Evaluación de resultados clínicos (y III): Índices de Cambio Fiable (ICF) como estimadores del cambio clínicamente significativo

Clinical outcomes evaluation (III): Reliable Change Index (RCI) as estimator of clinically significant change

Ioseba Iraurgi Castillo.
DeustoSalud. I+D+i en Psicología Clínica y de la Salud.
Universidad de Deusto, Bilbao.

Resumen: Como hemos visto en artículos previos (Iraurgi, 2009a, 2009b), la significación estadística muestra limitaciones para su uso en la investigación de resultados psicoterapéuticos porque se basa en las medias grupales, no proporcionando información sobre la variabilidad individual de los resultados, ni tampoco abordando la significación clínica. La significación clínica se refiere a la importancia práctica de cambio del paciente que se produce como consecuencia de la intervención psicoterapéutica. Jacobson y su grupo de trabajo (Jacobson, Follette y Revenstorf, 1984; Jacobson y Truax, 1991) han propuesto un método para evaluar la significación clínica que se basa en la validación normativa. El Índice de Cambio Fiable (ICF) de Jacobson y Truax es un procedimiento para la determinación de cambios clínicamente significativos que proporciona un medio adicional de análisis a las comparaciones de medias grupales en la investigación de resultados terapéuticos. Este artículo ofrece una visión general de este procedimiento y la metodología de validación clínica. También examina las críticas propuestas al método y las extensiones resultantes que han inspirado. Por último, se ofrece un ejemplo de la aplicación de los diferentes ICF revisados.

Palabras clave: Cambio terapéutico, Índices de Cambio Fiable, Significación estadística, Significación clínica.

Summary: As we have seen in previous papers (Iraurgi, 2009a, 2009b), the statistical significance is seen as limited for use in psychotherapy outcome research because it is based on group means and does not provide information on individual variability of outcome, nor does it address clinical significance. Clinical significance refers to the practical importance of patient change that occurs through psychotherapy. Jacobson and his working group (Jacobson, Follette and Revenstorf, 1984; Jacobson and Truax, 1991) have proposed a method for assessing clinical significance that is based on normative validation. The Reliable Change Index (RCI) of Jacobson and Truax is a procedure for determining clinically significant change that provides a supplemental means of analysis to group mean comparisons in outcomes research with clinical interventions. This paper provides an overview of this procedure and its underlying clinical validation methodology. It will also examine criticisms of the method and the resulting extensions these have inspired. Finally, an example of clinical application of the different ICF will be given.

Key Words: Therapeutic change, Reliable Change Index, Statistical significance, clinic significance.

Con el presente trabajo se completa la trilogía de artículos destinados a la evaluación de resultados terapéuticos. Hasta aquí se ha examinado el papel de la estadística para delimitar si una puntuación de cambio puede ser atribuida a un efecto real del factor de estudio o intervención, en nuestro caso un determinado tratamiento terapéutico, o ha de ser considerada como un posible desenlace debido al azar (Iraurgi, 2009a). También se ha valorado si el cambio observado es lo suficientemente importante en magnitud cómo para poder aceptar que tiene importancia empírica y supone un logro considerable (Iraurgi, 2009b). En el ámbito terapéutico, plantemos que tal logro -de magnitud aceptable y con pocas probabilidades de haberse producido por azar- es un cambio importante para la toma de decisiones; pero, ¿es clínicamente significativo?

Por otra parte, la mayor parte de los procedimientos que se han revisado se centran en la valoración de los cambios cuando se toma como referencia todo un grupo (Iraurgi, 2000; 2009a; 2009b). Un grupo de sujetos que recibe un tratamiento, o no, se compara consigo mismo para valorar si hay un cambio asociado al tratamiento; o, por el contrario, se produce un cambio espontáneo en el caso de que no lo reciban. O bien, dos grupos con tratamientos diferenciales se comparan entre sí para valorar cuál de ellos es más eficaz o efectivo en el logró de los objetivos terapéuticos. No obstante, en la práctica clínica habitual raramente se comparan grandes grupos, o al menos grupos con los suficientes efectivos como para lograr las condiciones óptimas de tratamiento estadístico, siendo más frecuente la valoración de casos individuales. Optamos por los tratamientos que han mostrado su eficacia y efectividad a través de estudios controlados (ensayos clínicos, estudios naturalísticos de intervención, etc.) y que, por tanto, se muestran de utilidad en la reducción de determinada sintomatología o logro de un mejor estado de salud. Pero no nos basta con saber que la intervención es eficaz, lo que interesa saber es si dicha intervención funciona en el caso particular de quién se está aplicando, si el receptor de la intervención logra mejorar su dolencia o estado de salud de una forma clínicamente relevante. Las preguntas que surgen a este respecto son del siguiente tipo: ¿Cuántos participantes modifican sus puntuaciones de forma relevante o clínicamente importante?, ¿cuántos mejoran de forma fiable hacia puntuaciones de no severidad?, ¿cuál es la diferencia de puntuaciones realmente aceptable? (Iraurgi, Trujols, Lozano y González, 2010; Christensen y Mendoza, 1986).

Para dar respuesta a estas cuestiones se han propuesto una serie de criterios y metodología específica basada en lo que se ha venido a llamar 'Índice de Cambio Fiable' (ICF, o en su acepción en inglés RCI: Reliable Change Index) o como -en una de sus primeras aproximaciones en castellano- 'Puntuación Precisa de Cambio' (Páez, Echeburua y Borda, 1993). Se han propuesto varias alternativas de esta metodología (Jacobson y Truax, 1991; Speer, 1992; Speer y Greenbaum, 1995; Hsu, 1999; Hageman y Arrindell, 1999), a las cuales nos aproximaremos en la presente sección. Todas ellas tienen como fundamento la propuesta inicial desarrollada por Jacobson y colaboradores (Jacobson, Follette y Reventorf, 1984, 1986; Jacobson y Truax, 1991; Jacobson, Robers, Berns y McGlinchey, 1999), el cual ha sido propuesto por Speer y Grennbaum (1995) como un buen método para calcular el número de puntos de cambio necesario que permitirían al clínico decidir con confianza qué sujetos han experimentado un cambio clínico relevante.

Índice de Cambio Fiable de Jacobson y Truax (ICFJT)

La primera aproximación a una medida del cambio clínicamente significativo fue propuesta por Jacobson, Follette y Reventorf en 1984, y posteriormente revisada por Jacobson y Truax (1991) en el ya célebre artículo *'Clinical significance: A statistical approach to defining meaningful change in psychotherapy research'*. La descripción del método que se realiza en este epígrafe sigue las indicaciones de estos autores en esta última obra, si bien nos permitiremos la licencia de adaptar algunas cuestiones a los intereses de la presente tesis.

El método propuesto por Jacobson y Truax, (en adelante nos referiremos a él como 'ICF_{JT}') consiste en dos partes diferenciadas: la primera consiste en la definición de los criterios que expresan qué ha de considerarse como 'cambio clínico', y la segunda la propuesta de un índice estadístico que valoré el grado de significación

en la precisión del cambio obtenido. El propósito de esta propuesta era: a) establecer un consenso sobre la definición de 'cambio clínicamente significativo' que sirviera para ser aplicado a cualquier trastorno clínico, b. determinar un punto de encuentro entre los profesionales y los usuarios respecto a lo que supone un cambio clínico en las expectativas de resultado en psicoterapia, y c. ofrecer un método preciso para la clasificación del resultado terapéutico alcanzado por los receptores de la intervención (mejoría, sin cambios, empeoramiento, etc.) en base a los criterios definidos como clínicamente significativos.

1.Criterios para la definición del cambio clínico

En una entrega anterior (Iraurgi, 2009a) hemos hecho alusión al concepto de significación clínica planteando que es un fenómeno que va más allá de cálculos aritméticos y está determinada por el juicio de valor de los distintos protagonistas del escenario sanitario, quienes pueden interpretar la relevancia clínica de forma diferente en tanto que cada uno de ellos pueden poner su(s) objetivo(s) de resultado(s) en opciones diversas (p.ej.: el tamaño de efecto, el alivio de la dolencia, los costes, la duración del tratamiento, la comodidad de la implementación, el mantenimiento de la mejora de salud y aceptación del tratamiento por el paciente, etc.).

De manera general, se puede plantear que la significación clínica de un tratamiento o intervención viene determinada por su capacidad para cumplir ciertas normas sobre su eficacia-efectividad establecidas por dichos protagonistas (pacientes, sanitarios, investigadores...). No existe un consenso al respecto, dado que las voces son múltiples; pero para su utilización práctica, y con aceptable aprobación por las distintas partes, se han aceptado algunas como las siguientes: a) un alto porcentaje de personas tratadas que hayan mejorado con el tratamiento; b) un cambio en el estado de salud que sea reconocible por las personas que están alrededor de la persona tratada (Kazdin, 1977; Wolf, 1978); c) la eliminación de los signos y síntomas que definen el problema de salud (Kazdin y Wilson, 1978); d) alcanzar un estado de salud normativo al finalizar la terapia (Kendall, Butcher y Holmbeck, 1999; Nietzel y Trull, 1988); o e) cambios que reduzcan significativamente el riesgo de presentar problemas de salud (Mavissakalian, 1986).

