrezca unas cifras muy altas en todos los casos, nos ayudaría en nuestra consideración positiva de los acuerdos estimados.

*Consideraciones sobre el empleo de índices de confiabilidad en la investigación aplicada*

Del anterior análisis en el ejemplo utilizado podrían desprenderse algunas consideraciones a tener en cuenta cuando utilizamos ciertos índices de confiabilidad en investigaciones aplicadas. Una de ellas es que consideramos más apropiados índices como el P*a*, que índices como el *kappa*, para hallar la confiabilidad, cuando trabajamos con sesiones de evaluación o entrenamiento formadas por ensayos discretos. No obstante, consideramos también que ambos tipos de índices podrían complementarse, en el sentido de que un *k* (o *phi*) bajo, pero con un elevado número de acuerdos presentes (por lo que el P*a* sería alto), serviría como una advertencia o aviso para el investigador educativo o clínico acerca de la posibilidad de que buena parte de sus acuerdos pudieran estar ocurriendo al azar. Que el P*a* obtenido sea alto, nos indicaría que los datos no deben ser desdeñados sin más, sino que, al contrario, dada la naturaleza de los mismos y la forma en que éstos han sido obtenidos, si la precisión de los observadores es correcta, los datos deben ser asumidos como reales. La advertencia de un *k* bajo y/o no significativo queda solventada si comprobamos que los observadores están funcionando en forma precisa (lo que se puede medir por su ajuste a un criterio).

Recordando las características de los diversos índices que se reflejan en los cuadros 1 y 2, podemos afirmar también que otros índices, como son el porcentaje de acuerdos sobre ocurrencias (P*o*) y sobre no ocurrencias (P*no*) parecen útiles, fundamentalmente cuando estimamos que las fuentes de variabilidad de los datos pueden tener cierta independencia. Es el caso de diferentes informantes (por ejemplo, padres y profesores), que en momentos y situaciones diferentes responden a una entrevista o cuestionario sobre la ocurrencia de ciertos comportamientos en un sujeto determinado. Ambos índices ayudan simultáneamente a detectar si existe o no tendencia a responder de alguna manera «socialmente bien vista» o, por contra, a infravalorar o supravalorar el comportamiento del sujeto.

La disponibilidad de una amplia oferta de índices de confiabilidad nos permite, en todo caso, adecuar la elección de índice a la naturaleza de nuestra investigación, puesto que no se debe olvidar que la confiabilidad de una observación (vale decir, de una investigación que utilice la observación en sus diversas posibilidades como la forma de obtener los datos)

no es un fin en sí mismo, sino tan sólo un indicador de que nuestros datos responden a las características del sujeto y no a las de los observadores, a las del cálculo o del instrumental empleado en la observación. Mientras que en el trabajo clínico y educativo cotidiano, la utilización de índices generales de cálculo sencillo pueden satisfacer las exigencias de confiabilidad, cuando el propósito es una investigación, la utilización añadida de índices corregidos al azar parece recomendable.

Una importante dificultad para el uso de los más sofisticados índices corregidos al azar (*k* y *phi*) puede provenir de su complejidad de cálculo. Si esto es así, una solución podría consistir en la aplicación de sencillos programas en lenguaje «Basic», que permitirán la obtención estandarizada de estos índices en forma automática, rápida y sencilla. Es nuestro propósito ofrecer aquí uno estos programas elaborados para el cálculo de los índices *kappa* y *phi*, tal como aparece en el anexo 1 de esta colaboración.

*Conclusiones*

A lo largo de esta colaboración hemos visto que la necesidad de objetividad y rigor en la investigación clínica y educativa puede quedar solventada (en lo que se refiere a la fiabilidad de los datos) con el empleo de observadores que registren los comportamientos de interés, cuando no disponemos de medios automáticos de registro o cuando la situación de observación no nos permite utilizarlos.

