



Universitat d'Alacant
Universidad de Alicante

Essays on Experimental Methods Applied
to Different Environments

Roberto Di Paolo



Tesis **Doctorales**

UNIVERSIDAD de ALICANTE

Unitat de Digitalització UA
Unidad de Digitalización UA



Universitat d'Alacant
Universidad de Alicante



SCHOOL
FOR ADVANCED
STUDIES
LUCCA

Departamento de Fundamentos del Análisis Económico

Facultad de Ciencias Económicas y Empresariales

IMT School for Advanced Studies in Lucca

Track in Economics, Networks and Business Analytics

Essays on Experimental Methods Applied to Different Environments

Roberto Di Paolo

Tesis presentada para aspirar al grado de

DOCTOR POR LA UNIVERSIDAD DE ALICANTE

Mención de Doctor Internacional

Doctorado en Economía

Dirigida por:

Prof. Giovanni Ponti

Prof. Ennio Bilancini

Acknowledgements

I want to thank my supervisor Giovanni Ponti for all his support during these years. Thanks to him, I have learnt many aspects of experimental economics that I didn't know. His advice, commitment, and patience have been a key part of my PhD.

I also want to thank Ennio Bilancini for his comments, suggestions, and support during my years in Lucca. Thanks to him, I have known an even bigger world in which to do research, combining work and pleasure. I am very grateful to have you as my co-supervisor.

I will always be indebted to both of you.

I also want to express my gratitude to Adam Sanjurjo and Iñigo Iturbe-Ormaeche for reading my thesis and helping me to improve it. My thanks are also intended to the director of the department María Dolores Collado. I want to thank all faculty members and classmates of the Quantitative Economics Doctorate Program at the University of Alicante. Your support and help during these years are unforgettable.

I am also thankful to all faculty members and new friends I have met in Lucca. Especially, I'm grateful to all the members of the B.E.E. group; they always helped me to improve my works and support me during these years. I would also thank Leonardo and Valerio; I learnt a lot from you during these years. I hope our collaboration will continue.

I owe a special thanks to my co-authors and "experimental-mates" Judit Alonso Berná and Tatiana Celadin. Their support was essential to grow, learn and move forward in hard times.

This achievement would not be possible without the support of my mom, Mirella. I thank you for believing in me even when I don't, for pushing me to keep going when things get hard, for teaching me that reality is harsh but feasible.

Last but not least, to my partner Barbara. Thanks for being by my side during these years, especially in these last months. You have been my strength in all the difficulties.

Contents

Acknowledgements	3
Introducción	9
Introduction	23
1 Experimental Analysis of the Efficiency in Multi-Attribute Procurement Auctions	29
1.1 Introduction	29
1.2 The Model	33
1.3 Experimental Design	37
1.3.1 Sessions & Matching	38
1.3.2 Debriefing	38
1.4 Results	39
1.4.1 Welfare analysis: a two-stage approach	41
1.5 Conclusion	44
Bibliography	45
Appendices	47
1.A Proof of Proposition 1	47
1.B Supplementary Statistical Evidence	50
1.C The Econometric Model	52
1.C.1 Stage 1	52
1.C.2 Stage 2	53
1.C.3 Empirical results	53
1.D Experimental Instructions	57

2 Promoting greater deliberation in the one-shot Stag-Hunt game played online	63
2.1 Introduction	63
2.2 Experimental Design	66
2.3 Results	68
2.4 Text Analysis	74
2.5 Discussion	76
Bibliography	80
Appendices	85
2.A Experimental Instructions	85
2.B Experimental Data & Analysis	91
2.B.1 Regression analysis	91
2.B.2 Tobit Regression on Beliefs	91
2.B.3 Figures and Tables	95
3 Game-based educational program promotes sustainable water use	99
3.1 Introduction	99
3.2 Material and Methods	101
3.2.1 The game-based educational program	101
3.2.2 Method and data	103
3.3 Results	105
3.3.1 Aggregated reported behavior	107
3.3.2 Disaggregated reported behaviors	111
3.4 Discussion	114
Bibliography	116
Appendices	119
3.A Material	119
3.A.1 Informed consent	119
3.A.2 The board game <i>Blutube</i>	121
3.A.3 The Survey	122
3.A.4 Participants and ranking	124
3.B Translated Survey	125
3.C Supplementary Analysis	128
3.C.1 Robustness check	128

3.C.2 Principal Component Analysis 130

3.C.3 Ordered Logit Regression Analysis 133



Universitat d'Alacant
Universidad de Alicante

Resumen en castellano

Introducción

El enfoque experimental es el corazón de algunos de los desarrollos más interesantes de la economía.

