

# EL DEBATE SOBRE LA VALIDEZ EXTERNA EN ECONOMÍA EXPERIMENTAL

## THE DEBATE ON EXTERNAL VALIDITY IN EXPERIMENTAL ECONOMY

RODRIGO MORO

*Departamento de Humanidades, Universidad Nacional del Sur (UNS), Bahía Blanca, ARGENTINA*  
*IIESS, Universidad Nacional del Sur-CONICET, Bahía Blanca, ARGENTINA*

rmoro@uns.edu.ar

MARCELO AUDAY

*Departamento de Humanidades, Universidad Nacional del Sur (UNS), Bahía Blanca, ARGENTINA*  
*IIESS, Universidad Nacional del Sur-CONICET, Bahía Blanca, ARGENTINA*

ccauday@criba.edu.ar

---

**Abstract.** We focus on the debate on the external validity of lab results in the area of experimental economics. More specifically, we focus on Colin Camerer's and John List's views, who postulate opposite positions on the issue. The problem is that they partially analyze the same empirical evidence and draw from it opposite conclusions. The goal of this work is to analyze such empirical evidence to try to explain the difference in interpretation of results. We argue that the dispute in interpretations can be expressed in terms of difference of criteria used to determine external validity. From our analysis we propose the formulation of different types of external validity.

**Keywords:** Experimental economics • lab experiments • field experiments • external validity • determination criteria

---

RECEIVED: 21/12/2018

ACCEPTED: 02/04/2019

## 1. Introducción

En el año 2015, Guillaume Fréchette y Andrew Schotter editaron y publicaron, a través de Oxford University Press, el *Handbook of Experimental Economic Methodology*. Es un gran libro que incluye los principales temas de interés en el área de economía experimental. El principal debate en juego es sobre la generalizabilidad o validez externa de los resultados de laboratorio del área. La gran pregunta es si los resultados de estudios de decisión hallados en el laboratorio son extrapolables o generalizables a situaciones de decisión fuera del laboratorio o *de campo*, como se las suele llamar. Los principales investigadores del área presentan sus posturas ante este problema. En los polos opuestos se encuentran Colin Camerer, quien defiende la validez externa de los resultados experimentales y John List –en colaboración con Omar Al-Ubaydli–,



quien critica fuertemente dicha postura. Desde el punto de vista epistemológico, el punto más interesante es que analizan parcialmente la misma evidencia empírica y, a partir de ella, sacan conclusiones opuestas. El objetivo de este trabajo será revisar dicha evidencia empírica e intentar explicar por qué se dan estas diferencias en la interpretación de resultados.<sup>1</sup>

## 2. Revisión de los artículos discutidos

A continuación analizaremos los tres artículos principales de los que Camerer y List sacan conclusiones opuestas. Esto no cubre la totalidad de estudios analizados por los autores. Hay estudios (ej. Stoop *et al.* 2012) sobre los que básicamente están de acuerdo en cómo interpretarlos y no se plantean como decisivos para ninguno de los bandos. Y cada autor analiza varios artículos que no son comentados por el otro. Por lo tanto, nuestro análisis no abarca toda la literatura discutida. Nuevamente, nos enfocaremos en los tres artículos que ambos analizan, sacando de conclusiones opuestas. En estos estudios se compara cómo se comportan las personas en una situación de laboratorio y cómo se comportan en una situación de campo análoga, donde los sujetos no saben que sus acciones están siendo observadas y analizadas. El punto global que intentaremos mostrar es que los datos empíricos de cada uno de los estudios no apuntan en una sola dirección. En cada uno de los artículos hay datos que parecerían apoyar la validez externa de los resultados y datos que parecerían ir en contra. Argumentaremos que el desacuerdo podría expresarse en términos de diferencias en los *criterios de determinación* de validez externa. En otras palabras, los autores parecen diferir en los criterios que implícitamente utilizan para establecer si los resultados de laboratorio son o no externamente válidos.

