

VALIDEZ DE LA INVESTIGACIÓN CAUSAL. TIPOLOGÍAS Y EVOLUCIÓN

Para alcanzar un necesario prestigio social de las Ciencias de la Educación resulta fundamental que las investigaciones educativas tengan un alto nivel de Validez, algo que no siempre se consigue. A pesar de que es unánimemente reconocida la labor de Campbell y sus colaboradores (Stanley, Cook, etc.) en el análisis del problema de la Validez de Investigación, una revisión de los textos de autores españoles escritos en los últimos años sobre Investigación Educativa revela que ninguno de ellos sigue los modelos de Validez de la Investigación expuestos por Campbell y sus colaboradores desde 1979, sino que en todo caso se refieren (o plantean diversas variaciones) a modelos más antiguos. Para intentar clarificar esta cuestión hacemos un repaso de la evolución y del estado de la cuestión en los últimos años, así como un análisis crítico de la situación actual del tema.

Es evidente que no todos los diseños de investigación educativa están orientados a determinar las causas de determinados fenómenos o las consecuencias de ciertas intervenciones. Sin embargo, en una disciplina aplicada como es la educación, este tipo de planteamientos tienen una importancia fundamental, hasta el punto de que en los últimos años se han desarrollado procedimientos diversos para intentar contrastar modelos causales a partir, incluso, de estudios de carácter descriptivo, tales como el análisis de vías o el análisis de relaciones estructurales lineales (LISREL) (ver Tuijnman y Keeves, 1997, para una revisión de la utilidad y limitaciones de estos procedimientos). Por tanto, resulta una de las principales áreas de interés de la investigación educativa (Cook y Campbell, 1979), por lo que ha venido prestándosele una notable atención.

Durante mucho tiempo ha sido una dolorosa evidencia que la investigación en Ciencias Sociales en general, y la educativa en particular, han estado, en general, muy mal realizadas, y que «contienen un número desproporcionado de malos estudios» (Cordray, 1986). No en balde, la decepción con los resultados obtenidos en

la investigación educativa son los que vinieron a fomentar la búsqueda de modelos y paradigmas nuevos que solventaran los problemas y limitaciones planteados (Wienert, 1997).

Entre los muchos ejemplos disponibles citaremos dos que nos parecen especialmente ilustrativos. La Oficina de Auditoría General de los Estados Unidos realizó en 1984 un estudio para evaluar el impacto de los programas de apoyo a niños, jóvenes y mujeres. Se revisaron sesenta y un trabajos de investigación sobre un tópico específico (la influencia de los programas sociales sobre el peso de los niños al nacer), de los que hubo que descartar cincuenta y cinco por problemas metodológicos y, consiguientemente, por su escasa credibilidad (citado por Cordray, 1986)

En otro ejemplo significativo de investigación educativa fue llevado a cabo a principios de esta década, el Departamento (Ministerio) de Educación de Estados Unidos solicitó al Consejo Nacional de Investigación que evaluara los principales estudios sobre los diversos tipos de educación bilingüe, a fin de determinar cuales eran los

mejores procedimientos disponibles. Los resultados de la evaluación (Meyer y Conrad, 1992) fueron frustrantes, destacando, entre otras conclusiones, las siguientes:

- «Los estudios no autorizan la conclusión de que ningún tipo de programa sea superior a ningún otro, ni que los programas sean igualmente efectivos. Incluso si alguno de los programas hubiera sido superior, los estudios, tal y como se planificaron y llevaron a cabo, podrían haber fracasado identificar tal circunstancia»: (p.104)
- «Los estudios (revisados) adolecen de una excesiva atención al uso de elaborados métodos estadísticos para intentar solventar las deficiencias en los diseños de investigación. Los métodos de análisis estadístico no pueden corregir los fallos en la conceptualización, diseño y ejecución de los estudios» (p. 104).
- «Dada la heterogeneidad de situaciones investigadas que hacen imposible realizar comparaciones adecuadas, el equipo evaluador recomienda que se lleven a cabo estudios mejor orientados y fundados teóricamente para analizar la interacción de diferentes planteamientos instruccionales en contextos de educación bilingüe de comunidades con características específicas» (p. 105).
- «Aunque el informe final (de uno de las principales investigaciones revisadas) afirma que los tres programas (tratamientos) representan tres modelos instruccionales distintos, los datos revelan que los programas no eran tan distintos. Fueron esencialmente diferentes versiones del mismo tratamiento; los programas de Inmersión lingüística y de Integración Temprana en ocasiones no podían distinguirse entre sí» (p. 103).

Cordray (1986) indica que esta mala calidad de la investigación amenaza la misma reputación del campo, lo que por ejemplo, podría justificar que, sin mayor problema, desde alguna instancia política se haya criticado recientemente el trabajo del INCE sobre evaluación de la ESO «por su escasa fiabilidad» (*sic*). Puede haber algunas razones que justifiquen esta mala calidad de las investigaciones, entre las que podríamos mencionar, siguiendo parcialmente a Cordray:

- Los métodos de investigación en Ciencias Sociales no son adecuados (o capaces) de explicar la compleja realidad social y humana. Planteamientos de este tipo han estado a la base, por ejemplo, de la creciente atención que se ha prestado en los últimos años a modelos alternativos, como la investigación denominada cualitativa o interpretativa (Keeves, 1997; Walker y Evers, 1997). Aunque los planteamientos puedan ser prometedores, la mejora que estos nuevos procedimientos puedan traer a la investigación educativa en su conjunto están aún por demostrar. Se han desarrollado incluso desde esta perspectiva cualitativa criterios de calidad de la investigación propios (Sandín, 2000), con una aceptación desigual, y que no revisaremos aquí porque su análisis excedería a los limitados objetivos de este trabajo (ver Aliaga, 1998 para un análisis más detallado de este tema). En cualquier caso conviene recordar aquí que difícilmente puede acusarse a las teorías clásicas sobre la validez de la investigación de «cuantitativas», ya que algunos de sus principales autores han manifestado su simpatía por los acercamientos cualitativos, e incluso en cierta medida han propiciado su crecimiento (Campbell, 1974; Cronbach, 1974).
- Los métodos de investigación en Ciencias Sociales están aún poco evolucionados y no se han desarrollado suficientemente como para ser capaces de detectar la eficacia de los distintos tratamientos. El desarrollo de una cierta proporción de investigaciones bien controladas y fiables (como, por ejemplo, en el caso de seis de los sesenta y un estudios revisados por la Oficina de Auditoría General de los Estados Unidos, 1984) nos permite descartar esta alternativa
- El trabajo empírico en Ciencias Sociales supone complicaciones específicas que hacen muy difícil llevar a cabo investigaciones válidas. Algunos de los problemas de este tipo han sido analizados en los últimos treinta años, básicamente gracias al trabajo de Campbell y sus colaboradores (Campbell y Stanley, 1966; Cook y Campbell, 1979; Campbell, 1986; Cook, Campbell y Peracchio, 1990), espolcados por planteamientos críticos como los de Cronbach (1982), de tal modo que se han desa-

rollado procedimientos específicos, como es el caso de los diseños cuasi—experimentales, que posibilitan, cuando menos, una comprensión global del proceso de investigación en las particulares circunstancias que se dan en las Ciencias Sociales y Humanas.

- Los profesionales e investigadores no hemos comprendido y/o aplicado adecuadamente los conocimientos desarrollados sobre diseños de investigación y los condicionantes a su validez que, conjuntamente con el desarrollo de los procedimientos de investigación (particularmente la cuasiexperimentación), se han desarrollado en las últimas décadas.

Existen algunos indicios que podrían permitirnos suponer que, al menos parcialmente, esta última posibilidad pudiera tener ciertos visos de credibilidad. Algunas revisiones sobre los procedimientos analíticos empleados en la investigación educativa en España (por ejemplo, Sáez et al., 1994; Suárez et al., 1995) revelan una mala utilización de los procedimientos más complejos de análisis estadístico. Ya Campbell y Stanley (1966) clamaban porque «una capacitación más a fondo de los investigadores educacionales en técnicas modernas de estadística experimental permitiría elevar la calidad de la experimentación pedagógica» (p. 14 de la edición en español).

Desde el punto de vista de los diseños, y más concretamente, respecto al tema de la validez de investigación, existe una heterogeneidad de concepciones tal que parece haber una diversidad de opiniones enorme. Revisando sólo los planteamientos mostrados en algunos de los principales textos sobre investigación educativa que han aparecido últimamente en nuestro país, diremos que Buendía, Colás y Hernández Pina (1997), cuando se refieren a la validez de la investigación experimental, distinguen dos grandes tipos, la Validez Interna y la Validez Externa, dentro de la cual engloban la Validez de Población, la Validez Ecológica, y la Validez de Constructo (p.82). Por su parte Latorre, Del Rincón y Arnal (1996) plantean que "hay que tener en cuenta tres tipos de validez: la interna, que exige el control; la externa (generalización) y la conceptual (coherencia)" (p. 127). Colás y Buendía (1994) señalan que los diseños experimentales deben responder a tres clases de validez (p. 111): Validez Interna, Validez Externa, y Validez Ecológica. Bisquerra (1989) distingue por un

lado entre Validez Interna y Validez Externa, añadiendo un apartado residual de «otros tipos de validez», en los que incluye la validez ecológica, la validez de población, la validez de conclusión estadística y la validez de constructo. López-Barajas, López López y Pérez Juste (1987) distinguen también entre Validez Interna y Validez Externa, dentro de la cual distinguen entre validez de muestras, de variables y ecológica. Pérez Juste (1985), en su revisión de los diseños experimentales, hace referencia sólo a la Validez Interna y a Validez Externa, siguiendo el primer modelo planteado por Campbell y Stanley (1966).