Para Jacobson y Truax, el punto de partida en la contextualización de la significación clínica tras una intervención terapéutica radica en el cambio hacia los valores de normalidad dentro del área clínica sometida a intervención. Es decir, el cambio clínicamente significativo se produce cuando se evidencia la recuperación de un funcionamiento 'normal' de salud, entendiendo 'normal' desde una conceptualización de distribución poblacional del fenómeno.

Tomemos como ejemplo el caso de la valoración de la calidad de vida como medida de resultado de una determinada intervención, donde se ha utilizado como instrumento de aproximación a dicho concepto el SF-36 (Ware y Sherbourne, 1992; Alonso et al, 1998). El SF-36 valora ocho dimensiones de calidad de vida relacionada con la salud con una puntuación teórica que oscila entre 0 (peor salud posible) y 100 (mejor estado de salud posible). De entre ellas, tomaremos la dimensión que hace referencia a la percepción de Salud General (SG - General Health) como marco de referencia para las explicaciones del procedimiento del ICF_{JT}. La Figura 1 presenta algunos gráficos de distribución poblacional que nos ayudarán en la descripción de los criterios propuestos por Jacobson y Truax en su modelo.

El gráfico F1A de la Figura 1 representa tres curvas de distribución poblacional. La curva en línea continua (en azul) representa el universo poblacional: todos los posibles elementos de un conjunto definido. Representaría el total de personas de una ciudad, una comunidad o un país, por ejemplo, en el caso de que estuviéramos realizando una evaluación sobre la salud percibida en ese universo. Esta es una distribución hipotética, dado que raramente se llega a poder capturar la valoración de todos los elementos de ese universo. A lo sumo se llega a una aproximación a través de selecciones muestrales -a las cuales llamaremos 'población general'-, que con un mínimo error de estimación asumible, llegan a ser una representación significativa de lo que ocurre en ese universo. Por la ley de los grandes números, se asume que la distribución de un determinado fenómeno, en nuestro caso la salud percibida, se aproxima a una curva normal;

habrá un porcentaje no muy grande de personas que referirán tener una salud pésima, una mayor proporción que tendrán una salud normal, adecuada, y de nuevo un pequeño porcentaje de personas que manifestarán un estado de salud excelente. Atendiendo a las puntuaciones en el SF-36, un porcentaje pequeño de personas presentaría muy bajas o muy altas puntuaciones, y la gran mayoría del universo mostraría puntuaciones en torno a un valor promedio más o menos centrado.

Si la extracción muestral se obtiene del colectivo de personas sanas, es decir, aquellas que no presentan ningún trastorno o enfermedad -al menos en lo que respecta al trastorno de valoración-, hablaríamos de una 'población funcional', representada en la Figura 1 por la curva de línea discontinua (en verde). Si por el contrario, el estudio se realizase exclusivamente con personas que presentan el trastorno o enfermedad de interés nos hallaríamos con una 'población disfuncional', representada en la Figura 1 por la curva de línea punteada (en rojo). En los estudios clínicos, como es obvio, se cuenta con muestras de enfermos -'poblaciones disfuncionales'- más o menos amplias y, en algunos casos, se cuenta con datos normativos (Iraurgi, Poo y Markez, 2004). Lo que resulta menos habitual es disponer de poblaciones funcionales, y en ese caso el recurso es acudir a muestras normativas de población general -como se ha comentado previamente obtenidas a través de encuestas de salud o estudios comunitarios-, las cuales incluyen a personas que pueden presentar el trastorno o enfermedad de interés a nivel clínico o subclínico.

Jacobson y Truax (1991) plantean que el cambio clínicamente significativo se produce cuando el enfermo retorna a un funcionamiento normal, es decir, cuando puede ser considerado como parte de la población funcional. De este modo, teniendo en cuenta los distintos tipos de poblaciones descritas y las características de su distribución teórica, estos autores propusieron tres criterios, y sus correspondientes puntos de corte -representados por las letras a, b y c en la Figura 1-, para determinar lo que hubiera de considerarse nivel funcional:

a. El nivel funcional tras el tratamiento, o puntuación post-test, debe situarse fuera de la amplitud de la población disfuncional en la dirección de la funcionalidad (área sombreada en rojo, más a la derecha, del gráfico F1B), entendiéndose por amplitud la distancia de dos desviaciones estándar (DE) desde la media (M) de la distribución. En el

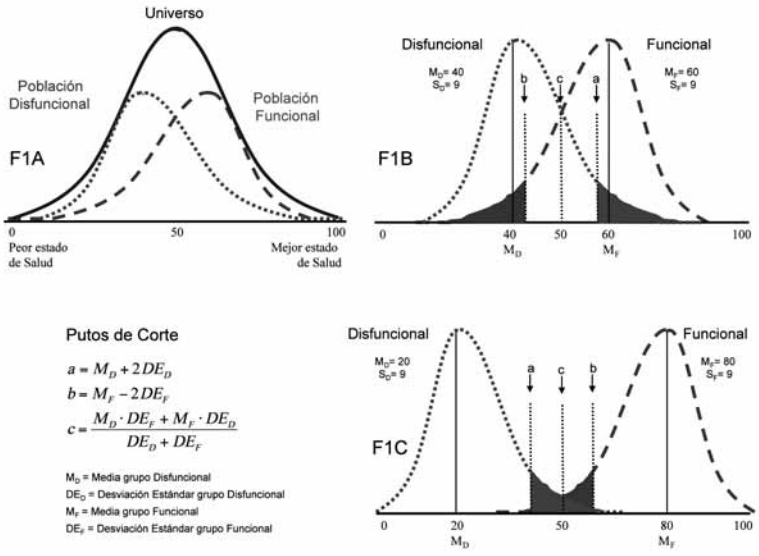

Figura 1. Estimación del Punto de Corte en el Modelo de Jacobson y Truax (1991).

caso que estamos tomando como ejemplo, la característica o variable valorada es positiva (calidad de vida, a mayor puntuación mejor estado) y, por tanto, para estimar el punto de corte 'a' -a partir del cual ha de considerarse alcanzado el nivel funcional- ha de sumarse dos desviaciones estándar a la media ($a = M_D+2DE_D$). En el caso de que la característica a valorar fuera negativa, por ejemplo, sintomatología depresiva (a mayor puntuación mayor expresión sintomática), la funcionalidad vendría expresada por puntuaciones bajas y, por tanto, el punto de corte se hallaría dos desviaciones estándar por debajo de la media ($a = M_D–2DE_D$).

b. El nivel funcional tras el tratamiento se sitúa dentro de la amplitud de la población normal (general) o funcional en la dirección de la disfuncionalidad, es decir, ha de situarse fuera del área sombreada en verde, más a la izquierda, del gráfico F1B. En nuestro ejemplo, al ser la calidad de vida una característica positiva, el punto de corte 'b' se logra como resultado de restar dos DE a la media de la población funcional ($b = M_F–2DE_F$), de modo que la puntuación a alcanzar por una persona tras el tratamiento para considerar que se halla en el nivel funcional ha de ser igual o superior a 'b'. Contrariamente, si la característica valorada fuera negativa (depresión) -las áreas funcional y disfuncional de la Figura 1 se hallarían intercambiadas de posición-, el punto de corte se calcularía mediante la suma ($b = M_F+2DE_F$), y, por tanto, la puntuación del sujeto para ser considerada como funcional hubiera de ser igual o inferior a 'b'.

   Se ha comentado que la disponibilidad de datos de distribución de poblaciones funcionales es poco usual y que en su defecto suelen utilizarse datos de población general. Como también ha sido expuesto, en las poblaciones generales podemos encontrar casos clínicos y/o subclínicos, por lo que el criterio de calcular la amplitud a partir de dos DE puede llegar a ser muy riguroso. A este respecto, se ha propuesto utilizar el criterio de calcular dicha amplitud a partir de la suma o la resta de una desviación estándar (Ware y Kosisnski, 2005), de modo que se obtuviera una mayor probabilidad de seleccionar el nivel de funcionalidad de las personas sin el problema o trastorno de interés, o, al menos, con puntuaciones menos disfuncionales.

c. El nivel funcional tras el tratamiento, o puntuación post-test, ha de ser más cercano a la media de la distribución de la población funcional que a la disfuncional. Este criterio de nivel funcional es el menos arbitrario de los tres y se basa en una probabilidad relativa, que se calcula, bajo el supuesto de que ambas distribuciones siguen una ley normal, a partir de la fórmula 'c' de la Figura 1, que en el caso de que las DE de ambas distribuciones fueran iguales se calcularía de forma más abreviada mediante la expresión: '$c' = (M_F+M_D)/2$. Atendiendo a este criterio, una persona que en su puntuación post-test vaya más allá que el valor obtenido de 'c' en la dirección de la funcionalidad, será más probable que pertenezca a la población funcional que a la disfuncional (el área a la derecha del punto c en el gráfico F1B es mayor en la curva de línea discontinua -funcional- que en la curva de línea punteada -disfuncional-). De este modo, como venimos especificando para los criterios previos, si el resultado valorado es de tipo positivo (calidad de vida), la persona en su post-test debe alcanzar puntuaciones iguales o superiores al valor 'c' obtenido; en el caso de variables que valoren características negativas (depresión), la persona debe puntuar por debajo del valor obtenido en el punto de corte 'c'.