### 2.1. Análisis de List (2006)

List (2006) reporta una serie de experimentos que permiten comparar decisiones en un contexto de laboratorio y en uno similar de campo. La situación de campo sobre la que se monta el estudio son ferias donde se venden cartas deportivas (*sport cards*) y otros ítems de colección relacionados. El punto central es que la calidad de estas cartas puede ser independientemente determinada por expertos en una escala ya establecida de 5 puntos. El foco de la investigación está en la decisión de los vendedores. La situación consiste en que un comprador se acerca y ofrece \$20 demandando una carta de calidad 4 puntos del beisbolista Frank Thomas y en otro momento (antes o después) otro comprador ofrece \$65, demandando una carta de calidad 5 puntos del mismo beisbolista. Lo que se evalúa es la calidad de la carta dada por el vendedor ante ambas ofertas. En la situación de laboratorio, los participantes sabían que

estaban participando de un experimento y que la calidad de la carta brindada iba a ser posteriormente evaluada. En la condición de campo, los participantes estaban en la feria y se les acercaban dos compradores (de nuevo, en momentos distintos), que les hacían una oferta específica (\$20 o \$65), con la demanda idéntica a la del laboratorio. En términos técnicos, el juego correspondería a un juego de intercambio de regalos (*gift exchange game*).

La pregunta es, entonces, si los participantes se comportaron de manera similar en uno y otro contexto. Argumentaremos que la respuesta depende de la manera en que evaluemos dicha similitud o, en términos técnicos, depende del *criterio que utilicemos para determinar validez externa*.

Analicemos primeramente los resultados globales (posteriormente nos enfocaremos en un análisis más detallado). Por un lado, los vendedores entregaron en promedio cartas en el laboratorio de calidad significativamente superior a las que entregaron en el campo (véase tabla 1). Como puede apreciarse en la tabla, para cada nivel de oferta, el promedio de calidad de cartas entregadas en el laboratorio es superior en aproximadamente 1 punto al promedio de cartas entregadas en el campo. Así, por ejemplo, para el nivel de oferta de \$20, en el laboratorio los vendedores entregan cartas de 3.1 de calidad promedio mientras que en el campo entregan cartas de 2.1 de calidad promedio. Y lo mismo ocurre para el nivel \$65, siendo estas diferencias estadísticamente significativas. La interpretación de List es que, por cuestiones de reputación (i.e., no quedar mal posicionados ante los experimentadores), las personas suelen ser mucho más pro-sociales en el laboratorio que en el campo. Los participantes, al saber que sus decisiones estarán bajo el escrutinio de los investigadores, tenderían a comportarse mejor de lo que lo harían cuando piensan que sus decisiones no serán registradas y analizadas. Al-Ubaydli y List (2015) toman esta diferencia significativa como evidencia negativa contra la validez externa de los datos de laboratorio. Consideremos, entonces, qué criterio se estaría implícitamente sosteniendo en dicha argumentación. El criterio en cuestión requeriría una *similitud de tipo cuantitativa de las variables de comportamiento de interés* (en este caso, la calidad de la carta entregada). Como consecuencia de este criterio, los datos del laboratorio deberían permitir realizar estimaciones adecuadas de valores numéricos de las variables de campo de interés. Nuevamente, en nuestro caso analizado, al utilizar los datos del laboratorio, los valores promedio de las cartas de campo resultarían sistemáticamente sobreestimados. Los autores argumentan que el fenómeno es robusto y señalan varias investigaciones que apoyarían la misma idea: en el laboratorio las personas suelen ser más pro-sociales de lo que son en el campo.

Por otra parte, hay más conclusiones para extraer de los datos de List (2006). En particular, la manipulación del precio ofrecido (\$20/\$65) produjo el mismo efecto global en ambos ambientes: cuando el precio ofrecido por el comprador aumentó, la calidad promedio de la carta entregada aumentó significativamente en ambos am-

bientes. Esto también puede apreciarse en la misma tabla 1 señalada anteriormente. Nótese que en cada uno de los ambientes, el paso de oferta de una oferta de \$20 a una de \$65 produce un aumento aproximado de 1 punto en el promedio de calidad de cartas entregadas. Estrictamente hablando, no importa demasiado que el aumento haya sido aproximadamente el mismo en ambos ambientes sino que el efecto ocurra en el mismo sentido, es decir, en ambos ambientes, al aumentar la oferta, aumenta la calidad de la carta. En la literatura del área se suele usar el concepto de *reciprocidad positiva*: si un agente recibe algo bueno de otro agente o es tratado bien por otro agente, tiende a reciprocitar. Así, en ambos ambientes se registraría el mismo fenómeno de reciprocidad positiva para el juego en cuestión. Camerer (2015) toma esto como evidencia positiva para la validez externa del estudio. Así, consideremos el criterio implícito que se estaría utilizando para establecer validez externa. En este caso, el criterio en cuestión requeriría una *similitud de tipo cualitativa de las variables de comportamiento de interés*. Como consecuencia de este criterio, las variables de interés que son manipuladas en un experimento deberían generar efectos en la misma dirección en ambos ambientes (laboratorio y campo). Camerer sostiene que la gran mayoría de los estudios donde se pueden comparar los resultados de laboratorio con una situación análoga de campo muestran lo mismo: las variables generan efectos en la misma dirección en ambos ambientes.

