anales de psicología, 2014, vol. 30, nº 2 (mayo), 756-771
http://dx.doi.org/10.6018/analesps.30.2.166911

ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

Validez Estructurada para una investigación cuasi-experimental de calidad. Se cumplen 50 años de la presentación en sociedad de los diseños cuasi-experimentales

Paula Fernández-García[1]*, Guillermo Vallejo-Seco[1], Pablo E. Livacic-Rojas[2] y Ellián Tuero-Herrero[1]

[1] Universidad de Oviedo, España
[2] Universidad de Santiago de Chile, Chile

Resumen: Investigación cuasi-experimental es aquella que tiene como objetivo poner a prueba una hipótesis causal manipulando (al menos) una variable independiente donde por razones logísticas o éticas no se puede asignar las unidades de investigación aleatoriamente a los grupos. Debido a que muchas decisiones a nivel social se toman en base al resultado de investigaciones con estas características, es imperativo que tengan una planificación exquisita de la aplicación del tratamiento, del control en el proceso de la investigación y del análisis de los datos. El pasado año 2013 los diseños cuasi-experimentales cumplieron 50 años, y este trabajo es un homenaje a Campbell y a todos los investigadores que día a día aportan ideas para mejorar el método cuasi-experimental en alguno de sus aspectos. De la mano de una revisión de las investigaciones cuasi-experimentales publicadas en un período de 11 años en tres revistas de Psicología destacamos algunos aspectos que se refieren al cuidado del método. Finalizamos el trabajo proponiendo el concepto de Validez Estructurada, que en resumen, es el hilo conductor que debe seguir la realización de toda investigación para poner a prueba con garantía las hipótesis que responden a los objetivos que en ella se plantean, concretamente, en las investigaciones cuasi-experimentales.
Palabras Clave: Calidad; planificación; evidencia científica; validez estructurada; revisión teórica y revisión sistemática no cuantitativa.

Title: Structured Validity for a quasi-experimental research of quality. They are fulfilled 50 years of the presentation in company of the quasi-experimental designs.

Abstract: Quasi-experimental investigation is that one that has as aim test a causal hypothesis manipulating (at least) an independent variable where for logistic or ethical reasons it is not possible to assign the units of investigation at random to the groups. Due to the fact that many decisions at the social level take on the basis of the result of investigations with these characteristics, it is imperative that have an exquisite planning of the application of the treatment, of the control in the process of the investigation and of the analysis of the data. Last year 2013 the quasi-experimental designs expired 50 years, and this work in an honoring to Campbell and to all the investigators who day after day contribute ideas to improve the quasi-experimental method in someone of his aspects. From the hand of a review of the quasi-experimental investigations published in a period of 11 years in three journals of psychology we distinguish some aspects that refer to the care of the method. We finished work by proposing the concept of Structured Validity, which in summary, is the thread that must follow all research to test with guarantee the hypothesis that respond to the objectives it raised, in particular, in quasi-experimental investigations.
Key words: Quality; planning, scientific evidence; structured validity; theoretical review and no quantitative systematic review.

Introducción

Caía la hoja en Central Park aquel 17 de octubre de 1956 cuando Robert James Fischer disputaba una partida de ajedrez frente a Donald Byrne. Negras frente a blancas que, hasta el movimiento 16 en que las blancas dan jaque a la dama de Bobby, la partida se tornaba sólo interesante. Bastaron dos movimientos más y definitivamente Byrne se decide por la dama negra. Bobby ha sacrificado a su reina, pero para entonces ya tenía la torre estratégicamente situada y también un caballo y un alfil. Desde ese momento, lentamente, como saboreando a su presa, las jugadas poliédricas de Bobby logran hacerse con un control absoluto del tablero, y en el movimiento 41 también de la partida. A quienes somos profanos en la materia nos cuesta comprender que aquellos movimientos, desde el 16 y hasta el 18, no fuesen movimientos improvisados de un adolescente de 13 años. A fuerza de perder, aprender y volver a perder entendemos de qué modo Bobby fue capaz de conseguir inmovilizar el rey blanco en una esquina, y en la esquina opuesta, movimiento a movimiento, restar todo el poder a la dama del mismo color a quien, por aquel entonces, era uno de los ajedrecistas más respetados de los Estados Unidos.

Al tiempo, Donald Campbell trabajaba en el Departamento de Psicología de la Universidad de Northwestern desde 1953 (en la que estuvo 26 años). Campbell se había doctorado en Berkeley con la fortuna de haber tenido dos grandes maestros que por entonces estaban trabajando la idea *representative design*, Egon Brunswik y Edward Tolman. Cuando Bobby Fischer jugaba la que fue bautizada como "la Inmortal del Siglo XX" se publicó el libro *Perception and the representative design of psychological experiments* (Brunswik, 1956). El concepto de diseño representativo surge como contrapunto al diseño clásico realizado en el laboratorio. Cuando un experimento se lleva a cabo en el ámbito ecológico en el que se manifiesta la conducta observada, entonces el diseño es representativo. No cabe duda, la investigación realizada en ambientes naturales entraña una enorme dificultad debido a que junto a las variables que se pretende estudiar pueden aparecer variables ajenas que es muy difícil, si no imposible, aislar o mantener constantes, y que provocan ambientes inciertos e inestables estableciendo un complejo entramado causal. Esta idea fue recogida por Campbell y al año siguiente, en 1957, publica en la revista *Psychological Bulletin* el trabajo titulado *Factors relevant to the validity of experiments in social settings*. Este artículo bien pudiera considerarse el punto de partida de la obra científica extraordinaria de un virtuoso del método, que, desde el capítulo 5 de la obra de Gage (1963) y junto con Julian Stanley (Campbel & Stanley, 1963), por primera vez ofrecen al mundo los diseños cuasi-experimentales comprensivamente explicados. Obras posteriores cuyo denominador común es la figura de Campbell (Cook & Campbell, 1979; Cook, Campbell & Peracchio, 1990; Shadish, Cook & Campbell, 2002, int.al.) abundarían

*** Dirección para correspondencia [Correspondence address]:**
Paula Fernández García. Facultad de Psicología. Plaza de Feijóo, s/n. 33003 Oviedo (España). E-mail: paula@uniovi.es

en lo que podríamos denominar *Mandamientos* que, a modo de brújula metodológica, nos ilustran y guían cómo realizar correctamente una investigación en las ciencias sociales.

Qué es un diseño cuasi-experimental: a vista de pájaro la trayectoria de los diseños cuasi-experimentales es obligada una definición. El diseño cuasi-experimental es un plan de trabajo con el que se pretende estudiar el impacto de los tratamientos y/o los procesos de cambio en situaciones donde los sujetos o unidades de observación no han sido asignados de acuerdo con un criterio aleatorio (ver Arnau, 1995). A veces incluso, la aplicación del tratamiento no la ejerce directamente el investigador, viene impuesta por una organización, por mandato gubernamental, etc., y si este es el caso, tampoco se tiene control sobre las circunstancias que rodean a la aplicación (Campbell & Stanley, 1963, p.34; Shadish et al*.,* 2002, p.14), en este caso con frecuencia se los denomina experimentos naturales o experimentos de campo (Kunstmann y Merino, 2008; Trochim, 2001, int.al.).

Cambiando el paso. Por su capacidad de movimiento, la dama es la pieza de mayor valor en el juego del ajedrez. No se conoce partida alguna disputada por jugadores de alto nivel, ni de medio o bajo nivel tampoco, que hayan perdido la dama por un descuido y hayan conseguido remontar la partida (J. Cordero Fernández, maestro de ajedrez y responsable de la página Web "Ajedrez de Ataque", comunicación personal, 14 de mayo de 2012). Tampoco sucedió en "la Inmortal del Siglo XX". Bobby Fischer había previsto y calculado cómo impulsar el ataque asumiendo al mismo tiempo una defensa difícil cuando sacrificó a la reina. Haciendo gala de una técnica excelente, granada de recursos estudiados para ser utilizados en situaciones complicadas, ganó la partida sumando su talento e imaginación. En este punto, podríamos decir que la dama es al ajedrez lo que la aleatorización al método científico. Ítem más, tan indiscutible es que la aleatorización y el control sobre la manipulación son recursos inestimables para poner a prueba las hipótesis de calado causal, como que las investigaciones realizadas en tan particulares circunstancias como las descritas al principio de este párrafo (frecuentemente con finalidad aplicada) son siempre previsibles por el mero hecho de que se conoce el cómo y cuándo de la sucesión temporal.

Antes de aplicar el tratamiento: la investigación cuasi-experimental es una herramienta poderosa para inferir relaciones causales, pero es una herramienta condicionada que requiere el mimo de un artesano. En 1979 Cook y Campbell presentan la primera obra dedicada íntegramente a los diseños cuasi-experimentales en la que se aprecia la profunda influencia de las ideas sobre causalidad de Kenny, y el feedback entre estos autores. No es casualidad, por tanto, que en ese mismo año Kenny publicase el magnífico trabajo *Correlation and causality* donde destaca los tres supuestos necesarios en que fundamentar su existencia, a saber, las variables de tratamiento y de resultado deben covariar, la relación entre ellas no debe ser espuria y la causa debe preceder al efecto. Estos tres supuestos son ávidamente buscados en la investigación cuasi-experimental. Así las cosas, antes de aplicar el tratamiento el investigador debe "profundizar" en el conocimiento de las condiciones particulares donde se va a realizar la intervención, de las complicaciones que puede conllevar, de los posibles efectos no deseados, de las personas a las que va dirigida la investigación, debe conocer si las variables dependientes pueden serlo también de otras variables alternativas a la variable independiente, etc., todo ello con la finalidad de prever y anticiparse a las adversidades y dificultades capaces de dar jaque a la prueba de su hipótesis. Esta tarea compleja es lo que Shadish et al. (2002, p.484) denominan *The Centrality of Fuzzy Plausibility*. Esto es, es necesario que el investigador tenga un juicio bien formado acerca de si las amenazas a la validez interna de su estudio son relevantes y las posibilidades que tiene de eliminarlas o de reducirlas (y hasta qué punto). Dicho de otro modo, si éstas sólo pueden ser controladas o ajustadas parcialmente deberá considerar si el sesgo que podrían producir pudiera ser mayor que el tamaño del efecto que espera encontrar. En definitiva, el investigador debe ponerse en guardia.

Este juicio que denominan *de plausibilidad* (*op.cit*) probablemente desemboque en la necesidad de tomar nota de variables del ambiente (en el que se realiza el experimento y ambiente adyacente también, características del responsable de la aplicación del tratamiento y del registro de los datos, condiciones en las que se lleva a cabo, cercanía entre los grupos control y experimental, etc.) y variables de sujeto (físicas y psicológicas, tanto actuales como históricas) que puedan estar relacionadas con la variable dependiente (y/o independiente) y ejercer un efecto activo no deseado (sesgo) moderador, o supresor, o mediador, incluso un efecto de confundido, siendo auténticas explicaciones alternativas del resultado hallado. La única posibilidad de conocer si es así o no, y de controlar el sesgo que son capaces de producir, es tomar nota de ellas y registrarlas antes de aplicar el tratamiento, "Better to have imprecise attention to plausibility than to have no attention at all paid to many important threats just because they cannot be well measured" (Shadish et al., 2002, p. 484). Esto tiene una importancia capital.

El tiempo y las medidas: estas investigaciones llevan tiempo, y como consecuencia, tanto los sujetos como las condiciones naturales donde es implementado el tratamiento y recogidos los datos pueden cambiar y evolucionar. Si esas variables no son constantes en su estado se debe considerar su variabilidad, lo que obliga, de una parte, a estudiar de qué depende, y de otra, a elegir cuándo y de qué modo es necesario tomar los registros para tenerla en cuenta.

Inadvertidamente, en los tres párrafos anteriores ya hemos sumado el registro de muchas medidas de variables de control y del efecto del tratamiento, así pues, es imperativo no descuidar la posible pérdida accidental de datos o de sujetos en determinadas ocasiones (puntuaciones faltantes, registros inservibles, errores del investigador, etc.), o peor aún, el abandono de los sujetos (por falta de adherencia al tratamiento, cambio de domicilio o condición social, incoherencia en la administración del tratamiento, etc.). El diagnóstico y conocimiento de la causa es determinante para poner solu-

ción a este problema que desafortunadamente es tan habitual (ver West et al., 2008).