Como puede apreciarse, esta propuesta requiere de datos normativos de la población funcional y disfuncional, las cuales pueden solaparse en mayor (gráfico F1B) o menor medida (gráfico F1C), pudiéndose llegar a producir distintos puntos de corte para cada uno de los tres criterios. Entonces, ¿cuál utilizar?. Jacobson y Truax (1991) hacen las siguientes recomendaciones:

1. Cuando se dispone de datos normativos de ambos tipos de población, funcional y disfuncional, los criterios 'b' y 'c' son preferibles al 'a', ya que permiten localizar a la persona evaluada dentro de la población funcional, lo cual corresponde con los objetivos

terapéuticos a alcanzar: "*el cambio clínicamente significativo se produce cuando el enfermo retorna a un funcionamiento normal*" (Jacobson y Truax, 1991; pp.13), es decir, cuando puede ser considerado como parte de la población funcional.

2. El criterio 'a' es aplicable cuando se carece de normas de población general o funcional. Por lo general, en las evaluaciones clínicas realizadas respecto a los resultados de una determinada intervención contamos con una acumulación de casos sometidos a la misma. Este grupo de casos, evaluados en el pre-test, constituyen una muestra que si resulta lo suficientemente amplia permite obtener datos de su distribución respecto a los cuales se podrá localizar el resultado de un sujeto en el post-test, adoptando el criterio de nivel funcional si su puntuación se desplaza dos desviaciones estándar respecto de la media del pre-test. No obstante, este criterio resulta muy estricto cuando las poblaciones se solapan y, además, puede ofrecer puntuaciones de corte fuera de rango cuando la desviación estándar resulta ser mayor que la puntuación media.

3. Por su parte, el criterio 'b' resultaría el de elección cuando no hay solapamiento con la población disfuncional (gráfico F1C de la Figura 1), como podría ocurrir con problemas de salud muy raros o muy graves. En el caso de solapamiento entre las distribuciones de la población funcional y disfuncional, resultaría ser el criterio más tolerante, pero en el caso de distribuciones no solapadas puede llegar a ser demasiado estricto, de ahí que algunos autores hayan sugerido que el punto de corte se situé a una desviación estándar de la media en lugar de a dos (Ware y Kosisnski, 2005).

4. Por último, el criterio 'c' sería el de elección cuando las distribuciones de la población funcional y disfuncional se solapan, situación que ocurre en la mayoría de fenómenos sociales y problemas de salud. En el caso de la valoración de la calidad de vida, las personas pertenecientes a la población funcional que puntúan más bajo (hacia la disfuncionalidad) presentan puntuaciones similares a las personas de la población disfuncional que puntúan más alto. Esto es debido a que la población funcional en ocasiones resulta ser una extracción de la población general, y las personas que puntúan de forma coincidente con las de la población disfuncional probablemente presenten niveles subclínicos del problema de interés.

Retomando el ejemplo propuesto al inicio sobre la valoración de la percepción de la salud general, consideraremos los datos de población funcional ofrecidos por Anaitua y Quintana (1999) para la población general de la Comunidad Autónoma Vasca (M_F= 67,6; DE_F= 19,6; n_F= 3.949) y los datos normativos ofrecidos por Iraurgi, Poo y Markez (2004) sobre una muestra de personas con adicción a opiáceos en tratamiento con metadona de la misma comunidad como representa-

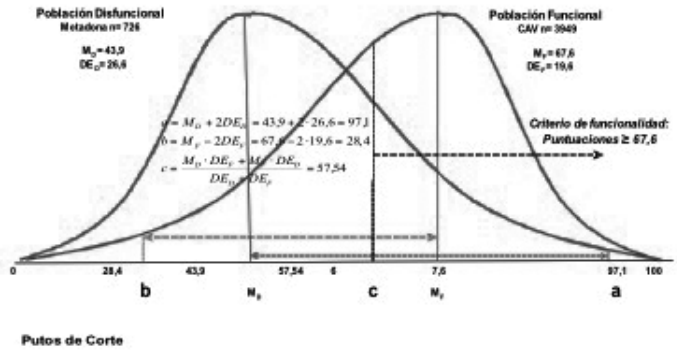

Figura 2. Distribución de poblaciones funcionales y disfuncionales y estimación de los Puntos de Corte para el caso de valoración de pacientes con adicción a opiáceos en tratamiento con metadona.

ción de la población disfuncional (M_D= 43,9; DE_D= 26,6; n_D= 726). Como se aprecia, las medias de ambas distribuciones se separan 23,4 puntos, presentando una peor salud general la población disfuncional -pacientes de metadona- que la funcional, siendo esta diferencia estadísticamente significativa (t= 23,9; p<0,001). Dada la considerable variabilidad de las poblaciones, se puede afirmar que ambas distribuciones se superponen (Figura 2), teniendo que decidirse qué criterio o punto de corte elegir para la asunción del nivel de funcionalidad. Como puede apreciarse en la Figura 2 el criterio 'a' resulta excesivamente estricto (se deberían obtener puntuaciones superiores a 97,1 puntos para cumplir criterios de funcionalidad), y el criterio 'b' resultaría demasiado laxo (bastaría con alcanzar puntuaciones superiores a 28,4 puntos, lo cual resulta paradójico al ser esta puntuación inferior a la media de la población disfuncional). Dado el solapamiento, entonces, se opta por el criterio 'c' como el más adecuado, de modo que para poder asumir que una persona con adicción a opiáceos ha alcanzado un nivel funcional tras ser tratado en un programa de metadona, debería alcanzar una puntuación igual o superior a 57,54 puntos en la escala de salud general del SF-36.

Ahora bien, alcanzar una puntuación post-test que se halle en el área de criterio de funcionalidad no implica necesariamente que se haya producido un cambio clínicamente significativo. Además de alcanzarse el nivel de funcionalidad que se ha descrito en este epígrafe, Jacobson y Truax (1991) proponen el cumplimiento de una segunda condición basada en el grado de cambio logrado tras el tratamiento.

2. Índice de Cambio Fiable (ICF_{JT})

La respuesta a un instrumento de valoración, como puede ser una escala de auto-informe como el SF-36, por ejemplo, está afectada de un cierto grado de imprecisión: la contestación a la escala puede variar por diversas circunstancias. Una puntuación en una escala rara vez es tan precisa que podamos estar seguros que una persona responda exactamente del mismo modo en dos momentos diferentes, o que la puntuación un punto más alto o más bajo sea sin duda un verdadero cambio en el estado de esa persona. De igual modo, la utilización de un esfigmomanómetro mal calibrado y/o utilizado por diferentes personas con diferente nivel de entrenamiento estaría produciendo una medición sesgada que no reflejaría el verdadero valor de la presión arterial. Al grado de error que comete cualquier instrumento se le denomina Error Estándar de Medida (EEM), y depende de la fiabilidad de la prueba (Anastasi y Urbina, 1997).

Así, cuando una persona tiene un cambio en la puntuación en una escala, como la puntuación en la dimensión de Salud General del SF-36, tenemos que evaluar en términos de la probabilidad que ese cambio sea un cambio real, y no un cambio debido a la inestabilidad intrínseca de la escala. Es por ello que resulta importante elegir instrumentos que hayan mostrado su fortaleza y precisión psicométrica. El uso de instrumentos con poca precisión llevaría a un mayor error de medida, de modo que a la variabilidad debida a los propios sujetos habría que añadir la cometida por el instrumento, es decir, la magnitud del cambio pudiera ser debido a un error de medición, y en ese caso los estimadores del tamaño del efecto estarían aportando una información sesgada.

Jacobson y Truax (1991) proponen, como segunda condición para la consideración de cambio clínicamente significativo, la utilización de un índice que contemple la precisión de la medida; el aludido Índice de Cambio Fiable (ICF_{JT}). Este índice fue propuesto inicialmente por Jacobson, Follette y Revenstorf (1984) y revisado posteriormente por Christensen y Mendoza (1986) y Jacobson y Revenstorf (1988). El algoritmo de cálculo se expresa del modo siguiente:

$$ICF_{JT} = \frac{X_{Post} - X_{Pre}}{EED} = \frac{X_{Post} - X_{Pre}}{\sqrt{2\left(DE_{Norma}\sqrt{1-r_{xx}}\right)^2}} \quad (1)$$

Donde $$EED = \sqrt{2(EEM)^2} = EEM\sqrt{2} \quad (2)$$

$$EEM = DE_{Norma}\sqrt{1-r_{xx}} \quad (3)$$

Siendo X_{Post} la puntuación Post-tratamiento, X_{Pre} la puntuación Pre-tratamiento, EED el Error Estándar de las Diferencias entre las dos medidas, DE_{norma} la Desviación Estándar de la muestra normativa y EEM el Error Estándar de Medida del instrumento.