Hasta aquí la cuestión puede dar la falsa impresión de ser relativamente simple: si bien la evidencia empírica no parecería respaldar una similitud de tipo *cuantitativa*, sí respaldaría una similitud de tipo *cualitativa* (Kessler y Vesterlund 2015). Sin embargo, la cuestión es más compleja. Más específicamente, los datos empíricos son mucho más ricos en información.

Notemos fundamentalmente que los análisis comentados hasta el momento, y los criterios resultantes, se basan en los resultados generales extraídos de la *muestra global* del estudio. ¿Pero qué encontramos si analizamos, no ya la muestra global, sino *sub-poblaciones* de la misma? Volviendo al primer estudio de interés, List (2006) divide la muestra entre vendedores locales, quienes regularmente asisten a la feria y vendedores no-locales o visitantes, quienes están de paso. Según List, aquí también hay una diferencia en la reputación en juego, ya que los vendedores locales probablemente se encuentren en el futuro cercano con los compradores y esto no valga para los vendedores visitantes, quienes probablemente no vuelvan a ver a sus compradores. El resultado hallado es que los vendedores locales tienden a responder con reciprocidad al aumento de oferta en ambos ambientes, pero los vendedores visitantes no lo hacen: muestran reciprocidad en el laboratorio pero no en el campo (véase Figura 1). Al-Ubaydli y List (2015) toman este dato como evidencia negativa para la validez externa. Consideremos, entonces, el criterio implícito utilizado en dicha argumentación. El criterio continuaría requiriendo una *similitud cualitativa* de las variables de interés pero ahora aplicado a *las subpoblaciones de interés (teórico)*.

El interés teórico de la subpoblación de vendedores visitantes estaría dado porque, justamente, en el ambiente de campo tendrían anuladas o muy disminuidas sus motivaciones de reputación. Y claramente, la similitud de comportamiento en ambos ambientes no se cumple para los vendedores visitantes. Nótese también que aquí la diferencia no es cuantitativa sino cualitativa: los vendedores visitantes son recíprocos en el laboratorio pero no así en el campo.

Camerer (2015) presenta una crítica realizando un análisis *intra*-sujeto, considerando aquellas personas que participaron en ambas condiciones, de laboratorio y de campo. Haciendo este recorte, las diferencias entre vendedores locales y visitantes ya no son estadísticamente significativas. Hay dos problemas con esta crítica. En primer lugar, no es claro, desde el punto de vista metodológico, que el análisis *intra*-sujeto sea el más adecuado. El problema es que las personas o bien pueden buscar consistencia o bien pueden buscar divergencia al enfrentarse a condiciones similares y no responder como naturalmente lo harían si esa similitud de condiciones estuviera ausente. Por lo tanto, lo más limpio metodológicamente suele ser usar distintos sujetos para distintos tratamientos y realizar el análisis *inter*-sujeto correspondiente. Y en segundo lugar, al realizar este recorte de la muestra, el número de participantes resulta extremadamente bajo y, por lo tanto, no es extraño que los resultados ya no sean estadísticamente significativos. Así, la crítica de Camerer no resulta satisfactoria.

Treatment Lab-Context	
\$20	3.1 (.9)
\$65	4.1 (.6)
Treatment Floor (Cards)	
\$20	2.1 (.9)
\$65	3.2 (1.0)

Figura 1: Promedio de calidad de carta entregada por vendedores en función de la oferta (\$20/\$65) y el contexto (laboratorio (lab) o feria (floor)); en paréntesis se señala la desviación estándar (extraído de List 2006).

