CIENCIA *ergo-sum*
Universidad Autónoma del Estado de México
ciencia.ergosum@yahoo.com.mx
E-ISSN: 2395-8782

La validez experimental en psicología: una revisión crítica de la literatura

Vera García, Fernando
La validez experimental en psicología: una revisión crítica de la literatura
CIENCIA *ergo-sum*, vol. 25, núm. 2, julio-octubre 2018 | **e14**
Universidad Autónoma del Estado de México, México

Vera García, F. (2018). La validez experimental en psicología: una revisión crítica de la literatura. *CIENCIA ergo-sum*, *25*(2). https://doi.org/10.30878/ces.v25n2a4

E-ISSN: 2395-8782

La validez experimental en psicología: una revisión crítica de la literatura

Experimental Validity in Psychology: A Critical Review of the Literature

Fernando Vera García
Universidad de Edimburgo, Reino Unido
fvera87@hotmail.com

Recepción: 15 de abril de 2016
Aprobación: 05 de mayo de 2017

Resumen:

A pesar de su importancia, el concepto de validez experimental ha recibido escaso desarrollo desde su concepción. Por esta razón, el objetivo es brindar un análisis crítico por medio de un método de análisis filosófico sobre tres ejes: la distinción entre hipótesis alternativas y artefactos experimentales, las listas de amenazas a la validez experimental y la supuesta tensión entre validez interna y externa. Se concluye que el desarrollo del concepto de validez experimental ha sido escaso debido al limitado tratamiento que han recibido tanto los supuestos causales como la incertidumbre, en el contexto experimental. Se finaliza con casos ilustrativos de dichos elementos en el ámbito de la psicología y se hacen recomendaciones metodológicas para mejorar la validez experimental.

Palabras clave: validez experimental, metodología, psicología experimental.

Abstract:

Despite its importance, the concept of experimental validity has received little development since its conception. The purpose of the present study is to provide a critical examination of it. Philosophical analysis of the concept was conducted along three lines: the distinction between alternative hypothesis and experimental artifacts, the use of lists of threats to validity, and the alleged tension between external and internal validity. It is concluded that the concept´s development has been restricted due to a lack of emphasis on the role of both, causal assumptions and uncertainty in experimental contexts. The study ends with case examples from within psychology, as well as methodological recommendations to improve experimental validity.

Keywords: experimental validity, methodology, experimental psychology.

Introducción

Desde su introducción en el clásico artículo de Campbell (1957), el concepto de validez ha pasado por un desarrollo ambivalente. Por una parte, se ha incorporado a buena parte de los libros de texto y enseñanza de metodología en psicología (Reis y Judd, 2014; Cook y Campbell, 1979; Shadish *et al.,* 2002; Kerlinger y Lee, 2002). Por la otra, a pesar del reconocimiento a su importancia por parte de académicos, los desarrollos que se han hecho en torno a las denominadas *amenazas* a la validez han sido limitados, y se ha puesto aún menos atención al desarrollo formal del concepto.

Que tal situación tenga lugar resulta problemático por varias razones. En primer lugar, en la medida en que la investigación científica busca emitir o refutar hipótesis empíricas, sería esperar que se prestara más atención a un concepto que denota la solidez de la investigación. En segundo lugar, la validez de la investigación, sea en la definición de Campbell o en las extensiones que del concepto se han hecho, permea la lista de problemáticas que se han reportado en la literatura psicológica, y entre las cuales figuran el muestreo por conveniencia (Henrich *et al.,* 2010), el error de variables omitidas (Mauro, 1990), el sesgo de confirmación (Nickerson, 1998), entre otros. Así pues, uno esperaría que un tratamiento sistemático del concepto sea benéfico para quienes se interesen por refinar los métodos utilizados en psicología.

Con lo expuesto, el objetivo del estudio es brindar un análisis del concepto de *validez experimental*, así como de las controversias que ha suscitado. Para ello se comienza por una revisión histórica, desde su formulación por parte de Campbell (1957), hasta sus desarrollos y controversias más recientes. Se podrá apreciar que no existen diferencias fundamentales en el abordaje utilizado históricamente, por lo que se le

denominará *enfoque tradicional*. Más adelante, se presenta un análisis teórico-filosófico que permite apreciar las limitaciones del enfoque tradicional al tiempo que aporta posibles respuestas a las controversias suscitadas y vías de enriquecimiento. Tras el análisis, se presentan y discuten casos reales de la literatura en psicología que ilustran las limitaciones del enfoque tradicional.

El artículo finaliza con una reflexión propositiva que permite dar un tratamiento más acertado al concepto de *validez* al tiempo que es congruente con las complejidades de la investigación que el abordaje tradicional no ha logrado capturar.

1. El enfoque tradicional

Como lo indica el título de su artículo seminal, Campbell (1957) propuso los conceptos de validez interna y externa en el contexto de factores (o amenazas) que inciden en los diferentes diseños de investigación. Revisó diversos diseños y se enfocó en los experimentales, donde delineó los factores a los cuales es propenso cada uno. Asimismo, acompañó su discusión de diversas anotaciones, las cuales se revisan en esta sección. Tras su publicación, las tesis de Campbell recibieron gran atención, lo cual es fácil de apreciar si se considera el atractivo que de manera intuitiva genera la noción, pues incorpora tanto la validez de la relación causal establecida en un experimento (validez interna) como la generalizabilidad a otros dominios (validez externa).