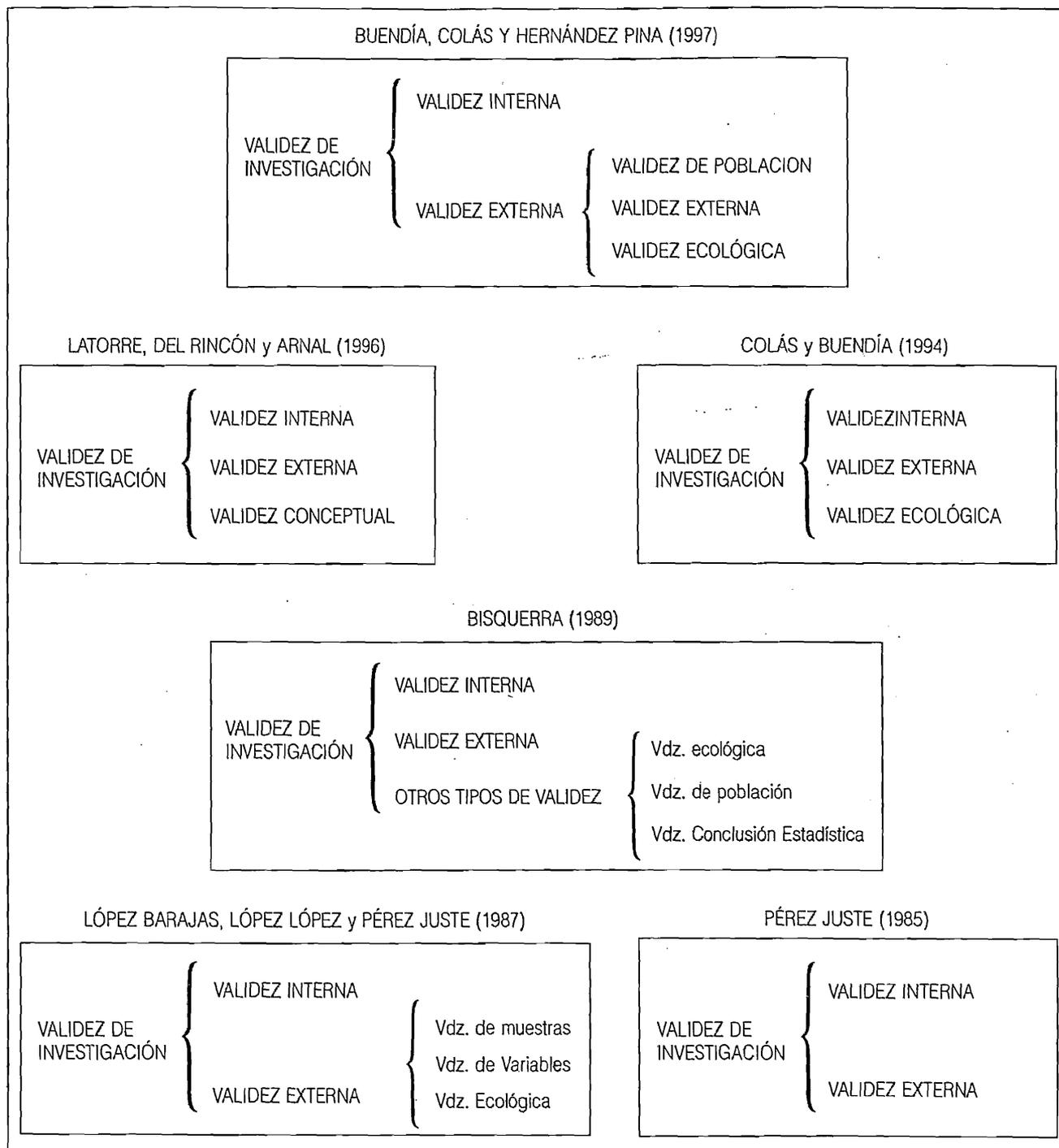
Como podemos comprobar, existe entre nosotros una enorme variedad de planteamientos, muchos de ellos basados parcialmente en el modelo original de Campbell y Stanley (1963; 1966) y al que cada autor ha añadido ciertas peculiaridades. Consideramos una medida interesante no acoger sin más los modelos planteados por otros autores, por prestigiosos que puedan ser, y por tanto creemos que cada una de las matizaciones aportadas en los textos pueden tener su interés. Sin embargo, llama poderosamente la atención que ninguno de los textos citados recoja, siquiera aproximadamente, los últimos modelos que sobre validez han ido surgiendo tanto desde las posturas de Campbell y sus colaboradores, como desde otras perspectivas alternativas. Nosotros pretendemos plantear en el presente trabajo algunas de las principales evoluciones en el campo que pueden ser de interés para poder entender el actual estado de la cuestión. Y buscar explicación a tal heterogeneidad de planteamientos.

Nos hemos centrado especialmente en el tema de la validez de investigación porque, como señala Mark (1986), el conocimiento de las tipologías de validez pueden ayudar a comprender de un modo más global las ventajas y desventajas de cada tipo de diseño a la hora de aplicarlo a una cuestión determinada. No se trata, por tanto, tan sólo de una disquisición teórica, sino que es una de esas ocasiones en las que como dijo Kurt Lewin, no hay nada más práctico que una buena teoría.

Teorías de la causación e implicaciones

Son diversos los modelos y concepciones que se tienen (o se han planteado) sobre la causación (Cfr. Cook y Campbell, 1979). Entre ellas tenemos posturas positivistas (en las que podríamos incluir desde Hume hasta

Modelos de Validez en textos españoles sobre Investigación Educativa



Pearson o Bertrand Russell), que rechazan el establecimiento de las relaciones causales, proponiendo sustituirlas por el estudio de las relaciones funcionales. Tam-

bién han tenido un importante predicamento las posturas que se han denominado *esencialistas*, según las cuales lo importante no es tanto encontrar vínculos em-

píricos entre las presuntas causas y consecuencias como la *comprensión* del grupo de factores que, necesaria, inevitable e infaliblemente, son determinantes causales de un determinado efecto. Esta postura esencialista supone la necesidad reduccionista de plantear un amplio programa de experimentación en el que los diversos constructos implicados como causas sean modificados y probados en diversos niveles, a fin de determinar la configuración exacta que determina cada consecuencia. Sin embargo, la comprensión de esa relación esencial no siempre tendría que ir vinculada a posibilidades empíricas de manipulación, ya que la comprensión va más allá de los meros hechos y su descripción. Esta concepción, más ambiciosa desde el punto de vista filosófico, choca con la realidad de las limitaciones que suponen las características concretas de la investigación empírica.

Otro modelo de particular importancia es el que se ha denominado teoría de la actividad de la causación (Cook y Campbell, 1979) o de la manipulabilidad de la causación (Cook y Shadish, 1994). Se trata de una concepción mucho más práctica del modelo de causa, basado en buena medida en el sentido común, aunque con un desarrollo racional inferior a los modelos citados anteriormente. Los experimentos no serían más que puestas en prácticas de este modelo, según el cual, si al modificar alguna variable se provocan cambios en otras variables tendremos un indicio de relación causal. Sin embargo, estos indicios no son suficientes, y suelen emplearse todavía los criterios planteados por John Stuart Mill para determinar una relación causal:

- a) *Precedencia temporal*: las causas han de acontecer antes que las consecuencias. La aplicación de este principio provocó, por ejemplo el rechazo de la teoría evolucionista de Lamarck, que planteaba una causación teleológica.
- b) *Covariación de causas y efectos*, es decir, que los cambios de una vayan acompañados de cambios en la otra. Esta condición es necesaria, pero no suficiente, por lo que las conclusiones causales de los diseños correlacionales, por sí solos, no están autorizadas.
- c) *Eliminación de otras explicaciones alternativas* a las relaciones causa-efecto. Esta es sin duda la

condición cuyo cumplimiento es más difícil de verificar, razón por la cual tuvo una repercusión tan amplia la obra de Campbell y Stanley (1963, 1966), que ofrecía un marco general que guiara en la consecución del cumplimiento de esta tercera condición.

A nuestro entender, aun no se han analizado en profundidad algunos de los modelos más actuales de causa (por ejemplo, los derivados de la Teoría General de Sistemas) que implican que sólo la adecuada interrelación compleja entre diversas «causas» o factores pueden resultar (se rechaza el término «determinar», ya que los resultados son probabilísticos) en un determinado efecto. La incompreensión de este tipo de *modelos probabilísticos* complejos (en los que la precedencia puede ser muy alejada en el tiempo, y la covariación tomada como un factor poco determinante) suele llevar consigo el no reconocimiento, al menos por parte de los legos —pero no sólo de ellos—, de las relaciones causales que parecen evidentes para los investigadores familiarizados con este tipo de concepto (esto ocurre frecuentemente, por ejemplo, con respecto a los avisos incluidos en los paquetes de tabaco: aunque hay pocas dudas de la relación causal entre fumar y, digamos, cáncer de pulmón, mucha gente, aplicando el más puro sentido común —y un modelo excesivamente simple de causación— rechazan la idea diciendo «pues yo conozco a alguien que fumaba dos paquetes diarios y no tuvo cáncer, muriendo de muy viejo»: es decir, no se considera causa porque no es una condición *suficiente*). Quizás el análisis más detallado de estos nuevos modelos pueda llegar a arrojar alguna luz nueva sobre el proceso de investigación en Ciencias Humanas.

Modelo original de Campbell y Stanley (1963; 1966)

Como hemos comentado anteriormente, la obra de Campbell y Stanley supuso un impulso notable a la investigación causal, quizás la más importante desde las aportaciones de Fisher en los años 20 y 30. De hecho su obra estuvo en buena medida motivada (Campbell, 1986) por el hecho de que se asumía de modo general que la mera asignación aleatoria de los casos a las distintas situaciones experimentales, según proponía el modelo fisheriano del Análisis de Varianza, permitía con-

trolar los sesgos de la investigación y llegar a conclusiones «sólidas». Por tanto, se suponía que el único control que había que tener a la hora de diseñar una investigación era el aleatorio. Sin embargo, Campbell y Stanley identificaron dos grandes grupos de sesgos posibles (Cfr. Campbell, 1986, p. 68), algunos de los cuales podían ser controlados mediante la asignación aleatoria, denominados Amenazas a la Validez Interna mientras que otro grupo de amenazas o sesgos no quedaban controladas aunque se utilizaran este procedimiento azaroso, las Amenazas a la Validez Externa (« *el fin principal que se persigue con la aleatorización (...) es la validez interna, no la externa*» p. 52¹). Por tanto, para el análisis de cualquier investigación se deberían separar ambos conjuntos de características o sesgos, de tal modo que pudiéramos identificar si cada uno de ellos estaba controlado o no, y por tanto si la investigación era válida o no.