## 2.2. Análisis de Benz y Meier (2008)

En este estudio, los autores obtuvieron datos de campo sobre donaciones y eso les permitió crear una situación análoga en el laboratorio para comparar la conducta de los participantes en ambos ambientes. Más específicamente, en la Universidad de

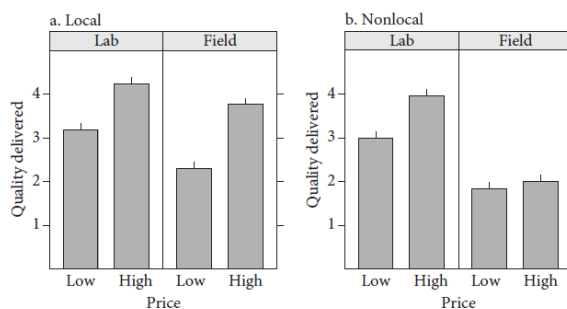


Figura 2: Promedio de calidad de cartas entregadas en función de la oferta (low= \$20 vs high= \$65), contexto (laboratorio (lab) vs. campo (field)) y tipo de vendedor (local vs. no local) (extraído de Camerer 2015, sobre datos provistos en List 2006).

Zúrich en Suiza los alumnos al inscribirse en cada cuatrimestre de cursado tienen la posibilidad de donar alrededor de FS\$ (francos suizos) 7 a un fondo de ayuda a alumnos con bajos ingresos y alrededor de FS\$ 5 a un fondo de ayuda a alumnos extranjeros. Esto lo realizan on-line, cliqueando en el espacio correspondiente. Benz y Meier convocaron a alumnos de dicha universidad a participar de un estudio experimental que consistía en darles una dotación inicial de unos FS\$ 12, con la misma división de 7 y 5 para cada uno de los fondos habituales, como en la situación de campo. Los participantes debían elegir en cada caso cuánto de esa dotación quedarse para ellos mismos y cuánto donar (i.e., juego del dictador). Así, en un primer experimento las organizaciones que recibirían la donación eran exactamente las mismas que en la donación de campo de cada semestre, con los mismos montos. En un segundo experimento, para examinar la robustez del resultado, cambiaron las entidades que recibirían las donaciones y las cantidades monetarias del juego (FS\$ 9 para cada una de las entidades en juego, que eran distintas de las dos habituales). Organizaron los datos, dividiendo a los participantes en tres grupos, de acuerdo a sus donaciones de campo: los que no donan ni una sola vez (realizan análisis separados considerando 2 años antes del experimento y considerando 2 años después al mismo), los que donan alguna vez pero no siempre y los que donan siempre.

Aquí lo que se ve nuevamente es que tanto Camerer como List encuentran datos que favorecen sus propias posiciones. Por una parte, el comportamiento global en el laboratorio correlaciona positivamente con el de campo: cuanto más se dona en el campo, más se dona en el laboratorio y viceversa. Así, el valor del laboratorio sería informativo del de campo, porque permitiría estimarlo, haciendo los ajustes adecuados. Véase la figura 2, que muestra gráficamente los datos del segundo experimento. Como puede apreciarse en la figura, aquellos que no donan nunca en el campo son

los que menos donan en promedio en el laboratorio. Asimismo, los que donan de manera intermedia en el campo (a veces pero no siempre) son los que donan de manera intermedia en el laboratorio. Y finalmente, los que donan siempre en el campo son los que más donan en promedio en el laboratorio. Crucialmente, se realiza un análisis estadístico de los datos y se concluye que hay una correlación positiva significativa entre el comportamiento de campo y el del laboratorio. Este es el argumento que Camerer toma como evidencia positiva para la validez externa del estudio. Así, nuevamente, el criterio implícito requeriría *similitud cualitativa* en el comportamiento, analizando la *muestra global* del estudio. Particularmente, cuando las variables no se manipulan sino que se miden, tal criterio tendría como consecuencia que las variables de laboratorio deberían correlacionar positivamente con las variables de campo.

Por otra parte, hay un dato extremadamente llamativo que parece ir en dirección contraria: aquellos que no hacen ninguna donación en el campo (antes/después del experimento), donan entre un 50 y un 67% de la dotación dada en el laboratorio. En la figura 2 señalada anteriormente puede verse que los que no donan nunca en el campo donan aproximadamente 9 de FS\$ 18, o sea, aproximadamente el 50% de su dotación. Nuevamente, eso indicaría que en el laboratorio la gente es más pro-social de lo que es en el campo (apoyando la tesis de List). Ese dato es marcado por Al-Ubaydli y List (2015) como evidencia negativa en contra de la validez externa del resultado de laboratorio. Nuevamente, el criterio implícito utilizado requeriría que las *subpoblaciones* de interés se comporten de manera *cualitativamente* similar en los ambientes de laboratorio y de campo. Claramente, la subpoblación de los que no donan nada en el campo *no* se comporta de manera similar en el laboratorio.