Ahora bien, desde el abordaje inicial de Campbell, ha habido contados esfuerzos por desarrollar más a fondo la noción. Estos desarrollos han sido de dos tipos: *a*) por una parte se encuentra la extensión de la lista de amenazas propuesta por Campbell y *b*) por la otra se encuentran las categorizaciones que sobre dicha lista se han realizado. En la primera categoría se encuentran extensiones que van de las siete amenazas originalmente concebidas por Campbell, con incrementos en el camino (Huck y Sandler, 1979; McMillan, 2000), hasta llegar a un estimado de 70 potenciales amenazas propuestas por Onwuegbuzie (2000). A la segunda categoría pertenecen reformulaciones como la de McMillan (2000), quien introdujo subcategorías dentro de la de validez interna y externa, y la de Onwuegbuzie (2000), quien las distinguió según se presenten en la fase de diseño y recolección de datos, análisis o interpretación de resultados. Asimismo, Campbell (Cook y Campbell, 1979) introdujo después los conceptos de validez de constructo y estadística, como categorías diferentes a las dos ya propuestas. Nótese, sin embargo, que la lista de autores que han contribuido al desarrollo del concepto es limitada. En la tabla A1 del anexo se presenta la lista propuesta por Onwuegbuzie (2000), la cual incluye notación especial para diferenciar las contribuciones de autores previos. Esta tabla permite apreciar la creciente cantidad, complejidad y heterogeneidad de amenazas a la validez postuladas.

En lo que respecta a su incorporación a la enseñanza y literatura convencional en la materia, en varios casos el tratamiento que ha recibido el tema es escaso. Por ejemplo, en la *Investigación del comportamiento. Métodos de investigación en ciencias sociales* de Kerlinger y Lee (2002) únicamente dedican una sección de capítulo al tema de la validez, en la cual se limitan a presentar la tesis original de Campbell. De igual modo, en *Handbook of research methods in social and personality psychology* de Reis y Judd (2014) asignan un capítulo independiente a la validez con una limitada discusión de las tesis de Campbell. La misma situación se repite en Coolican (2014).

En el marco de las pocas controversias a que la noción ha dado lugar, el debate se ha centrado exclusivamente sobre dos tesis de Campbell (1957): *a)* que existe una tensión inherente entre la validez interna y la externa y *b)* que el investigador debe siempre dar prioridad a la primera por encima de la segunda. Ante esto, la mayor parte de los revisores se han limitado a explicitar que ambas son importantes (Persson y Wallin, 2012; Tebes, 2000), [1] lo cual sacrifica la necesidad de un análisis riguroso a favor de una recomendación diplomática. Por último, vale la pena mencionar que el concepto de validez también ha sido reformulado para entenderse como uno que alude a la existencia de hipótesis rivales, ante lo cual Campbell propuso más bien pensarlo en términos de artefactos (o interferencia). Sin embargo, esto se ha reducido más bien a una cuestión de estilo.

2. Problemas con el enfoque tradicional

Esta sección presenta un análisis crítico del enfoque tradicional a partir de un sencillo aparato teórico basado en la tesis de Duhem (1962) sobre la refutación de hipótesis experimentales. Para conseguirlo se comienza con un análisis que permite situar debidamente el concepto de validez en la situación experimental, después se analizan las implicaciones de los desarrollos hechos sobre el concepto en términos de listas de amenazas a la validez y finaliza con un replanteamiento de la supuesta tensión entre validez interna y externa.

2. 1. Artefactos o hipótesis alternativas

Una vez concediendo, como lo planteó Duhem (1962), que una hipótesis nunca puede ser puesta a prueba de manera aislada, sino que siempre es en conjunto con una serie de supuestos auxiliares, se puede proceder con un sencillo aparato teórico para representar la situación experimental:

$$[(H \cdot Aux) \rightarrow O] \quad (1)$$

$$[(H \cdot Aux) \rightarrow (E \rightarrow O)] \quad (2)$$

La fórmula 1 afirma que "si la hipótesis (H) y sus auxiliares (Aux) son ciertos, entonces, se observará un determinado fenómeno (O)". Ésta es intercambiable con la fórmula 2 para denotar más explícitamente la situación experimental, donde se plantea que "si la hipótesis (H) y sus auxiliares (Aux) son ciertos, entonces si se realiza un experimento (E), se observará determinado fenómeno (O)". Ahora bien, el término Aux se refiere al conjunto de supuestos que acompañan a la hipótesis y que pueden entenderse de múltiples formas. Por ejemplo, se puede distinguir entre demostrados y no demostrados (la línea divisoria claramente es arbitraria), es decir, entre el supuesto de que la resonancia magnética mide de igual forma los diferentes lóbulos del cerebro (demostrado) y uno que afirma que los niños con autismo comprenden las instrucciones de una prueba (no demostrado). También pueden pensarse en términos de relevancia: entre el supuesto de que la altura de la Ciudad de México no altera la concentración en pruebas de inteligencia (irrelevante) de aquel que afirma que el estado de ayuno sí altera la concentración en pruebas de inteligencia (de nueva cuenta la relevancia también se traza de modo arbitrario). Quizás la mejor forma de caracterizar los supuestos sea distinguiendo entre aquellos que de manera explícita asume el investigador y aquellos que se encuentran implícitos, aunque sobre estos últimos deba notarse que a pesar de ser implícitos, figuran en la estructura lógica del experimento.