Esta circunstancia dio lugar a que se creara el concepto de «diseño cuasi-experimental», para referirse a aquellas investigaciones en las que, por diversas razones (ver Coyle, Boruch y Turner, 1991 para una revisión de las mismas) las amenazas a la validez interna no se hubieran controlado mediante la aleatorización. Los diseños cuasi-experimentales son, por tanto, entendidos como una aproximación imperfecta (por lo que hace referencia a su validez interna) a los estudios experimentales, pero ambos diseños, experimentales y cuasi experimentales, carecen, en principio, de validez externa, que habría que planificar separadamente.

La Validez Interna de una investigación causal vendría a ser interpretada, por tanto, como el nivel de se-

guridad en que el estímulo experimental aplicado provoca algún cambio significativo en la variable dependiente (Campbell, 1957) «*en este caso experimental concreto*» (Campbell y Stanley, 1966, p. 16). Dicho de otro modo, y basándonos tanto en el principio de falsación popperiano como en el tercer criterio de causalidad expuesto por Stuart Mill, puede entenderse la Validez Interna como la inversa de la «*admisibilidad relativa de las teorías rivales*» (p. 74), de tal modo que cuanto menos probables sean esas otras explicaciones alternativas, mayor será la Validez Interna del estudio. Hemos de tener muy presente que, de ningún modo, podemos llegar a alcanzar una Validez Interna plena, es decir, no podemos *confirmar* (o *demostrar*) una determinada hipótesis (y la existencia de una relación causal es sólo uno de los posibles tipos de hipótesis) ya que «*cuando por conveniencia utilizamos ese término queremos significar, más bien, que la hipótesis fue expuesta a refutación y salió airosa de ella*» (p. 73). Sin embargo, se considera la asignación aleatoria como un método seguro para crear grupos no-sesgados, aunque ello no es lo mismo que garantizar la igualdad entre dichos grupos (algo que así sólo se logra de modo «*muy imperfecto*», en palabras de Campbell y Stanley, 1966, p. 35. «*No obstante, es la única forma práctica de hacerlo*»). Por ello, los procedimientos cuasi-experimentales han de prestar particular atención a la eliminación de las explicaciones alternativas (por usar los términos de Krathwohl, 1985): «*precisamente porque se carece de control experimental total, es imprescindible que el investigador tenga un conocimiento a fondo de cuales son las variables específicas que su diseño particular no controla*» (Campbell y Stanley, 1966, p. 71)

¹ Las páginas a las que hacemos referencia al citar la obra de Campbell y Stanley (1966) se refieren a las de la edición española de 1973, más asequible para el lector, aunque respetamos la fecha del original para mejor resaltar la influencia del paso del tiempo en los distintos modelos sobre validez.

Tabla 1. Tipos de Validez y sus amenazas según Campbell y Stanley (1966)

VALIDEZ INTERNA	Historia: <i>acontecimiento específico entre el pretest y el post-test</i>
	Maduración: <i>cambios debidos al mero paso del tiempo e independientes del tratamiento</i>
	Administración de tests: <i>efecto de la administración de un test sobre los resultados en otro posterior</i>
	Instrumentación: <i>cambios en los criterios de evaluación o achacables a las características del instrumento</i>
	Regresión Estadística: <i>tendencia a aproximarse a la media que ocurre en los grupos seleccionados por sus valores extremos</i>
	Selección: <i>Diferencias entre los grupos provocadas por una asignación sesgada de los sujetos a los tratamientos</i>
	Mortalidad experimental: <i>pérdida diferencial de participantes entre los grupos de comparación</i>
	Interacción entre selección y otras amenazas
VALIDEZ EXTERNA	Interacción entre las pruebas y el tratamiento: <i>cambio en la sensibilización de los sujetos experimentales provocado por el pretest</i>
	Interacción entre selección y tratamiento: <i>se refiere a aquellos casos en los que los resultados observados pudieran ser específicos de la situación investigada</i>
	Otras interacciones con el tratamiento: <i>referidas a los factores que podían afectar a la Validez Interna, tales como maduración, historia, mortalidad, etc.</i>
	Efectos reactivos de los dispositivos experimentales: <i>cambios en la reacción de los sujetos como consecuencia de las condiciones experimentales</i>
	Interferencia entre tratamientos múltiples: <i>efectos que pueden ejercer antiguos tratamientos aplicados al grupo experimental cuando se le aplican varias intervenciones</i>

El segundo concepto fundamental que se plantea en la obra de Campbell y Stanley con respecto a la Validez de la investigación causal es el de Validez Externa, término éste que hace referencia a la posibilidad de generalización de los resultados obtenidos en un estudio determinado a otras condiciones (poblaciones, situaciones, variables de tratamiento y variables de medición) diferentes. Puesto que esa generalización siempre parte de unos datos concretos, se basa en un proceso inductivo, procedimiento éste denunciado desde hace siglos como particularmente poco fiable. Como indican

Campbell y Stanley (1966, p. 39), «Desde el punto de vista lógico no podemos generalizar más allá de dichos límites (las características específicas de la investigación original); es decir, no podemos generalizar en absoluto». Si embargo, la generalización de conocimientos se ha basado, generalmente, en la experiencia y el sentido común, y en lo que los autores denominaron una «ley general empírica»: «cuanto más cercanos se hallan dos acontecimientos en tiempo, espacio y valor (...) más tienden a ajustarse a las mismas leyes» (Campbell y Stanley, 1966, p. 40). Téngase en cuenta que para resolver

los problemas de Validez Externa no siempre es posible acudir a procedimientos probabilísticos, ya que los mismos sólo son aplicables a algunos de sus componentes (por ejemplo a la muestra de casos), pero no a todos (por ejemplo, el instrumento de medida o el tiempo).

La concepción de Validez Externa planteada por Campbell y Stanley es menos homogénea que la de Validez Externa, lo que ha llevado a muchos autores a subdividirla en diversos componentes independientes entre sí (Ver Aliaga, 1993). Así, entre dichos componentes de la Validez Externa se ha hecho referencia a la Validez de Población (representatividad de casos o sujetos), Validez ecológica (representatividad de la situación), etc.. Para Snow (1963), la Validez Externa es el resultado conjunto de la validez de población, de la validez ecológica y de la por él denominada «generalidad referente». Para Kratochwill (1978) se incluirían como componentes de la Validez Externa la validez de población, la generalización de grupos a individuos y la validez ecológica). Cone, (1977), siguiendo a Wiggin (1973), considera generalización de las puntuaciones, de los elementos, temporal, de las situaciones, del método y de las dimensiones. Otros autores han hablado de Validez de Variables, Generalización del instrumento de medida, Validez de constructo, Validez de tratamiento, etc.

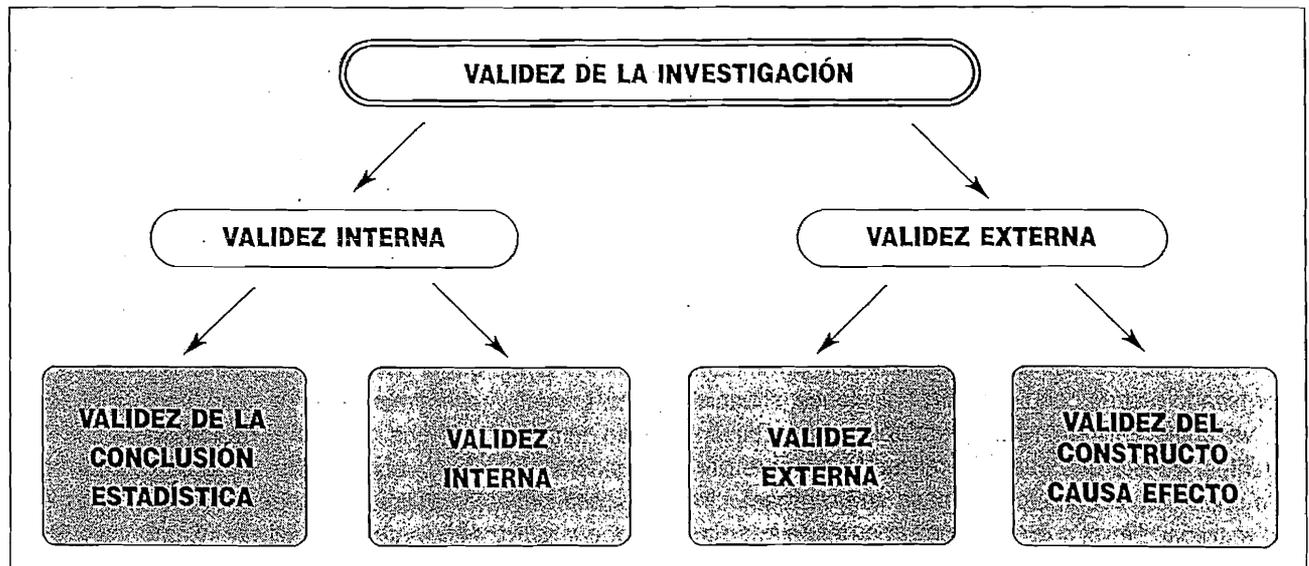
Para verificar la generalización de las conclusiones causales de un determinado estudio, algo de particular importancia para disciplinas aplicadas como es el caso de la Educación, Campbell y Stanley proponen dos vías. Por un lado la *representatividad* de las muestras (aunque la capacidad de hacer muestreos representativos, como hemos dicho, está limitada sólo a algunos de los componentes de la Validez Externa). Se trata de un método que plantea diversos problemas prácticos, por lo que ha de ser considerado más bien como un *desideratum* (p. 43). El otro procedimiento, por lo demás propio de ciencias más avanzadas, es la *replicación* de los estudios (p. 40)