Los grupos de control: no sólo de múltiples medidas de múltiples variables se alimentan los diseños cuasi-experimentales, también necesitan grupos de control bien formados (mejor múltiples) que sean lo más homogéneos posible a los grupos experimentales con ánimo de tener capacidad para controlar el sesgo de selección, y otros, capaces de confundir la acción del tratamiento. Aunque se debate las ventajas y desventajas de utilizar grupos de control activos o pasivos (e.g., Datta, 2007; Donaldson & Chistie, 2005; Cook, 2006), parece existir consenso acerca de que los primeros son mejores. Además, siempre que sea posible se debe evitar la autoselección (Shadish et al., 2002), la participación voluntaria, la selección arbitraria, o que los pacientes con peor pronóstico (Kunz & Oxman, 1998) sean quienes formen el grupo de control. Precisamente, y a colación de lo comentado en el párrafo anterior, son las investigaciones donde se presta escasa atención a la formación de los grupos (control y tratamiento) las que sufren mayor desgaste de muestra (Heisman & Shadish, 1996; Shadish & Ragsdale, l996).

Sobre el diseño: hemos decidido qué variables controlar, cuántos registros efectuar y qué grupos van a participar en la investigación, pero aún resta determinar de qué modo vamos a aplicar el tratamiento y recoger los datos, es decir, qué diseño de investigación elegimos utilizar. Campbell y colaboradores han destacado que los distintos diseños cuasi-experimentales varían entre sí con respecto a la transparencia y la capacidad para poner a prueba la hipótesis causal. De hecho, han ordenado deliberadamente los capítulos de sus libros para reflejar el aumento de la potencia inferencial que supone pasar de diseños sin una medida previa o sin un grupo de control, diseños pre-experimentales, a aquellos con ambas divisas, los diseños cuasi-experimentales. Sucintamente, estos diseños se pueden organizar en dos grandes bloques, transversales y longitudinales (ver Arnau, 1995). Los primeros se clasifican a su vez en dos tipos, diseños de grupo control no equivalente y diseños de discontinuidad en la regresión (ver Ato, 1995). De estos, los últimos tienen mayor potencia inferencial porque controlan la variable de selección utilizándola como criterio de formación de los grupos. Los primeros no deben olvidar que una de las medidas pretratamiento debe referirse a la posible variable de selección para así ser eficazmente controlada estadísticamente. Los diseños longitudinales ocupan una gran variedad de diseños de series temporales (ver Vallejo, 1995), y, como Kenny argumentaba en su obra de 1979, en estos es más fácil verificar los tres supuestos de causalidad. Volviendo a los capítulos de sus libros. Dentro de cada uno de ellos también ilustran cómo la inferencia se puede mejorar mediante la adición de elementos de diseño (añadiendo más puntos de observación pre-test, aplicando y retirando el tratamiento en varias ocasiones, registrando variables dependientes no equivalentes, utilizando más grupos tanto de control como de tratamiento y poniendo esmero en su elección y/o formación, etc.). La capacidad para poner a prueba las hipótesis del investigador y descartar hipótesis alternativas es distinta en función de todas estas características.

Sobre el análisis de los datos: Mimando la planificación de la investigación en los matices anteriormente comentados ya solo resta analizar los datos. Son varias las alternativas de análisis que existen, pero no se trata de elegir la más cómoda, sino la más oportuna en función de nuestras hipótesis, del modelo que estimamos explica los resultados, y de las características de nuestros datos. Más adelante nos extenderemos en este asunto, hasta entonces, sólo un apunte. Campbell y colaboradores siempre han defendido que los ajustes estadísticos sólo se deben utilizar cuando se han extremado todos los controles anteriormente comentados con la finalidad de reducir todo lo posible la no equivalencia entre los grupos "In this book, we have advocated that statistical adjustments for group nonequivalence are best used after design controls have already been used to the maximum in order to reduce nonequivalence to a mínimum" (Shadish, et al*.,* 2002, p.503).

Evolución del diseño cuasi-experimental: nada hemos leído de Campbell que pudiera presagiar la idea de que estas investigaciones fuesen otra cosa que investigaciones primarias. Ahora, cuando se cumple medio siglo de historia, existe una segunda derivada, investigaciones cuasi-experimentales realizadas mediante técnicas de minería de datos. Esta idea ha sido desarrollada en la Universidad de Amherst por un grupo de investigadores liderados por David Jensen. En 2001 fundaron Knowledge Discovery Laboratory con la finalidad de desarrollar técnicas innovadoras para el descubrimiento de conocimiento que denominan *software de proximidad*. La materia prima con la que trabajan se extrae de las enormes bases de datos públicas que prácticamente existen en todos los ámbitos sociales (salud, educación, etc.), y que son alimentadas de modo regular con el registro temporal de una abundante cantidad de variables que guardan relación entre sí. En el año 2008 Jensen, Fast, Taylor and Maier presentaron el algoritmo AIQ (*Automated Identification of Quasiexperiments*). Es el primer sistema automatizado que identifica lo que "denominan" diseños cuasi-experimnetales (es la primera versión, y sólo identifica Diseños de grupo control no equivalente).

Evidencia científica: no existe objetivo más codiciado en cualquier disciplina de las Ciencias Sociales y de la Salud que buscar y encontrar evidencia científica en la que fundamentar decisiones clínicas, sanitarias, educativas, etc., y los diseños cuasi-experimentales alcanzan evidencia científica cuando el investigador cuida los aspectos anteriormente mencionados (e.g., Avellar & Paulsell, 2011 y Cook & Gorard, 2007) en aras de conseguir con ellos las garantías que ofrecen las investigaciones experimentales. Así es, a pesar de que muchos científicos comparten la idea de que las investigaciones experimentales constituyen el mejor modo de alcanzar evidencia científica (e.g., Cook, 2000; Nezu & Nezu, 2008), quienes más saben de ésto están convencidos de que "in the best of quasi-experiments, internal validity is not

much worse than with the randomized experiment " (Shadish et al., 2002, p.484).

Al hilo, y en este empeño, en 1968 en la Universidad de Northwestern se fundó *The Institute for Policy Research* (IPR) al abrigo de Campbell. IPR se gestó con la misión de estimular y apoyar la investigación de excelencia en las ciencias sociales sobre importantes cuestiones de política pública. Una visita a su página Web no dejará indiferente a ningún investigador interesado en estos diseños, http://www.ipr.northwestern.edu/index.html. Desde el año 2008 se celebran en la citada Universidad, y en colaboración con el *Institute of Education Sciences* (IES), importantes Workshops sobre cómo realizar una investigación cuasi-experimental de calidad en el ámbito de Educación (diseño y análisis) organizados y dirigidos por dos de los grandes discípulos de Campbell, Thomas D. Cook y William Shadish. Desde la página anterior podemos conocer todo lo que en ellos se ha expuesto y debatido.

Si nuestra investigación tiene calidad metodológica suficiente para garantizar evidencia científica quizá tenga el privilegio de formar parte de alguna revisión sistemática realizada mediante técnicas de meta-análisis (Bai, Shukla, Bak & Wells, 2012; Centre for Review and Dissemination, 2010), hasta el momento, considerados los mejores estudios para sintetizar la evidencia científica capaces de responder a cuestiones específicas. En lo que a revisiones sistemáticas respecta, dos asociaciones, *The Cochrane Collaboration* en el ámbito de la salud y *The Campbel Collaboration* en el ámbito de las ciencias sociales (en http://www.cochrane.org y http://www.campbellcollaboration.org/, respectivamente) son el referente. Desde que se constituyeron cada año celebran un Colloquium en algún lugar del mundo. Si nos introducimos en sus páginas Web y visitamos los programas de los Colloquia ya realizados podemos advertir, además de su estrecha colaboración, cómo una parte importante de los trabajos presentados están relacionados con la incorporación de las investigaciones no aleatorizadas en las revisiones sistemáticas y las características de calidad que deben tener para ser incorporadas.

Práctica basada en la evidencia, evaluación y aplicación de programas de intervención: la evidencia científica certificada mediante revisiones sistemáticas realizadas con investigaciones primarias de alta calidad metodológica con frecuencia tiene buena acogida para dar respuesta a necesidades de políticas gubernamentales o a intereses de determinadas empresas que deciden transformar el conocimiento científico en práctica efectiva realizando investigación aplicada.

Cuando se utiliza la expresión "programa de intervención" se está haciendo referencia a tareas múltiples y complejas que requieren contemplar múltiples aspectos de diseño, de implementación y de evaluación en contextos de intervención, que no son otros que contextos naturales, que, además, están sometidos a cambio continuo (e.g., Rossi & Freeman, 1985 y Trochim, 1984). No es intención de este trabajo abundar en la aplicación y evaluación de programas, pero sí destacar tres aspectos fundamentales:

Primero: ejecutar de facto la implementación de un programa requiere evaluaciones constantes antes, durante y después de la aplicación de la intervención (Avellar & Paulsell, 2011; Brandy & Moore, 2011) por dos razones fundamentalmente. Además de que los motivos económicos no son baladí y es preciso asegurar que el programa se implanta correctamente y que no tiene consecuencias indeseadas, porque los objetivos del programa siempre son varios (prácticos, económicos, teóricos, etc., donde la aplicación del tratamiento es sólo uno de ellos) y nunca podrían abordarse todos desde una única aproximación metodológica. En los últimos 20 años el concepto de Metodología Mixta está cobrando un auge extraordinario (se ocupa de integrar datos de una multiplicidad de fuentes para captar la diversidad de los fenómenos de modo profundo e integral), como muestra, la prestigiosa revista *New Directions for Evaluation* dedica el segundo número del pasado año 2013 a este tema que titula *Mixed Methods and Credibility of Evidence in Evaluation.*

Segundo: implantar un programa a nivel social requiere tener en cuenta que las personas no somos compartimentos estancos, sino que de modo natural estamos organizados en sociedad en unidades jerárquicamente superiores que nos imprimen características de grupo y diferencias entre los mismos que interesa conocer, controlar y tratar. La evaluación o implantación de un programa demanda por tanto realizar un muestreo de estas unidades de agregación porque interesa conocer lo que ocurre en cada nivel, pero también interesa conocer el flujo de relaciones entre los distintos niveles con ánimo de examinar si la intervención ha tenido su impacto sobre los resultados previstos. El análisis de datos tradicional con técnicas contenidas en el Modelo Lineal General es inerme a dos particularidades, entre otras, que tienen los datos agregados, la existencia de más de un término aleatorio, y el hecho de que las unidades pertenecientes a una misma unidad de agregación de nivel emiten respuestas correlacionas. En los últimos 20 años ha tenido lugar el desarrollo de las técnicas de análisis de datos más potentes, versátiles y con mayores prestaciones de cuantas existen para poner en evidencia de qué modo, en qué medida y de qué depende que un programa de intervención tenga efectos (positivos, negativos, directos, inversos, etc) sobre las diferentes unidades de análisis en que están inmersas las personas. Nos estamos refiriendo a los procedimientos estadísticos enmarcados en los denominados Modelos Mixtos, Modelos Jerárquicos o Modelos Multinivel.

Tercero: La meta última de toda disciplina científica es disponer de teorías causales capaces de explicar todo el universo de sea lo que sea que abarca esa disciplina. Elaborar un modelo causal capaz de explicar todo lo que sucede y por qué en el proceso de la intervención no es tarea fácil. El concepto de causalidad es complejo y multifacético (e.g., Funnell & Rogers, 2011), los agentes causales son muchos, a veces complementarios y a veces en conflicto, a veces la relación entre causa y efecto no es lineal, a veces incluso es discontinua, incluso a veces la causa tiene que alcanzar un umbral determinado antes de producir un efecto, añadiendo

que los individuos a quien va dirigida la evaluación del programa lejos de ser agentes pasivos ejercen un papel activo en la determinación de los efectos del programa. El modelo causal casi nunca es unidireccional, pueden existir bucles de feedback, peor aún, incluso ese feedback puede ser incluso retardado complicando el análisis de los modelos causales. A la par que el desarrollo de los Modelos Mixtos en los últimos 20 años se han desarrollado los Modelos de Ecuaciones Estructurales que permiten poner a prueba relaciones de causalidad y elaborar teorías causales basadas en la evaluación.