Veamos como es el desarrollo del cómputo, a partir de los datos de nuestro ejemplo, valorando el caso particular de una persona que en su pre-test obtuvo una puntuación de 43,9 en la escala de Salud General (SG) del SF-36, y de 67,6 en el post-test, lo que supone una mejora de 23,7 puntos. El primer paso consiste en calcular el Error Estándar de Medida, para lo cual es necesario contar con el coeficiente de fiabilidad de la escala (·= 0,83) y su desviación estándar (DE= 26,6), habiéndose tomado para este ejemplo los valores correspondientes a la población disfuncional. Por lo demás, el procedimiento sólo requiere hacer los cómputos a partir de las fórmulas (1) a (3), obteniéndose que EEM= 10,9; EED= 15,4 y ICF_{JT}= 1,539. Si el ICF_{JT}, en valor absoluto, es igual o superior a 1,96 (valor de las puntuaciones Z que equivale a dos desviaciones estándar), entonces podremos asegurar que el cambio en la puntuación en SG lograda por ese sujeto resulta fiable. Nótese que los valores pre y post-test asignados al sujeto del ejemplo corresponden con la media de la población disfuncional y la de la población funcional, respectivamente. La diferencia de medias entre ambos grupos había resultado estadísticamente significativa (t= 23,9; p<0,001), pero la misma diferencia para el caso de nuestro ejemplo no alcanza el valor de fiabilidad. Es decir, nuestro caso alcanza un cambio de 23,7 puntos, logra incluso superar la puntuación que le ubica dentro del criterio de funcionalidad (c= 57,4), pero el cambio que ha registrado no resulta con la fiabilidad exigida.

Otro modo de expresar el ICF_{JT} es estimar la puntuación mínima a lograr para que el cambio sea real, fiable. Con una confianza del 95%, se precisaría alcanzar una diferencia de 30,18 puntos entre el pre-test y el post-test ($Z_{0,025}$ x EED = 1,96 x 15,4 = 30,18).

3. Clasificación de los resultados de cambio

Para Jacobson y Truax la combinación de estas dos condiciones son las que definen la existencia de un cambio clínicamente significativo: cuando la magnitud del cambio puede ser considerada estadísticamente fiable y, simultáneamente, alcanza un criterio que localiza la puntuación obtenida en un nivel de funcionalidad. Sin embargo, como señalan algunos autores (Kazdin, 1999) también se puede hablar de significación clínica cuando se alcanza un cambio suficientemente relevante, aunque no alcance el criterio de funcionalidad. Este es el caso de algunas enfermedades graves o crónicas como la esquizofrenia, ciertos traumatismos o algunas personas con adicción a drogas con múltiples patologías orgánicas que se han acogido a un plan de mantenimiento paliativo con agonistas opiáceos hasta que se acaben sus vidas. Conservar e incluso incrementar ciertos niveles de funcionalidad, aunque no sean plenos, y mejorar su calidad de vida son algunos de los objetivos de las políticas de reducción de riesgos y daños para este tipo de colectivos (Iraurgi, 2010).

Pues bien, del cumplimiento o no de los dos criterios planteados pueden diferenciarse varias posibilidades de clasificación del resultado, las cuales pueden asociarse a los tipos de respuesta al proceso terapéutico (Figura 3) propuestos por Kupler (1991):

a. Diremos que una persona se ha '*Recuperado*' cuando el cambio evidenciado sea significativamente fiable (ICF_{JT} ? 1,96) y su puntuación final se sitúe dentro de la distribución normal o funcional. En el modelo de Kupler aparecen dos conceptos equivalentes al de recuperación: el de '*Remisión*', que haría alusión a una reducción significativa de los signos y síntomas de la enfermedad con retorno al nivel funcional, y el de '*Recuperación*', que supondría una remisión mantenida durante un periodo de 6-12 meses.

b. Se clasificará como '*Mejorado*' cuando se evidencie una mejora significativa a partir del ICF_{JT}, aunque no se llegue a alcanzar el nivel funcional. El equivalente en la propuesta de Kupler sería el concepto de '*Respuesta*', reducción de los signos y síntomas en al menos el 50% de los presentados al inicio del tratamiento (Kupler, 1991). Este sería, por ejemplo, el caso de una reducción importante de la sintomatología depresiva en una persona que pasa de un episodio de melancolía con necesidad de ingreso hospitalario a un estado distímico con tratamiento ambulatorio.

c. Se valorará como '*No cambio*' cuando no se produzca un cambio significativo en el ICF_{JT}, independientemente de la posición que

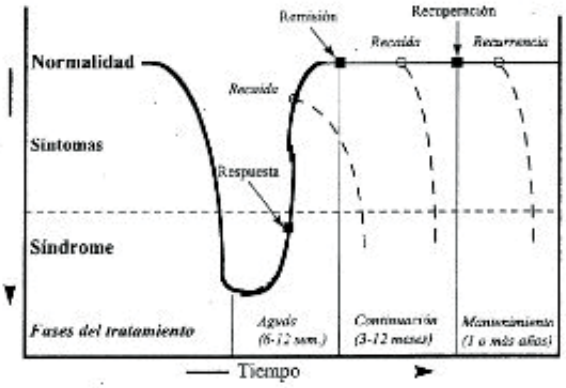

Figura 3. Modelo de Kupler (1991) sobre las fases del tratamiento. (Tomado de Echeburua y Corral, 2001)

ocupe la puntuación del post-test en la distribución poblacional funcional o disfuncional.

d. Se catalogará como '*Deteriorado*' aquellos casos en los que el cambio sea significativo según el ICF_{JT} y se produzca en el sentido contrario al esperado tras la intervención, de modo que sus puntuaciones tornen hacia valores de mayor disfuncionalidad.

e. Cabría una quinta posibilidad, aunque constituiría un subtipo de la clasificación de deterioro, y haría referencia a un cambio desde una situación funcional -una vez lograda tras el tratamiento- a una posición de disfuncionalidad, en cuyo caso estaríamos ante una '*Recaída*' (aparición de sintomatología durante la fase de remisión o recuperación) y/o de una '*Recurrencia*' (reaparición del problema de salud después de la recuperación).

En la Tabla 1 se presentan de modo prototípico una serie de circunstancias que especificarían situaciones posibles de cambio atendiendo a los criterios propuestos por Jacobson y Truax (1991). En la segunda y tercera columnas se presentan las puntuaciones de partida (pre) y finales (post), y en la cuarta columna la diferencia entre ambas (post–pre). Las puntuaciones positivas indican que la puntuación final es mayor que la inicial y, por tanto, ha habido una progresión de la funcionalidad, indicativo de una mejoría clínica; el valor cero revela que ambas puntuaciones son la misma y, por tanto, no ha habido cambio alguno siendo indicativo de estabilidad; por último, puntuaciones negativas son reflejo de una menor puntuación final respecto a la inicial y, por tanto, expresión de un empeoramiento de la situación clínica. La simple sustracción de las puntuaciones nos orienta en la decisión del sentido del cambio. En la sexta columna se presenta el punto de corte, ya calculado previamente, para delimitar el nivel funcional a partir del cual se considerará el logro alcanzado como clínicamente relevante. El valor 'c' se sitúa en 57,4 puntos en la escala SG del SF-36. Puntuaciones iguales o superiores a este valor clasificará el estado del individuo como funcional, puntuaciones inferiores como disfuncional. En la sétima columna aparecen los valores del ICF_{JT} y en la octava columna su interpretación: valores superiores a 1,96 serán consideradas como estadísticamente significativos. Por último, la novena columna ofrece la clasificación de cada caso en función de la combinación de los criterios de Jacobson y Truax (1991). Se plantean hasta un total de 15 situaciones, nueve de las cuales representan un resultado de 'No cambio', si bien se especifican diferentes contextos en los que se evidencia ofreciendo una mayor riqueza descriptiva. También se recogen las posibles situaciones de cambio fiable, entre las que se produce un cambio clínicamente significativo (una respuesta de 'Recuperación' -caso10-), o bien de cambio sin alcanzar la funcionalidad (dos situaciones de 'Mejoría' -casos 11 y 12- y tres de 'Deterioro' -casos 1 a 3-).