Otro aspecto de gran relevancia en el modelo de validez planteado por Campbell y Stanley se refería a la prioridad dada por ellos a la Validez Interna (considerada como una condición *sine qua non*) sobre la Validez Externa, ya que sin seguridad en la existencia de una determinada relación causal no tiene sentido considerar su posible generalización, aunque esta concepción ha sido duramente criticada (por ejemplo, Cronbach, 1982)

Modelo de Cook y Campbell (1979)

Fruto de un análisis más minucioso de las circunstancias y conceptos del modelo planteado por Campbell y

Evolución desde Campbell y Stanley (1966) a Cook y Campbell (1979)



Stanley, se reformuló (más bien habría que decir que se especificó) dicho modelo al cabo de casi tres lustros en un libro de gran importancia que apareció en 1979 (Cook y Campbell). La aportación principal de este nuevo planteamiento es que, a fin de lograr una mayor claridad, se dividió cada uno de los dos tipos de Validez propuestos anteriormente en otros dos nuevos subtipos, al darle una mayor entidad a alguna de las características que constituían los conceptos originales, lo que produjo una nueva tipología de cuatro tipos de validez:

- A) *Validez de la conclusión estadística*: puesto que la covariación es una condición necesaria para poder determinar una relación causal, el nivel de seguridad que podamos tener en que tal variación conjunta existe es de considerable importancia, lo que justifica que este concepto se desligase del resto del concepto de «Validez interna», tal y como había sido definido por Campbell y Stanley (referente a si la relación entre dos variables, tal y como se habían operacionalizado, era causal). De hecho, los problemas de la validez de la conclusión estadística se habían considerado previamente como amenazas a la validez interna. Sin embargo, debido al hecho de que la covariación no siempre implica relación causal y de que los procedimientos para determinar si podemos determinar que tal covariación existe y en qué magnitud suelen ser estadísticos, e independientes del diseño en cuanto tal, Cook y Campbell decidieron añadir esta categoría como diferenciada del resto del concepto de la Validez interna, a la que, como veremos, le dan un significado más específico. Las amenazas a la Validez de la Conclusión estadística están relacionadas fundamentalmente con las fases de medición y análisis de datos
- B) *Validez Interna*: El concepto de Validez Interna en el modelo de Cook y Campbell (1979) se centra específicamente en determinar si la relación encontrada entre las variables (suponiendo que tal relación tenga una buena Validez de la Conclusión estadística) es de tipo causal, y en tal caso, cual es la dirección de dicha relación causal. Ha de tenerse presente que las conclusiones sobre esta posible relación causal se establece exclusivamente, como ya mantenía el modelo anterior, entre las variables tal y como han sido operati-

zadas en la práctica de la investigación («*en este caso experimental concreto*», en palabras de Campbell y Stanley, 1966, p. 16). A las amenazas a la validez interna ya conocidas en el modelo anterior se suman algunas más, tales como la *ambigüedad sobre la dirección de la relación causal*, la *difusión o imitación de tratamientos*, la *igualación compensatoria de tratamientos* (cuando dejar «sin nada» al grupo control parece inadecuado y se les concede algún bien o servicio), la *rivalidad compensatoria de los sujetos* (esto es, el efecto competencia que se puede desatar en el grupo control) o la *desmoralización o resentimiento de los sujetos que reciben menor cantidad de tratamiento deseable*. Curiosamente, algunas de estas nuevas amenazas introducidas por Cook, con la aprobación de Campbell, en el nuevo modelo —particularmente la desmoralización— no cumplen con la característica original que había definido los sesgos de la denominada Validez Interna, es decir, que no pueden ser controlados mediante la asignación aleatoria. Esto hace que posteriormente Campbell (1986, p. 67) considere que alguna de estas amenazas (incluso puede que todas) no sean específicamente de la validez interna, sino de la de constructo.

- C) *Validez de constructo de las causas y los efectos*: Cuando Campbell y Stanley (1966) se refieren a la validez externa afirman que esta se ocupa de la posibilidad de generalizar a «otras poblaciones, situaciones, variables de tratamiento y variables de medición» (p. 16). En la reformulación planteada por Cook y Campbell (1979) —aunque el propio Campbell (1986) confiesa que realmente la propuesta es de Cook— las dos últimas generalizaciones se independizan y toman entidad propia. En efecto, ya Campbell y Stanley habían insistido repetidamente en que las relaciones causales encontradas en una investigación determinada se limitaban, exclusivamente, a las operacionalizaciones específicas de la Variable Dependiente y la Variable Independiente «*en este caso experimental concreto*» (Campbell y Stanley, 1966, p. 16). Por tanto, la generalización desde esas operacionalizaciones concretas a otras formas de entender (operacionalizar) dichas variables era un proceso que no debía asumirse sin

Tabla 2. Amenazas a la Validez de la investigación en Cook y Campbell (1979)

VALIDEZ CONCLUSIÓN ESTADÍSTICA	<p>Baja potencia estadística</p> <p>Violación de los supuestos de las pruebas estadísticas</p> <p>Violación del error tipo I</p> <p>Fiabilidad de la medición de la respuesta</p> <p>Fiabilidad de la administración de los tratamientos</p> <p>Irrelevancias aleatorias del contexto de investigación</p> <p>Heterogeneidad aleatoria de las unidades de respuesta</p>
VALIDEZ INTERNA	<p>Historia</p> <p>Maduración</p> <p>Administración de pruebas</p> <p>Instrumentación</p> <p>Regresión estadística</p> <p>Selección</p> <p>Mortalidad Selectiva</p> <p>Interacciones con la selección</p> <p>Ambigüedad en la dirección de la relación causa-efecto</p> <p>Difusión e imitación de tratamientos</p> <p>Igualación compensatoria de tratamientos</p> <p>Rivalidad compensatoria de los sujetos</p> <p>Desmoralización de los sujetos</p>
VALIDEZ DE CONSTRUCTO	<p>Explicación preoperacional inadecuada.</p> <p>Sesgos derivados del uso de una operación única</p> <p>Sesgos derivados del uso de un solo método</p> <p>Adivinación de hipótesis</p> <p>Recelo de evaluación</p> <p>Expectativas del experimentador</p> <p>Confusión de constructos y niveles de constructo.</p> <p>Interacción de tratamientos intrasujeto</p> <p>Interacción de administración de pruebas y tratamiento</p> <p>Generalidad restringida entre constructos</p>
VALIDEZ EXTERNA	<p>Interacción selección-tratamiento (validez de población)</p> <p>Interacción contexto-tratamiento (validez ecológica)</p> <p>Interacción historia-tratamiento (validez histórica)</p>

Tomado de Ato en Anguera et al. (1995)

más de modo automático. Debe resaltarse que mientras que la «*validez interna tiene que ver con la eliminación de explicaciones alternativas de la supuesta relación causal entre X tal y como es manipulada y O tal y como es medida, la Validez de Constructo se ocupa de eliminar las explicaciones alternativas de las supuestas categorías de X y O*» (Cook, Campbell y Peracchio², 1990, p. 495).

Esta idea de Validez de Constructo está íntimamente relacionada con el concepto de «*Confundido*», que hace referencia a la circunstancia de que un mismo hecho, una misma operacionalización, puede ser entendida como perteneciente a dos categorías o constructos teóricos diferentes, de tal modo que las variables que realmente mantienen una relación causal pueden estar enmascaradas o confundidas con otras que no tienen ese tipo de relación, pero que son coherentes con las operacionalizaciones utilizadas en la investigación. Shadish, Cook y Houts (1986) informan, por ejemplo, de los resultados obtenidos por Coleman (1972) en su análisis sobre la igualdad de oportunidades educativas. Dependiendo del modo en que se hubiera operacionalizado en cada investigación la variable «calidad escolar» (como tamaño de la clase, gastos por alumno, antigüedad de las instalaciones y el material, aislamiento acústico, etc.) encontraba resultados diferentes de dicha variable. Otros ejemplos del efecto *Confundido* son el efecto *Hawthorne*, el efecto *Pigmalión* o el efecto placebo (Ball, 1997). Para solventar este problema de generalización desde las causas o las consecuencias tal y como las hemos medido y/o manipulado a los conceptos de los enmarcan Cook y Campbell (1979) proponen un doble proceso, basado en la matriz multirrasgo-multimétodo de Campbell y Fiske (1959):

- a) comprobar la *convergencia* a través de diferentes medidas o manipulaciones del mismo constructo. Es la parte más importante (y empíricamente a veces casi la única: Cook y Campbell, 1979, p. 61).
- b) comprobar la *divergencia* respecto de los resultados obtenidos con medidas o manipulaciones de constructo parecidos pero distintos.