El Fichero adicional 5 (ver Adenda) contiene referencias destacadas sobre la implantación y evaluación de programas, la Metodología Mixta, los Modelos Mixtos o Multinivel y sobre los Modelos de Ecuaciones Estructurales.

Así comenzaron Campbell y Stanley el capítulo 5 de la obra de Gage (1963):

> McCall as a model. In 1923, W.A. McCall published a book entitled How to Experiment in Education …. In this preface said: "There are excellent books and courses of instruction dealing with the statistical manipulation of experimental data, but there is little help to be found on the methods of securing adequate and proper data to which to apply statistical procedure". This sentence remains true enough today to serve as the leitmotif of this presentation also. (Campbell & Stanley, 1963, p.171).

Este trabajo es un tributo a esta idea, y lo llevamos a cabo realizando una revisión de las investigaciones cuasi-experimentales y pre-experimentales publicadas en tres revistas españolas de Psicología evaluando algunos aspectos metodológicos relacionados con la planificación del diseño, la recogida y análisis de datos, la exposición de resultados y la elaboración de conclusiones. El objetivo que perseguimos es saber cuáles son los diseños más utilizados y qué características tienen, pero también, y sobre todo, conocer si los investigadores toman las precauciones necesarias para que la investigación ofrezca confianza en sus resultados, saber cuáles son sus errores y sus vicios, y si son los mismos que los reportados en otros trabajos similares a éste.

El pasado año 2013 los diseños cuasi-experimentales cumplen 50 años. Este trabajo es un homenaje a la persona que revolucionó los principios fundamentales de la investigación científica en las Ciencias Sociales, Campbell, y a sus discípulos Stanley, Cook, Shadish, Trochim, Myers, etc., que tanto están trabajando esta idea que tan buenos resultados aporta. Sirva para convulsionar a investigadores, revisores y editores para que (se) exijan, más si cabe, el cuidado del método. Va por ellos, de quienes hemos aprendido tanto, y va para todos aquellos que gusten de hacer bien las cosas.

Método

Hemos realizado una revisión sistemática no cuantitativa (Shadish & Myers, 2004 a) de las investigaciones pre-experimentales y cuasi-experimentales, en adelante Cx. y Px., contenidas en tres revistas españolas de Psicología durante el período temporal de 11 años entre 1999 y 2009, ambos incluidos.

Materiales

El criterio para la selección de las revistas fue cumplir con tres requisitos, ser revistas de temática general en el campo de la Psicología, aparecer en el *Journal Citation Reports (JCR)* con factor de impacto en el año 2009, y constar en el IN-*RECS* con índice de impacto dentro del primer cuartil en el mismo año. Estos criterios los satisfacían las revistas *Psicothema*, *Internacional Journal of Clinical and Health Psychology* y *Psicológica*.

Unidad de análisis

La unidad de análisis ha sido el estudio, considerando unidades independientes cada uno de los estudios Cx. y Px. publicados en un mismo artículo.

Diseño y procedimiento

Una vez identificados los artículos que contenían investigaciones Cx. y Px. aleatoriamente fueron repartidos en dos grupos y cada uno de ellos fue asignado aleatoriamente a dos expertos en metodología de investigación que, de modo independiente entre e intra-grupo, examinaron las variables de interés. La información de cada estudio fue extraída de los apartados introducción, método, resultados y conclusiones. Posteriormente los expertos de cada grupo compararon sus datos. El porcentaje de acuerdo fue del 94%. Los casos en que hubo discrepancia fueron revisados por un tercer experto. El contenido del estudio se discutía hasta que desembocara en una conclusión compartida.

Variables registradas

Las variables que hemos observado han sido consideradas de importancia capital en destacados trabajos sobre calidad de las investigaciones no aleatorizadas en el ámbito médico (e.g., Li, Moja, Romero, Sayre & Grimshaw, 2009; Shahar & Shahar, 2009, int al.), educativo (e.g., Cook, Cook, Landrum & Tankersley, 2008; Cook, Tankersley & Landrum, 2009, int al.), de las organizaciones (e.g., Creswell, 2009; Gibbert & Ruigrok, 2010; Paluck & Green, 2009, int.al.), y en el ámbito de las ciencias sociales en general (e.g., Shadish & Myers, 2004 a). Anidadas en cinco bloques, de modo resumido son:

1.- Características molares de la investigación Cuasi y Pre-experimental: Prevalencia, área de conocimiento, finalidad de la investigación, y definición del diseño y metodología utilizada.
2.- Aspectos determinantes de la planificación para defender las inferencias sustantivas y estadísticas: Composición y tamaño de la muestra, cálculo del tamaño, composición y

formación de los grupos de control, número de variables dependientes y cantidad de registros efectuados.

3.- Examen exploratorio de los datos y precauciones tomadas antes de decidir qué estadístico utilizar para poner a prueba las hipótesis: Tamaño de los grupos y equilibrio entre ellos, relación entre el tamaño de los grupos y el tamaño de las varianzas, examen de outliers, evaluación de datos perdidos y comprobación de asunciones sobre los datos.

4.- Análisis de los datos: Si comprueban o no la igualdad de los grupos antes de la aplicación del tratamiento y cómo, qué modelo asumen los investigadores que explica sus resultados, qué estadístico inferencial utilizan, si examinan el tamaño del efecto, los intervalos confidenciales y la potencia de la prueba empírica.

5.- Exposición y redacción de las conclusiones: Hemos examinado si se realizan comentarios de autocrítica sobre el tamaño y composición de la muestra y sobre amenazas a la validez interna, externa y de conclusión estadística.

Una exposición detallada y sucintamente justificada de todas las variables registradas está expuesta en el Fichero Adicional 1.

Análisis de los datos

Los estadísticos utilizados son frecuencias, porcentajes y razones (de modo testimonial también medias y desviaciones típicas). Los cálculos se han realizado mediante el paquete estadístico SPSS 19.

Resultados y discusión

Utilizando como modelo el trabajo de Keselman et al., (1998) los resultados se exponen acompañados de una breve discusión. Sólo se comenta una selección de todos los análisis realizados, sin embargo, nos van a permitir evaluar la coherencia metodológica interna de las investigaciones en todo su recorrido, desde la gestación del diseño hasta la elaboración de las conclusiones (en el Fichero Adicional 2 se exponen, libres de discusión, los resultados de todos los análisis llevados a cabo). En el texto hacemos referencia a unas Tablas donde se detalla la información que describimos. Todas ellas están contenidas en el Fichero Adicional 3.

En el período de 11 años comprendido entre 1999 y 2009 las investigaciones Cx. y Px. contenidas en las tres revistas examinadas apenas suponen un 4% en el volumen de sus publicaciones. En este período de tiempo, por cada 10 investigaciones experimentales sólo se han publicado 1.2 Cx. o Px.

El debate en torno al valor de las investigaciones cuya muestra está compuesta por universitarios es un clásico en vanguardia (Cooper, McCord & Socha, 2010; Highhouse & Gillespie, 2010; McNemar, 1946; Wiecko, 2010, int al.). En nuestro caso el porcentaje no es muy alto (21.21%), aunque en absoluto desdeñable. Las personas con algún problema que resolver o con alguna necesidad especial (que hemos denominado muestra específica enferma y no enferma respectivamente) están presentes en el 53% de las investigaciones. Si sumamos a las anteriores los niños menores de 12 años y a las personas con más de 65 alcanzarían un porcentaje del 68.15% (Tabla 3). Este es el motivo por el cual la investigación con finalidad aplicada es muy superior (68.2%) a la básica, y la realizada en el área clínica superior a la realizada en el resto de categorías contempladas. Aunque en la introducción hemos destacado que la investigación Cx. y la evaluación del impacto de programas van de la mano, no hemos encontrado ningún trabajo que se refiera a la evaluación del impacto de ningún programa de trascendencia social relevante.

Ubicar la investigación que publicamos en su metodología correspondiente y denominar correctamente el diseño utilizado para recoger los datos no es sinónimo de calidad del trabajo realizado, pero tampoco es baladí. Más al contrario, es de importancia destacada para que los lectores y consumidores de los productos de investigación ajusten sus lentes y puedan juzgar su validez y determinar la confianza que les merece (Wilkinson & Task Force on Statistical Inferencie, 1999). Que únicamente el 43.9% de los D.Cx. clasifiquen (en metodología) y definan el diseño correctamente en el trabajo que defienden no es, por escaso, un buen dato (Harris et al., 2005 hallaron resultados similares), aunque peor es saber que el 29.2% no definen la metodología o lo hacen incorrectamente, y que el 51.2% no definen el diseño o no lo hacen bien (ver Tabla 2).

Sin embargo, el diseño "concreto" utilizado sí es sinónimo de confianza en la capacidad para poner a prueba las hipótesis del investigador. Abundando es este aspecto, Cook and Campbell (1979), Heisman and Shadish (1996), Marcantonio and Cook (1994), Orwin (1997), Shadish and Myers, (2004 a y b), Shadish and Ragsdale (1996), Shadish et al. (2002) y Trochim (1984) int.al., han destacado que los D. Cx. de Discontinuidad en la Regresión y de Series Temporales son más potentes que los diseños Cx. de Grupo Control no Equivalente (GCNE), incluso algunos investigadores piensan que son tan potentes como los diseños experimentales (Avellar & Paulsell, 2011; Cook, 2008; Cook, Scriven, Coryn & Evergreen, 2010; Cook, y Wong, 2008; Shadish & Cook, 2009; Shadish, Galindo, Wong, Steiner, & Cook, 2011). De otra parte, en lo que a los D.Cx.-GCNE respecta, muchos investigadores han destacado que existe una jerarquía en la que los diseños con intercambio de tratamiento, aquellos que tienen más de un grupo experimental (G.E), más de un grupo de control (G.C), más medidas post que una y más medidas pre que una son mucho más potentes por dos razones. La primera, porque tienen más capacidad para poner a prueba nuestra hipótesis. A la sazón existe una gran cantidad de trabajos meta-analíticos que han concluido que diseños Cx. distintos alcanzan tamaños del efecto (TE) distintos (e.g., Heisman & Shadish, 1996; Lipsey & Wilson, 1993; MacLehose et al., 2000; Sacks, Chalmers & Smith, 1982, 1983; Schochet 2009a; Schochet et al., 2010; Shadish

et al., 2002; Shadish & Heinsman, 1997; Shadish & Ragsdale, 1996; Swaen, Teggeler & Van Amelsvoort, 2001; Weisburd, Lum & Petrosino, 2001). La segunda es porque son más capaces de liberar nuestros datos de posibles amenazas a la validez interna (Deeks et al*.,* 2003; Johnston, Ottenbacher, y Reichardt, 1995; Shadish et al., 2002; Shadish & Myers, 2004 b).

A pesar de lo comentado en el párrafo anterior, hemos visto que el Diseño Cx. de GCNE clásico (2x2) y el Diseño Px. de un solo grupo con medidas pre y post son los más utilizados, suponen casi el 50% de todos los de su clase (Tabla 2). Estos resultados no divergen de los hallados por otros autores en ambientes disciplinares tan distintos como intervención temprana (Snyder, Thompson, Mclean & Smith, 2002), educación (Gersten, Baker, Smith-Johnson, Flojo & Hagan-Burke, 2004; Hsieh et al., 2005; Seethaler & Fuchs, 2005), Psicología de las organizaciones (Grant & Wall, 2009; Scandura & Williams, 2000), o enfermedades infeccionas (Harris, Lautenbach & Perencevich, 2005; Harris et al*.,* 2006; Shardell et al., 2007). Solo hemos contado una investigación Cx. (2.4%) que realiza un diseño de GCNE con intercambio de tratamiento, y únicamente 7(17.1%) D. Cx. que "podrían ser" de Discontinuidad en la Regresión (DR). Escribimos "podrían ser", porque ninguno ha sido denominado con tal por sus autores, que sería lo de menos, lo de más es que tampoco han sabido aprovechar la potencia que tienen por el hecho de conocer qué variable es la que divide a los grupos experimental y control (Lesik, 2006; Reardon & Robinson, 2012; Reichardt & Henry, 2012; Shadish et al., 2002, int.al. En los trabajos revisados ha sido más habitual contar con más de un G.E (sucede en el 24.39% de los D. Cx. y en el 48% de los D. Px.) que con más de un G.C (sucede en el 12.19% de los D. Cx. y en el 16% de los D.Px.). Este aspecto es de sumo interés. Los grupos experimentales tienen el poder de replicar el efecto del tratamiento y también de controlar el efecto de historia. Los grupos de control tienen el poder extraordinario de poner a prueba los efectos de sesgo de selección, historia, maduración y regresión a la media entre otros.