Tabla 1. Toma de decisiones en la valoración del cambio clínico a partir de las puntuaciones obtenidas en dimensión de Salud General (SG) del SF-36 (datos ficticios). Combinación de los métodos basados en el Sentido del Cambio (SC), la Puntuación de Corte (PC) y el Índice de Cambio Fiable (ICF_{JT}) caso Puntuación HG del SF-36 Puntuación de Cambio Sentido Cambio Punto de Corte

1	55,2	25,0	-30,2	Empeora	Disfuncional	1,96	Cambio Fiable	Deterioro dentro de la severidad
2	64,2	28,2	-36,0	Empeora	Disfuncional	2,33	Cambio Fiable	Recaida hacia la disfuncionalidad
3	89,9	57,6	-32,3	Empeora	Funcional	2,09	Cambio Fiable	Empeora dentro de la funcionalidad
4	43,5	39,8	-3,7	Empeora	Disfuncional	0,24	No Cambio	Sin cambio dentro de la disfuncionalidad
5	60,2	50,1	-10,1	Empeora	Disfuncional	0,65	No Cambio	Sin cambio, hacia la disfuncionalidad
6	67,5	63,2	-4,3	Empeora	Funcional	0,28	No Cambio	Sin cambio dentro de la funcionalidad
7	39,2	39,2	0,0	Estable	Disfuncional	0,00	No Cambio	Sin cambio dentro de la disfuncionalidad
8	57,3	57,3	0,0	Estable	Disfuncional	0,00	No Cambio	Sin cambio dentro de la disfuncionalidad
9	67,2	67,2	0,0	Estable	Funcional	0,00	No Cambio	Sin cambio dentro de la funcionalidad
10	26,1	59,1	+33,0	Mejora	Funcional	2,14	Cambio Fiable	Recuperado
11	14,3	44,9	+30,6	Mejora	Disfuncional	1,98	Cambio Fiable	Mejora dentro de la disfuncionalidad
12	57,4	87,6	+30,2	Mejora	Funcional	1,96	Cambio Fiable	Mejora dentro de la funcionalidad
13	34,5	48,1	+13,6	Mejora	Disfuncional	0,88	No Cambio	Sin cambio dentro de la disfuncionalidad
14	50,2	58,9	+8,7	Mejora	Funcional	0,56	No Cambio	Sin cambio hacia la funcionalidad
15	60,1	73,1	+13,0	Mejora	Funcional	0,84	No Cambio	Sin cambio dentro de la funcionalidad

Resulta apreciable que este tipo de metodología permite una clasificación más específica y ajustada a los logros individualizados obtenidos por las personas que han recibido tratamiento por un problema de salud, más allá de la respuesta promedio que ofrecería un grupo tratado con dicha terapéutica. Esta característica presenta especial interés para el clínico, que tiene que tomar decisiones cotidianas en relación a la respuesta al tratamiento de sus pacientes, y si bien la estadística no es una de las herramientas habituales de este colectivo, los cálculos para clasificar a una persona en su respuesta al tratamiento no son complicados una vez establecida la medida de resultado y los estándares poblacionales. En el caso de nuestro ejemplo, bastaría con que el clínico restase la puntuaciones obtenidas en la escala SG del SF-36 antes y después de la intervención y lo dividiese por el valor del error estándar de las diferencias (EED= 15,4); si el resultado fuera mayor de 1,96 y la puntuación post-test superase el valor de 57,4, entonces se encontraría ante un cambio fiable clínicamente significativo.

No obstante, el ICF_{JT} presenta algunas limitaciones. Por una parte, es un estimador basado en el error estándar de medida y, por lo tanto, depende de forma directa de la fiabilidad y dispersión de la distribución, de modo que una baja fiabilidad o una alta variabilidad convierten el índice en muy estricto. Asimismo, el índice puede verse afectado en la designación de un cambio fiable cuando las puntuaciones del pre-test son moderadamente funcionales o próximas al límite de la funcionalidad, pues aun cuando puede producirse un cambio quizá no sea de la magnitud que permita considerarlo como fiable al alcanzar el efecto techo o suelo según la característica del fenómeno valorado. Algunos autores (Maassen, 2004) han criticado el método utilizado por el ICF_{JT} para la estimación del error estándar de las diferencias, y otros han advertido la posible influencia del sesgo de regresión a la media en el resultado del índice (Hsu, 1989, 1995, 1999; Speer, 1992; Speer y Greenbaum, 1995).

Métodos alternativos y desarrollos a partir del ICFJT

El interés por la evaluación de los resultados terapéuticos y las limitaciones atribuidas al ICF_{JT}, han suscitado un especial interés por la evaluación del

cambio clínicamente significativo y motivado una prolífica búsqueda del método más apropiado para su estimación. A este respecto, han surgido varias propuestas alternativas de ICF de las que expondremos cuatro, las más representativas.

1. El ICF_{GLN}: Propuesta de Hsu (1989, 1995)

Hsu es el primer investigador que propone un ICF alternativo al del grupo de Jacobson (1984, 1991) basado en los trabajos sobre metodología estadística de Gulliksen en 1950, y los de Lord y Novick en 1969, de ahí las siglas que definen su método (GLN). Hsu, junto a Speer (1992), fueron los primeros autores que hicieron notar cómo el ICF_{JT} podría verse afectado por el fenómeno de regresión a la media, de modo que la interpretación de las puntuaciones del post-test se vería afecta por este efecto. De modo general, la regresión a la media hace alusión a la tendencia de una medición extrema a presentarse más cercana a valores medios en subsiguientes mediciones.

El método GLM implica el control de este potencial efecto de confusión introduciendo en la fórmula de cálculo (4) los valores hipotéticos (media y desviación estándar) de la población de referencia sobre las cuales regresarían las puntuaciones del sujeto o grupo. En el caso de no poder disponer de datos de la población de referencia, Hsu (1999) sugiere utilizar las puntuaciones medias pre-tratamiento del conjunto de participantes en la evaluación terapéutica.

El algoritmo de cálculo propuesto por Hsu para la estimación del ICF_{GLM} es el siguiente:

$$ICF_{GLN} = \frac{(X_{post} - \overline{X}_{pob}) - r_{xx}(X_{pre} - \overline{X}_{pob})}{DE_{pob} \cdot \sqrt{1 - r_{xx}^2}} \quad (4)$$

Siendo X_{post} la puntuación Post-tratamiento, $_{pob}$ la media de la población, r_{xx} el coeficiente de fiabilidad del instrumento de medida, X_{pre} la puntuación Pre-tratamiento, y DE_{pob} la desviación estándar de la población.

2. El ICF_{EN}: Propuesta de Speer (1992)

Speer, además de advertir que el ICF_{JT} podría verse afectado por la regresión a la media, en su primera propuesta de modificación del ICF sugirió la posibilidad de calcular el intervalo de confianza del 95% (IC-95) del índice y la valoración de la puntuación post-test en relación a dicho intervalo. Su método, y el índice correspondiente (ICF_{EN}) se conoce como Edwards-Nunnally, debido a que Speer basa su propuesta en los trabajos de estos dos metodólogos.

Speer (1992) desarrolla su propio método para hacer frente al efecto de la regresión de la media y, como decimos, le incorpora la estimación de un IC-95%, planteando el siguiente algoritmo de cálculo:

$$ICF_{EN} = \left[r_{xx} \cdot (X_{pre} - \overline{X}_{pre}) + \overline{X}_{pre}\right] \pm 2 \cdot DE_{pre} \cdot \sqrt{1 - r_{xx}} \quad (5)$$

Siendo r_{xx} el cociente de fiabilidad del instrumento de medida, X_{pre} la puntuación Pre-tratamiento, $_{pre}$ la media del pre-test y DE_{pre} la desviación estándar del pre-test.

El cálculo ofrece un intervalo de confianza del 95% que indica que en dicho intervalo se halla la posible variación de las puntuaciones del pre-test debidas a la variabilidad causada por el instrumento de medida. La puntuación post-test alcanzada se pone en relación con los valores de este intervalo para tomar la decisión: si la puntuación post tratamiento se halla dentro del intervalo ha de concluirse que no se ha logrado un cambio significativo ya que dicha puntuación está en el rango de posibles puntuaciones que se hallaban en el pre-test; si por el contrario, la puntuación post-test se halla fuera de los márgenes del IC-95%, entonces el cambio alcanzado será estadísticamente fiable.

3. El ICF_{HA}: Propuesta de Hagerman y Arrindell (1999a, 1999b)

Hagerman y Arrindell (1999a, 199b) plantean la conveniencia de revisar el ICF_{JT} por dos motivos: 1) la necesidad de diferenciar entre el cambio que se produce a nivel individual y el que se ocasiona a nivel grupal, que a su juicio requieren diferentes procedimientos de cálculo, y 2) al igual que Hsu y Speer, la necesidad de controlar el efecto de la regresión a la media. Basándose en los trabajos de Cronbach y Gleser de 1959, el método de H-A presenta dos nuevos índices para evaluar la significación clínica del cambio -denominado por ellos como CS_{INDIV}- y el cambio fiable -RC_{INVID}-. El índice CS_{INDIV} supone un intento de proporcionar un punto de corte más preciso a través de una serie de modificaciones como el uso de la puntuación media del pre-test y post-test y los coeficientes de

fiabilidad mostrados por el instrumento de medida en ambos momentos de valoración.