La validez de constructo tiene un amplio listado de amenazas identificadas (ver tabla 2), que Ato (en Anguera et al., 1995) ha dividido en dos grandes grupos, a las que nosotros hemos asignado todas ellas (Ato sólo clasifica siete de las diez amenazas citadas por Cook y Campbell, 1989):

- a) *Sesgos de operacionalización de las causas y los efectos:*
 - 1) *explicación preoperacional inadecuada:* esta amenaza puede estar presente cuando la elección de la definición operativa de una variable no se realiza a partir de un cuidadoso análisis conceptual de la misma, y por tanto no se adecua exactamente a la definición constitutiva (Kerlinger, 1975) del constructo.
 - 2) *sesgos por el empleo de una única operación:* puesto que definir operativamente un constructo o variable puede ser inadecuado (no tener validez de contenido o ser incompleta, estar sesgada, ser incorrecta) siempre es preferible utilizar diversos procedimientos, lo que nos permitirá triangular los datos y captar mejor el significado de la variable. Esta es una amenaza derivada de no comprobar la *convergencia* a la que hacíamos referencia anteriormente.
 - 3) *sesgos derivados de utilizar un solo método:* esta amenaza puede presentarse cuando, independientemente de que se operacionalice de una o varias formas, todas ellas utilizan un mismo procedimiento (sólo técnicas de lápiz y papel o sólo procedimientos para evaluar la reactividad fisiológica). En tales casos, el propio método puede conllevar un sesgo sistemático que nos impida aprehender adecuadamente el constructo que nos interesa.
 - 4) *Confusión entre constructos y niveles de constructo:* ocurre en aquellos casos en que un determinado constructo sólo se relaciona con otro cuando alcanza un determinado nivel, pero el in-

² Hemos sustituido los términos A y B con los que en el original se refieren a las variables dependiente e independiente por los términos X y O, respectivamente, mucho más cercanos a la tradición experimental y a la notación utilizada por Campbell y colaboradores.

- investigador no es capaz de descubrirlo porque no llega a sobrepasar dicho umbral. Un ejemplo simplificado podría ser el de un presunto investigador que concluyese que la temperatura no afecta al estado físico del agua, porque lo ha modificado desde los 10° centígrados hasta los 90° sin obtener resultados. En el mismo sentido, puede haber problemas para generalizar los efectos de la utilización de un sistema de educación compensatoria (o la ausencia de los mismos) si éstos consistían exclusivamente en una hora semanal de apoyo, cuando hubiera sido necesaria una mayor intensidad del tratamiento para conseguir algún resultado apreciable.
- 5) *interacción entre diferentes tratamientos*: se refiere a la posibilidad de que los efectos de un determinado tratamiento aplicado con anterioridad al grupo experimental interfieran con los del tratamiento actual, por lo que malinterpretemos los efectos producidos por este último, que no podrán generalizarse a otras condiciones.
 - 6) *interacción entre la administración de pruebas y el tratamiento*: se refiere a los sesgos que la aplicación de un pretest puede provocar en los sujetos (sensibilización), consiguiendo efectos diferentes a los que se conseguirían en una situación natural, diferente de la investigación en que carece de esos controles previos
 - 7) *generalizabilidad a través de los constructos*: se refiere a los posibles efectos, no considerados en el diseño previo, que un tratamiento puede tener en variables o constructos distintos a la variable dependiente.
- b) *Sesgos de reactividad de las unidades de respuesta (sujetos o casos)*
 - 1) *adivinación de la hipótesis*: los sujetos que participan en una investigación con frecuencia pretenden *adivinar* qué es lo que el investigador pretende de ellos, y en un proceso que la psicología social ha denominado *deseabilidad social*, adaptarse a dichas expectativas, más que a responder directamente al tratamiento que, realmente, se le administra.
 - 2) *aprensión a la evaluación (o efectos reactivos de la observación)*: indica la propensión de los sujetos a cambiar su comportamiento cuando saben que están siendo evaluados u observados (es la aplicación *blanda* en las Ciencias Humanas del Principio de Incertidumbre de Heisenberg en la Física).
 - 3) *expectativas del experimentador*: también conocido con el nombre de *efecto Rosenthal* (por su descubridor), *efecto Pigmalión* o, en sentido inverso, *efecto Golem* (Cfr. Ball, 1997). Aunque su auténtico alcance ha sido puesto en duda, está relacionado con los sesgos que de modo más o menos inconsciente puede intentar auto-cumplir sus predicciones.
 - D) *Validez Externa*: El concepto de validez externa en Cook y Campbell (1979) ha quedado como una categoría residual, referida a la generalización de los resultados a través de diversos componentes, los no incluidos en la *Validez de Constructo de causas y efectos*, básicamente los sujetos y las situaciones, a los que se ha añadido la dimensión temporal. Consecuentemente, las amenazas también han quedado reducidas sustancialmente, quedando fundamentalmente en tres:
 - a) *interacción de selección y tratamiento*: esta amenaza se refiere a la posible limitación de la generalización de resultados a sujetos o poblaciones no directamente incluidas en la investigación, es decir, mantienen presente la posibilidad de que las relaciones causales identificadas sean específicas de los sujetos directamente analizados.
 - b) *interacción de situaciones y tratamiento*: En este caso las limitaciones a la generalización de los resultados de la investigación puede venir de la especificidad de la relación causal en la situación específicamente considerada.
 - c) *Interacción de historia y tratamiento*: en este caso se considera que las limitaciones a la generalización de las relaciones causales pueden estar directamente vinculadas con el momento temporal, de modo que suponer que tal relación causal existió en el pasado o se mantendrá en el futuro, al ir más allá de lo expresamente descubierto, tiene una alta dosis de suposición no contrastada.

Son varios los procedimientos que Cook y Campbell (1979) analizan para procurar aumentar la Validez Externa de una investigación. En primer lugar se consideran diversos tipos de muestreo, que incumben tanto a sujetos como a situaciones, pero que no considera, obviamente, la dimensión temporal:

- a) *muestreo aleatorio*: este muestreo aleatorio es el que mejor procura la representatividad de los resultados, y por tanto, el que nos ofrece mayores garantías para generalizar los resultados a otras situaciones o sujetos pertenecientes a la población global. Sin embargo, plantea serios problemas prácticos, fundamentalmente debido a que las circunstancias no siempre permiten hacer un muestreo de este tipo, o a que cuando se realiza las muestras suelen estar muy dispersas.
- b) *muestreo heterogéneo deliberado*: consiste en procurar incluir en el estudio un amplio rango de sujetos y circunstancias de tal modo que abarquen la mayor amplitud posible, aunque esta idea no debe entenderse como selección exclusiva de los casos extremos, sino que han de incluirse también casos «típicos». Este procedimiento es mucho más asequible que el aleatorio, y aunque técnicamente no se puede generalizar a partir de sus resultados, la amplia gama considerada nos permite hacernos una idea global de las relaciones estudiadas en las diversas modalidades posibles.
- c) *muestreo impresionista de categorías o modos*: Consiste en determinar las clases o categorías a las que queremos generalizar (cursos o municipios) y elegir al menos un elemento de cada una de esas modalidades, a fin de obtener una impresión general. Este tipo de procedimiento es útil en aquellos casos en los que las pretensiones de generalización son limitadas, y particularmente en investigación aplicada, ya que no es útil para establecer un conocimiento teórico ampliamente generalizable. Es el procedimiento menos poderoso de los tres tipos de muestreo, pero tiene como ventajas principales su simplicidad y economía.

Hay, sin embargo, un último procedimiento para asegurar la generalización de los resultados: la replicación. Mediante este método conseguimos establecer si los datos avalan la generalización de los resultados cuando algunas de las condiciones (sujetos, situaciones o tiempo) es modificada. Por tanto, es un procedimiento con respaldo empírico, en lugar de acudir a la justificación teórica o racional más propia de los tipos de muestreo.

Modelos alternativos

Aunque, como hemos repetido, los modelos más influyentes sobre validez de investigación han sido los creados por Campbell y sus colaboradores, diversos modelos alternativos han sido planteados por otros autores. De ellos el que ha tenido un mayor eco en la comunidad científica ha sido el de Cronbach, un modelo de notable complejidad de análisis que vino a aportar una visión algo diferente de la validez de las investigaciones. Otras aportaciones interesantes, aunque han tenido un menor predicamento, son las de Krathwoll (1985), Kruglanski y Kroy (1976) o el de Reichardt (1986, citado en Mark, 1986), que es una variación del de Cronbach (Ver Mark, 1986 para una análisis e integración de las diversas propuestas).

Por problemas de extensión de este trabajo vamos a centrarnos fundamentalmente en el trabajo de Cronbach (1982), que es probablemente el modelo sobre validez más elaborado y de mayor interés entre los que se han construido como alternativa a los de Campbell y sus colaboradores (Campbell y Stanley, 1966; Cook y Campbell, 1979). En su modelo Cronbach distingue cuatro tipos de elementos diferentes en una investigación:

- *Unidades*: suele referirse a los sujetos experimentales, ya sean estos individuos, grupos, centros escolares, etc.
- *Tratamiento*: es la presunta causa, es decir, aquello que manipula (o selecciona) el investigador.
- *Observación*: de manera general se refiere al presunto efecto en la relación causal, que no se manipula sino que únicamente se mide.