Probablemente, cuando un investigador lleva a cabo una investigación Cx., más importante que cualquier otro aspecto de la planificación, es la elección adecuada de un G.C que sea comparable al G.E para evitar el riesgo de posibles amenazas a la validez interna. En este sentido, se debe considerar que la proximidad de los grupos no implica necesariamente que sean comparables, que características de los sujetos que se antojan accidentales (e.g., unos eligieron cursar ética otros religión) pueden conllevar diferencias importantes entre los grupos, que no basta con estimar que los grupos son parecidos (hay que demostrarlo), que los sujetos voluntarios nunca forman un buen G.C., y por supuesto, no decir nada de cómo es el G.C es enfatizar que nada se sabe de estos diseños (ver ICH Expert Working Group, 2000). En esta revisión, el 69.69% de los D. Cx. de GCNE han formado el G.C. de alguna de estas maneras (ver Tabla 5). Por el contrario, toda técnica utilizada para formar los grupos que trate de emular la asignación no condicionada que supone la asignación aleatoria, utilizando la técnica del apareamiento (*matching*) por ejemplo, o que permita considerar una continuidad entre los sujetos de los grupos experimental y control como por ejemplo dejar que el azar decida cuál de los grupos incidentales (comparables, que no exclusivamente próximos) sea G.E y cuál G.C, que el G.C sea aquel que está en lista de espera, que el G.C esté compuesto por aquellas personas deseosas de participar en la investigación porque tienen un problema que resolver pero por algún motivo no pueden acudir a la terapia (e.g., coincide en horario laboral; no les van bien las fechas, etc.) "garantiza" un G.C válido (Marcus, Stuart, Wang, Shadish & Steiner, 2012; McCaffery et al., 2011; Walter, Turner, Macaskill, McCaffery & Irwig, 2012), y de éstos hemos visto sólo el 30% en los D. Cx. de GCNE. [En el Fichero Adicional 5 se exponen referencias destacadas sobre el concepto, control y corrección del sesgo de selección].

Además de la función de "control", otro aspecto que se debe valorar del G. C es la potencia que puede aportar al diseño para poner a prueba las hipótesis del investigador. Los resultados meta-analíticos constatan que los TE son menores cuando el G.E se compara con G.C activos (e.g., placebo, tratamiento de costumbre, tratamiento alternativo) que cuando se compara con G.C pasivos (e.g*.,* sin tratamiento, en lista de espera) (Heisman & Shadish, 1996; Shadish & Ragsdale, 1996) (ni que decir tiene que el investigador también debe cuidar que no surjan efectos reactivos, ni suyos ni de los sujetos, capaces de sesgar seriamente los resultados, y éstos es más fácil evitarlos cuando los G.C son activos). También se ha demostrado que el TE es mayor, y más preciso cuando se emplea la técnica de apareamiento anteriormente citada en la formación de los grupos (Shadish & Myers, 2004 a; Shadish & Ragsdale, 1996).

Cuidar el G.C no tiene sentido si no cuidamos también el G.E. Es imperativo considerar si el tratamiento se administra de modo individual o en grupo, y sea de un modo u otro es preciso asegurar que la administración sea homogénea para todos los sujetos y que la adherencia al tratamiento sea satisfactoria. Estos aspectos y más tienen que ver con la integridad del tratamiento (Devito et al., 2011; Higgins & Green, 2011; McLeod & Islam, 2011; Perepletchikova, Hilt, Chereji & Kazdin, 2009). Nosotros no hemos revisado estos matices. Algunas revisiones y otras referencias destacadas sobre esta temática las exponemos en los Ficheros Adicionales 2 y 5.

El número de medidas registradas, tanto antes de la intervención como después de ella es extraordinariamente importante (Mara & Cribbie, 2012; Mara et al., 2012; Rausch, Maxwell & Kelley, 2003; Schulz, Czaja, McKay, Ory & Belle, 2010; Shadish et al*.,* 2011; Steiner, Cook, Shadish & Clark, 2010; Venter, Maxwell & Bolig, 2002). Las efectuadas antes de la intervención son un bastión esencial para defender la no existencia de regresión a la media, de error en el registro, de efectos de maduración de los sujetos (evolución o involución emocional, conductual, intelectual, etc.), y para defen-

der la estabilidad y confiabilidad en las medidas. Llevar a cabo varios registros post tratamiento nos permitirán analizar la evolución de la conducta estudiada, si se mantiene el resultado esperado, hasta cuándo se mantiene, si el efecto es inmediato o retardado, o si no hay efecto. Hemos comprobado que en ninguna investigación de esta revisión se efectúa más de una medida pre-tratamiento, y sólo en el 24.39% de los diseños Cx. y en el 36% de los Px. se registra más de dos medidas post-tratamiento (Tabla 2).

La planificación del tamaño de la muestra en base a un TE deseado tiene una importancia capital. Implica tener en cuenta tanto la sensibilidad de la investigación como la precisión de la misma, y ambos aspectos determinan la posibilidad de tener suficiente potencia de prueba. Ninguna de las investigaciones revisadas calcula a priori el tamaño de la muestra (Tabla 4), con el añadido de que los tamaños de muestra utilizados tampoco son elevados (Tabla 6). Sánchez Meca, Valera, Velandrino y Marín (1992), Valera, Sánchez Meca y Marín (1998) y Valera, Sánchez Meca, Marín y Velandrino (2000) hallaron resultados similares. Así pues, sea grande o no el tamaño de la muestra, lo que no podemos saber es si fue suficiente para detectar el TE que deseaban encontrar. De otra parte, siendo el tamaño de muestra pequeño y recordando cómo fueron formados los grupos, es probable que algunas muestras tampoco fuesen representativas, con el problema que eso supone respecto a la trasferibilidad de los resultados y a la validez externa de población (Burchett , Umoquit & Dobrow, 2011; Cambon , Minary, Ridde & Alla, 2012; Ferguson, 2004; Glasgow & Emmons, 2007; Green & Glasgow, 2006; Steckler & McLeroy, 2008; Thomson y Thomas , 2012).

Tan preocupante como lo anteriormente expuesto es, de una parte, el desequilibrio en el tamaño de los grupos, y de otra, que a esto se sume la tan temida relación directa o inversa entre el tamaño de los grupos y el tamaño de las varianzas. El examen de la potencia y robustez de los procedimientos analíticos frente a estos problemas es un tema candente en la actualidad tanto por lo habitual del problema como por las consecuencias indeseables que tiene. Es numerosa la investigación que ha puesto en evidencia la inflación del error de Tipo I y de Tipo II cuando la relación es negativa y positiva respectivamente. En esta revisión hemos encontrado un porcentaje de diseños no balanceados altísimo (82.69%) donde la diferencia media entre los tamaños de los grupos es de 28.6 (*DT* = 49.93) (Tabla 6). Ya se observe el desequilibrio entre el tamaño de los grupos en función de diferencia o de razón es notable el fuerte sesgo positivo, lo que quiere decir, que, aunque no demasiadas, existen investigaciones con una diferencia mucho mayor aún entre los tamaños de los grupos (Sánchez Meca et al., 1992, Valera et al., 1998 y Valera et al., 2000 en revistas de Psicología editadas en España, y Keselman et al., 1989 y Ruscio & Roche, 2012 en revistas editadas fuera de nuestras fronteras han encontrado parecidos resultados). Añadido a esto, observamos dos problemas más. Uno, que solamente hemos encontrado una razón entre el tamaño de los grupos y el tamaño de las varianzas nula, en el resto, o es positiva o es negativa casi al 50%. Otro, que esto sólo lo hemos podido observar en el 50% de los D.Cx. no balanceados debido a que en la otra mitad no se exponen las varianzas (Keselman et al., 1989 hallaron resultados similares). Este último dato no es inocuo, aquellos que realizan investigaciones meta-analíticas conocen como nadie la importancia de contar con los estadísticos descriptivos de las variables (Higgins & Green, 2011).

Registrar la medida pre y otras variables relacionadas con la variable dependiente antes de aplicar el tratamiento tiene una importancia crucial dado que nos van a permitir poner a prueba la igualdad de los grupos antes de aplicar el tratamiento. En 27 (84.37%) de las investigaciones Cx.-GCNE se comprueba de algún modo la igualdad de los grupos antes de aplicar el tratamiento, pero existen 5 (15.62%) en que no se hace. De mayor envergadura es el valor que tiene la medida pre-tratamiento y el registro inicial de otras variables para aparear a los sujetos de los grupos antes de decidir qué grupo va a ser el experimental y cuál va a ser control y evitar así de modo mucho más eficiente el sesgo de selección (entre otros). En esta revisión esto sólo se hace en 4(12.5%) investigaciones Cx.-GCNE, luego en el 87.5% de ellas ambos registros se han efectuado después de decidir qué grupo es el experimental y qué grupo es el control. Sólo en 3 (42.85%) investigaciones Cx.-DR se aparearon los sujetos de los grupos en base a la información de variables relevantes. Todos estos resultados están expuestos en la Tabla 8 a.

Todas las precauciones anteriores son necesarias pero no son suficientes para demostrar la veracidad de nuestra hipótesis cuando la aleatorización no es posible. Son imprescindibles, al menos, dos condiciones más. Una, debemos demostrar que otras variables no son responsables de los resultados hallados, y para eso debemos de eliminar o tener controladas variables extrañas capaces de ser explicaciones alternativas a nuestra hipótesis. En este caso sólo en 22(68.75%) investigaciones Cx.-GCNE y en 5 (71.42%) Cx.-DR se registran variables consideradas "sospechosas" para examinar la equivalencia de los grupos en ellas o para formar los grupos en base a que esas características permanezcan constantes, pero sólo en 4 (12.5%) investigaciones Cx.-GCNE (y en ninguna Cx.-DR) se lleva a cabo control estadístico (se introducen en el modelo matemático) de las mismas (ver Tabla 8 a). Dos, debemos tener meridianamente claro qué modelo explica mejor los resultados encontrados (Ato y Vallejo, 2011), y esto porque nuestra hipótesis quizás sea cierta con matices (puede haber variables mediadoras o moderadoras o supresoras de la relación causal). Que el investigador avance un modelo explicativo de sus resultados es de importancia capital en las investigaciones no aleatorizadas. En ninguna investigación de esta revisión los autores exponen que "asumen un modelo de cambio" o un "modelo mediacional", o simplemente "un modelo causal directo". En su lugar hemos encontrado que "escriben" que deciden analizan los datos mediante las puntuaciones de cambio, o que están interesados en la interacción, o en ambas (sin saber que el resultado es el mismo estadísticamente), o que só-

lo les interesa analizar las diferencias post-tratamiento, o que les interesa todo lo anterior (ver Tabla 8 a). Precisamente son aquellos que están interesados en las diferencias post los que únicamente analizan los datos mediante el análisis de la covarianza si es que hallaron diferencias iniciales entre los grupos en la medida pre u en otras. La medida pre u otras registradas antes de recibir el tratamiento nunca se utilizaron como covariables cuando las diferencias iniciales entre los grupos no fueron estadísticamente significativas, y debiera hacerse si se asume un modelo mediacional. En fin, los autores parecen haber olvidado que en la investigación cuasi-experimental el análisis de la covarianza no sólo se hace para reducir variabilidad e incrementar la precisión, sino también, porque la covariable puede ser una variable explicativa del resultado. En el Fichero Adicional 5 se exponen referencias destacadas sobre los efectos, la detección y el análisis de fenómenos como mediación, moderación, etc.

Antes de exponer qué estadístico escogen utilizar los investigadores para poner a prueba "las puntuaciones de cambio, o la interacción, o la simple diferencia entre las medidas post" hacemos un inciso para destacar de nuevo que, si la planificación de la investigación es deficitaria poco sentido tiene ocuparse del análisis de los datos. Campbell y Stanley (1963, p.22) ya hicieron hincapié en este aspecto:

> Good experimental design is separable from the use of statistical tests of significance. It is the art of achieving interpretable comparisons and as such would be required even if the end product were to be graphed percentages, parallel prose case studies, photographs of groups in action, etc. In all such cases, the interpretability of the results depends upon control oven the factors we have been describing.