Una gran cantidad de experimentos ha establecido diferencias en la toma de decisiones individual (Thaler and Ganser, 2015) y las evidencias de experimentos sobre coordinación, subastas, toma de decisiones, el bien público siguen creciendo (Plott and Smith, 2008; Kagel and Roth, 2016). Además, durante los últimos 20 años se han publicado libros dedicados a métodos experimentales (Davis and Holt, 1993; Plott and Smith, 2008), algunos de los cuales dan seguimiento a los problemas metodológicos en diferentes formas de experimentos (Guala et al., 2005; Sugden, 2005; Caplin and Schotter, 2008).

Básicamente, los experimentos se utilizan para generar datos controlados. El término “datos controlados” se refiere al hecho de que la mayoría de los factores que influyen en los comportamientos se mantienen constantes, y solo un factor de interés (el “tratamiento”) cambia a la vez. Este es el punto crítico para hacer una inferencia causal. A veces, este proceso de generación ocurre de forma natural (es decir, un “experimento natural”). Sin embargo, la mayoría de las veces, el investigador es el encargado de desarrollar y controlar el proceso de generación. Los experimentos económicos están diseñados para responder preguntas económicas. Las características comunes de los experimentos económicos son: controlar lo que el tomador de decisiones puede hacer, decidir la información que conoce y el incentivo monetario que puede ganar (Smith, 1982). Como resultado, un experimento puede lograr las características más importantes de las teorías económicas. Pero, al igual que los modelos teóricos, los experimentos son simplemente observaciones del mundo real. El entorno experimental es a menudo (no siempre) inexacto en el contexto, las instrucciones y la configuración. Sin embargo, gracias a esta metodología, es posible aportar evidencias sobre las preferencias de los individuos, testear teorías o comprender fenómenos económicos.

Esta tesis se compone de ensayos en diferentes temas utilizando tres metodologías experimentales diferentes: un entorno en el laboratorio, un entorno online y un cuasi-experimento en el campo. Cada

capítulo tiene el mismo propósito: comprender los comportamientos de las personas en una situación específica. En el primer capítulo, un trabajo conjunto con Gianluigi Albano, Angela Cipollone, Giovanni Ponti y Marco Sparro, presentamos los resultados de un experimento de laboratorio donde los sujetos compiten por contratos de adquisición que se asignan mediante una regla de puntuación. Dada la creciente importancia de los mecanismos competitivos de atributos múltiples en los mercados de contratación pública y privada, uno podría preguntarse qué tan bien los postores se enfrentan al entorno estratégico posiblemente más sofisticado de las subastas de puntuación. En los tratamientos diseñados, el comprador se preocupa por los aspectos financieros y no financieros presentados por los vendedores. Específicamente, el comprador solicita presentar una oferta bidimensional que incluye una oferta de calidad (que afecta los costos de producción) y una oferta financiera, que es un descuento sobre el precio de reserva anunciado. Las dimensiones de precio y calidad se convierten luego en una puntuación unidimensional y el contrato se adjudica al licitador con la puntuación más alta. En nuestro experimento de adquisición, la calidad se determina exógenamente. En cada período, cada vendedor está dotado de un nivel de calidad fijo, que es un sorteo independiente (sin reemplazo) de 11 valores diferentes (de 0 a 10). Esto es interesante porque hay muchos entornos diferentes donde la calidad ya está establecida antes del comienzo de la subasta de puntuación. En el laboratorio, los participantes imitan una subasta de adquisiciones en la que un comprador hipotético solicita una oferta a 5 vendedores. Se realizan cuatro sesiones experimentales (entre sujetos) en las que los participantes son asignados aleatoriamente a uno de nuestros tratamientos: i) el comprador prefiere más la calidad que el precio o, ii) la dimensión del precio es más importante que la calidad. Cada participante juega 11 rondas donde observar todos y cada uno de los niveles de calidad (aleatorizados y sin reemplazo) y presentar el descuento correspondiente. No se dan comentarios hasta el final del experimento.