Con este aparato teórico se puede dar respuesta a la cuestión de si la validez (y sus amenazas) deban concebirse en términos de *artefactos* o de *hipótesis alternativas*. Un *artefacto*, o variable de confusión, figura en la porción Aux del conjunto, mientras que una hipótesis alternativa figura en la porción H del conjunto. En estricto sentido, lo que se pone a prueba es el conjunto, no sólo la hipótesis, siendo crítico conceder, como hizo Duhem (1962), que la falta de resultados esperados no refuta la hipótesis, sino el conjunto entero. Queda incierto si se trata de un artefacto o de falsedad de la *hipótesis*.

La situación se agrava cuando se consideran aquellos casos en donde los resultados observados son los esperados por la hipótesis, lo cual requiere hacer ajustes al aparato teórico, tanto en H (3) como en Aux (4 y 5):

(3) H puede ser parcialmente correcta. Por ejemplo, pueden existir variables intermitentes mediando el efecto, o una versión más compleja de H (H').

(4) H obtiene O a pesar del factor (F) que opera en su contra.

(5) *H* obtiene *O* con un efecto mayor gracias a un factor (*F*) que magnifica su efecto.

Asimismo, nótese que (4) y (5) pueden pensarse términos probabilísticos como se muestra en (4′) y (5′), o en términos de tamaño del efecto (sobre este punto se volverá más adelante).

(4′) La probabilidad de *O* dado *H* y el auxiliar α (*Aux* α) es menor que la probabilidad de *O* dado *H* únicamente: $P(O/H \cdot Aux\,\alpha) < P(O/H)$.

(5′) La probabilidad de *O* dado *H* y el auxiliar α (*Aux* α) es mayor que la probabilidad de *O* dado *H* únicamente: $P(O/H \cdot Aux\,\alpha) > P(O/H)$.

Vale la pena notar que son este tipo de casos los que llevaron a Cartwright a afirmar que vale la pena sustituir el discurso sobre validez externa por uno de capacidades fijas, en donde los resultados de experimentos son interpretados en dichos términos y concede que el fracaso en la aplicación o replicación de una investigación en otra población se puede deber a factores desconocidos operando en contra (Cartwright, 2010; Cartwright y Munro, 2010); un argumento que ya ha sido aplicado a nivel teórico en el ámbito de la psiquiatría por Hubbeling (2012).

Este análisis sirve más para reconocer el papel de la incertidumbre en el diseño experimental que para la simple respuesta de la disputa entre aparatos e hipótesis alternas, toda vez que hace necesario conceder las fuentes de error tan amplias que pueden incidir en el diseño experimental. Esto se verá reforzado en el siguiente apartado.

2. 2. Las listas enumerativas de amenazas a la validez experimental

Una revisión más detallada de la tabla A1 del anexo hace patente la creciente ambigüedad en la lista de potenciales amenazas, así como la falta de un criterio homogéneo que las unifique. Un análisis permite catalogarlas tentativamente en tres tipos de amenazas: *a*) aquellas que se derivan de métodos de investigación defectuosos, *b*) las que atañen a la estructura del modelo estadístico utilizado para la prueba de significancia de hipótesis nula (NHST en lo sucesivo, por sus siglas en inglés) y *c*) por último aquellas que se relacionan con hechos empíricos que se han descubierto.

Al primer tipo corresponden las supuestas amenazas derivadas del Error de tipo VII, propuesto por Onwuegbuzie (2000) para denotar la falta de verificación de supuestos estadísticos previa a la realización de análisis de varianza y covarianza; de igual modo, pertenece a esta categoría la *confusión en niveles de constructo*. Siguiendo esta línea argumentativa, uno podría proponer como amenazas a la validez errores tan absurdos como captura errónea de los datos o categorización errónea de variables continuas como discretas. Todos estos son casos de errores u omisiones del investigador en tanto que persona, más que de amenazas reales a la validez, por lo que se podrían añadir cuantos errores humanos se pueda concebir.

En cuanto a la segunda categoría, a esta pertenecen las supuestas amenazas como los Errores de tipo I y II, el tamaño del efecto, entre otras. Ahora, que éstas sean fuentes de preocupación legítimas no es lo que se discute, sino más bien el considerarlas como amenazas a la validez. Todas ellas se refieren a la estructura del modelo estadístico utilizado para analizar los datos, y en última instancia son descriptivas por naturaleza. Cómo el investigador las utiliza, es un tema separado. Por ejemplo, que una prueba de hipótesis nula arroje resultados estadísticamente significativos, es sólo indicativo de algo en la medida en que el investigador toma una decisión respecto a la probabilidad de Error del tipo I asociado.

Por último, aquellas amenazas derivadas de fenómenos descubiertos incluyen casos como el efecto Hawthorne, la ansiedad de evaluación, entre otras. Respecto a este tipo de amenazas, hay dos puntos a resaltar: en primer lugar es erróneo suponer que tal lista pueda ser algún día exhaustiva; por ejemplo, considérese el caso del rango de atención de neonatos o fenómenos como la mentira o el engaño. Asimismo, ¿cómo se puede lidiar con ellas? Los intentos por controlarlas, como son incentivos para decir la verdad, o las técnicas para reducir la ansiedad de evaluación acarrean con ellas toda una serie de supuestos (*Aux*); más aún, nada impide que estas mismas amenazas sean manipuladas de manera funcional, por ejemplo, mediante el engaño

de adolescentes de quienes se sospecha conducta rebelde en el contexto experimental. A ambas objeciones subyace el reconocimiento de que la validez siempre depende del contexto del que se trate.