Cuarenta años más tarde se vuelve a insistir en ello:

> In this book, we have advocated that statistical adjustments for group nonequivalence are best used after design controls have already been used to the maximum in order to reduce nonequivalence to a minimum. So we are not opponents of statistical adjustment techniques such as those advocated by the statisticians and econometricians described in the appendix to Chapter 5. Rather we want to use them as the last resort. (Shadish et al., 2002, p.503).

Este pensamiento no es exclusivo del ámbito de las Ciencias Sociales, Deeks et al*.,* (2003, p.91) en el ámbito médico expresan que "*Statistical methods of analysis cannot properly correct for inadequacies of study design*". En definitiva, es notorio en todos los campos de investigación ver expresado que por muy elegante, novedoso, robusto y poderoso que sea el procedimiento de análisis utilizado, nunca va a corregir ni las deficiencias del método ni la mala planificación del diseño elegido para recoger los datos (e.g., Des Jarlais, Lyles, Crepaz & TREND Group, 2004; Eastmond, 1998; Harris et al., 2004; Scheirer, 1998; Shadish & Myers, 2004 a; Shardell et al., 2007; Sridharan & Nakaima, 2011; Stone et al., 2007; Valentine & Cooper, 2003, int. al).

La probabilidad de acertar en la elección del estadístico (conveniente y correcto) para poner a prueba nuestra hipótesis es muy pequeña si antes no se realiza un examen exploratorio de los datos (Cohen, 1992, 1994) para determinar por ejemplo, si algunos sujetos tienen puntuaciones extremas (valores atípicos debidos a un defectuoso instrumento de medida, a fallos en el registro, a una mala aplicación del tratamiento, etc.), si tenemos pérdida de datos ocasionales (no acudieron ese día por algún motivo ajeno a la investigación, no se registró el dato debido a un fallo del experimentador, etc.) o si lo que sucede es que hay sujetos que abandonan (debido a la no adherencia al tratamiento, de modo circunstancial se han trasladado a otro lugar, una enfermedad nos les ha permitido continuar, etc.). Es necesario también conocer si se satisfacen o no las asunciones de normalidad y homoscedasticidad, que con frecuencia sucede cuando los grupos están desequilibrados como en este caso. Y si procede, es imprescindible valorar la independencia entre la(s) covariada(s) y el tratamiento, si las pendientes en los grupos son paralelas, si existe o no multicolinealidad, si es posible que alguna variable esté mediando en la relación causal, etc. Las causas capaces de provocar estos problemas pueden ser varias y requerirían tratamientos distintos.

La existencia de puntuaciones atípicas sólo se examina en una investigación. En ninguna se comenta si falta algún dato, a pesar de lo habitual que es este problema en los diseños Cx. (Shadish et al*.,* 2002). Sólo en 14(21.2%) investigaciones se reconoce haber perdido sujetos (Tabla 7) pero en ningún caso los investigadores se plantean por qué se ha producido, cual ha sido el patrón de pérdida ni qué consecuencias puede tener para la inferencia. Investigaciones meta-analíticas han demostrado que el TE es sensible al abandono de los sujetos (Shadish & Ragsdale, 1996) y por lo tanto es preciso tener en cuenta el tamaño de la muestra al principio y al final de la investigación para valorar el efecto del desgaste de muestra (Shadish & Myers, 2004 a). Sin excepción, ya sean puntuaciones atípicas ya sea pérdida de sujetos, en las investigaciones afectadas se afronta el problema eliminando el problema, esto es, eliminando del análisis a los sujetos que tienen puntuaciones atípicas y a los que abandonan (Tabla 7). Una excelente exposición sobre estos problemas y el modo de hacerlos frente la realizan West et al. (2008).

En el 87.9% (Tabla 7) de las investigaciones no se comprueba si se satisfacen las asunciones de normalidad ni de homogeneidad, tampoco las asunciones añadidas a éstas cuando se hace análisis de la covarianza. En las pocas veces que se utiliza el análisis de la regresión o el análisis multivariado no se examina la existencia de multicolinealidad, en el primero tampoco mediación y en el segundo nunca se completa con el análisis discriminante. Así las cosas, no se ha utilizado ningún estadístico robusto (e.g., *F* de Brown Forsythe en ausencia de homogeneidad) y las pruebas no paramétricas se utilizan en pocas ocasiones, y cuando se utilizan se justifican por ser escaso el tamaño de muestra. Por lo tanto, el análisis de los datos realizado para poner a prueba las hipótesis es un fiel precipitado del diagnóstico que no se ha hecho de los datos.

Aunque en las Tablas 8 b y 8 c podemos observar que casi cada diseño Cx. y Px. ha sido analizado de un modo dis-

tinto, y que los diseños que comentamos "podrían ser" de DR ninguno se analiza como tal, sin duda alguna la prueba *t* de Student, estresándola hasta la saciedad, es la prueba estrella. Con frecuencia no se tiene una visión conjunta del diseño y, desmembrando los grupos y las ocasiones temporales, los autores ponen a prueba las hipótesis estudiando las diferencias dos a dos en cada grupo (entre dos momentos temporales) y dos a dos entre grupos (en cada momento temporal), y esto para cada una de las variables dependientes que, del mismo modo que otros autores han encontrado (e.g., Scandura & Williams, 2000; Snyder et al*.,* 2002), aquí también son muchas.

Es indiscutible que analizando así los datos los investigadores aseguran tener una alta probabilidad de rechazar la Ho, sin embargo no lo hacen cuidando la planificación de la investigación, sino que lo consiguen debido al nivel de significación acumulado tras realizar una gran cantidad de contrastes, y no debemos olvidar que el estrés sometido al análisis de los datos no convierte en válidas las comparaciones efectuadas "Use of significnce tests presumes but does not prove or supply the comparability of the comparison groups or the interpretability of the difference found" (Campbell & Stanley, 1963, p.22). De esta parte, por tanto, cometer un elevado error de Tipo I está garantizado. De otra parte, párrafos atrás expusimos que el tamaño de los grupos es pequeño por lo general, en los diseños balanceados y en los no balanceados también. En este caso, los investigadores trabajan con escasa potencia de prueba para detectar, quizás, un TE relevante, que nunca sabremos si lo era porque pocas veces se calcula, y en aquellas que se calculó no hubo comentario alguno sobre su relevancia a nivel sustantivo. De este modo, por tanto, los trabajos están destinados a tener una alta probabilidad de cometer error de Tipo II. Esta conducta polarizada convierte la inferencia estadística en un ejercicio abandonado al capricho del azar. Desafortunadamente sigue siendo actual el comentario de Campbell y Stanley "The tests of significance used with it are often wrong, incomplete, or inapropiate" (*op.cit,* pp. 21-22).

Finalmente, aunque el valor empírico del estadístico de contraste y los *gl.* están reflejados prácticamente en la totalidad de los trabajos, el valor exacto de *p* se expone en 41(62.1%) investigaciones y el TE en 13(19.7%). Sin embargo, la potencia de prueba empírica sólo se calcula en una y los intervalos confidenciales en ninguna (Tabla 11). Este panorama también se ha detectado en otras revisiones sobre investigaciones Cx. (Hsieh et al., 2005; Snyder et al., 2002).

En el Fichero Adicional 5 se exponen excelentes referencias sobre cómo analizar los D. Cx. de GCNE, de DR y de series temporales.

Una vez redactados los resultados corresponde elaborar una discusión y concluir. Cuando se ha realizado una investigación Cx. o Px. es esperable, en mayor medida que cuando se lleva a cabo una investigación experimental, que el investigador extreme precauciones al extender sus conclusiones. Es el momento de reconocer las debilidades de la investigación, y de poner en tela de juicio si la pretendida relación causal ha podido resultar afectada por alguna de las posibles amenazas a la validez interna. Es el momento de valorar si la relación estadísticamente significativa encontrada es espuria en lugar de causal. O por el contrario, esa relación estadística que no hemos encontrado ha sido debida a la presencia de variables que estaban oscureciendo la relación o a la existencia de una incontrolada varianza del error por ejemplo, entre otras razones de cariz estadístico. Esto es, el investigador debe valorar si también ha cuidado la validez de conclusión estadística. Sólo si ambas se han cuidado tiene sentido examinar la validez externa y ecológica. Examinamos los apartados "discusión" y "conclusiones" con ánimo de observar si los investigadores manifiestan la sospecha, o explicitan la certeza de que alguna amenaza a la validez pudiera estar presente, y si así es, saber de qué modo la ponen a prueba y de qué modo solventan el problema. En la Tabla 12 se exponen los resultados.

Advertimos que en 31(77.5%) investigaciones Cx. no se comenta ninguna posible amenaza a la validez interna. Del resto, sólo en un trabajo el hallazgo de diferencias estadísticamente significativas entre las medidas pre de los grupos se denominó sesgo de selección, pese a que en 33(84.61%) se pone a prueba la igualdad inicial de los grupos de alguna manera y que en 6 (18.8%) se efectúa un análisis de la covarianza con la medida pre o con otras (ver Tabla 8 a) justificando que se hace debido a que ha habido diferencias entre los grupos en ellas y que así ya estarían controladas. Por lo tanto, existe la creencia generalizada de que si los grupos experimental y control son iguales estadísticamente con respecto a la medida pre, y/o con respecto a alguna otra, no existe sesgo de selección, lo escriban así o no, y además, que basta el análisis de la covarianza para controlar esas diferencias iniciales, de haberlas. Nada más lejos de la realidad. No basta con eso. El que no haya diferencias en la medida pre o el que las haya, ni es suficiente ni es necesario para que exista o no sesgo de selección. El sesgo de selección es un activo tóxico complejo responsable de las diferencias sistemáticas entre los grupos de comparación en la respuesta al tratamiento o en su pronóstico, y se produce cuando variables relacionadas con la variable dependiente existen en grado distinto entre los grupos sometidos a comparación (debido a distorsiones en los procedimientos utilizados para seleccionar los sujetos, debidos a factores que influyen en la participación en el estudio, etc.) y pueden ser otras distintas a las registradas, esto es, quizás variables como la edad, el sexo, etc*.,* no sean causa de sesgo en según qué muestra, pese a ser las que habitualmente se registran. Esta falsa creencia explicaría por qué las variables registradas antes recibir la intervención nunca se han introducido como covariables cuando han resultado estadísticamente no significativas entre los grupos, cuando de hecho debiera hacerse si se asume un modelo mediacional (ver Judd & Kenny, 1981 y Huitema, 2011).

En 7 (17.5%) investigaciones Cx. se reconoce haber perdido muestra debido al abandono de sujetos aunque sólo en una se denomina *mortandad* experimental. Denominarlo así o

no pensamos que no tiene importancia, lo que sí la tiene es que en ninguna se ha examinado a qué ha podido deberse y sólo en 4 de ellas se estudia de algún modo el posible efecto que esto haya podido causar poniendo a prueba si los sujetos que abandonan tienen las mismas características que tiene el grupo al que pertenecían (Tabla 12). En las 7 investigaciones realizan los análisis inferenciales posteriores prescindiendo de los sujetos que han abandonado. Sin saberlo han realizado lo que se denomina análisis por Protocolo (*per-protocol-analysis*), pero podrían haber procedido de modo más saludable realizando el análisis por Intención de Tratar (*Intention-to-Treat Analysis*). Para conocer las particularidades y ventajas de uno y otro modo de proceder, y tomar el pulso al debate que sobre este tema está teniendo lugar se puede consultar Anderson (2012), Sedgwick (2011), Shah (2011), White, Carpenter & Horton (2012), int. al.