Hagerman y Arrindell (1999b) proponen un sofisticado sistema de algoritmos para el cálculo de sus índices, que se plantean del modo siguiente. El índice de cambio fiable individual viene dado por la expresión:

$$RC_{INDIV} = \frac{(X_{post} - X_{pre})\, r_{dd} + (\overline{X}_{post} - \overline{X}_{pre})(1 - r_{dd})}{\left(\sqrt{r_{dd}}\right)\left(\sqrt{2 \cdot EEM^2}\right)} \quad (6)$$

siendo X_{post} y X_{pre} las puntuaciones post y pre-tratamiento, $_{post}$ y $_{pre}$ medias post y pre-tratamiento, EEM es el error estándar de medida (formula 3); y donde aparece un nuevo estadístico, r_{dd} que es el coeficiente de fiabilidad de la diferencia de puntuaciones postest-pretest, cuyo algoritmo de cálculo viene dado por:

$$r_{dd} = \frac{DE_{pre}^2 \cdot r_{xx-pre)} + DE_{post}^2 \cdot r_{xx-post)} - 2 \cdot DE_{pre} \cdot DE_{post} \cdot r_{pre-post}}{DE_{pre}^2 + DE_{post}^2 - 2 \cdot DE_{pre} \cdot DE_{post} \cdot r_{pre-post}} \quad (7)$$

siendo DE_{pre} y DE_{post} las desviaciones estándar pre y post-test, $r_{xx\text{-}pre}$ y $r_{xx\text{-}post}$ los coeficientes de fiabilidad pre-test y post-test, y $r_{pre\text{-}post}$ el coeficiente de fiabilidad test-retest. Y donde Hagerman y Arrindell proponen los siguientes algoritmos para el cálculo de los coeficientes de fiabilidad en el pre-test y post-test:

$$r_{xx-pre} = \frac{\left(DE_{pre}^2 - DE_E^2\right)}{DE_{pre}^2} \quad \text{y} \quad r_{xx-post} = \frac{\left(DE_{post}^2 - DE_E^2\right)}{DE_{post}^2} \quad (8)$$

Por otra parte, el segundo índice propuesto por Hagerman y Arrindell, la significación del cambio, vendría dada por la siguiente expresión:

$$CS_{INDIV} = \frac{\overline{X}_{post} + (X_{post} - \overline{X}_{post})\, r_{xx-post} - TRC}{\sqrt{r_{xx-post}} \cdot DE_E} \quad (9)$$

donde TRC (True Cutoff) es el Punto de Corte que delimita el nivel funcional y se calcula a partir del siguiente algoritmo:

$$TRC = \frac{\left(DE_{norm}\sqrt{r_{xx(norm)}}\right)\overline{X}_{norm} + \left(DE_{pre}\sqrt{r_{xx(pre)}}\right)\overline{X}_{pre}}{DE_{norm}\sqrt{r_{xx(norm)}} + DE_{pre}\sqrt{r_{xx(pre)}}} \quad (10)$$

Para la interpretación de RC_{INDIV}, Hagerman y Arrindell (1999a, 1999b) proponen localizar la puntuación obtenida en el índice respecto al valor 1,65 (puntuación z que deja por debajo o por encima de si el 5% de las puntuaciones de la distribución), de modo que en el caso de estar valorándose una característica positiva (calidad de vida), puntuaciones inferiores a -1,65 indicarían un 'Deterioro' significativo, valores RC_{INDIV} entre -1,65 y 1,65 expresarían una situación de 'No cambio' fiable, y valores por encima de 1,65 harían referencia a una mejoría relevante. La interpretación inversa se realizaría si la característica de valoración fuera de tipo negativo (por ejemplo, sintomatología depresiva). El CS_{INDIV} proporciona los puntos de corte fiables, de modo que su combinación con los valores del RC_{INDIV} permite clasificar a los individuos evaluados en cuatro grupos: a) Deteriorado, b) sin cambio fiable, c) Mejorado pero no recuperado, y d) Recuperado.

Este conjunto de algoritmos son los especificados por Hagerman y Arrindell para la estimación de las puntuaciones individuales, existiendo todo un desarrollo de cálculo para el caso del nivel grupal que permite, análogamente, la obtención de dos índices: el RC_{Grup} y el CS_{Grup} (Hagerman y Arrindell (1999a, 1999b).

4. El ICF_{HLM}: Propuesta de Speer y Greenbaun (1995)

Esta propuesta, planteada inicialmente por Speer y Greenbaun (1995), constituye una variante metodológica muy distinta a la planteada en los supuestos previos. El método se fundamenta en modelos de curva de crecimiento, basados en modelos jerárquicos lineales, donde se precisan al menos tres medidas temporales de cada participante. La determinación de los cambios de un individuo se estima mediante un procedimiento bayesiano y permite hacer estimaciones incluso en casos donde existen datos ausentes al tener en cuenta para dicha estimación toda la información contenida (individual y grupal). Los detalles de estos cálculos exceden los propósitos de esta tesis y en la práctica son estimados mediante programas específicos de cálculo. Los defensores de este método consideran que permiten una mayor flexibilidad que los ICF clásicos de la información ofrecida resulta más valiosa.

5. Evidencias sobre la comparación de los ICF

Los cinco métodos descritos para la valoración del cambio clínicamente significativos han sido utilizados de forma variable por distintos autores. En algunos estudios, se han puesto a prue-

ba simultáneamente tratando de diferenciar sus potencialidades y determinar cuál sería el más óptimo en su utilización.

El primer intento de examinar las características de estas propuestas de ICF se debe a uno de los grupos mencionados (Speer y Greenbaum, 1995). Estos autores comparan los métodos de Jacobson y Truax (JT), Hsu (GLN), Speer (EN), Hagerman y Arrindell (HA) y el suyo propio planteado en ese trabajo (HLM). En tanto que estos autores configuran el artículo como la presentación de un nuevo método alternativo a los ya existentes, sus conclusiones derivan en la recomendación del método HLM como el más específico y altamente sensible para clasificar los sujetos como 'recuperados' o 'mejorados sin recuperación'. No obstante, el método HLM no resulta tan recomendable para detectar los casos que presentan un 'deterioro' de la respuesta de salud (Speer y Greenbaum, 1995).

En un trabajo posterior, McGlinchey, Atkins y Jacobson (2002) replican la estructura de estudio de Speer y Greenbaum con una muestra de sujetos con depresión mayor sometidos a psicoterapia. Específicamente, este grupo valoró los métodos JT, GLN, EN y HA; no así el HLM. Los resultados muestran como los diferentes métodos difieren en su capacidad para clasificar los casos cuyos síntomas depresivos remiten, mejoran o empeoran.

No obstante, estos autores concluyen que a pesar de que los métodos difieren en su capacidad para detectar los cambios, las diferencias no son relevantes mostrando todos ellos su potencial para predecir las recaídas dos años después de concluirse la terapia. En definitiva, McGlinchey y colaboradores (2002) concluyen que, a pasar de las mejoras que supuestamente realizaron los métodos GLN, EN y HA a la evaluación del cambio clínicamente significativo, no hubo pruebas de que cualquiera de estos métodos fuera mejor que el propuesto originariamente por Jacobson.

En un estudio de simulación, Atkins, Bedics, McGlinchey y Beauchaine (2005) comparan de nuevo los cuatro tipos de métodos (JT, GLM, EN y HA). Los resultados mostraron que JT y GLN eran casi idénticos, mientras que HA fue significativamente más conservador. El método EN era el que presentaba un mayor número de casos clasificados como 'recuperados' o 'deteriorados', y menos como 'no cambio'. De forma coincidente con el estudio previo, los resultados no mostraban evidencias claras que desaconsejasen el uso del método de Jacobson y Truax, si bien los autores insisten en la conveniencia de seguir investigando sobre las ventajas diferenciales de cada uno de estos métodos en la valoración de los cambios durante el tratamiento.

Ejemplo de la aplicación de las diferentes alternativas de cálculo del ICF

En la Tabla 2 se presenta un resumen de los algoritmos de cálculo de los Índices de Cambio Fiable correspondientes a cada una de los métodos propuestos. Asimismo, se acompaña de un ejemplo de un caso particular para ejemplificar el procedimiento de cómputo y la interpretación del valor de los índices. En la parte inferior de la Tabla se presentan los datos necesarios para la estimación; por una parte, los que corresponden al grupo, al cual puede pertenecer el sujeto de valoración, o bien pueden ser obtenidos a partir de muestras normativas si las hubiera; por otra parte, los que corresponden al instrumento algunos de los cuales han sido calculados mediante las fórmulas 2, 3, 7 y 8; y, finalmente, los datos obtenidos por el sujeto respecto al cual se pretende valorar el cambio producido.

La persona de muestro ejemplo ha pasado de una puntuación inicial de 43,9 puntos en la escala de percepción general de salud del SF-36 a una puntuación tras el tratamiento de 67,6 puntos, lo cual supone un incremento de 23,7 puntos en la dirección de un cambio funcional. Si atendiéramos a la magnitud del efecto logrado (Iraurgi, 2009b) observamos que el valor d de Cohen es de 0,89 lo cual supone alcanzar un tamaño del efecto alto, que es interpretado por Testa (1987) como un cambio clínico sustancial. Como se ha propuesto en un epígrafe anterior las puntuaciones individuales de este sujeto correspondían con los valores medios de dos grupos -uno funcional y otro disfuncional- cuya prueba de significación estadística resultaba significativa ($t= 23,9$; $p<0,001$). Ambos datos son interpretables como interesantes respecto a la significación del cambio, pero ¿resulta ser un cambio clínicamente fiable?