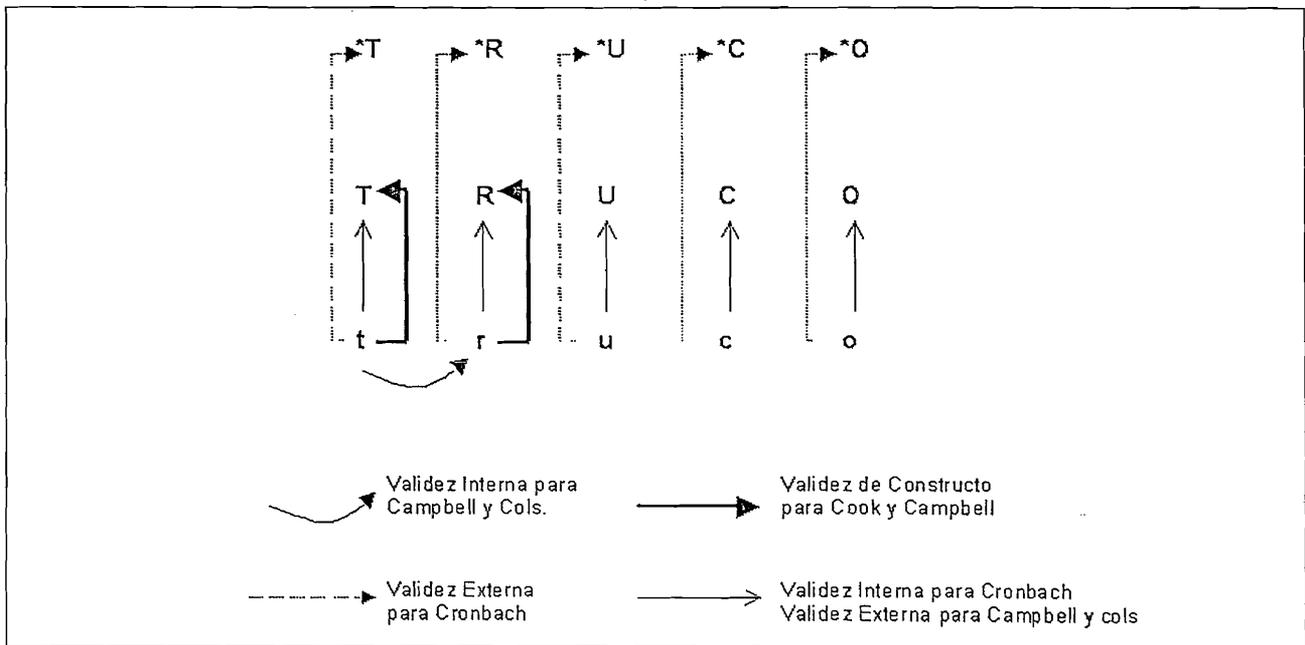
— **Situación:** se refiere al «contexto social, entendido en un sentido amplio, en el que el estudio tiene lugar» (Cronbach, 1982, p. 79). Por tanto, incluye desde el clima social, económico, ideológico o político hasta el momento temporal específico. Algunos autores (por ejemplo, Cook, Campbell y Peracchio, 1990, Ato en Anguera *et al.*, 1995) han dividido este concepto en dos diferentes, uno sobre las condiciones en las que se realiza propiamente la investigación (el contexto o situación) y otro referido específicamente al momento temporal, en el sentido dado por Cook y Campbell (1979) a este concepto, cosa que compartimos, ya que puede darse el caso de que una misma investigación se realice en un mismo contexto (por ejemplo, en un mismo colegio público entre alumnos de la misma asignatura) pero que los resultados sean distintos debido a los cambios más intangibles que puedan deberse al paso del tiempo (cambios en la ideología, la legislación, etc.).

por ejemplo, utiliza el término **UTOS**, en mayúsculas, para referirse a los constructos que son objeto de investigación, bien entendido que en un nivel abstracto (al añadir el momento temporal en uno de los elementos relevantes de la investigación Cook, Campbell y Peracchio, 1990 convierte esas iniciales en **UTOSCH**, al representar el tiempo como CH, del griego *chronos*; Ato, en Anguera *et al.*, 1995, ha traducido ingeniosamente los términos al castellano dando lugar al acrónimo **TRUCO** —Tratamiento, Respuesta, Unidades, Contexto y Ocasión o momento temporal— que facilita el proceso mnemotécnico).

Cook y Campbell (1979) habían puesto de relieve las dificultades de interpretación que pueden presentarse al operativizar, en la práctica, los distintos constructos. Cronbach (1982) profundiza en esta idea y utiliza el término **utos** para referirse a los elementos concretos incluidos en la investigación. Así, la *u* hace referencia a la muestra de unidades (generalmente se refiere a los sujetos concretos) que son directamente analizados en la investigación; la *t* se refiere al modo específico en que se ha operacionalizado la variable tratamiento y la *o* a

Cronbach utiliza las iniciales de estos elementos de distintas maneras según el grado de especificidad. Así,

Ilustración 2. Relación entre los modelos de Campbell y Stanley (1966), Cook y Campbell (1979) y Cronbach (1982)



Adaptado a partir de Cronbach (1982) y Ato, en Anguera *et al.* (1995)

la forma en que se ha medido la variable dependientes. Ha de hacerse notar que Cronbach (1982) siempre utiliza la *S* de la situación en mayúsculas ya que, atendiendo a la clasificación fisheriana de los términos de un análisis de varianza, la situación (y habrían que incluir también el tiempo) es un factor fijo, en el sentido de que sólo hay uno en cada estudio, mientras que los demás términos pueden ser factores aleatorios (p. 86).

En un tercer nivel de especificidad, Cronbach utiliza los ***UTOS** para referirse a aquellas condiciones a las que podríamos estar interesado en generalizar los resultados obtenidos al investigar los *utoS* específicos de una investigación, pero que difieren de las poblaciones de referencia (los *UTOS*) en alguna o algunas características (por ejemplo, otras poblaciones, contextos diferentes, efectos distintos, o tratamientos parecidos).

Puesto que existen tres niveles diferentes de concreción-generalización, Cronbach (1982) plantea dos tipos deferentes de inferencia, según las relaciones que deseamos establecer a partir de las posibles relaciones causales identificadas en una investigación concreta. En primer lugar, el establecimiento de relaciones que puedan existir entre los *utoS* y los *UTOS* (es decir, las operaciones o muestras con los constructos o poblaciones que pretenden representar) son denominadas por Cronbach *Inferencias Internas*, y la seguridad que podemos tener en tal relación es denominada *Validez Interna*. En segundo lugar, la relación entre los *utoS* y los ***UTOS** (o sea, entre los casos específicos y los constructos o poblaciones ligeramente diferentes a aquellos que pretendían representar) recibe el nombre de *Inferencias Externas*, y el grado de credibilidad o seguridad en tal relación es denominada *Validez Externa*. Puesto que Cronbach está especialmente interesado en la utilidad práctica de las investigaciones, concede una importancia prioritaria a estos dos tipos de inferencias, particularmente en las inferencias externas, ya que como señala «normalmente sólo una pequeña fracción de la audiencia está interesada en los *UTOS* que define el estudio» (p. 99). En efecto, para ser aplicado, los lectores de la investigación normalmente habrán de preguntarse cosas como «¿funcionará este programa de lectoescritura en nuestro idioma (*S), con nuestros alumnos (*U), tal y como nosotros evaluamos el rendimiento (*O) y adaptado para que lo apliquen nuestros profesores (*T)?». Es decir, las aplicaciones de los resultados han de

hacerse con frecuencia a condiciones diferentes a las estudiadas y, en el caso de la investigación aplicada, este proceso es el prioritario, ya que los responsables de tales tomas de decisión no suelen ser tan puntillosos respecto a la seguridad con que se haya establecido la relación causal, sino que están más bien interesados en la aplicabilidad de los descubrimientos.

Como podemos comprobar, Cronbach utiliza algunos conceptos acuñados por Campbell y sus colaboradores (fundamentalmente el término de Validez Interna) en un sentido distinto, lo que ha producido cierta confusión, lo cual, añadido a la complejidad del modelo ha propiciado que no se haya difundido demasiado, excepto entre los especialistas.

Otra de las críticas que se han realizado contra este modelo se refiere al excesivo esquematismo dado al tema de la generalización (Mark, 1986). Es bien cierto que Cronbach tiene razón al plantear que las aplicaciones prácticas de las investigaciones con frecuencia requieren su generalización a situaciones o condiciones que varían, en mayor o menor grado, de las poblaciones o condiciones analizadas inicialmente. Sin embargo, no hay razón alguna para mantener que esa mayor diversidad deba plantearse en sólo dos niveles (*UTOS* y ***UTOS**), sino que, en nuestra opinión, más bien debe ser concebida como un continuo (Mark, 1986, p. 52).

Evolución posterior de las tipologías de validez

A pesar de la complejidad del modelo de Cronbach (1982) y de su limitada difusión, su contribución a la aclaración del tema fue importante, básicamente porque obligó a reflexionar sobre los modelos existentes y a introducir algunos cambios que permitiesen aclarar el campo. Así, por un lado, ya hemos citado la consideración que la dimensión temporal de las conclusiones causales fue tomando, lo que hizo que se incluyese incluso en el modelo ampliado que se deriva del propuesto inicialmente por Cronbach (1982) y que hemos representado en la ilustración 2.

Entre las reacciones más importantes que produjo el modelo de Cronbach (1982), aunque hubo algunas otras razones que justificaron el cambio (ver Cook y Shadish

1994), cabe citar la reformulación que hizo Campbell (1986) de algunos de los conceptos incluidos en sus propios modelos. Así, como una reacción al hecho de que muchos autores solían considerar, erróneamente, (a pesar de lo indicado por él de modo explícito) que el término de Validez Interna se refería al nivel de seguridad que podíamos tener en que existía una relación causal entre dos conceptos (por ejemplo «un Programa de Enriquecimiento Instrumental mejora el rendimiento académico»), Campbell propone cambiar el nombre de Validez Interna por otro que especifique, de modo más detallado, las características que tiene dicho concepto. La alternativa que ofrece es la de pasar a denominarla *Validez Causal Local Molar*. Con esta nueva denominación pretende hacer hincapié en algunos aspectos esenciales, representados por cada uno de los términos empleados:

- *Causal*: se refiere a que este tipo de Validez ha de centrarse fundamentalmente en que las relaciones analizadas sean realmente de tipo causal.
- *Local*: hace referencia exclusivamente a que el nivel de relación causal se establece exclusivamente entre la *t* y la *o* del modelo de (la *t* y la *r* en el modelo traducido del esquema *truco*). Recordemos que ya en Campbell y Stanley (1966), se referían a la Validez Interna como los cambios que introducían «los tratamientos empíricos en este caso experimental concreto» (p. 16; el subrayado es nuestro). Existe en esta nueva conceptualización, por tanto, un rechazo implícito a la postura de Cronbach (1982) que no sólo incluye en el concepto de Validez Interna los Constructos *T* y *O*, sino incluso los de *U* y *S*)
- *Molar*: implica que las relaciones causales identificadas en el nivel más bajo de operacionalización han de entenderse como el resultado de la interacción global entre todos los elementos implicados (sujetos, contexto, forma de manipular el tratamiento, modo de observar los resultados, etc.) más que como el efecto de alguno o algunos de dichos microelementos aisladamente. Dicho de otro modo, lo que se determina no es si esta causa provoca esta consecuencia, sino si tal relación acontece para estos sujetos determinados en estas circunstancias concretas y en este preciso momento temporal.