En dos investigaciones los autores sospechan que la cantidad de tiempo existente entre los registros de las medidas pudiera ser un problema. En ambas, y sólo en el G. C, examinan si estas diferencias (post-pre) son estadísticamente significativas y en las dos resultó no serlo. En ambas concluyeron entonces que la distancia temporal no influía en el efecto del tratamiento. Esta conclusión es errónea, pero sin saberlo estaban poniendo a prueba el efecto de la maduración, pero también de la historia, de la regresión a la media (si las medidas hubiesen estado más próximas), pero de modo tan débil que, en ningún caso, esas no diferencias estadísticamente significativas pueden eliminar la sospecha de que estas amenazas pudieran estar presentes, máxime conociendo el escaso cuidado que se ha puesto en la formación de los grupos. De ahí la necesidad de contar con más de un G. C como anteriormente hemos comentado. De otra parte, los autores podrían haber observado lo mismo teniendo también en cuenta al G. E. Si hubiesen tenido la sospecha de un cambio maduracional, examinar las puntuaciones de cambio tiene sentido. De haber resultado el análisis estadísticamente significativo pensaríamos que ambos grupos evolucionan con un ritmo distinto, esto es, es posible que pudiera haber cierto efecto de maduración pero sería posible concluir que el tratamiento ejerce un efecto añadido en el grupo que lo recibe. Si no hubiese sido estadísticamente significativo (y además observamos el gráfico de medias) tal vez podríamos concluir que los dos grupos evolucionan al mismo ritmo. Esta podría haber sido una forma sencilla de proceder. Ahora bien, cuando la distancia entre las medidas es mucha (quizás debido a que el tratamiento tarda un tiempo en ser efectivo y eso justificaría esa distancia), o sea por otro motivo, hay que pensar que en mucho tiempo caben muchas circunstancias, y alguna es esperable que afecte a las variables dependientes que se toman. De ahí la necesidad de registrar variables dependientes no equivalentes y de contar con más grupos tanto de tratamiento como de control y ponerlos a prueba en situaciones y momentos temporales distintos. En fin, necesidades distintas requieren controles distintos.

De otras amenazas a la validez se comenta sólo su sospecha pero no se ponen a prueba, como el posible efecto de regresión a la media que se comenta en dos investigaciones, y los posibles efectos de orden que se comentan sólo en una.

Otras amenazas ni siquiera se sospechan. A saber, en la Tabla 5 pudimos ver que en 12 (30.7%) ocasiones los grupos experimental y control están muy próximos entre sí y en estas situaciones puede ocurrir rivalidad compensatoria, difusión del tratamiento, etc., (Cook & Campbell , 1979) que en ningún caso se comentan.

Sólo en una investigación Px. los autores calculan el tamaño de muestra, pero lo hacen una vez que ya estaban formados los grupos con ánimo de saber si sería suficiente para estimar un tamaño del efecto medio según Cohen (1988). Lo fue. Así las cosas, teniendo en cuenta que en ninguna investigación se planificó el tamaño de la muestra esperábamos leer algún comentario referente a ella tanto con respecto a su composición como con respecto a su tamaño. Advertimos que en 30(75%) investigaciones no se hace ningún comentario explícito acerca de la muestra, del problema que pueda suponer tener poco tamaño, o tener diferente número de sujetos en los grupos, si su composición no es enteramente adecuada, etc. Únicamente en 2 (5%) los autores advierten que podría estar sesgada, y, sólo en 8 (20%) investigaciones se comenta que el tamaño de la muestra *puede ser* escaso. Como anécdota, en una investigación los autores escriben que su *tamaño de muestra es suficiente porque es mayor que el empleado en otras investigaciones*. Ninguna investigación cuestiona la representatividad de la muestra utilizada.

Finalmente estábamos interesados en saber si los autores que habían realizado una investigación con finalidad aplicada 45(68.2%), sobre todo cuando la composición de la muestra fue "específica enferma" y habían concluido que la intervención fue efectiva, manifestaban explícitamente el propósito de aplicar el tratamiento en los grupos que habían sido de control. En ningún caso se expresó explícitamente esto cuando de hecho debiera de hacerse por razones éticas, sin embargo, expresiones como "sería deseable" o "sería conveniente" fueron habituales.

Es un clásico concluir los informes de investigación comentando limitaciones en el sentido de que "manipulando otros niveles de variables…, aplicándolo a otras muestras distintas…, etc, los resultados pudieran ser otros". Esto también se hace en los trabajos que revisamos, pero no es suficiente. A la misma conclusión han llegado otros investigadores que también han realizado revisiones sobre la calidad de las investigaciones Cx. A saber, Scandura & Willians (2002) manifiestan que tienen una baja validez de constructo, interna y externa. Castro & Mastropieri (1986), Deeks et al. (2003), Dunst & Rheingrover (1981), Farran (1990, 2000), Grant & Wall (2009), Higghouse (2009) y Snyder et al. (2002) lamentan el poco rigor en la planificación de estas investigaciones (diseño pobre, medidas imprecisas e inservibles, etc.) e inadecuados análisis estadísticos. Harris et al. (2005) manifiestan su enorme preocupación por el hecho de que muy pocos autores describen las limitaciones de los diseños que utilizan en las investigaciones Cx. y por lo tanto no pueden esbozar las posibles amenazas a las conclusiones.

Ya se había avisado "Encouraging good design and logic will help improve the quality of conclusions." (Wilkison & TFSI, 1999).

Conclusiones y recomendaciones

Hemos examinado minuciosamente las 66 investigaciones Cx. y Px. contenidas en las revistas *Psicothema*, *International Journal of Clinical and Health Psychology* y *Psicológica* en el periodo comprendido entre 1999 y 2009 (11 años). Debido a que son el total de las de su clase en el período temporal examinado constituyen un censo.

A colación de la cantidad. Es un hecho que conforme el total de casos se aleja de 100 en sentido negativo más grande nos parece el porcentaje de una cantidad dada, del mismo modo que cuando de 100 se aleja en sentido positivo nos parece más pequeño el de la misma cantidad. Por esta razón, contar con tan pocas investigaciones ha supuesto una virtud para el propósito que pretendíamos, ya que el cálculo de porcentajes ha hecho las veces de tinta de contraste que hemos aprovechado para destacar las debilidades de estas investigaciones y ahondar en la enorme importancia que tiene cuidar el método.

Consideramos de justicia, antes de nada, realizar una digresión para destacar tres aspectos. Primero. Estas 66 investigaciones sólo suponen el 4.02%, 4.47% y 2.2% de los trabajos publicados por las tres revistas (respectivamente, en el orden anteriormente citado, ver también Tabla 1) y en absoluto son representativas del conjunto de trabajos publicados en ellas. Segundo. Entre las 66 investigaciones sometidas a revisión hay algunas extraordinarias, y salvo alguna que requeriría una revisión en mayor profundidad, el resto adolecen de una u otra deficiencia, pero no de todas, hecho que en absoluto va a suponer que pierdan su envergadura y su sentido y sirvan como fuente de tensión en su campo de estudio sea el que sea. Tercero. Huelga comentar que la excelencia de las tres revistas sometidas a examen es incuestionable, no sólo por su trayectoria, sino porque las tres han sido y son objeto de deseo de grandes científicos en el campo de la Psicología o afines que deciden publicar en ellas sus trabajos. Proseguimos.

En el proceso de investigación desde que comienza hasta que se redactan los resultados existe un orden lógico en el que se debe considerar cada aspecto y cada matiz de la misma. Hemos tratado de ordenar el discurso del apartado anterior en ese mismo orden, y lo hemos hecho así para destacar que la planificación de una investigación, sea la que sea, (cualitativa o cuantitativa, experimental o no experimental, etc.) es en sí un ejercicio estructurado que exige considerar la validez en toda su diversidad, aunque con distinta profundidad en función de la finalidad perseguida.

Así pues, para comenzar a planificar una investigación cuasi-experimental es preciso reconocer que las condiciones son SIEMPRE adversas debido a la ausencia de aleatorización. Esta falta de control exige que el investigador se ponga en alerta y comience a pensar que los supuestos que debe considerar son muchos y que ponerlos en evidencia no va a ser tarea fácil. Lo primero que debe hacer es familiarizarse con el contexto en el que se va a celebrar la investigación y conocer en profundidad las características de la población en la que desea poner a prueba su hipótesis. Evidentemente a estas alturas las hipótesis las debe tener perfectamente definidas y el modelo explicativo también. En este momento se produce uno de los pasos más delicados, que es traducir o interpretar las piezas de las hipótesis para hacerlas operativas. Es el momento de cuidar la validez de constructo (referente a sujetos, medidas, momentos, tratamiento, variables dependientes, etc.) porque será la que nos va a permitir cuidar la formación de los grupos, aplicar el tratamiento "con integridad", y observar sus efectos en la extensión deseada para que la validez interna en la parrilla de salida sea impecable. El modelo explicativo de sus hipótesis y las posibilidades con las que cuenta para ponerlas a prueba van a determinar el diseño de recogida de datos que garantizará que la validez interna no se malogre. En este punto la validez de la conclusión estadística ya está casi en su totalidad sentenciada. Si en lo que resta el investigador cuida que el modelo explicativo de sus hipótesis también sea el que dirija la elección del modo de análisis, y en este punto usa los estadísticos de prueba del modo que sea correcto y pertinente, el error de Tipo I y de Tipo II los tendrá controlados. A esto es lo que nosotros denominamos *Validez Estructurada* y condición *sine qua non* para que TODA investigación alcance niveles óptimos de calidad. Conforme nos alejamos de la libertad que permite la asignación aleatoria es preciso molestarse más para hacer el mismo recorrido.

En resumen, la validez siempre va a venir explicada por un modelo no aditivo en el que aparecen enhebrados los cuatro tipos de validez del modelo capitaneado por Campbell. Es la Validez Estructurada. Una investigación tiene Validez Estructurada cuando, en primer lugar se ha cuidado la validez de constructo. Cuidada ésta, es posible la validez interna. Satisfechas ambas posibilitan, hacen eficaz la validez de conclusión estadística. La validez de conclusión estadística no es validez interna, pero si el modelo que explica las hipótesis ha sido el correcto, la validez de conclusión estadística consolidaría la validez interna. La validez externa y ecológica en parte están delimitadas por la validez de constructo inicial, pero no tienen sentido si la validez interna no se logra. Cook & Campbell (1979, p.80-85) ya habían comentado la relación que había entre validez interna y de conclusión, de una parte, y validez de constructo y validez externa de otra. Nosotros hemos querido hacer una modesta aportación a su modelo de validez.

> In science we are like sailors who must repair a rotting ship while it is afloat at sea. We depend on the relative soundness of all other planks while we replace a particularly weak one. Each of the planks we now depend on we will in turn have to replace. No one of them is a foundation, for point of certainty, no one of them is incorrigible. (Campbell, 1969, p. 43).

Recomendaciones prácticas que nos permitirán salir a flote:

1.- Al principio de este trabajo redactamos que la dama es al ajedrez lo que la aleatorización al método científico, y pensamos que así es, sin embargo, es la posición del rey la que gana o pierde la partida. Los maestros del método nos han inculcado que la aleatorización no hace bueno el método, el buen método se hace gestionando bien los recursos.

2.- Sólo es posible gestionar bien los recursos trabajando con Validez Estructurada. Es el único modo de evitar la acumulación de amenazas, activos tóxicos venidos desde distintos frentes, siendo uno de los más vulnerables el que a los grupos de control se refiere. Si éstos no están bien formados, o aún estándolo, no son adecuados para poner a prueba la hipótesis que se pretende, lejos de ser un valor, se convierten en testigos incómodos capaces de arruinar la investigación.

3.- Nos guste o no, las hipótesis se ponen a prueba mediante alguna técnica estadística, y como antes hemos comentado, las ideas excelentes (hipótesis) no hacen válidos ni siquiera los más potentes procedimientos de análisis de datos. Si la validez de conclusión estadística está desnuda de masa sustantiva habrá una conclusión estadística que se mantendrá dentro de unos márgenes de error (cuando por ejemplo tenemos mucho tamaño de nuestra pero sin considerar el tamaño del efecto), pero la probabilidad de no ser válida (validez interna) será muy alta.

4.- Resulta fatigoso leer el apartado de resultados en muchos trabajos publicados. Es imperativo embridar el derroche en el análisis de los datos. Los análisis no se hacen bien, más que por no elegir correctamente el estadístico, por el modo de hacerlo. Por ejemplo, es cierto que el análisis de la varianza es imperfecto en algunas situaciones, pero repetirlo múltiples veces conduce a generar un error de tipo I insostenible. Esta conducta ha sido y sigue siendo criticada hasta la saciedad en múltiples revisiones en cualquier área de investigación (*e.g.,* Schochet, 2009b).