En el segundo capítulo miro cómo la cognición puede afectar la disposición a colaborar con extraños. Antes, solo un estudio (Belloc et al., 2019) ha analizado el efecto de inducir menos reflexión sobre comportamientos colaborativos en un experimento de laboratorio. En contraste con ellos, intento incitar a las personas a reflexionar más sobre su toma de decisiones. Para hacerlo, ejecuto un experimento online con diseño mixto en el que se juega una serie de juegos de Stag-Hunt one-shot con diferentes estructuras de pago (dentro de los sujetos) mientras se manipula la cognición de los participantes (entre los sujetos). En el entorno online, los sujetos se asignan aleatoriamente a tres condiciones diferentes diseñadas para diferenciar el alcance de la reflexión de los participantes sobre sus elecciones: en el *control* no hay restricciones para la toma de decisiones, en el tratamiento

de *retraso de tiempo* los participantes se ven obligados a esperar 40 segundos antes de elegir una acción, y en el tratamiento de *retraso motivado*, los participantes también se ven obligados a esperar 40 segundos y, además, deben escribir una motivación para su elección antes de elegir una acción. Aunque el tratamiento de retraso motivado es eficaz para inducir una mayor reflexión (Bilancini et al., 2019b, 2020a, 2021), todavía no se ha establecido cómo se compara con el retraso de tiempo y no se ha establecido si pedir una motivación tiene efectos adicionales y cualitativamente diferentes. Por lo tanto, una contribución adicional de este capítulo es proporcionar nuevos conocimientos sobre estos dos enfoques metodológicos y comprender si y cómo promueven una mayor reflexión en un entorno experimental en línea. Además, este último enfoque metodológico, tiene oportunidad de estudiar los componentes emocionales que presentan las motivaciones escritas. La literatura reciente en economía examina la inducción de un estado de ánimo positivo o negativo en los juegos de economía one-shot (Kirchsteiger et al., 2006; Capra, 2004). Luego, siguiendo Proto et al. (2019), analizo los textos escritos de los participantes, aplicando lo que se llama “análisis de sentimientos”, y miro cómo las emociones de los individuos se correlacionan con sus elecciones.

El tercer capítulo es un trabajo conjunto con Ennio Bilancini y Leonardo Boncinelli donde estimamos el impacto de un programa educativo basado en juegos dirigido a promover el uso sostenible del agua entre los estudiantes de 2do a 4to año de escuela primaria y sus familias que viven en el municipio de Lucca, Italia. Proporcionar oportunidades para que los niños pequeños desarrollen comportamientos pro-sociales es un objetivo fundamental para padres y maestros (Copple et al., 2013). Educar a los niños sobre el uso pro-social del agua, hacerles comprender la importancia del tema y las implicaciones para las generaciones futuras es un tema importante. Se han implementado con éxito una variedad de métodos para aumentar la pro-socialidad en los niños. Estos métodos incluyen espacio de juego, juguetes de usos múltiples, libros para niños y juegos y juegos en grupo (Orlick, 1983). En particular, el alto grado de interacción que proviene de los juegos y juegos grupales para los niños y el tiempo que ocupan los juegos en sus vidas pueden desarrollar un comportamiento pro-social en sus actividades diarias. Para ello, el Ayuntamiento de Lucca, Italia, junto con Lucca Crea ¹ Y GEAL ², ha lanzado recientemente un innovador programa educativo en diferentes escuelas primarias generales. El proyecto *Blutube* se basa en un compromiso lúdico para enseñar a los niños cómo funciona el ciclo del agua, con el objetivo de promover la conciencia sobre el desperdicio de agua y el uso eficiente del agua. En este artículo, dado que no podemos intervenir directamente en la organización del programa, basamos nuestro enfoque de la evaluación empírica del impacto del programa

¹La empresa que organiza Lucca Comics & Games, uno de los espectáculos transmedia más grandes del mundo.