Tabla 2. Alternativas de cálculo del Índice de Cambio Fiable (ICF) y aplicación práctica a partir de los datos de un caso

Método	Algoritmo	Ejemplo
ICF_{JT} JT - Jacobson y Truax (1991)	$\frac{(X_{post} - X_{pre})}{\sqrt{2 \cdot (DE_{pre} \cdot \sqrt{1 - r_{xx}})}}$	$\frac{(67,6 - 43,9)}{\sqrt{2 \cdot (26,6 \cdot \sqrt{1 - 0,83})}} = 1,539$
Punto de Corte	$PC = \frac{(X_{post} \cdot DE_{pre}) + (X_{pre} \cdot DE_{post})}{DE_{pre} + DE_{post}}$	$\frac{(67,3 \cdot 26,6) + (43,9 \cdot 19,6)}{26,6 + 19,6} = 57,37$
ICF_{GLN} Gulliksen-Lord-Novick (GLN), (Hsu, 1998, 1999)	$\frac{(X_{post} - \bar{X}_{pre}) - r_{xx}(X_{pre} - \bar{X}_{pre})}{DE_{pre} \cdot \sqrt{1 - r_{xx}^2}}$	$\frac{(67,6 - 58,3) - 0,83(43,9 - 58,3)}{26,6 \cdot \sqrt{1 - 0,83^2}} = 1,43$
ICF_{EN} Edwards-Nunnally (EN) - (Speer, 1992)	$\left[r_{xx} \cdot (X_{pre} - \bar{X}_{pre}) + \bar{X}_{pre}\right] \pm 2 \cdot DE_{pre} \cdot \sqrt{1 - r_{xx}}$	$\left[0,83 \cdot (43,9 - 58,3)\right] + 58,3 \pm 2 \cdot 26,6\sqrt{1 - 0,83} =$ $= 24,4 \; a \; 68,3$
RC_{INDIV} y CS_{INDIV} HA - Hageman y Arrindell (1999)	$RC_{INDIV} = \frac{(X_{post} - X_{pre}) \, r_{dd} + (\bar{X}_{post} - \bar{X}_{pre})(1 - r_{dd})}{\left(\sqrt{r_{dd}}\right)\left(\sqrt{2 \cdot EEM^2}\right)}$	$\frac{(67,6 - 43,9) \cdot 0,56 + (67,3 - 58,3) \cdot (1 - 0,56)}{\left(\sqrt{0,56}\right)\left(\sqrt{2 \cdot 10,96^2}\right)} = 1,48$
	$TRC = \frac{\left(DE_{norm}\sqrt{r_{xx(norm)}}\right)\bar{X}_{pre} + \left(DE_{pre}\sqrt{r_{xx(pre)}}\right)\bar{X}_{norm}}{DE_{norm}\sqrt{r_{xx(norm)}} + DE_{pre}\sqrt{r_{xx(pre)}}}$	$\frac{\left(19,6\sqrt{0,83}\right)58,3 + \left(26,6\sqrt{0,83}\right)67,3}{19,6\sqrt{0,83} + 26,6\sqrt{0,83}} = 63,49$
	$CS_{INDIV} = \frac{\bar{X}_{post} + (X_{post} - \bar{X}_{post})\, r_{xx-post} - TRC}{\sqrt{r_{xx-post}} \cdot EEM}$	$\frac{67,3 + (67,6 - 67,3) \cdot 0,69 - 63,49}{\sqrt{0,69} \cdot 10,96} = 0,441$

Puntuaciones del Grupo	
Media pre-tratamiento de la muestra	$\bar{X}_{pre} = 58,3$
Desviación Estándar pre-tratamiento de la muestra	$DE_{pre} = 26,6$
Media post-tratamiento de la muestra / Media normativa	$\bar{X}_{post} = \bar{X}_{norm} = 67,3$
Desviación Estándar post-tratamiento de la muestra / DE normativa	$DE_{post} = DE_{norm} = 19,6$

Puntuaciones del Individuo	
Puntuación individual pre-tratamiento	$X_{pre} = 43,9$
Puntuación individual post-tratamiento	$X_{post} = 67,6$
Diferencia Post-Pre Tto	$Dif = 23,7$

Puntuaciones del Instrumento	
Fiabilidad población normativa	$r_{xx} = 0,83$
Fiabilidad test-retest	$r_{pre\text{-}post} = 0,64$
Fiabilidad pre-tratamiento	$r_{xx(pre)} = 0,83$
Fiabilidad post-tratamiento	$r_{xx(post)} = 0,69$
Fiabilidad de las diferencias	$r_{dd} = 0,56$
Error Estándar de Medida	$EEM = 10,96$

Acudiremos, en primer lugar, a los resultados obtenidos a partir del método de Jacobson y Truax (1991): un punto de corte (PC) fijado en 57,37 puntos y un ICF_{JT}= 1,54. La puntuación obtenida en el post-test por nuestro sujeto ha sido de 67,6 lo que supone localizarse 10,23 puntos por encima del PC y, por tanto, se halla dentro del área de funcionalidad. Ahora bien, el valor logrado por el ICF_{JT} no alcanza el valor mínimo de 1,96 representativo de un cambio fiable, lo cual implica que el sujeto sería clasificado como 'No cambio', aunque sí le sitúa en una orientación hacia la funcionalidad ('Mejoría sin cambio'). Dos razones pueden dar respuesta a esta falta de precisión: la fiabilidad del instrumento o bien la amplitud de la dispersión de las puntuaciones del grupo en el instrumento. Un coeficiente de fiabilidad de 0,83 puede ser considerado como adecuado, si bien para las estimaciones individuales suele requerirse coeficientes con valores superiores a 0,90 (Nunnally y Beristaín, 1995). Por ello, en este caso, estaríamos más acordes en aceptar que la limitación en la precisión se debe a la alta dispersión, lo cual podría redundar en una falta de sensibilidad al cambio por parte del instrumento.

El método de Hsu (1998, 1999), el cual introduce una corrección para controlar el efecto de regresión a la media, alcanza un valor de ICF_{GLN}= 1,43. Como puede apreciarse resulta ligeramente inferior al valor del ICF_{JT} (1,43‹····1,54), lo que índica que ese método es más conservador que el preferente ya que se aleja en mayor medida del valor de significación de cambio fiable (1,96). Como sabemos, el valor de probabilidad asociado a 1,96 es de 0,025 (0,025 ? 2= 0,05); el asociado al valor 1,53 obtenido con el método de Jacobson y Truax es de 0,063 (?2= 0,126) y el asociado al valor logrado por el método de Hsu es de 0,076 (?2= 0,152). Al igual que con el método previo, en este caso también conduce a considerar la nueva situación del sujeto como 'No cambio fiable'.

El tercer método, propuesto por Speer (1992) consiste en calcular el intervalo de confianza del 95% de la puntuación lograda por el sujeto en el pre-test. El supuesto de cambio radica en considerar que si la puntuación del sujeto en el post-test excede los límites de disco intervalo, entonces, se considerará que se ha producido un cambio significativo, que será clínicamente relevante si el cambio se produce en la dirección de las puntuaciones de funcionalidad. De este modo, el IC-95% del pre-test se sitúa entre los valores 24,4 y 68,3; la puntuación alcanzada en el post-test se sitúa dentro de dicho intervalo (67,3‹····68,3), por lo que es interpretada como una puntuación que podría hallarse entre las fluctuaciones de la puntuación pre-test debidas al azar. De modo análogo a los métodos previos, éste también clasificaría a nuestro sujeto como 'No cambio'.

Por último, el método de Hageman y Arrindell (1999) nos ofrece tres índices diferentes a partir de su propuesta. El índice de cambio individual (RC_{INDIV}) alcanza un valor de 1,48 que, aunque un poco más cercano al alcanzado por el ICF_{JT}, sigue siendo un valor conservador indicativo de que nos se ha producido un cambio fiable. El segundo indicador que nos ofrece este método es la estimación del verdadero valor del punto de corte que es estimado no sólo como un promedio de las puntuaciones medias en función de la dispersión de las muestras, sino también en función de la precisión del instrumento. El punto de corte obtenido por el método de JT ha sido de 57,37 y en el caso del método de HA es de 63,49 lo cual hace este PC más riguroso que el de JT para poderse alcanzar la condición de funcionalidad, para poder lograr la significación clínica. Por su parte, el valor del cambio significativo individual (CS_{INDIV}) resulta ser de 0,441 de modo que lleva a la conclusión, una vez más, de que el sujeto no logra presentar un cambio suficientemente amplio como para poder ser considerado como un cambio fiable.

Como vemos, los métodos alternativos al originariamente planteado por Jacobson y Truax tienden a ser más restrictivos y conservadores, de modo que puntuaciones de cambio que pudieran ser consideradas de forma intuitiva como relevantes, resultan ser poco concluyentes.

En cualquier caso, todos estos métodos constituyen una importante herramienta de clasificación y decisión respecto a la valoración del alcance de los efectos de los tratamientos, y precisan de mayor utilización en los ámbitos clínicos y de mayor desarrollo por parte de investigadores y estadísticos.