Es menester mencionar las objeciones que presenta Camerer sobre tal divergencia encontrada, porque son planteos razonables. Una de las potenciales causas de la divergencia de conducta en el caso de los que no donan nada en el campo sería la siguiente. Ya se ha mostrado que en el juego del dictador las personas son menos generosas cuando a la dotación “se la ganan”, por ejemplo, completando una tarea inicial (Cherry *et al.* 2002). Esto, unido a la literatura de *mental accounting* (Thaler 1985), sugeriría que las dotaciones de laboratorio puedan considerarse de la cuenta “dinero para donar”. Otro aspecto de divergencia es que en el laboratorio se pueden realizar donaciones parciales, con incrementos de 50 centavos. En cambio, en la decisión de campo, la donación es de todo o nada. Así, si las situaciones no son completamente análogas, la diferencia hallada en el comportamiento de los que no donan nunca en el campo pueden deberse a esos factores mencionados y no al contexto laboratorio/campo.

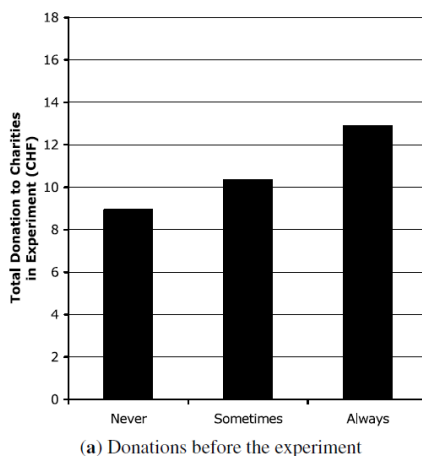


Figura 3: Promedio de donaciones de laboratorio para cada grupo definido de acuerdo a sus donaciones de campo previas, 2do experimento (extraído de Benz & Meier 2008).

### 2.3. Análisis de Palacios-Huerta & Volij (2008)

Los autores analizan juegos de suma 0, donde el equilibrio está en estrategias mixtas y conviene aleatorizar (*randomize*). Se pretende representar en forma abstracta la situación de los penales en fútbol. Los datos de campo se extraen de registros de eventos en partidos de fútbol y parecen apoyar la idea que los jugadores profesionales aleatorizan cuando patean/atajan penales, tal como es predicho por la Teoría de Juegos. Volviendo al estudio de laboratorio, los autores encuentran que los profesionales del fútbol deciden en el juego abstracto de manera muy similar a la predicción de teoría de juegos, mientras los estudiantes no lo hacen. De los estudiantes no se recolectaron datos de campo de pateo/atajada de penales, por lo que no es posible realizar la comparación laboratorio-campo. Camerer argumenta que el paralelismo de habilidades de jugadores profesionales entre el contexto de campo y de laboratorio va a favor de la extrapolación de resultados. Aquí el criterio de validez externa requeriría la similitud *cualitativa* de comportamiento de cierta *subpoblación* de interés, a saber, personas con cierto nivel de experiencia en la situación de decisión en cuestión. Por otra parte, List realiza dos críticas. Por un lado, sostiene que el fenómeno reportado en este estudio no es robusto, porque varios estudios van en dirección opuesta. Esta crítica merece atención pero excede nuestro análisis. Por otro lado, argumenta que el hecho de que los estudiantes jueguen de manera distinta que los profesionales es la clave, porque justamente la muestra típica de los experimentos de laboratorio está formada por estudiantes universitarios. Así, los resultados de este estudio irían



en contra de la validez externa del área. El criterio implícito que se estaría utilizando, entonces, requeriría que las distintas *subpoblaciones* deberían comportarse de manera *cualitativamente* similar a como lo hacen participantes de la *muestra típica de los estudios de laboratorio*, es decir, deberían comportarse como lo hacen estudiantes universitarios de grado. En principio, este criterio parece demasiado exigente. La razón es que parece sensato asumir que la experiencia repetida en cierta situación genera patrones de conducta distintos y, particularmente, más eficientes que los patrones de las personas que se inician en la situación en cuestión. Es cierto que tal vez puede no ser el mejor ejemplo para defender la validez externa de toda el área, pero ciertamente, el resultado hallado parece positivo: en la única subpoblación que pueden establecerse comparaciones empíricas, los agentes parecen comportarse de la misma manera en el campo que en el laboratorio.