²La sociedad que gestiona el sistema integrado de agua de la ciudad.

en la metodología de cuasi-experimento (Campbell and Stanley, 2015): utilizamos un diseño simple de dos grupos (tratamiento y control) y tres medidas distintas de las variables objetivo desde enero hasta noviembre de 2019. Específicamente, la conciencia de los estudiantes sobre el desperdicio de agua y el uso eficiente del agua se identificó con un cuestionario recopilado antes del programa, justo después y seis meses después del programa. Les hacemos siete preguntas sobre sus comportamientos de consumo de agua en circunstancias familiares, hasta qué punto hablan sobre el agua con sus padres y hasta qué punto comen alimentos que contienen agua (frutas y verduras). Codificamos estas informaciones en una medida sintética de la conducta reportada agregada sumando las puntuaciones de las 7 preguntas (cada respuesta está en una escala Likert de 1 a 5, donde 1 corresponde a la conducta menos virtuosa y 5 a la más virtuosa).

Resumen y discusión de resultados

Capítulo 1

Durante las dos últimas décadas, la contratación pública ha experimentado cambios profundos. Tanto los responsables de la formulación de políticas como los académicos y los profesionales comparten la visión amplia de que la contratación pública ha pasado de ser un conjunto de actividades basadas en la aprobación administrativa a una herramienta estratégica para mejorar la eficiencia en las organizaciones públicas, regular los mercados y promover el desarrollo sostenible. Gracias a una profunda reformulación de las regulaciones de contratación pública a nivel mundial, promovida por los responsables de la formulación de políticas con visión de futuro, y al surgimiento de una fuerza laboral de contratación más calificada, así como de organizaciones de contratación especializadas, la contratación pública se está utilizando cada vez más para perseguir objetivos más allá de los simples adquisición de obras / productos / servicios. De manera coherente con estos objetivos, se insta a las organizaciones públicas a realizar procesos competitivos mediante la evaluación de una amplia gama de características, que comprenden dimensiones tanto financieras como no financieras. Por ejemplo, la Directiva de contratación pública de la UE 2014/24 / UE prevé que "... los poderes adjudicadores basarán la adjudicación de los contratos públicos en la oferta económicamente más ventajosa".³ Esto implica que, en circunstancias normales, las organizaciones públicas deben considerar las dimensiones de precio y no precio en la adjudicación de contratos públicos, aunque la adjudicación de

³Directiva 2014/24 / UE, art.67 (1).

precio más bajo sigue siendo un criterio de adjudicación admisible.⁴

Las subastas de puntuación (o de atributos múltiples) se encuentran entre los mecanismos competitivos más extendidos para evaluar ofertas heterogéneas. En una subasta de puntuación, el comprador se compromete con un mecanismo de puntuación, que asigna los atributos financieros y no financieros de cada oferta en una puntuación unidimensional.⁵ En una subasta con la puntuación más alta, la oferta que obtenga la puntuación más alta se considera la ganadora y recibe un pago financiero equivalente a la oferta presentada.⁶ A pesar de la relevancia práctica en los mercados de adquisiciones reales, las subastas de puntuación solo han atraído una investigación teórica limitada. Che (1993) proporciona la primera caracterización completa de las estrategias óptimas de los licitadores con una elección de calidad endógena. En su modelo, los licitadores observan en privado su nivel de eficiencia (es decir, sus costos para producir calidad) y luego, simultáneamente, presentan un par calidad-precio. En este marco, puede demostrar que los licitadores se enfrentan a una decisión precio / calidad que puede reducirse a un problema unidimensional. La razón es que los licitadores racionales siempre presentarán el nivel de calidad socialmente eficiente, independientemente de su comportamiento de licitación. En este problema unidimensional reducido, los licitadores pueden clasificarse de acuerdo con su “ potencial productivo ”, definido como pseudo-tipo, es decir, el nivel más alto de bienestar social que pueden producir. También resulta que si los pseudo-tipos del Che son monótonos en los niveles de eficiencia, entonces las subastas de puntuación pueden asimilarse a las subastas de primer precio y, por lo tanto, los resultados bien conocidos en las subastas de solo precio se pueden aplicar para derivar el comportamiento óptimo de los licitadores.⁷