Campbell (1986) no sólo propuso cambiar el nombre de la Validez Interna, sino que su afán clarificador de conceptos le hizo proponer igualmente un cambio de denominación para la Validez Externa. Como respuesta a los dos niveles de inferencia (interna y externa) planteados por Cronbach (1982), Campbell propone modificar el concepto de Validez Externa y sustituirlo por el de *Gradiente de Semejanza (Proximal Similarity)*, como parte de un proceso de reflexión que le lleva a alejarse cada vez más de los procedimientos basados en procurar la representatividad a través del muestreo. Por un lado debido a que tal tipo de muestreo es impracticable para algunos componentes de la investigación, como el tiempo. Por otro porque esa misma dimensión temporal, aparte de otras, afecta a la representatividad del resto de componentes: ¿un muestreo aleatorio de sujetos —o de escuelas, o de aulas— puede seguir considerándose representativo de cómo es la población de referencia diez o quince años después de realizado el estudio?. Por último, razones de tipo práctico (económicas, geográficas, etc.) recomiendan la utilización de muestreos ilustrativos (impresionistas, tal y como los denominaban Cook y Campbell, 1979) orientados por alguna teoría específica en lugar de los muestreos aleatorios de poblaciones amplias, cuyo abandono formalmente recomienda Campbell (1986, p. 73). Mark (1986) y, de modo especial, Shadish, Cook y Houts (1986) abogan por una planteamiento semejante, en el que las inferencias se construyan no a través de procedimientos de estadística inferencial sino a través de la integración de la información proveniente de diversas fuentes (p. 43).

Por tanto, y como resumen, Campbell reformula en 1986 su tipología de la validez de la investigación causal de tal modo que decanta su antigua Validez Interna (ahora denominada *Validez Causal Local Molar*) hacia una concepción más empírica y menos teórica, mientras que impulsa su antigua Validez Externa (ahora *Gradiente de Semejanza*) hacia un modelo más guiado por la teoría específica de cada estudio y menos por procedimientos aleatorios más o menos automatizados.

Otros planteamientos han puesto de manifiesto algunos problemas con los que se enfrenta la investigación causal. Por ejemplo, Mark (1986) en su excelente trabajo integrador de las diversas tipologías de validez, señala que el intentar asegurar una alta validez en el establecimiento de las relaciones causales puede carecer

de utilidad práctica, tal y como ya había señalado anteriormente la postura de Cronbach, mucho más interesado por la trascendencia práctica de las investigaciones que por el nivel de seguridad en las conclusiones.

Cook y Shadish (1994) señalan que las tipologías de validez no están (ni probablemente puedan llegar a estar) completas. Siguiendo a Dunn (1982) reconocen que la cantidad de tipos de validez es ilimitada: ¿porqué separar la validez de constructo de la validez externa, como hacen Cook y Campbell en 1979,

y no hacer lo propio con la validez de población, o la ecológica o la temporal, por ejemplo?. Cook, Campbell y Peracchio (1994) presentan un listado de amenazas a la validez de investigación que, aunque basada en las propuestas anteriores de Campbell y Cook (1979), que a su vez se basaba en la de Campbell y Stanley (1966), presenta ciertas modificaciones, fundamentalmente respecto a la categoría a la que se asigna cada tipo de amenaza, lo que nos sugiere que tampoco es éste un campo completamente cerrado (Ver Tabla 3).

Tabla 3. Evolución de las clasificaciones de amenazas a la validez en los diversos modelos de Campbell y Cols.

	CAMPBELL Y STANLEY (1966)	COOK Y CAMPBELL (1979)	COOK, CAMPBELL Y PERACCHIO (1990)
VALIDEZ INTERNA	<p>Historia Maduración Administración de tests Instrumentación Regresión Selección Mortalidad experimental Interacción entre selección y otras amenazas</p>	<p>Historia Maduración Administración de pruebas Instrumentación Regresión estadística Selección Mortalidad Interacciones con la selección Ambigüedad en la dirección de la relación causa-efecto Difusión o imitación de tratamientos Igualación compensatoria de tratamientos Rivalidad compensatoria de los sujetos Desmoralización de los sujetos</p>	<p>Historia Maduración Administración de pruebas Instrumentación Regresión estadística Selección Mortalidad Interacciones con la selección Ambigüedad en la dirección de la relación causa-efecto</p>
VALIDEZ DE CONCLUSION ESTADISTICA		<p>Baja potencia estadística Violación de los supuestos de las pruebas estadísticas Violación del error tipo I Fiabilidad de la medición de la respuesta Fiabilidad de la administración de los tratamientos Irrelevancias aleatorias del contexto de investigación Heterogeneidad aleatoria de las unidades de respuesta</p>	<p>Baja potencia estadística Violación del error tipo I Fiabilidad de las medidas Fiabilidad de la administración de los tratamientos Irrelevancias aleatorias del contexto Heterogeneidad aleatoria de las unidades de respuesta</p>
VALIDEZ EXTERNA	<p>Efecto reactivo o de interacción de los tests Interacción selección-tratamiento Efectos reactivos de dispositivos experimentales Interacción entre tratamientos múltiples</p>	<p>Interacción selección-tratamiento Interacción contexto-tratamiento Interacción historia-tratamiento</p>	<p>Interacción entre tratamientos múltiples Interacción medición-tratamiento Interacción selección-tratamiento Interacción contexto-tratamiento Interacción historia-tratamiento</p>
VALIDEZ DE CONSTRUCTO DE CAUSAS Y EFECTOS		<p>Explicación preoperacional inadecuada. Sesgos derivados del uso de una operación única Sesgos derivados del uso de un solo método Adivinación de hipótesis Recelo de evaluación Expectativas del experimentador Confusión de constructos y niveles de constructo. Interacción de tratamientos intrasujeto Interacción de administración de pruebas y tratamiento Generalidad restringida entre constructos</p>	<p>Explicación preoperacional inadecuada. Sesgos derivados del uso de una operación única Sesgos derivados del uso de un solo método Interacción entre el procedimiento y el tratamiento Difusión o imitación de tratamientos Igualación compensatoria de tratamientos Rivalidad compensatoria de los sujetos Desmoralización de los sujetos Adivinación de hipótesis en de las condiciones exper. Recelo de evaluación Expectativas del experimentador Confusión de constructos y niveles de constructo.</p>

47

Conclusiones sobre el estado actual de la cuestión

La utilidad de los modelos de Validez ha sido, a pesar de la polémica, indudable, y ha permitido a los investigadores en Ciencias Sociales centrarse en aspectos que en modelos anteriores pasaban desapercibidos. Sin embargo, la creciente complejidad y fragmentación de los modelos, así como la confusión con respecto al vocabulario empleado (con distinto significado según los autores, y con diversidad terminológica para un mismo concepto), puede estar a la base de la actitud conservadora que detectábamos a la hora de revisar los textos españoles (y no sólo los españoles) más recientes. Asumir sin más los cambios que en cada momento se van proponiendo podría denotar que se está al día, pero también falta de criterio. Por el contrario, el hecho de que tantos autores, cuya capacidad profesional no puede ser puesta en duda, se resistan a utilizar los modelos más actualizados quizás sea una señal de que éstos deberían ser reconsiderados nuevamente, quizás en la línea que ya hiciera Campbell (1986) de aclarar los conceptos originales para evitar malentendidos (aunque su propuesta del cambio terminológico nos parece excesiva), y evitar sucesivas y potencialmente infinitas subdivisiones de cada concepto o modelo (Cook y Shadish, 1994, reconocen que «no hay límites al número de posibles tipos de validez» -p.553-), que tan sólo dificultan la comprensión global del problema

Por otro lado, y centrándonos en la delimitación de los posibles problemas de la validez de investigación, el intento de crear un modelo excesivamente simple, con un listado de amenazas supuestamente cerrado ha conducido a ciertos excesos. Así, por ejemplo, Cronbach ha señalado que muchas investigaciones, incluso las no causales, se han visto innecesariamente complicadas al plagarse de controles para algunas amenazas. Reichardt (1985), por su parte, señala algunos ejemplos en los que algunas amenazas no se han detectado por acogerse demasiado rígidamente a los esquemas de validez planteados. Estos excesos no son, sin embargo, achacables a los modelos planteados, sino más bien a la utilización ciega (o excesivamente rígida) que se ha realizado de los mismos. De hecho casi todos los modelos sobre tipologías de validez iban acompañados de una advertencia sobre sus

limitaciones o su provisionalidad. Campbell (1986) afirma que su nuevo planteamiento «es una reacción dialéctica o una sobre-reacción. Confiemos en que la reformulación global conduzca hacia la convergencia» (p. 76); Cronbach (1982) dice que el suyo es «uno de los muchos esquemas posibles» (p. 77), y Cook y Campbell (1979) mantienen que su modelo está guiado por consideraciones prácticas, «lo cual no determina necesariamente razones lógicas para la clasificación» (p.39). Por último, y de modo más claro todavía, Campbell y Stanley (1966) en su capítulo final previenen sobre la utilización rígida de los modelos o tablas por ellos planteados: «más bien habría que estimular una actitud abierta hacia la indagación de los nuevos mecanismos de obtención de datos, y un nuevo análisis acerca de algunas de las imperfecciones que acompañan a la aplicación rutinaria de los (diseños) tradicionales (...) hemos llamado la atención sobre la posibilidad de utilizar creativamente las características peculiares de cualquier situación concreta de investigación al diseñar pruebas específicas de hipótesis causales» (pp. 138-139).

La complejidad, e incluso confusión, del estado actual de la cuestión justifican una actitud conservadora, al menos con intenciones didácticas, ante los modelos de Validez que hemos de plantear, de modo introductorio, a los alumnos, tal y como hacen los textos de autores españoles que hemos analizado en la presentación y que están, en general, orientados a ese tipo de público.