### 3. Análisis global de los estudios analizados

De la sección anterior parece razonable concluir los autores están utilizando distintos criterios para evaluar la similitud de resultados entre el campo y el laboratorio. Así, la cuestión de la validez externa es ambigua y, planteada de manera general, tal vez no tenga demasiado sentido. ¿Hay similitud entre los resultados de campo y los de laboratorio? Una respuesta adecuada debe contemplar que hay distintas maneras en que dos resultados pueden ser similares. Sólo cuando se ha logrado consensuar el criterio, es posible extraer conclusiones aceptables.

La pregunta es, entonces, si es posible extraer de todo lo anterior cierto resumen ordenado preliminar de los *tipos* de validez externa que es posible evaluar. De manera general, pudimos ver que hay dos ejes que permiten realizar comparaciones. Por un lado, está el *tipo de similitud* a determinar, que puede ser de tipo *cuantitativa* o *cualitativa*. Por otra parte, está el *rango de análisis* a considerar, es decir, si el análisis se enfoca en la *muestra global* del estudio o si se enfoca en las *subpoblaciones* de interés. El cruce de estos factores daría lugar a cuatro tipos de potencial validez externa: similitud cuantitativa global, similitud cuantitativa de subpoblaciones, similitud cualitativa global y similitud cualitativa de subpoblaciones.

Así, con respecto al tipo de validez denominado *similitud cuantitativa global*, el estudio de List (2006) parece *no* respaldar este tipo de validez, ya que los vendedores tienden a dar en el laboratorio cartas sistemáticamente superiores a las que dan en el campo. Los resultados de Benz y Meier (2008) parecen ir en la misma dirección. No obstante esto, también es cierto que este tipo de validez externa no es considerada por algunos como fundamental (Kessler y Vesterlund 2015). En otras palabras, siendo el comportamiento humano fuertemente contextual, es bastante esperable que los resultados de laboratorio no ofrezcan estimaciones cuantitativas extremadamen-

te cercanas al valor de campo buscado, y tal vez requerir eso sea exigir demasiado. Dado esto, es aún mucho más difícil que se registre el tipo de validez denominado *similitud cuantitativa de subpoblaciones*, ya que requeriría que las subpoblaciones de interés tengan aproximadamente los mismos valores en las variables de comportamiento de interés.

En cuanto al tipo *similitud cualitativa global*, en cambio, tanto el estudio de List (2006) como el de Benz y Meier (2008) parecerían respaldar la presencia de este tipo de validez. El estudio de List (2006) mostraría que el aumento de cantidad de dinero ofrecida por el comprador genera en el vendedor en ambos ambientes un aumento en la calidad de carta entregada. Asimismo, el estudio de Benz y Meier (2008) muestra que hay una correlación positiva entre las donaciones de laboratorio y las de campo.

Pero como vimos, no son todas buenas noticias. Si se considera el tipo de validez denominado *similitud cualitativa de subpoblaciones*, el dato principal es del estudio de Benz y Meier (2008) donde la subpoblación muestral de los que no donan absolutamente nada en el campo dona entre el 50 y el 67% de su dotación en el laboratorio. De la misma manera, en el estudio de List (2006), el dato clave es que la subpoblación de los vendedores visitantes muestran reciprocidad positiva en el laboratorio pero no lo hacen en el campo. Así, ambos estudios muestran que los resultados cualitativos globales hallados no son tan robustos, ya que dejan de valer en ciertas subpoblaciones de interés. En consecuencia, ninguno de estos dos estudios parece respaldar este último de tipo de validez externa. Por otra parte, el estudio de Palacios-Huerta y Volij (2008) sí parecería poseer este tipo de validez externa, ya que la subpoblación de interés (jugadores de fútbol) parece comportarse de manera cualitativamente similar en ambos ambientes.