2. 3. La supuesta tensión entre validez interna y externa

Campbell (1957) fue explícito al proponer que existe una tensión inherente entre la validez interna y la externa. Propuso como norma que en casos de duda, siempre habría de ser sacrificada la validez externa en favor de la interna. Ambas tesis son intuitivamente atractivas, pues por una parte conciben a la situación experimental como una de grados de especificidad, en donde resultados aislados pero capaces de emitir juicios causales difieren de resultados amplios pero con considerable varianza de error; por la otra, considerado así el problema, es evidente que uno prefiere emitir juicios causales limitados que juicios inservibles pero amplios. Sin embargo, las implicaciones de dicha tesis son cuestionables, como se verá a continuación.

Para sostener que existe una tensión inherente entre validez interna y externa es necesario aceptar dos proposiciones como condiciones necesarias:

(6) Tesis positiva: los controles que resultan en un incremento en un tipo de validez dan lugar a un decremento en el otro tipo.

(7) Tesis negativa: ausencia de un control para un tipo de validez resulta en un incremento en el otro tipo.

Con las anteriores proposiciones en mente, considérese el caso de fenómenos empíricos como son el efecto Hawthorne o la ansiedad de evaluación. Por ejemplo, un estudio (sea experimental o correlacional) sobre las percepciones de género en donde los participantes responden un cuestionario diseñado para detectar sesgos al respecto. Conscientes de encontrarse en un experimento psicológico, los participantes buscan dar una impresión positiva de sí mismos, la cual difiere de su comportamiento regular. Los resultados, sean positivos o no respecto a una hipótesis dada, no son generalizables a la población de la que se extrajo la muestra, pues más bien constituyen el resultado de un *artefacto experimental*. Inversamente, si el investigador implementa controles que permitan eliminar este artefacto (y concediendo que eso logran), los resultados sí serán generalizables. En este caso, tanto (6) como (7) prueban ser falsas, toda vez que validez interna y externa aumentan o disminuyen de manera simultánea.

En segundo lugar considérese algunos casos de problemas con el uso de pruebas estadísticas (teniendo en mente que, como ya se argumentó, se trata de problemáticas diferentes en su totalidad). Un estudio realizado con una muestra pequeña de pacientes con trastorno obsesivo compulsivo encuentra diferencias estadísticamente significativas entre quienes reciben terapia cognitivo conductual y quienes reciben terapia psicoanalítica. Tal estudio está sujeto en especial a cometer Error del tipo II, así como a la posibilidad de una posterior regresión estadística a la media en cuyo caso se trataría más bien de un *accidente* que no se replicará en el grueso de la población. De nueva cuenta, validez externa e interna van de la mano.

En lugar de evaluar todas las posibles combinaciones y escenarios, tómese ahora un caso aún más crítico: considérese el caso de un investigador que busca poner a prueba la hipótesis de que los videojuegos violentos reducen la sensibilidad ante la agresión. El investigador utiliza un diseño de *pretest/postest* donde la condición experimental es el juego de tales videojuegos, y la variable dependiente es medida mediante el ritmo de EEG durante exposición a estímulos violentos. Ahora bien, si el investigador espera demasiado entre el *pretest* y el *postest*, su estudio se ve amenazado por *historia* y *maduración*, mientras que si espera poco entre una exposición y otra se ve amenazado por la *instrumentación* o el *testeo*. Con este sencillo experimento se ha logrado exponer una verdadera tensión, sólo que ésta se refiere a dos tipos de amenazas a la validez *interna*.

Por último, existe un caso que a primera vista pareciera justificar la tesis de una tensión inherente entre un tipo de validez y el otro: los estudios aleatorios, controlados (RCT en lo sucesivo, por sus siglas en inglés). Este tipo de estudios, considerados como el estándar en investigación clínica y medicina basada en evidencia, cuentan con la siguiente estructura: miembros de una población clínica *P* participan en un experimento en el que son asignados de manera *aleatoria* a uno de dos grupos (experimental o control), siendo los investigadores

ciegos respecto a tal asignación. Después los miembros de un grupo reciben un *placebo*, mientras que los miembros del otro grupo reciben el tratamiento cuya eficacia se desea comprobar. Lo interesante de estos estudios, asumiendo que se realicen de manera apropiada, es que es justamente debido a las tres características descritas que logran establecer un vínculo causal entre el tratamiento y los resultados. La ceguera permite eliminar los sesgos del investigador, el placebo es elegido de tal forma que no tenga efecto causal sobre el padecimiento y la aleatorización permite homogeneizar el resto de los factores causales que podrían actuar, de tal forma que ambos grupos tienen la misma probabilidad de verse afectados por ellos. El problema radica, como lo ha expuesto Cartwright (2010), en la generalización, pues el estudio ha demostrado que el tratamiento funciona en *alguna población*, pero al desconocer la estructura causal que subyace a la elección de poblaciones, es incierto que los resultados se sostengan en otras. Ante esto, parece tentador suponer que sí existe una tensión entre validez interna y externa; después de todo, al homogeneizar factores de confusión entre ambos grupos, se ha perdido la posibilidad de asegurar que la población a la cual se desea generalizar tiene la misma estructura causal que la muestra experimental. Sin embargo, esta conclusión no es asegurada por la evidencia, pues la tesis negativa (7) sigue siendo falsa: eliminar la aleatorización o la ceguera no resulta en una mayor generalizabilidad. Considerados así, los RCT no deben su capacidad de emitir juicios causales a una menor generalizabilidad, sino más bien a su manejo de la incertidumbre (aquella a la cual ya se aludió al presentar el aparato teórico).