Dada la importancia cada vez mayor en los mercados de contratación pública y privada del mecanismo competitivo de atributos múltiples, uno puede preguntarse hasta qué punto los licitadores son capaces de hacer frente al entorno estratégico posiblemente más sofisticado de las subastas de puntuación. Esta pregunta se vuelve aún más convincente ya que existe una evidencia experimental sustancial de que, incluso en subastas simples de precio único, el comportamiento real puede diferir

⁴“Los Estados miembros pueden proporcionar que los poderes adjudicadores no pueden utilizar solo el precio o el coste como único criterio de adjudicación ni restringir su uso a determinadas categorías de poderes adjudicadores o determinados tipos de contratos. ” (Directiva 2014/24 / UE, artículo 67 (2))

⁵Un mecanismo similar es llamado subasta de adquisiciones determinada por el comprador, que puede ser considerada como una subasta multidimensional en la que la regla de puntuación es una información privada. En una subasta de adquisiciones determinada por el comprador, el comprador simplemente establece el precio de reserva y una lista de condiciones sobre la calidad del bien/servicio. Una vez que los vendedores han presentado sus ofertas, el comprador es libre de asignar el contrato cuando lo desee (Santamaría, 2015).

⁶Este es posiblemente el mecanismo más utilizado en la familia de las subastas de puntuación.

⁷Asker and Cantillon (2008) generaliza y amplía los resultados mostrados en Che (1993) al permitir un espacio de tipo multidimensional.

sistemáticamente de lo que predice la teoría (ver, por ejemplo, Kagel and Levin, 2002, 2008).⁸ Desafortunadamente, probar estas predicciones teóricas en el campo es difícil dada la heterogeneidad de los contratos de adquisiciones en una base de datos típica. Sin embargo, estos mecanismos se han estudiado tanto de forma experimental como en un entorno controlado. Engelbrecht-Wiggans et al. (2007) compara, teórica y experimentalmente, un mecanismo de múltiples atributos determinado por el comprador y uno basado en el precio. Bajo ambos mecanismos, los licitadores están dotados de ingenio h un nivel de calidad exógeno y presentar solo una oferta financiera. Los autores muestran que el mecanismo determinado por el comprador puede aumentar el bienestar del comprador cuando un gran número de proveedores compite por el contrato. Por otro lado, el comprador está mejor con el uso del mecanismo basado en el precio cuando el número de proveedores es bajo y la correlación entre el costo y la calidad es baja.

Shachat and Swarthout (2010) compare la subasta de oferta sellada determinada por el comprador con una subasta inglesa con créditos de licitación (EBC). En este último mecanismo, el comprador otorga a cada vendedor una cierta cantidad de créditos de licitación, que dependen de la calidad de los bienes ofrecidos. Demuestran formalmente que el mecanismo determinado por el comprador es menos eficiente que el EBC. En el escenario experimental, sin embargo, muestran que el comportamiento real se aparta de la predicción teórica. Más precisamente, encuentran que, en la subasta de oferta sellada determinada por el comprador, los compradores y proveedores se desempeñan mejor debido a la oferta no equilibrada y los créditos de oferta demasiado generosos. Strecker (2010) estudia el efecto de revelar información en una subasta inglesa inversa de atributos múltiples con un comprador y cinco vendedores. En su entorno, las ofertas comprenden un atributo financiero y dos no financieros. Sus hallazgos sugieren que la eficiencia es mayor cuando se revela la regla de puntuación que cuando solo se proporciona información limitada a los vendedores; sin embargo, el excedente del comprador no se ve afectado significativamente por la naturaleza de la política de revelación de información.