Antes de pasar a la última sección, es menester reconocer las limitaciones de nuestro trabajo. Particularmente, queremos destacar que la tipología sugerida de criterios de validez externa proviene del análisis de pocos estudios. Por lo tanto, con el análisis de más estudios, los tipos propuestos podrían cambiarse y sobre todo, ampliarse. No solo consideramos este avance posible sino también deseable. De todas maneras, también queremos destacar que esto no afectaría en absoluto el punto central de nuestro artículo: cuando hay una disputa acerca de la validez externa de cierto resultado de laboratorio, es necesario comenzar clarificando qué criterio de validez externa se va a utilizar. Sin eso, es imposible mantener un diálogo racional entre las distintas posturas.

#### **4. Sobre la cuestión global de validez externa de toda el área**

Resta preguntarse por la cuestión global de la validez externa de toda el área. Es menester comenzar dejando en claro que la cuestión global excede el objetivo de

este artículo, ya que habría que analizar toda la literatura relevante del área para pronunciarse con fundamento sobre el tema. El análisis de los estudios descriptos parecería apoyar cierto tipo de validez externa, aunque también ser incapaz de respaldar otros. Sin embargo, no parece sensato generalizar a partir de pocos casos. Hechas estas salvedades, realizaremos para finalizar un par de comentarios sobre el tema.

Hay un punto correcto en la posición de John List y compañía (véase Levitt y List 2007). Muchas veces se presenta como principal hallazgo de la literatura del área que las personas son altruistas, cooperativas y tienden a mostrar reciprocidad positiva. Parece razonable aceptar que las condiciones de laboratorio -el escrutinio de las decisiones, entre otras- favorezcan este tipo de conductas pro-sociales. Y estando ausentes estas condiciones en situaciones de campo de interés, probablemente encontremos conductas mucho más egoístas y anti-sociales. Así, parece razonable la sugerencia de sospechar de la extrapolación de resultados de laboratorio “optimistas” sobre la conducta humana.

Aceptada esta postura, es necesario establecer varios puntos que muestran que el problema puede no ser tan grave como parece inicialmente.

En primer lugar, hay estudios *de laboratorio* que muestran que, cambiando las condiciones típicas, surgen ya resultados poco optimistas de la conducta humana. Por ejemplo, considérese el estudio sobre el juego del dictador de Cherry *et al.* (2002), mencionado anteriormente: cuando la dotación es ganada por el participante (por realizar una tarea inicial) y el anonimato es llevado al extremo (formato doble-ciego), casi ningún participante decide realizar donaciones positivas al receptor (Cherry *et al.* 2002).

En segundo lugar, y obviando los estudios recién mencionados, hay que considerar que, en muchas situaciones de campo, el factor reputación puede ser tan grande o mayor que en el laboratorio. Nótese, sin ir más lejos, el estudio de List (2006). Los vendedores locales se comportaron de la misma manera en el laboratorio que en el campo, probablemente porque la demanda de reputación haya sido similar en ambos ambientes.

En tercer lugar, hay muchas situaciones y estudios experimentales sobre ellas, donde existen demandas sociales contrapuestas, por lo que ya deja de ser claro qué cuenta cómo conducta pro-social. Un ejemplo de esto son los estudios experimentales de corrupción (véase un compendio en Abbink y Serra 2012). Allí los participantes tienen la posibilidad de ofrecer o aceptar “coimas”, habiendo terceros perjudicados como consecuencia de las transacciones corruptas. En estos estudios, si un participante establece relaciones de reciprocidad positiva con el compañero asociado, perjudica a uno o varios terceros, sean ONGs u otros participantes pasivos. Y si uno se abstiene de perjudicar a terceros, renuncia a establecer relaciones recíprocas con el compañero asociado.

Finalmente, una gran porción de juegos no involucran demandas sociales, por lo que el problema directamente no surge (ej. los estudios sobre funcionamiento de subastas o los juegos de suma cero, como en Palacios-Huerta y Volij 2008). O incluso en otros juegos la demanda puede ir en sentido contrario: ante un juego donde las personas pueden destruir parcialmente la dotación del otro jugador –aún debiendo pagar para ejercer tal posibilidad y sin beneficios propios–, también tienden a hacerlo (Abbink y Herrmann 2011).