Conclusiones

Más allá de los métodos de contraste de las hipótesis estadísticas y del logro de los valores de significación estadística (Iraurgi y Markez, 2009; Iraurgi, 2009a), en la valoración de los resultados terapéuticos se hace pertinente y necesario obtener el valor de los efectos alcanzados (Iraurgi, 2009b) en tanto que éstos representan una mayor aproximación a la relevancia clínica obtenida. La síntesis de ambos procedimientos se alcanza mediante las técnicas de estimación del cambio clínicamente significativo, llamadas índices de cambio fiable (ICF, o RCI en inglés -Reliable Change Index-), las cuales añaden información más precisa sobre los cambio terapéuticos observados. Además de informar sobre el cambio alcanzado por un grupo intervenido, ayudan a determinar el cambio en el nivel individual: si un sujeto dado ha variado de forma relevante su situación de salud tras el tratamiento y si ese cambio tiene implicaciones en el reajuste del diagnóstico.

Añadir información sobre la significación clínica en la investigación de resultados terapéuticos sería del todo deseable. Deseable para los investigadores, que obtendrían información más certera sobre el alcance de los tratamientos; y deseable también, y sobre todo, para los clínicos que les ayudaría en la toma de decisiones sobre los procesos diagnósticos y terapéuticos de las personas atendidas. Desde nuestro punto de vista, la utilización de las metodologías de valoración del cambio clínicamente significativo debieran ser consideradas como una forma prometedora de interpretar los datos en la investigación de los resultados de salud.

Correspondencia:
Ioseba Iraurgi
DeustoSalud. Universidad de Deusto • Avda. de las Universidades, 24 • 48007 Bilbao.
correo electrónico: ioseba.iraurgi@deusto.es

Referencias bibliográficas

Alonso, J., Prieto, L. y Antó, J.M. (1995). La versión española del SF-36 Health Survey (Cuestionario de Salud SF-36): un instrumento para la medida de los resultados clínicos. *Medicina Clínica (Barcelona), 104*, 771-776.

Alonso, J., Regidor, E., Barrio, G., Prieto, L., Rodríguez, C. y de-la-Fuente, L. (1998). Valores poblacionales de referencia de la versión española del Cuestionario de Salud SF-36. *Medicina Clínica (Barcelona), 111*, 410-416.

Anaitua, C., Quintana, J.M. (1999). Valores poblacionales del índice de salud SF-36 en el País Vasco: importancia y aplicación en la práctica clínica. *Osasunkaria,* 17, 10-17.

Anastasi, A. y Urbina, S. (1997). *Psychological testing*. 7ª Ed. Upper Saddle River, New Jersey; Prentice Hall.

Atkins, D.C., Bedics, J.D., McGlinchey, J.B. y Beauchaine, T.P. (2005). Assessing Clinical Significance: Does It Matter Which Method We Use? *Journal of Consulting and Clinical Psychology,* 73, 5, 982-989.

Christensen, L. y Mendoza, J.L. (1986). A method of assessing change in a single subject: An alteration of the RC Index. *Behavior Therapy*, 17, 305-308.

Echeburua, E y Corral, P. (2001). Eficacia de las terapias psicológicas: de la investigación a la práctica clínica. *Revista Internacional de Psicología Clínica y de la Salud*, 1, 1, 181-204.

Hageman, W.J. y Arrindell, W.A. (1999a). Establishing clinically significant change: increment of precision and the distinction between individual and group level of analysis. *Behaviour Research and Therapy,* 37, 1169-1193.

Hageman, W.J. y Arrindell, W.A. (1999b). Clinically significant and practical! Enhancing precision does make a difference. Reply to McGlinchey and Jacobson, Hsu, and Speer. *Behavior Research and Therapy*, 37, 1219-1233.

Hsu, L.M. (1989). Reliable change in psychotherapy: Taking into account regression toward the mean. *Behavioural Assessment,* 11, 459-467.

Hsu, L.M. (1995). Regression toward the mean associated with measurement error and identification of improvement and deterioration in psychotherapy. *Journal of Consulting and Clinical Psychology,* 63, 141-144.

Hsu, L.M. (1999). A comparison of three methods of identifying reliable and clinically significant client changes: commentary on Hageman and Arrindell. *Behaviour Research and Therapy, 37*, 1195-1202.

Iraurgi, I. (2000). Cuestiones metodológicas en la evaluación de programas terapéuticos. *Trastornos Adictivos,* 2, 2, 99-113.

Iraurgi, I. (2009a). Evaluación de resultados clínicos I: Entre la significación estadística y la relevancia clínica. *Norte de Salud Mental, 33*, 94-108

Iraurgi, I. (2009b). Evaluación de resultados clínicos II: Las medidas de la significación clínica o los tamaños del efecto. *Norte de Salud Mental, 34*, 94-110.

Iraurgi I. (2010). Reducción de Daños y Riesgos: Lecciones aprendidas y retos futuros. En T. Laespada y I. Iraurgi (Eds). *El Modelo de la Reducción de Daños: Lo aprendido de la heroína*. Bilbao; Publicaciones de la Univ. de Deusto.

Iraurgi, I. y Markez, I. (2009). La investigación, la estadística y la tiranía de los valores-p. *Norte de Salud Mental, 33*, 6-8.

Iraurgi, I., Poó, M. y Markez, I. (2004). Valoración del índice de salud SF-36 en usuarios de Programas de Metadona. Valores de referencia para la Comunidad Autónoma Vasca. *Revista Española de Salud Pública*, 78, 609-621.

Iraurgi, I., Trujols J., Lozano-Rojas, O. y González-Saiz, F. (2010). Valoración del cambio clínicamente significativo en el tratamiento de trastornos adictivos: Una aplicación con la Escala de Severidad de la Dependencia (SDS). *Trastornos Adictivos*, en valoración.

Jacobson, N.S., Follette, W.C. y Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*, 336-352.

Jacobson, N.S., Follette, W.C. y Revenstorf, D. (1986). Toward a standard definition of clinically significant change. *Behavior Therapy, 17*, 308-311.

Jacobson, N.S. y Revenstorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems, and new developments. *Behavioral Assessment, 10*, 133-145.

Jacobson, N.S., Roberts, L.J., Berns, S.B. y McGlinchey, J.B. (1999). Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. *Journal of Consulting and Clinical Psichology,* 67, 300-307.

Jacobson, N.S. y Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psichology*, 59, 12-19.

Kazdin, A.E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification, 1,* 427-452.

Kazdin, A.E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 332-339.

Kazdin, A. E., y Wilson, G. T. (1978). *Evaluation of behavior therapy: Issues, evidence, and research strategies. C*ambridge, MA: Ballinger.

Kendall, P.C., Butcher, J.N, y Holmbeck, G.N. (1999). *Handbook of research methods in clinical psychology*. New York; Wiley.

Kendall, P.C., Marrs-García, A., Nath, S.R. y Sheldrick, R.C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psichology,* 67, 3, 285-299.

Kupler, D.J. (1991). Long-term of treatment of depression. *Journal of Clinical Psychiatry*, 52, (suppl.5), 28-34.

Lord, F.M. y Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maassen, G.H. (2004). The standard error in the Jacobson and Truax reliable change index: The classic approach to the assessment of reliable change. *Journal of International Neuropsychological Society*, 10, 6, 888-893.

Mavissakalian, M. (1986). Clinically significant improvement in agoraphobia research. *Behaviour Research and Therapy, 24,* 369-370.

McGlinchey, J.B., Atkins, D.C. y Jacobson, N.S. (2002). Clinical Significance Methods: Which One to Use

and How Useful Are They? *Behavior Therapy*, 33, 529-550.

Nietzel, M.T, y Trull, T.J. (1988). Meta-analytic approaches to social comparisons: A method for measuring clinical significance. *Behavioral Assessment, 10,* 146-159.

Páez, D., Echeburua, E. y Borda, M. (1993). Evaluación de La eficacia de los tratamientos psicológicos: una propuesta metodológica. *Revista de Psicología General y Aplicada*, 46, 2, 187-198.

Speer, D.C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology, 60*, 402-408.

Speer, D.C. y Greenbaum, P. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology*, 63, 1044-1048.

Testa, M. (1987). Interpreting quality of life clinical trial data for use in the clinical practice of antihypertensive therapy. *Journal of Hypertension*, 5, S9-S13.

Ware, J.E. y Kosinski, M. (1997). SF-36 Physical and Mental Health summary scales: A Manual for Users of Version 1. 2nd ed. Lincoln, RI; QualityMetric Inc.

Ware, J.E., Kosinski, M., Bayliss, M.S., et al. (1995). Comparison of methods for scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. *Medical Care,* 33, AS264-AS279.

Ware, J.E. y Sherbourne, C.D. (1992). The MOS 36-item short form health survey (SF-36):1 Conceptual framework and item selection. *Medical Care*, 30, 6, 473-483.

Wolf, M.M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis. 11,* 203-214.

• Recibido: 28-11-2009.