En este capítulo, presentamos los resultados de un experimento de subasta de adquisiciones estilizado en el que un comprador simulado tiene que seleccionar al contratista de un grupo de cinco proveedores potenciales mediante un mecanismo competitivo. El comprador se preocupa por los aspectos financieros y no financieros de las ofertas presentadas. Más específicamente, el comprador solicita ofertas bidimensionales que comprenden una oferta de calidad (que afecta los costos de producción) y una oferta financiera, un descuento con respecto a un precio de reserva (base) anunciado públicamente. Las dimensiones de precio y calidad se mapean luego en una puntuación unidimen-

⁸Véase también (Bichler, 2000; Chen-Ritzo et al., 2005; Chang et al., 2015, 2016).

sional y el contrato se adjudica al postor con la puntuación más alta. Como se explica en Seshadri et al. (1991), la mayoría de los modelos de ofertas asumen que los vendedores basan sus ofertas en algunas características conocidas de forma privada.

En nuestro experimento de períodos múltiples, la calidad se determina exógenamente, en el sentido de que cada participante, al comienzo de cada período, está dotado de un nivel de calidad fijo, un sorteo independiente (sin reemplazo) de una cuadrícula finita. Por lo tanto, nuestro entorno pertenece a la clase de modelos de subasta de valor privados independientes. Hay varias razones para diseñar un marco de selección adversa de este tipo. En primer lugar, existen muchos entornos de adquisiciones en los que se toman decisiones de calidad antes del diseño de la subasta puntuable o independientemente del mismo. Este suele ser el caso en la adquisición de equipos médicos, donde las decisiones de las empresas sobre las características de calidad de, por ejemplo, un ultrasonido o una máquina de resonancia magnética (MRI) se toman considerando el impacto en las ventas globales en lugar de los procesos competitivos llevados a cabo por un solo hospital en un país específico. Esta situación también se aplica a la adquisición de equipos de TI como fotocopiadoras o computadoras portátiles. En segundo lugar, una subasta de puntuación con niveles de calidad fijos da lugar a un entorno estratégico menos complejo para los participantes en el experimento. Dado que los participantes conocen la regla de puntuación antes de pujar, cada postor, dotado de un cierto nivel de calidad, se da cuenta de inmediato de su puntuación técnica. Por lo tanto, su problema estratégico se reduce a calcular el descuento óptimo para maximizar las ganancias esperadas, donde el evento de ganar coincide con el evento de que el mismo postor tiene la puntuación más alta. Por último, pero no menos importante, al proporcionar a cada postor una gama completa de posibles cualidades (sin reemplazo) podemos obtener una función de licitación completa para cada participante (ver Grimm et al., 2008).

En la sección de teoría, modelamos nuestro mecanismo competitivo como una subasta de puntuación lineal con niveles de calidad exógenos. El “tipo” de cada participante (el nivel de calidad asignado) está asociado con un pseudo-tipo, que representa la capacidad del postor para satisfacer la preferencia de precio/calidad del comprador, expresada por la regla de puntuación. Nuestras dos condiciones de tratamiento están especialmente diseñadas para que los pseudo-tipos puedan o no aumentar monótonamente con la calidad. Esto depende del peso relativo del atributo financiero en la regla de puntuación. En un tratamiento, el peso de la calidad es lo suficientemente alto como para que el entorno estratégico sea compatible con los supuestos del modelo de Che y la distribución de pseudo-tipos aumente de manera monótona en el nivel de calidad. Al revés, en el otro tratamiento

el peso de la rebaja es suficientemente alto para que la distribución de pseudo-tipos se convierta en una forma de U inversa, que, a su vez, implica que el vendedor con los pseudo-tipos más altos se encuentra en el interior del soporte de los niveles de calidad posibles. Por lo tanto, cuando la regla de puntuación pone un peso relativamente alto en el precio, no solo se proporciona a los licitadores un incentivo para ofertar de manera más agresiva, sino que también la distribución no monótona resultante de pseudo-tipos altera dramáticamente el problema estratégico que enfrentan los licitadores. La proposición 1 recoge las principales características de estas dos configuraciones de equilibrio, que dependen del peso relativo de la calidad vs. reembolso. Nuestro análisis teórico requiere un diseño experimental que se basa en dos condiciones (entre sujetos), dependiendo del peso relativo de la calidad vs. precio. Grupos fijos de cinco licitadores juegan repetidamente durante 11 rondas, donde cada licitador se asigna a todos y cada uno de los niveles de calidad dentro de la cuadrícula. Los participantes no reciben comentarios hasta el final del experimento, donde un sorteo selecciona la subasta relevante para el pago.