En conclusión, se ha visto cómo es discutible que exista una tensión entre la validez interna y la externa. Tal apariencia parece ser más bien el resultado de una concepción errónea sobre la situación experimental que contrapone lo específico de los vínculos causales demostrables de lo general de los vínculos no demostrables.

3. Discusión de casos ilustrativos

En la sección anterior se ha presentado la crítica al enfoque tradicional en los siguientes términos: la variabilidad y grado de incertidumbre que operan en la situación experimental impiden la aplicación de una lista enumerativa y exhaustiva de amenazas a la validez. Más bien, ésta depende siempre del contexto del cual se trata, y de los supuestos causales, implícitos y explícitos, que acompañan a la hipótesis en cuestión. A fin de apreciar esto, a continuación se presentan casos reales a manera de ejemplos que permitan apreciar las problemáticas. Esta lista no pretende ser exhaustiva, precisamente en virtud de dicha variabilidad.

3. 1. El papel del contexto en el diseño: el caso de la hipótesis de teoría de la mente (ToM)

La investigación en el ámbito de teoría de la mente en niños con autismo comenzó con un estudio publicado por Baron-Cohen *et al.* (1985), en el cual los autores compararon los resultados de niños con autismo, síndrome de Down y sanos en una prueba denominada *Sally-Anne*. A grandes rasgos, la prueba pretende determinar si una persona tiene la capacidad de atribuir estados mentales (creencias) a otros. Para el caso de este artículo, no es relevante la validez de constructo de la prueba, ni el hecho de que se trata de un muestreo por conveniencia, más bien considérense las siguientes características de la muestra de niños con autismo: se trató de un grupo de 20 niños, con una media de edad de 11;11, una desviación estándar de 3;0, y un rango de 6;1 a 16;6. La razón de que una muestra tan pequeña y con un rango de edad tan variable sea de interés radica en la cantidad de investigaciones en torno al desarrollo cognitivo infantil y adolescente, así como en torno al tratamiento de personas con autismo. En una muestra tal existe una considerable amenaza de factores como el desarrollo cognitivo u hormonal, la exposición a tratamiento, o el simple hecho de que algunos de los niños eran adolescentes y otros niños pequeños.

Ahora bien, se podría argumentar que la muestra fue de la población con autismo en general, no sólo infantil. Sin embargo, esto pospone el problema, pues hace que la población sea causalmente inapropiada (este

argumento sería equivalente a defender la aplicación de encuestas de preferencias electorales a una población de todos los ciudadanos, incluyendo niños).

Las conclusiones relevantes a extraer del presente caso son dos: por una parte este tipo de controles no serían necesarios en todos los dominios por lo que no pueden ser capturados por una lista enumerativa de amenazas, sino que más bien dependen del contexto (por ejemplo, en una investigación donde un rango de edades similar no se caracteriza por cambios cognitivos u hormonales). En segundo lugar, que tales factores, como ya se argumentó, pueden operar a favor o en contra de la hipótesis: en el presente caso se puede considerar que los adolescentes en el extremo superior de edad pasen la prueba afectando negativamente una hipótesis correcta, o puede ser que la refuten por completo en conjunción con los resultados de niños pequeños.

3. 2. Instrumentación: los casos de diferencias de género y sueños

En un controversial estudio sobre diferencias de género, conducido por Connellan *et al.* (2000), se expuso a 102 bebés con una edad promedio de un día y medio a dos estímulos: un rostro humano y un objeto móvil. Los investigadores midieron el tiempo que los bebés atendían a cada estímulo para probar si había diferencias entre hombres y mujeres. La idea era que quienes dedicaban más tiempo al rostro humano, prefieren o tienen más capacidad de *empatizar*, mientras que el pasar mayor tiempo con el objeto móvil indica preferencia o mayor capacidad para *sistematizar*. Se encontró que los hombres dedicaban más tiempo al objeto y las mujeres al rostro humano. Ahora bien, tanto Fine (2010) como Nash y Grossi (2007) ya se han encargado de criticar este estudio en términos metodológicos como no haber controlado claves de género como globos de colores o juguetes en el cunero o la neutralidad de los mismos investigadores (a diferencia de otros estudios que sí han controlado estos factores). Sin embargo, y a pesar de lo fundadas de sus críticas, éstas omiten abordar los problemáticos supuestos causales respecto a las cualidades de los estímulos, ya que es por pura especulación que los investigadores asumieron la existencia de una cualidad *sistematizadora* inherente a los objetos móviles (cuando bien pudo haber sido el color, el tamaño, etcétera), o *empatizadora* correspondiente al rostro humano.