La primera condición que deben cumplir los observadores para que los datos obtenidos por ellos sean útiles es que observen con precisión, esto es, que se adecuen a los criterios definicionales y de codificación previamente estipulados. Una segunda condición es que sus datos sean confiables, esto es, que obtengan un elevado porcentaje de acuerdos sobre la ocurrencia y la no ocurrencia de los comportamientos de interés.

Hemos expuesto diversos índices, que representan distintas maneras de calcular el grado en que la precisión y la confiabilidad de una observación permite considerar sus datos como objetivos. Dichos índices representan alternativas de cálculo que, desde puntos de vista diferentes, pretenden el mismo resultado con una mayor o menor sofisticación y complejidad. De todos ellos algunos resultan extremadamente sencillos de aplicar, pero el grado de información que ofrecen es dudoso en bastantes casos; otros son más complejos y sofisticados, pero añaden un punto de

información importante. Si se revisan con atención los cuadros anteriores, fácilmente se puede llegar a la conclusión de que existe un considerable solapamiento en el número de índices que se ofrecen, y que con frecuencia «nuevos» índices sólo han representado diferentes formas de decir lo mismo.

Entre los índices expuestos, vimos que para ciertos tipos de trabajos clínicos y educativos (los más frecuentes en el contexto profesional en que nos movemos) deberían utilizarse conjuntamente, siempre que sea posible, un índice general de acuerdo y un índice de acuerdos corregido al azar. En general, las características y las ventajas y desventajas que se señalan para cada tipo de índices, permite deducir fácilmente qué tipos de índices se adecuan mejor a qué tipos de trabajos.

Otro de los aspectos considerados es la dificultad de cálculo (que se traduce en un uso más bien escaso) de los índices que denominamos corregidos al azar. Ofrecemos un programa sencillo en lenguaje «Basic» —y aplicable a cualquier tipo de ordenador personal—, que soluciona dicho problema y permite un uso cotidiano —rutinario casi— de este tipo de índices.

En posteriores trabajos abordaremos otros aspectos capitales del empleo de observadores humanos en la toma de datos, como la cuestión del entrenamiento de los observadores y las posibles fuentes de error que amenazan la fiabilidad y validez de los datos obtenidos en una observación. De esta forma, creemos haber elaborado una información de gran interés, tanto para aquellos profesionales e investigadores que deseen buscar fórmulas para objetivar algo más su trabajo, como para quienes tienen la responsabilidad directa en la formación de los profesionales vinculados al campo de la salud mental y la educación, sea ésta regular o especial, en los que la aplicación de los observadores resulta absolutamente imprescindible.

## ANEXO 1

### *Programa en «Basic» para el cálculo de los índices* kappa *y* phi

El programa que se presenta ha sido realizado con un ordenador personal «Sinclair ZX 81», con ampliación de 16 *k*. Su objetivo es la obtención de los índices de acuerdo corregidos al azar *kappa* y *phi*. Tiene como entradas, los datos en el orden correspondiente a los casilleros de

la matriz de datos (B, A, C, D), y como salidas, la aparición en pantalla de la tabla de doble entrada, los valores de los índices de acuerdo (*k* y *phi*) y el valor crítico Z para el índice *kappa*. El listado del programa se muestra en la figura 2.

El valor crítico Z del índice *kappa* es directamente contrastable con los valores de la tabla normal de probabilidad.

FIGURA 3.—*Listado en Basic del programa para la obtención de los índices de acuerdo Kappa y phi, y el valor crítico para Kappa.*

```
 10 PRINT «KAPPA Y PHI»
 20 PRINT
 30 INPUT B
 40 INPUT A
 50 INPUT D
 60 INPUT C
 70 PRINT »
 80 PRINT »
 90 PRINT »
100 PRINT »
110 PRINT »
140 PRINT AT 3,1; B
150 PRINT AT 3,6; A
160 PRINT AT 5,1; D
170 PRINT AT 5,6; C
180 LET T = A + B + C + D
185 LET PO = (B + C)/T
190 LET PC = (((A + B) . (B + D))/T . .2) +
    (((A + C) . (C + D))/T . . 2)
200 LET K = (PO — PC)/(1 — PC)
205 PRINT
207 PRINT «PO = »; PO
210 PRINT «KAPPA = »; K
220 LET F = (B . C — A . C)/((A+B) .
    (C + D) . (A + C) . (B + d)) . . 0.5
230 PRINT
240 PRINT «PHI = »; F
250 LET S=SQR (PC/(T . (1 — PC)))
260 LET Z = K/S
270 PRINT
280 PRINT «VALOR CRITICO=»; Z
```