Creemos, sin embargo, que hemos de concebir, y enseñar a los investigadores actuales y a los futuros, que el problema de la validez de una investigación no puede resolverse con un libro de recetas (Trochim, 1986), automáticamente, que el tema implica un amplio proceso de reflexión sobre las circunstancias de cada caso. Por ello es interesante la concepción de De la Orden (1998) de que nos enfrentamos, al realizar una investigación en Ciencias Humanas, con un problema de falta de control de los factores que pueden afectar a la misma, de tal modo que cada caso habría de plantearse como una situación de resolución de los problemas específicos. En este sentido, es particularmente ilustrativa la imagen, planteada por Shadish, Cook y Houts (1986), de que debemos concebir la investigación y su validez como el aje-

drez, en el que difícilmente encontraremos dos partidas iguales, y en las que a pesar de que hay ciertas reglas particulares (cómo se utiliza y para qué sirve cada pieza) el resultado final (en el caso de la investigación, sus conclusiones y su posibilidad de generalización) será el resultado de un esfuerzo com-

binado y de una organización conjunta de todas las piezas a disposición del investigador. Sólo esta flexibilidad, basada en el conocimiento y en el sentido común, nos permitirá mejorar sustancialmente la investigación educativa y, de paso, la deteriorada imagen de la misma.

Referencias bibliográficas

- Aliaga, F. (1993). *Estrategias de automanejo en situaciones aversivas: un estudio comparativo*. Valencia: Servicio de Publicaciones de la Universidad de Valencia.
- Aliaga, F. (1998). *Proyecto Docente*. Universidad de Valencia. Documento no publicado.
- Anguera, M.T., Arnau, J., Ato, M. Martínez Arias, R., Pascual, J. y Vallejo, G. (1995). *Métodos de investigación en psicología*. Madrid: Síntesis.
- Ball, S. (1997). Unintended Effects in Educational Research. En J.P. Keeves (Ed.), *Educational Research, Methodology and Measurement. An International Handbook*. 2ª Edición. Londres: Pergamon.
- Bisquerra, R. (1989). *Métodos de investigación educativa. Guía práctica*. Barcelona: Ceac.
- Buendía, L.; Colás, P y Hernández Pina, F. (1997). *Métodos de investigación en Psicopedagogía*. Madrid: McGraw-Hill.
- Campbell, D.T. y Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 56, pp. 81-105.
- Campbell, D.T. (1957). Factor Relevant to the Validity of Experiments in Social Settings. *Psychological Bulletin*, 54 (4), pp. 297-312.
- Campbell, D.T. (1974). Qualitative knowing in action research. Comunicación presentada en la *Reunión anual de la American Psychological Association*, Los Angeles California.
- Campbell, D.T. (1986). Relabeling Internal and External Validity for Applied Social Scientist. En W. Trochim (Ed.), *Advances in Quasi-Experimental Design and Analysis*. San Francisco: Jossey-Bass.
- Campbell, D.T. y Overman, E.S. (1988). *Methodology and epistemology for social science: Selected papers*. Chicago: University of Chicago Press.
- Campbell, D.T. y Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. En N.L. Gagé (Ed.) *Handbook for research on teaching*. Chicago: Rand McNally.
- Campbell, D.T. y Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally (Traducción española: Diseños experimentales y cuasiexperimentales en la investigación social. Buenos Aires: Amorrortu. 1973).
- Colás, Mª P. y Buendía, L. (1994). *Investigación educativa*. 2ª edición. Sevilla: Alfar.
- Coleman, F. (1972). The evaluation of Equality of Educational Opportunity. En F. Mosteller y D.P. Moynihan (Eds.), *On Equality of Educational Opportunity*. Nueva York: Random House
- Cone, J.D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, 8, 411-426.
- Cook, T. D. y Campbell, D. T. (1979). *Quasi-Experimentation Design and Analysis Issues for Fields Settings*. Chicago: Rand McNally.
- Cook, T. D. y Shadish, Jr., W. R. (1986). Program Evaluation: The World Science, *Annual Review of Psychology*, 37, pp. 139-232
- Cook, T. D. y Shadish, W.R. (1994). Social experiments: some developments over the past fifteen years. *Annual Review of Psychology*, 45, pp. 545-580.
- Cook, T. D.; Campbell, D.T. y Peracchio, L. (1990). Quasi Experimentation. En M.D. Dunnette y L.M. Hough (Eds.), *Handbook of industrial and Organizational Psychology*, Vol 1, pags. 491-576. 2ª Edición. Palo Alto (CA) Consulting Psychologist Press.
- Cordray, D.S. (1986). Quasi-Experimental Analysis: A Mixture of Methods and Judgment. En W. Trochim (Ed.), *Advances in Quasi-Experimental Design and Analysis*. San Francisco: Jossey-Bass.
- Coyle, S.; Boruch, R. y Turner (Eds.). (1991). *Evaluating AIDS prevention programs: expanded edition*. Washington DC: National Academy Press.
- Cronbach, L. J. (1974). Beyond the two disciplines of the scientific psychology. Comunicación a la asamblea de la A.P.A., 2 de Septiembre. (Traducción española en F. Alvira, M.D. Avia, R. Calvo y F. Morales (Eds.) (1979): *Los dos métodos de las Ciencias Sociales*. Madrid: C.I.S.).
- Cronbach, L.J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.
- De la Orden, A (1998). Comunicación personal.
- Keeves, J.P. (Ed.) (1997). *Educational Research, Methodology and Measurement. An International Handbook*. 2ª Edición. Londres: Pergamon.
- Krathwohl, D. R. (1985). *Social Behavioral Science Research*. Beverly Hills: Sage.
- Kratochwill, T.R. (1978). *Single Subject Research. Strategies for evaluating change*. Nueva York: Academic Press.
- Kruglanski, A.W. y Kroy, M. (1976). Outcome validity in experimental research: A reconceptualization. *Representative Research in Social Psychology*, 7, pp. 166-178.
- López-Barajas, E; López, E. y Pérez Juste, R (1987). *Pedagogía experimental I*. Madrid: UNED.
- Mark, M. (1986). Validity Typologies and the Logic and Practice of Quasi-experimentation. En W. Trochim (Ed.), *Advances in Quasi-Experimental Design and Analysis*. San Francisco: Jossey-Bass.
- Meyer, M. y Conrad, S. (1992). *The Case of Bilingual Education Strategies*. Washington, D.C.: National Academy Press.
- Pérez Juste, R. (1985). Diseño experimental. En A. De la Orden (Dir.), *Investigación educativa. Diccionario de Ciencias de la Educación*. Madrid: Anaya

Validez de la investigación causal.
Tipologías y evolución

- Reichardt, C.S. (1985). Reinterpreting Seaver's Study of Teacher Expectancies as a Regression Artifact. *Journal of Educational Psychology*, 77, pp. 231-236.
- Reichardt, C.S. (1986). *Estimating Effects*. Manuscrito no publicado. Departamento de Psicología, Universidad de Denver.
- Sáez, A., Suárez, J., Aliaga, F. y Bo, R. (1994). La utilización de los procedimientos de comparaciones múltiples en la investigación educativa en España. *Revista de Investigación Educativa*, Nº 23, 396-404.
- Sandín, M.P. (2000). *Criterios de validez en la investigación cualitativa: de la objetividad a la solidaridad*. *Revista de Investigación Educativa*, 18 (1), 223-242.
- Shadish, W.; Cook, T.D. y Houts, A. (1986). Quasiexperimentation in a Critical Multiplist Mode. En W. Trochim (Ed.), *Advances in Quasi-Experimental Design and Analysis*. San Francisco: Jossey-Bass.
- Snow, R.E. (1979). Diseños representativos y cuasirepresentativos para la investigación en la enseñanza. En F. Alvira, M.D. Avia; R. Calvo y F. Morales (Eds.) *Los dos métodos de las Ciencias Sociales*. Madrid: Centro de Investigaciones Sociológicas.
- Suárez, J.; Aliaga, F.; Orellana, N.; Salavert, L.; Sáez, A.; Belloch, C. y Bo, R. (1995). La utilización de los procedimientos de medida del cambio en la investigación educativa en España. En AIDIPE, *Estudios de investigación Educativa en intervención Psicopedagógica*. Valencia: AIDIPE.
- Trochim, W. (Ed.) (1986). *Advances in Quasi-Experimental Design and Analysis*. San Francisco: Jossey-Bass.
- Tuijman, A.C. y Keeves, J.P. (1997). Path analysis and Linear Structural Relations Analysis. En J.P. Keeves (Ed.), *Educational Research, Methodology and Measurement. An International Handbook*. 2ª Edición. Londres: Pergamon.
- U.S. General Accounting Office (1984). *WIC evaluations provide some favorable but not conclusive evidence*. Washington D.C.: U.S. General Accounting Office
- Walker, J.C. y Evers, C.W. (1997). Research in Education: Epistemological Issues. En J. Keeves (Ed.) *Educational Research, Methodology and Measurement. An International Handbook*. 2ª Edición. Londres: Pergamon
- Wienert, F.E. (1997). Translating Research into Practice. En J.P. Keeves (Ed.), *Educational Research, Methodology and Measurement. An International Handbook*. 2ª Edición. Londres: Pergamon.
- Wiggins, J.S. (1966). Substantive dimensions of self-report in the MMPI item pool. *Psychol. Monograph*, 80, 630.

Abstract

To obtain social prestige, the Educational Research need achieve a high grade of Validity but this is not ever present. For this reason, is important clarify properly the subject.

Despite the unanimous importance conferred to seminal work of Campbell and his collaborators (Stanley, Cook, etc.) in the field of Research Validity, a review of most re-

cent texts published in Spain about Educational Research shows that none follow the models of Research Validity exposed by Campell and collaborators since 1979, but in all cases they refer (or uses diverse variations) to older models. To clarify the situation we make a review of the evolution and state, of art over the last few years, as well as a critical analysis of the actual situation of the subject