Ahora bien, es necesario reconocer una crítica de Fine (2010) que sí es acorde con la consideración de supuestos causales. La autora cuestiona el valor de aquellas investigaciones que muestran la presencia de más mujeres en posgrados en psicología (a diferencia de física por ejemplo), como evidencia a favor de esta dicotomía de género. Después de todo, argumenta, el laboratorio de psicología está igualmente cargado por estudios que se valen de diseños experimentales, análisis estadístico y una serie de elementos que sólo una conceptualización sesgada de la disciplina podría oponer a la supuesta *sistematización*.

Por último, considérese el caso de la teoría de simulación de amenazas (TST, por sus siglas en inglés). Revonsuo (2000) propuso la hipótesis de que los sueños cumplen la función adaptativa de ensayar respuestas a amenazas con el fin de mejorar el desempeño ante éstas. Años después, Valli *et al.* (2005) estudiaron varias características de los reportes de sueños (frecuencia, temas, etcétera) en dos poblaciones: una de personas traumatizadas provenientes de Irak y una de personas provenientes de un país europeo. De nuevo, se trata de un muestreo por conveniencia, toda vez que los ciudadanos de Irak tenían mayor probabilidad de ser elegidos en la muestra que los de otros países en guerra. En este caso el problema plateado consiste en asumir que es el antecedente traumático o bélico lo que determina los sueños. De modo plausible se podría asumir que la religiosidad de los sujetos iraquíes, en contraposición a la secularización de los países europeos, es la razón de dichos resultados (considerando el papel tan extenso que los sueños juegan en la religión). Lo importante de este caso es lo siguiente: el muestreo por conveniencia no es, en sí mismo, la raíz del problema, sino las variables causales que le afectan. Se puede concebir una situación en que se muestree de manera aleatoria de todos los países en guerra en un momento dado, pero si todos esos países son religiosos, la variable de riesgo seguiría operando sin ser reconocida por los investigadores. La lección a extraer de estos tres casos corresponde a la

necesidad de hacer explícitos los supuestos causales a favor de un determinado diseño, muestra, o intervención en investigación.

Recomendaciones y conclusiones

En las secciones anteriores se han revisado las limitaciones del abordaje tradicional a la validez. Sin embargo, las críticas basadas en la dependencia del contexto y la incertidumbre constante sobre los factores de confusión acarrean siempre un riesgo de caer en el nihilismo metodológico; después de todo, ¿qué sentido tiene hablar de validez experimental cuando se postula la imposibilidad de prescribir recomendaciones metodológicas específicas? Acaso valdría más la pena seguir recurriendo a los listados de amenazas de Campbell y sus contemporáneos que, a pesar de ser limitados, cuando menos brindan un punto de partida. En esta sección se presentan una serie de recomendaciones metodológicas que permiten preservar el concepto de validez al tiempo que hacen un mejor trabajo en reflejar las complejidades de la investigación del comportamiento y después se abordan las limitaciones del estudio, así como vías de posterior desarrollo.

En primer lugar se recomienda un mayor énfasis en la incorporación de conocimiento previo en el diseño experimental. A pesar de parecer evidente, los casos expuestos como el de teoría de la mente y estudios de género en neonatos prueban hasta qué punto el conocimiento previo sólo es considerado en la formulación de la hipótesis y no en el diseño experimental. Esto también se ha reportado en el ámbito de la psicología laboral, donde las técnicas de muestreo más utilizadas son inapropiadas por no considerar factores como la industria, el nivel jerárquico, o el área de los integrantes de la muestra (Fisher y Sandell, 2015; Landers y Behrend, 2015). [2]

En segundo lugar se recomienda el uso de una medición poco utilizada en el ámbito de prueba de hipótesis nula, concerniente a la estimación del tamaño del efecto (ES, por sus siglas en inglés). Esta medida, a pesar de ser parte indispensable del método de prueba de hipótesis (Cohen, 1988), es reportada con poca frecuencia en la literatura, y aún menor es la frecuencia con la cual se interpreta (Sun *et al.* 2010). Su utilidad radica en permitir una estimación del grado en el que los resultados de la muestra se distancian de los esperados en la hipótesis nula. En la práctica permite incorporar el conocimiento previo en una línea de investigación a la hora de formular hipótesis o de conducir meta-análisis. Cabe mencionar que la estimación del tamaño del efecto es uno de los factores que han contribuido al sesgo de confirmación ya aludido (Nickerson, 1998), toda vez que un resultado estadísticamente significativo como los que se reportan de manera constante en psicología, puede no tener relevancia práctica una vez que se considera el ES.

En tercer lugar, y este es quizás la recomendación más difícil de seguir, dada su variabilidad, los investigadores harían bien en poner más atención a los supuestos causales que subyacen a un determinado experimento o estudio. Por ejemplo, considérense los supuestos detrás de la relación entre psicología y feminidad o entre un objeto móvil y la sistematización. En ambos casos se trata de tesis con una fuerte carga especulativa y deben entenderse en sus contextos específicos con el fin de determinar qué potenciales factores de confusión pueden operar en el estudio.