## BIBLIOGRAFIA


ARRINGTON, R. E.: «Some technical aspects of observer reliability as indicated in studies of the 'talkies'», *The American Journal of Sociology,* 38 (1932), 407-417.

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

— «Back to basics: percentage agreement measures are adequate, but there are easier ways», *Journal of Applied Behavior Analysis,* 12 (1979b), 535-543.

COHEN, J.: «A coefficient of agreement for nominal scales», *Educational and Psychological Measurement,* 20 (1960), 37-46.

HARTMANN, D. P.: «Considerations in the choice of interobserver reliability estimates», *Journal of Applied Behavior Analysis,* 10 (1977), 103-116.

HAWKINS, R. P., y DOTSON, V. A.: «Reliability scores that delude: an Alice in Wonderland trip throught the misleading of characteristics of inter-observer agreement scores in interval recording». Paper presentado en el *Tercer Symposium Anual sobre Análisis de Conducta en Educación,* Kansas, 1972.

HAWKINS, P. P., y FABRY, B. D.: «Applied Behavior Analysis and Interobserver Reliability», *Journal of Applied Behavior Analysis,* 12 (1979), 545-552.
HAYNES, S. N.: *Principles of Behavioral Assessment,* New York, Gardner Press, 1978.
KAZDIN, A. E.: *Behavior Modification in Applied Settings,* Homewood (Illinois), The Dorsey Press, 1975. Traducido al castellano por C. PARTIDA: *Modificación de la conducta y sus aplicaciones prácticas,* México, El Manual Moderno, 1978.
KELLY, M. B.: «A review of the observational data collection and reliability procedures reported in the JABA», *Journal of Applied Behavior Analysis,* 10 (1977), 97-101.
KENT, R. N., y FOSTER, S. L.: «Direct observation procedures: methodological issues in naturalistic settings», en CIMINERO, CALHOUN y ADAMS (eds.): *Handbook of Behavioral Assessment,* New York, Wiley, 1977, págs. 279-328.
SCOTT, W. A.: «Reliability of content analysis: The case of nominal scale coding», *Publics Opinion Quarterly,* 19 (1955), 321-325.
THOMAS, D. S.; LOOMIS, A. M., y ARRINGTON, R. E.: *Observational studies of social behavior,* vol. 1, Yale University: Institute of Human Relations, 1933.
WALLS, R. T.; WERNER, T. J.; BACON, A., y ZANE, T.: «Behavior Checklist», en J. D. CONE y R. P. HAWKINS: *Behavioral Assessment: new directions in clinical psychology,* New York, Brumer/Mazel, 1977.
WILDMAN, B. G., y ERICKSON, M. T.: «Methodological problems in behavioral observation», en J. D. CONE y R. P. HAWKINS: *Ob. cit.,* 1977.

## RESUMEN

El artículo trata de la importancia de emplear observadores para la obtención de datos en investigaciones aplicadas clínicas y educativas. Asimismo, se discuten diversos índices de confiabilidad que garantizan que los observadores registran los datos en forma precisa y confiable. Se concluye señalando la utilidad específica de los diferentes índices, y ofreciendo un programa en lenguaje «Basic» para el cálculo de los índices más complejos.

## SUMMARY

This paper deals with the importance of using observers to record data in applied clinical and educational research. Additionally, various indices of agreement which guarantee that data are recorded with accuracy and reliability are discussed. To conclude, different specific applications of the respective indices are indicated and computer program in «Basic» is described which can be used to calculate the more complex indices.