Por todas estas razones, si bien es siempre aconsejable ser cuidadoso al realizar extrapolaciones, el problema planteado por List y compañía no nos parece tan grave como puede parecer inicialmente. Más allá de eso, conviene finalizar resaltando uno de los pocos puntos en común en el que coinciden Camerer y List: hay que continuar realizando estudios comparativos de campo/laboratorio para mejorar nuestro entendimiento de los paralelismos y sus falencias.

## Referencias

- Abbink, K. & Herrmann, B. 2011. The moral costs of nastiness. *Economic Inquiry* 49(2): 631-633.
- Abbink, K. & Serra, D. 2012. Anticorruption Policies: Lessons from the Lab. In: D. Serra & L. Wantchekon (ed.), *New advances in experimental research on corruption*, p.77-115. Bingley, UK: Emerald Group Publishing Limited.
- Al-Ubaydli, O. & List, J. 2015. On the generalizability of experimental results in economics. In: G. Frechette & A. Schotter (ed.) *Handbook of Experimental Economic Methodology*, p.420-462. New York, NY: Oxford University Press.
- Benz, M. & Meier, S. 2008. Do people behave in experiments as in the field?—evidence from donations. *Experimental Economics* 11(3): 268-281.
- Bhaskar, R. 2013. *A realist theory of science*. New York, NY: Routledge.
- Camerer, C. 2015. The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. In: G. Frechette & A. Schotter (ed.) *Handbook of Experimental Economic Methodology*, p.249-295. New York, NY: Oxford University Press.
- Cherry, T.; Frykblom, P. & Shogren, J. 2002. Hardnose the dictator. *The American Economic Review* 92(4): 1218-1221.
- Fontaine, P. & Leonard, R. 2005. *The experiment in the history of economics*. New York, NY: Routledge.
- Fréchette, G. & Schotter, A. (ed.) 2015. *Handbook of experimental economic methodology*. New York, NY: Oxford University Press.
- Hands, D. 2001. *Reflection without rules: economic methodology and contemporary science theory*. Cambridge: Cambridge University Press.
- Kessler, J. & Vesterlund, L. 2015. The external validity of laboratory experiments: The misleading emphasis on quantitative effects. In: G. Frechette & A. Schotter (ed.) *Handbook of Experimental Economic Methodology*, p.391-406. New York, NY: Oxford University Press.

- Levitt, S. & List, J. 2007. What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives* 21(2): 153-174.
- List, J. 2006. The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy* 114(1): 1-37.
- Palacios-Huerta, I. & Volij, O. 2008. Experientia docet: Professionals play minimax in laboratory experiments. *Econometrica* 76(1): 71-115.
- Radder, H. 2009. The philosophy of scientific experimentation: a review. *Automated Experimentation* 1(2): 1-8.
- Stoop, J.; Noussair, C. & Van Soest, D. 2012. From the lab to the field: Cooperation among fishermen. *Journal of Political Economy* 120(6): 1027-1056.
- Thaler, R. 1985. Mental Accounting and Consumer Choice. *Marketing Science* 4(3): 199-214.

## Notes

<sup>1</sup>Deseamos en este punto agradecer a ambos revisores por señalarnos el vínculo de nuestro trabajo con algunos problemas filosóficos relevantes dentro de la filosofía de la ciencia. Dichas cuestiones remiten, entre otras, al problema general de la objetividad científica en el ámbito de la investigación social; en particular, a las complejidades adicionales que involucra la experimentación en ciencias sociales respecto de las ciencias naturales, debido al doble rol del experimentador como sujeto y objeto del proceso de investigación experimental. También es cierto que nuestro trabajo se vincula a problemáticas dentro del área general de filosofía de la experimentación, como la de sistematizar los roles que cumplen los experimentos en ciencia y establecer las diversas relaciones entre teoría y experimento (Radder, 2009). Sin embargo, queremos dejar en claro que el no tratar dichos temas obedece, por una parte, a un criterio metodológico de ser prudentes en el alcance de nuestro trabajo (no abarcar temas o problemas que no pudiéramos analizar con la profundidad que se merecen). Por otra parte, y más específicamente, nuestro trabajo se inscribe en una discusión ya establecida y, en tal sentido, tomamos como marco de referencia dicha discusión en los términos en los que ha sido originalmente planteada. Más allá de estas consideraciones, agradecemos y recomendamos la lectura de Bhaskar (2013), Hands (2001) y Fontaine & Leonard (2005).