5.- La conducta compulsiva anterior nos lleva a pensar que muchos investigadores quizás tengan la creencia errónea de que tener validez es equivalente a rechazar la Ho, con lo importantes que son las Ho cuando de significación práctica se trata, o simplemente para conocer qué camino nos lleva a ninguna parte, que no es poco. Rechazar la hipótesis nula (y su correspondiente análisis forense) sólo es un punto de llegada (y siempre provisional) si antes hemos demostrado que las hipótesis alternativas (que hasta el momento conocemos o podemos poner a prueba) son falsas.

6.- Es imperativo reconocer que no somos autosuficientes, y aunque estemos versados en el tema que investigamos conviene recurrir a un experto en metodología y análisis de datos que nos ayude a planificar la investigación que nos permita poner a prueba del mejor modo posible nuestras hipótesis y a analizar correctamente los datos (Altman, 2006; Harris et al., 2004, 2005; Shadish et al., 2002; Shardell et al*.,* 2007, int.al).

Bobby Fisher reconocía que había aprendido a jugar al ajedrez observando el juego de los clásicos, de los jugadores románticos del Siglo XIX y principios del XX que gustaban de comienzos difíciles, con sacrificios de piezas que ponían a prueba el talento e imaginación de quien defendía uno y otro color en el tablero. Pensamos que para investigar "bien y mejor" es absolutamente necesario saber lo que nos falta por aprender, y ahora ya lo sabernos, al menos parte. Es bueno también tener algún modelo de referente. En este sentido, y para abrir boca, recomendamos leer el trabajo de una de las investigaciones cuasi-experimentales realizadas de modo magistral por Campbell y Ross (1968), *The Connecticut crackdown on speeding: Time-series data in quasiexperimental analysis.*

Así fue cómo en plena guerra fría se produjo la chispa que impulsaría, en el mundo del ajedrez, la carrera de uno de los más grandes genios de la historia, y en la ciencia, el punto de partida de una obra metodológica extraordinaria.

Agradecimientos.- Este trabajo ha sido apoyado por una beca que el Ministerio Español de Ciencia e Innovación (Ref: PSI-2011-23.395) otorgó a los autores.

Adenda

A lo largo de artículo se ha hecho referencia a cinco ficheros adicionales. Todos ellos se encuentran en la página http://www.unioviedo.es/dise_investigacion/. Los ficheros son:

Fichero Adicional 1: descripción de todas las variables registradas en todas sus categorías de clasificación.

Fichero Adicional 2: redacción de los resultados obtenidos libres de discusión.

Fichero Adicional 3: exposición de los resultados obtenidos en 12 Tablas.

Fichero Adicional 4: referencias bibliográficas seleccionadas sobre investigación Cx. (y otras investigaciones no aleatorizadas), algunas de carácter general y otras más específicas referidas a distintos ambientes disciplinares (educativo, clínico, etc.), también sobre causalidad.

Fichero Adicional 5: referencias bibliográficas seleccionadas sobre aspectos concretos y específicos de las investigaciones Cx. (y en general de las investigaciones no aleatorizadas): análisis de datos (para D-GCNE, D-DR y de series temporales), sesgo de selección, mediación, pérdida de datos, evaluación de programas, Método mixto, Modelos Multinivel, Modelos de Ecuaciones Estructurales, etc.

Referencias

Altman, D. G. (2006). *Practical statistics for medical research*. New York: Chapman & Hall.

Anderson, R. A. (2012). Reply: Intention-to-treat and per-protocol analyses. *Human Reproduction, 27*, 3118-3119.

Arnau, J. (1995). Metodología de la investigación en psicología. En M. T. Anguera, J. Arnau, M. Ato, R. Martínez, J. Pascual, J., y G. Vallejo, G. (Eds.), *Métodos de investigación en Psicología* (Cap. 1). Madrid: Síntesis.

Ato, M. (1995). Tipología de los diseños cuasi-experimentales. En M. T. Anguera, J. Arnau, M. Ato, R. Martínez, J. Pascual, J., y G. Vallejo, G. (Eds.). (1995). *Métodos de investigación en Psicología* (Cap. 9). Madrid: Síntesis.

Ato, M., y Vallejo, G. (2011). Los efectos de terceras variables en la investigación psicológica. *Anales de Psicología, 27*, 550-561.

Avellar, S., & Paulsell, D. (2011). *Lessons Learned from the Home Visiting Evidence of Effectiveness Review*. Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Washington, DC. Recuperado de http://homvee.acf.hhs.gov/Lessons_Learned.pdf.

Bai, A., Shukla, V. K., Bak, G., & Wells, G. (2012). *Quality assessment tools project report.* Ottawa: Canadian Agency for Drugs and Technologies in Health.

Brunswik, E. (1956). *Perception and the representative design of psychological e*xperiments (2nd. Ed.). Berkeley, CA: University of California Press.

Burchett, H., Umoquit, M., & Dobrow, M. J. (2011). How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. Journal of *Health Services Research & Policy, 16*, 238-44. .

Cambon, L., Minary, L., Ridde, V., & Alla, F. (2012). Transferability of interventions in health education: a review. *BMC Public Health, 12*(497).

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Campbell, D.T. (1969). A phenomenology of the other one: Corrigible, hypothetical and critical. En T Mischel (Ed.), *Human Action: conceptual and empirical issues* (pp. 41-69). New York: Academic Press.

Campbell, D. T., & Ross, H. L. (1968). The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis. *Journal of the Law and Society Association, 3*, 33-54.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally. (Reprinted as Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966).

Castro, G., & Mastropieri, M. (1986). The efficacy of early intervention programs: A meta-analysis. *Exceptional Children, 52*, 417-424.

Centre for Reviews and Dissemination, CRD. (2010). *Finding studies for systematic reviews: A resource list for researchers*. New York: University of York. Recuperado de http://www.york.ac. uk/inst/crd/revs.htm [accessed 13 Sep 2012].

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49,* 997-1003.

Cohen, J. (1992). Cosas que he aprendido (hasta ahora). *Anales de Psicología, 8*, 3-17.

Cook, T.D. (2000). The false choice between theory-based evaluation and experimentation. *New Directions for Evaluation, 87*, 27-34.

Cook, T. D. (2006) Describing what is special about the role of experiments in contemporary educational research: putting the "Gold Standard" rhetoric into perspective. *Journal of MultiDisciplinary Evaluation, 3*, 1-7.

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics, 1422,* 636-54.

Cook, T. D., & Campbell, D. T. *(*1976*).* The design and conduct of quasi-experiments and true experiments in fields settings. *En* M. Dunnette *(Ed.), Handbook of industrial and organizational psychology (pp.* 223-326*).* Chicago*:* Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi-experimentation. En M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 491-576). Palo Alto, CA: Consulting Psychologists Press.

Cook, L., Cook, B. G., Landrum T. J., & Tankersley, M. (2008). Examining the role of group experimental research in establishing evidenced-based practices. *Intervention in School and Clinic, 44*, 76-82.

Cook, T. D., & Gorard, S. (2007). Where does good evidence come from? International *Journal of Research and Method in Education, 30*, 307-323.

Cook, T. D., Scriven, M., Coryn, C. L. S., & Evergreen, S. D. H. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. American Journal of Evaluation *March, 31*, 105-117.

Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining Evidence-Based Practices in Special *Education. Exceptional Children, 75*, 365-383.

Cook, T. D., & Wong, V. C. (2008). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique, 91-92,* 127-150.

Cooper, C.A., McCord, D. M., & Socha, A. (2010). Evaluating the college sophomore problem: The case of personality and politics. *The Journal of Psychology: Interdisciplinary and Applied, 145*, 23-37.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Datta, L. E. (2007). *What are we, chopped liver? or why it matters if the comparisons are active and what to do*. The Evaluation Center, The Evaluation Café. Recuperado de http://www.wmich.edu/evalctr/wp-content/uploads/2010/05/chopped-liver.pdf.

Devito, D. A., Song, M. K., Hawkins, R., Aubrecht, J., Kovach, K., Terhorst, L., … Callan, J. (2011). An Intervention Fidelity Framework for Technology-Based Behavioral Interventions. *Nursing research, 60*, 340-347.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., ... Altman, D.G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27), 1-173.

Des Jarlais, D. C., Lyles, C., Crepaz, N., & TREND Group (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: (2004). The TREND Statement. *American Journal of Public Health, 94*, 361-366.

Donaldson, S., & Christie, C. (2005). The 2004 Claremont debate: Lipsey vs. Scriven. Determining causality in program evaluation and applied research: Should experimental evidence be the gold standard? *Journal of Multidisciplinary Evaluation, 3*, 60-77.

Dunst, C. J., & Rheingrover, R. (1981) An analysis of the efficacy of early intervention programs with organically handicapped children. *Evaluation and Program Planning, 4*, 87-323.

Eastmond, N. (1998). Commentary: when funders want to compromise your design. *American Journal of Evaluation, 19*, 392-395.

Farran, D. C. (1990). Effects of intervention with disadvantaged and disabled children: A decade review. En S. J. Meisels & J. P. Shonkoff (Eds.), *Handbook of early childhood intervention* (pp. 501-539). New York: Cambridge University Press.

Farran, D. C. (2000). Another decade of intervention for children who are low income or disabled: What do we know now? En J. P. Shonkoff & S. J. Meisels (Eds.), *Handbook of early childhood intervention* (2nd ed., pp. 510-548). New York: Cambridge University Press.

Ferguson, L. (2004). External validity, generalizability, and knowledge utilization. *Journal of Nursing Scholarship***,** *36*(1)**,** 16-22.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23,* 132-138.

Funnell, S. C., & Rogers, P. J. (2011). *Purposeful program theory: Effective use of theories of change and logic models (Research Methods for the Social Sciences)*. San Francisco: John Wiley & Sons.

Gage, N. L. (Ed.) (1963). *Handbook of research on teaching*. Chicago: Rand McNally.

Gersten, R., Baker, S. K., Smith-Johnson, J., Flojo, J. R., & Hagan-Burke, S. (2004). A tale of two decades: Trends in support of federally funded