Los resultados experimentales muestran que nuestras dos condiciones producen una marcada diferencia en el comportamiento: cuando el peso relativo en el descuento es alto, los sujetos pujan de manera más agresiva y se acercan al equilibrio. Como era de esperar, cuando la puntuación final depende más de la rebaja, los sujetos compiten más que cuando depende de la calidad. También detectamos una diferencia en términos de eficiencia entre los dos tratamientos, donde la eficiencia se mide por la probabilidad con la que el sujeto con el pseudo-tipo más alto dentro del grupo coincidente gana la subasta. En concreto, encontramos que, en el caso en el que el peso sobre la calidad es mayor, las subastas se adjudican, en el 95% de los casos, al individuo con el pseudo-tipo más eficiente. Este porcentaje desciende al 43% cuando la bonificación tiene un peso mayor.

Esta notable diferencia en la eficiencia probablemente está debida a más factores, que incluyen, entre otros, las características de las subastas y el impacto de estas últimas en el comportamiento de licitación, así como los efectos en el comportamiento debido a características específicas de cada individuo. Esto sugiere un ejercicio econométrico más sofisticado cuyo objetivo es desenredar el efecto de eficiencia “directo” de un cambio de tratamiento (es decir, el que se debe solo a la diferencia en las características estratégicas de los dos mecanismos alternativos) del efecto “indirecto” (es decir, el que depende del nivel de las desviaciones del equilibrio que también pueden estar influenciadas por el tratamiento). Nuestra “mediation analysis” (Imai et al., 2011) arroja dos conclusiones principales. Primero, los efectos directos e indirectos son significativos y apuntan en direcciones opuestas, favoreciendo (dificultando, respectivamente) la eficiencia en el alto (bajo, respectivamente) peso sobre la

calidad del tratamiento. En segundo lugar, el efecto directo supera al indirecto, lo que justifica la diferencia global de eficiencia a favor del tratamiento de alta

Capítulo 2

El hecho de que el comportamiento en dilemas sociales puede verse afectado por el alcance de la reflexión ha sido objeto de un animado debate en los últimos años (ver Capraro, 2019, para una encuesta reciente). Si bien el debate en curso se ha centrado principalmente en el comportamiento cooperativo (Rand, 2016, 2017; Bouwmeester et al., 2017; Kvarven et al., 2020; Alós-Ferrer and Garagnani, 2020) y, en menor medida, en la disposición a donar (Achtziger et al., 2015; Rand et al., 2016; Merkel and Lohse, 2019; Bago et al., 2020; Fromell et al., 2020),⁹ solo un artículo (Belloc et al., 2019) ha explorado el efecto de la poca reflexión sobre el comportamiento colaborativo. Belloc et al. (2019) estudian el efecto de inducir menos reflexión sobre el comportamiento colaborativo en una serie de juegos de Stag-Hunt. En el juego Stag-Hunt, un individuo tiene que elegir entre una acción más eficiente pero arriesgada, es decir, colaborar para cazar un *Ciervo*, y una acción más segura con una recompensa máxima menor, es decir, ir solo a cazar un *Liebre* (Skyrms, 2004). A diferencia del juego Prisoner's Dilemma, donde cada jugador puede incurrir en un costo personal para generar un beneficio mayor para el otro, el Stag-Hunt es un juego de coordinación en el que un individuo enfrenta el compromiso entre una colaboración arriesgada que puede proporcionar la mayor recompensa y un comportamiento no colaborativo más seguro que proporciona una recompensa menor mas certa. La compensación entre eficiencia y seguridad se ha investigado de diferentes formas. Schmidt et al. (2003) encuentran que los cambios en el dominio del riesgo afectan significativamente los comportamientos de los sujetos, mientras que los cambios en el nivel de dominio del pago no lo hacen. Capraro et al. (2020) han explorado los motivos para colaborar en el juego Stag-Hunt y han descubierto que está impulsado principalmente por preferencias por la eficiencia, más que por preferencias morales. Sin embargo, a lo mejor de mi conocimiento, solo Belloc et al. (2019) han intentado ver cómo la manipulación de la cognición puede afectar la colaboración, y lo hicieron en un experimento de laboratorio en el que se tomaron decisiones intuitivas imponiendo una restricción de tiempo de 10 segundos (la condición que se llama "presión de tiempo", consulta Spiliopoulos and Ortmann, 2018) para elegir una acción en el juego Stag-Hunt. Sus datos experimentales muestran que la probabilidad de ser más colaborativos es mayor para los participantes en el tratamiento de presión de tiempo.