Antes de introducir la última recomendación relativa a la NHST es necesario situarla en su debido contexto. Tras la acumulación de críticas a la práctica de NHST, la Asociación Psicológica Americana (APA) (2010) se dio a la tarea de investigar la situación y emitir una serie de recomendaciones en la sexta edición de su manual. Estas incluyen reportar la potencia estadística, el tamaño del efecto, entre otras. Si bien estas recomendaciones son un paso en la dirección correcta, no libran a los investigadores de amenazas (o errores) a la validez. Como se vio con los casos de teoría de la mente, es el contexto de investigación el cual debe guiar los controles experimentales, y el simple reporte de resultados estadísticos no es suficiente. Se recomienda que tanto jueces como investigadores evalúen los reportes de investigación de manera crítica haciendo explícitos los supuestos o problemáticas encontradas. Y si esto es considerado una cuestión de sentido común, sólo es

necesario volver a los ejemplos mencionados, pues la lista de artículos publicados en estos ámbitos que no cuentan con tales controles hacen evidente la falta de dicha evaluación crítica.

Ahora bien, dada la revisión que el este estudio realizó del enfoque tradicional, queda claro que se pone más énfasis en la crítica que en la propuesta alternativa. Esto es necesario toda vez que el abordaje tradicional permea libros y enseñanza, pero es de esperar que nuevas propuestas metodológicas puedan complementar a las actuales a fin de robustecer el alcance del concepto.

Por otra parte, la principal limitación del método teórico-filosófico utilizado radica en que debe ser complementado con investigación empírica. Por ejemplo, lo ideal sería realizar estudios de seguimiento sobre los ejemplos ilustrativos utilizados, ya sea mediante replicación independiente o por medio de análisis sobre los datos crudos obtenidos en la investigación original de que se trate. Esto puede ser muy valioso para investigadores que realicen revisiones de la literatura, pues pueden considerar la totalidad de variables contextuales que este estudio apenas ha podido describir.

Por último, el artículo se ha limitado al ámbito de la psicología por tratarse de lugar donde el concepto se concibió y se aplica con más frecuencia. Lo anterior no obsta que sea necesario evaluar el estado de disciplinas vecinas como la sociología, la pedagogía, etcétera, y que también puedan verse beneficiadas de estas consideraciones.

Referencias

APA (American Psychological Association) (2010). *Publication Manual of the APA* (*6th edition*). Washington DC.

Baron-Cohen, S., Leslie, A. y Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37-46. Disponible en https://doi.org/10.1016/0010-0277(85)90022-8.

Campbell, D. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297-312. Disponible en http://dx.doi.org/10.1037/h0040950.

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies*, *147*, 59-70. http://dx.doi.org/10.1007/s11098-009-9450-2.

Cartwright, N. y Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice*, *16*(2), 260-266. Disponible en http://dx.doi.org/10.1111/j.1365-2753.2010.01382.x.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (second edition). New York: Taylor & Francis Group.

Connellan, J., Baron-Cohen, S., Wheelwright, S., Batki, A. y Ahluwalia, J. (2000). Sex differences in human neonatal social perception. *Infant Behavior and Development*, *23*(1), 113-118. Disponible en http://dx.doi.org/10.1016/S0163-6383(00)00032-1.

Cook, T. y Campbell, D. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Boston: Houghton Mifflin.

Coolican, H. (2014). *Research methods and statistics in psychology*. East Sussex: Psychology Press.

Duhem, P. (1962). *The aim and structure of physical theory*. New York: Atheneum.

Fine, C. (2010). *Delusions of gender. How our minds, society, and neurosexism create difference*. New York: Norton.

Fisher, G. y Sandell, K. (2015). Sampling in industrial-organizational psychology: Now what? *Industrial and Organizational Psychology*, *8*(2), 232-237. Disponible en http://dx.doi.org/10.1017/iop.2015.31.

Henrich, J., Heine, S. y Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2), 61-83. Disponible en http://dx.doi.org/10.1017/S0140525X0999152X.

Hubbeling, D. (2012). The application of Cartwright´s concept of capacities to complex interventions in psychiatry. *Journal of Evaluation in Clinical Practice*, *18*(5), 1013-1018. Disponible en http://dx.doi.org/10.1111/j.1365-2753.2012.01909.x.

Huck, S. y Sandler, H. (1979). *Rival hypotheses: alternative interpretations of data based conclusions.* New York: Harper & Row.

Kerlinger, F. y Lee, H. (2002). *Investigación del comportamiento. Métodos de investigación en ciencias sociales*. México: McGraw-Hill.

Landers, R. y Behrend, T. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology*, *8*(2), 142-164. Disponible en http://dx.doi.org/10.1017/iop.2015.13.

Mauro, R. (1990). Understanding L.O.V.E. (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, *108*(2), 314-329. Disponible en http://dx.doi.org/10.1037/0033-2909.108.2.314.

McMillan, J. (2000). Examining categories of rival hypothesis for educational research. *Reporte: Annual Meeting of the American Educational Research Association.* Disponible en http://eric.ed.gov/?id=ED447194.

Mook, D. (1983). In defense of external invalidity. *American Psychologist*, *38*(4), 379-387. Disponible en http://dx.doi.org/10.1037/0003-066X.38.4.379.

Nash, A. y Grossi, G. (2007). Picking BarbieTM´s brain: Inherent sex differences in scientific ability? *Journal of Interdisciplinary Feminist Thought*, *2*(1), 5. Disponible en http://digitalcommons.salve.edu/jift/vol2/iss1/5.

Nickerson, R. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220. Disponible en http://dx.doi.org/10.1037/1089-2680.2.2.175.