experimental research in special education. *Exceptional Children, 70,* 323-332.
Gibbert, M., & Ruigrok, W. (2010). The "What" and "How" of Case Study Rigor: Three Strategies Based on Published Work. Organizational Research Methods, 13, 710-737.
Glasgow, R. E., & Emmons, K. M. (2007). How can we increase translation of research into practice? Types of evidence needed. *Annual Review of Public Health, 28*, 413-33.
Grant, A. M., & Wall, T. D. (2009). The Neglected Science and Art of Quasi-Experimentation Why-to, When-to, and How-to Advice for Organizational Researchers. *Organizational Research Methods, 12*, 653-686.
Green, L. W., & Glasgow, R. E. (2006). Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. *Evaluation & the Health Professions, 29*(1), 126-53.
Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research, 2*, 109-112.
Harris, A. D., Bradham, D. D., Baumgarten, M., Zuckerman, I. H., Fink, J. C., & Perencevich, E. N. (2004). The use and interpretation of quasi-experimental studies in infectious diseases. *Clinical Infectious Diseases, 38*, 1586-91.
Harris, A. D. Lautenbach, E., & Perencevich, E. (2005). A systematic review of quasi-experimental study designs in the fields of infection control and antibiotic resistance. *Clinical Infectious Diseases, 41*, 77-82.
Harris, A. D. McGregor, J. D., Perencevich, E. N., Furuno, J. P., Zhu, J., Peterson, D. E., & Finkelstein, J. (2006). The Use and Interpretation of Quasi-Experimental Studies in Medical Informatics. *Journal of the American Medical Informatics Association, 13*, 16-23.
Highhouse, S. (2009). Designing experiments that generalize. Organizational Research Methods. *Organizational Research Methods*, *12*, 584-566.
Highhouse, S., & Gillespie, J. Z. (2010). Do samples really matter that much? En C. E. Lance & R. J. Vandenberg (Eds.) *Statistical and methodological myths and urban legends* (pp. 247-262). New York: Taylor & Francis Group.
Higgins, J. P. T., & Green, S. (Eds.) (2011). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0. The Cochrane Collaboration. Recuperado de http://www.cochrane.es/files/handbookcast/Manual_Cochrane_510.pdf.
Heisman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1*, 154-169.
Hsieh, P., Acee, T., Chung, WH., Hsieh, YP., Kim, H., Thomas, G.D., … Robinson, D.H. (2005). Is Educational Intervention Research on the Decline? *Journal of Educational Psychology, 97*, 523-529.
Huitema, B. E. (2011).*The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies* (2da. Ed.). New York: John Wiley.
ICH Expert Working Group. (2000). Choice of control group and related issues in clinical trials. ICH Harmonised Tripartite Guideline. Current Step 4. E10. Recuperado de http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf.
Jensen, D. D., Fast, A. S., Taylor, B. J., & Maier, M. E. (2008). *Automatic Identification of Quasi-Experimental Designs for Discovering Causal Knowledge*. 14th ACM SIGKDD, International Conference on Knowledge and data mining. August 24–27, Las Vegas, NV, USA.
Johnston, M. V., Ottenbacher, K. J., & Reichardt, C. S. (1995). Strong quasi-experimental designs for research on the effectiveness of rehabilitation. *American Journal of Physical Medicine Rehabilitation, 74*, 383-92.
Judd, C.M., y Kenny, D.A. (1981). *Estimating the effects of social interventions*. Cambridge, England: Cambridge University Press.
Kenny, D. A. (1979). *Correlation and causality*. New York: John Wiley.
Keselman, H. J., Huberty, C.J., Liz, L.M., Olejnik, S., Cribbie, R., Donahue, B., … Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.
Kunstmann, L. N., y Merino, E. J. M. (2008). El experimento natural como un nuevo diseño cuasi-experimental en investigación social y de salud. *Ciencia y Enfermería XIV* (2), 9-12.
Kunz, R., & Oxman, A. D. (1998). The unpredictability paradox: review of empirical comparisons of randomised and nonrandomised trials. *British Medical Journal, 317,* 1185-1190.
Lesik, S. (2006). Applying the regression-discontinuity design to infer causality with non-random assignment. *Review of Higher Education*, *30*, 1-19.
Li, L. C., Moja, L., Romero, A., Sayre, E. C., & Grimshaw, J. M. (2009). Nonrandomized quality improvement intervention trials might overstate the strength of causal inference of their findings. *Journal of Clinical Epidemiology, 62*, 959-66.
Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational,and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181-1209.
MacLehose, R. R., Reeves, B. C., Harvey, I. M., Sheldon, T. A., Russell, I. T., & Black, A. M. (2000). A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment, 4,* 1-154.
McCaffery, K. J., Turner, R., Macaskill, P., Walter, S.D., Chan, S. F., & Irwig L. (2011). Determining the impact of informed choice: separating treatment effects from the effects of choice and selection in randomized trials. *Medical Decision Making, 31*, 229-236.
McLeod, B. D., & Islam, N. Y. (2011). Using Treatment Integrity Methods to Study the Implementation Process. *Clinical Psychology. Science and practice, 18*, 36-40.
McNemar, Q. (1946). Opinion attitude methodology. *Psychological Bulletin*, 43, 289-374.
Mara, C., & Cribbie, R. A. (2012). Paired-samples tests of equivalence. *Communications in Statistics: Simulation and Computation, 41*, 1928-1943.
Mara, C., Cribbie, R. A., Flora, D., LaBrish, C., Mills, L., & Fiksenbaum, L. (2012). An improved model for evaluating change in randomized pretest, posttest, follow-up designs. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8*, 97-103.
Marcantonio, R. J., & Cook, T. D. (1994). Convincing quasi-experiments: The interrupted time series and regression-discontinuity designs. En J. S.Wholey, H.P. Hatry, & K.E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 133-154). San Francisco: Jossey-Bass.
Marcus, S. M., Stuart, E. A., Wang, P., Shadish, W. R., & Steiner, P. M. (2012). Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychological Methods, 17*, 244-54.
Nezu, A. M., & Nezu, C. M. (Eds.) (2008*). Evidenced-based outcome research: A practical guide to conducting randomized clinical trials for psychosocial interventions.* New York: Oxford Press.
Orwin, R. G. (1997). Twenty-one years and counting: The interrupted time series design comes of age. En E. Chelimsky & W. R Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp.443-465)**.** Thousand Oaks, CA: Sage.
Paluck, E. L., & Green, D. P. (2009). Prejudice Reduction: What Works? A Review and Assessment of Research and Practice. *Annual Review of Psychology, 60*, 339-367.
Perepletchikova, F., Hilt, L. M., Chereji, E., & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: Survey of treatment outcome researchers. *Journal of Consulting and Clinical Psychology, 77*, 212-218.
Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic Methods for Questions Pertaining to a Randomized Pretest, Posttest, Follow-Up Design. *Journal of Clinical Child and Adolescent Psychology, 32*, 467-486.
Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, *5*, 83-104.
Reichardt, C. S., & Henry, G. T. (2012). Regression-discontinuity designs. En H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf & K J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 511-526). Washington, DC US: American Psychological Association.
Rossi, P. H., & Freeman, H. (1985). *Evaluation: A systematic approach*. Beverly Hills, CA: Sage Publications.
Ruscio, J., & Roche, B. (2012). Variance heterogeneity in Published Psychological Research: A Review and New Index. *Methodology, 8*, 1-11.

Sacks, H. S., Chalmers, T. C., & Smith, H. (1982). Randomized versus historical controls for clinical trials. *The American Journal of Medicine, 72,* 233-240.

Sacks, H. S., Chalmers, T. C., & Smith, H. (1983). Sensitivity and specificity of clinical trials: Randomized v historical controls. *Archives of Internal Medicine, 143*, 753-755.

Sánchez, J., Valera, A., Velandrino, A. P., y Marín, F. (1992). Un estudio de la potencia estadística en Anales de Psicología. *Anales de Psicología, 8*, 19-32.

Scandura, T. A., & Williams E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal, 43*, 1248-1264.

Scheirer, M. A. (1998). Commentary: Evaluation Planning is The Heart of the Matter. *American Journal of Evaluation, 19*, 385-391.

Schochet, P. Z. (2009a). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics*, *34*, 238-266.

Schochet, P. Z. (2009b). An approach for addressing the multiple testing problem in social policy impact evaluations. *Evaluation Review, 33,* 539-567.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). *Standards for regression discontinuity designs*. Washington DC: U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. Recuperado de http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Schulz, R., Czaja, S. J., McKay, J. R., Ory, M. G., & Belle, S. H. (2010). Intervention Taxonomy (ITAX): Describing Essential Features of Interventions (HMC). *American journal of health behavior, 34,* 811-821.

Sedgwick, P. (2011). Analysis by per protocol. *British Medical Journal, 342*:d2330

Seethaler, P. M., & Fuchs, L. S. (2005). A drop in the bucket: Randomized controlled trials testing reading and math interventions. *Learning Disabilities Research & Practice, 20,* 98-102.

Shadish, W. J., & Cook, D. T. (2009). The renaissance of experiments. *Annual Review of Psychology,* 60, 607-29.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Shadish, W. R., & Heinsman, D. T. (1997). Experiments versus quasi-experiments: Do you get the same answer? En W. J. Bukoski (Ed.), *Meta-Analysis of Drug Abuse Prevention Programs* (pp. 147-164). Washington DC: Superintendent of Documents.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random to cutoff-based assignment. *Psychological Methods, 16*, 179-191.

Shadish, W., & Myers, D. (2004a). *Research Design Policy Brief*. Recuperado de http://www.campbellcollaboration.org/artman2/uploads/1/C2_Research_Design_Policy_Brief-2.pdf.

Shadish, W., & Myers, D. (2004b). *How to make a Campbell Collaboration Review: The Review*. Documento elaborado para Nordic Campbell Center (The DNIHSR). Recuperado de http://www.sfi.dk/graphics/Campbell/Dokumenter/For_Forskere/guide_3_review_samlet20DEC04.pdf.

Shadish, W. R., & Ragsdale, K. (1996). Randomized versus quasi-experimental designs: A metaanalysis investigating the relationship between random assignment and effect size. *Journal of Consulting and Clinical Psychology, 64,* 1290-1305.

Shah, P. B. (2011). Intention-to-treat and per-protocol analysis. *Canadian Medial Association Journal, 183*, 696.

Shahar, E., & Shahar, D. J. (2009). On the causal structure of information bias and confounding bias in randomized trials. *Journal of Evaluation in Clinical Practice, 15*, 1214-1216.

Shardell, M., Harris, A. D., El-Kamary, S.S., Furuno, J. P., Miller, R. R., & Perencevich, E. N. (2007). Statistical Analysis and Application of Quasi Experiments to *Antimicrobial Resistance Intervention Studies. Antimicrobial Resistance, 45,* 901-907.

Snyder, P., Thompson, B., Mclean, M. E., & Smith, B. J. (2002). Examination of Quantitative Methods Used in Early Intervention Research: Linkages With Recommended Practices**.** *Journal of Early Intervention, 25*, 137-150.

Sridharan, A., & Nakaima, A. (2011). Then steps to making evaluation matter. *Evaluation and Program Panning, 34*, 135-146.

Steckler, A., & McLeroy, K. R. (2008). The importance of external validity. *American Journal of Public Health, 98*, 9-10.

Steiner, P. M., Cook; T. D., Shadish, W. R., & Clark M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods. 15*, 250-67.

Stone, S. P., Cooper, B. S., Kibbler, C. C., Cookson, B. D., Roberts, J. A., Medley, G. F., … Davey, P. G. (2007). The ORION statement: guidelines for transparent reporting of outbreak reports and intervention studies of nosocomial infection. *The Lancet Infectious diseases, 7*, 282-288.

Swaen, G., Teggeler, O., & van Amelsvoort, L. (2001) False positive outcomes and design characteristics in occupational cancer epidemiology studies. *International Journal of Epidemiology, 30*, 948-954.

Thomson, H. J., & Thomas, S. (2012). External validity in healthy public policy: application of the RE-AIM tool to the field of housing improvement. *BMC Public Health, 12*, 633.

Trochim, W. M. K. (1984). *Research Design for Program Evaluation: The Regression- Discontinuity Approach*. Beverly Hills: Sage.

Trochim, W. M. K. (2001). *The research methods knowledge base*. Cincinnati: Atomic Dog Publishing.

Valentine, J. C., & Cooper, H. (2003). *What Works Clearinghouse Study Design and Implementation Assessment Device* (Version 1.0). Washington, DC: U.S. Department of Education. Recuperado de http://www.w-w-c.org/standards.html.

Valera, A., Sánchez Meca, J., y Marín, F. (2000). Contraste de hipótesis e investigación psicológica española: análisis y propuestas. *Psicothema, 12*, 549-552.

Valera, A., Sánchez Meca, J., Marín, F., y Velandrino, A. P. (1998). Potencia estadística de la revista de Psicología General y Aplicada (1990-1992). *Revista de Psicología General Aplicada, 51*, 233-246.

Vallejo, G. (1995). Diseños de series temporales interrumpidas. En M. T. Anguera, J. Arnau, M. Ato, R. Martínez, J. Pascual, J., y G. Vallejo, G. (Eds.), *Métodos de investigación en Psicología* (Cap. 12). Madrid: Síntesis.

Venter, A., Maxwell, S. E., & Bolig, E. (2002). Power in randomized group comparisons: The value of adding a single intermediate time point to a traditional pretest-posttest design. *Psychological Methods, 7,* 194-209.

Walter, S. D., Turner, R. M., Macaskill, P., McCaffery, K. J., & Irwig, L. (2012). Optimal allocation of participants for the estimation of selection, preference and treatment effects in the two-stage randomised trial design. *Statistics in Medicine, 31*, 1307-1322.

Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *Annals of the American Association of Political and Social Science, 578*, 50-70.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., … Mullen, P. D. (2008). Alternatives to the Randomized Controlled Trial. *American Journal of Public Health*, *98*, 1359-1366.

Whitea, I. R., Carpenterb, J., & Hortonc, N. J. (2012). Including all individuals is not enough: Lessons for intention-to-treat analysis. Clinical Trials, 9, 396-407.

Wiecko, F. M. (2010). Research Note: Assessing the validity of college samples: Are students really that different? *Journal of Criminal Justice, 38*, 1196-1190.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

(Artículo recibido: 22-1-2013; revisado: 20-5-2013; aceptado: 22-7-2013)