Onwuegbuzie, A. (2000). Expanding the framework of internal and external validity in quantitative research. *Reporte: Annual Meeting of the American Educational Research Association.* Disponible en http://eric.ed.gov/?id=ED448205.

Persson, J. y Wallin, A. (2012). Why internal validity is not prior to external validity. *Philosophy of Science Association*, 23rd Biennial Meeting. Disponible en http://philsci-archive.pitt.edu/9171/

Reis, H. y Judd, C. (2014). *Handbook of research methods in social and personality psychology.* Cambridge: Cambridge University Press.

Revonsuo, A. (2000). The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and Brain Sciences*, *23*(6), 877-901. Disponible en http://dx.doi.org/10.1017/S0140525X00004015.

Shadish, W., Cook, T. y Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Belmont: Wadsworth, Cengage Learning.

Sun, S., Pan, W. y Wang, L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*(4), 989-1004. Disponible en http://dx.doi.org/10.1037/a0019507.

Tebes, J. (2000). External validity and scientific psychology. *American Psychologist*, *55*(12), 1508. Disponible en http://dx.doi.org/10.1037/0003-066X.55.12.1508.

Valli, K., Revonsuo, A., Pälkäs, O., Ismail, K.H., Ali, K.J. y Punamäki, R. (2005). The threat simulation theory of the evolutionary function of dreaming: Evidence from dreams of traumatized children. *Consciousness and Cognition*, *14*(1), 188-218. Disponible en http://dx.doi.org/10.1016/S1053-8100(03)00019-9.

Anexo

TABLA A1
Lista de fuentes de invalidez según la taxonomía de Onwuegbuzie (2000)

Amenaza	Tipo	Diseño/recolección de datos	Análisis de datos	Interpretación de datos
Historia[1, 2]	I	X		
Maduración[1, 2]	I	X		
Testeo[1, 2]	I	X		
Instrumentación[1, 2]	I	X		
Regresión estadística (a la media)[1, 2]	I	X	X	X
Sesgo de selección[1, 2]	I	X		
Mortalidad[1, 2]	I	X	X	
Efectos de interacción en la selección[1, 2]	I	X		
Sesgo en la implementación	I	X		
Sesgo en aumento de la muestra	I	X		
Sesgo del comportamiento	I	X		
Sesgo del orden[2]	I & E	Xx		
Sesgo de la observación[2]	I	X	X	
Sesgo del investigador[2]	I & E	Xx	Xx	
Sesgo en el balanceo[2]	I & E	Xx	Xx	
Error en la replicación del tratamiento[2]	I	X	X	
Ansiedad de evaluación[2]	I	X		
Interferencia de tratamientos múltiples[2]	I & E	Xx		
Acomodos reactivos[1]	I & E	Xx		
Difusión del tratamiento	I & E	Xx		
Interacción tiempo x tratamiento	I	X		
Interacción historia x tratamiento[2]	I	X		
Rango restringido[2]	I		X	
Sesgo de búsqueda de "no-interacción"	I		X	
Errores tipo 1 a 10[2]	I		X	
Multicolinearidad[2]	I		X	
Error en la especificación	I &E		Xx	
Violación de supuestos[2]	I		X	
Tamaño del efecto[2]	I			X
Sesgo de confirmación	I			X
Gráficas distorsionadas[2]	I			X
Correlaciones ilusorias	I			X
Crud factor	I			X
Positivo en muchos sentidos	I			X
Error causal	I			X
Validez de población	E	X	x	x
Validez ecológica	E	X		x
Validez temporal	E	X		x
Especificidad de las variables	E	X	x	
Interacción *pretest* / tratamiento	E	X		
Interacción selección / tratamiento[2]	E	X		

Fuente: elaboración propia a partir de Onwuegbuzie (2000), Campbell (1957) y McMillan (2000).
Nota: la primera columna describe la amenaza y los numerales 1 o 2 en superíndice indican si fue postulada previamente por Campbell (1957) o McMillan (2000). La segunda columna indica si se trata de una amenaza a la validez interna (I), externa (E), o ambas (I&E). Las columnas tres a cinco indican la fase del proceso de investigación en la cual se presenta esta amenaza, según el autor. La "X" se utiliza para amenazas a la validez interna y la "x" para amenazas a la validez externa con el fin de poder identificar la fase en que incide cada una, en aquellos casos en que una amenaza afecta a ambos tipos de validez. Dos fuentes de validez externa que propuso Campbell (1957) no permiten una homologación precisa con las amenazas propuestas por Onwuegbuzie (2000). Se trata de *a)* la generalización a otras condiciones y *b)* la generalización a otras poblaciones. Por otra parte, la propuesta de McMillan (2000) cuenta con amenazas que no figuran explícitamente en la de Onwuegbuzie (2000), las cuales son heterogeneidad aleatoria de los respondientes, *outliers*, datos curvilíneos, homogeneidad de los respondientes, correlación y causalidad, ambigüedad sobre la dirección de la causa, variables no consideradas, rivalidad compensatoria, desmoralización con resentimiento, efecto Hawthorne, entre otras.

Notas

[1] Una excepción notable se puede encontrar en Mook (1983), quien acepta que exista dicha tensión, pero no ve como problemático priorizarla por encima de la externa.

[2] Si bien los anteriores ejemplos se centran en el muestreo, esta recomendación es aplicable al resto del diseño experimental, como es el caso del error por variables omitidas (Mauro, 1990).