⁹Consulte Hallsson et al. (2018) para conocer la encuesta sobre los comportamientos de enfoque en el juego Ultimatum.

Voy a contribuir a esta última línea de investigación con un experimento online pre-registrado donde, al contrario de Belloc et al. (2019), me centro en la condición de deliberación aplicando dos manipulaciones distintas de la cognición destinadas a inducir una reflexión mayor. Esto es potencialmente importante porque existen algunas críticas sobre el uso de limitaciones de tiempo. Primero, la restricción de tiempo debe establecerse en la pantalla de decisión y esto implica que los sujetos tienen todo el tiempo para pensar en la decisión en la pantalla de instrucción. En segundo lugar, el cumplimiento de la condición de presión de tiempo es imposible de forzar, lo que genera una cuestión metodológica. Además, también es interesante ver si y en qué medida los hallazgos previos obtenidos en el laboratorio por (Belloc et al., 2019) continúan en el entorno en línea, también si no directamente comparables.

En el experimento, intento inducir a las personas a reflexionar más sobre su toma de decisiones. En particular, confío en dos métodos para inducir una mayor reflexión en los participantes. El primero es un tratamiento de *retraso de tiempo*, que es lo opuesto a un tratamiento de presión de tiempo: los participantes deben esperar por lo menos 40 segundos antes de poder elegir una acción. El segundo método es un tratamiento de *retraso motivado* (Bilancini et al., 2017): los participantes deben esperar por lo menos 40 segundos (como en el tratamiento de retraso de tiempo) y escribir una motivación para su decisión antes de poder elegir una acción. Evidencia reciente sugiere que el tratamiento de retraso motivado es eficaz para inducir una mayor reflexión (Bilancini et al., 2019b, 2020a, 2021), aunque aún no se ha establecido cómo se compara con el retraso de tiempo y si pedir una motivación tiene efectos adicionales y cualitativamente diferentes. Para que los dos tratamientos sean comparables, requiero que los participantes esperen la misma cantidad de segundos. Por lo tanto, una contribución adicional de este trabajo es proporcionar nuevos conocimientos sobre estos dos enfoques metodológicos y cómo podrían promover una mayor reflexión en un entorno experimental en línea. Para estar en línea con Belloc et al. (2019), he medido también la aversión al riesgo y la confianza. Estas dos medidas están directamente relacionadas con el juego Stag-Hunt: por un lado, elegir *Stag* es una opción más arriesgada y confiable, mientras que, por otro lado, elegir *Hare* es más seguro.

Finalmente, de acuerdo con la literatura existente (psicología, neurociencia, economía y administración), los estados de ánimo y las emociones pueden afectar sistemáticamente el comportamiento del individuo. Muchos trabajos reconocen los vínculos importantes entre la emoción y la toma de decisiones (Loewenstein and Lerner, 2003; Rick and Loewenstein, 2008) y la emoción y la interacción social (Heilman et al., 2010). Además, la literatura experimental reciente también investiga la inducción de un estado de ánimo positivo o negativo en la interacción económica única (Kirchsteiger

et al., 2006; Capra, 2004). Ellos encuentran que un estado de ánimo más positivo induce comportamientos más altruistas y confiados. Para aprovechar al máximo el tratamiento de *retraso motivado*, pude realizar un análisis de texto destinado a comparar las características de los textos escritos por los participantes. El propósito es utilizar el análisis de sentimientos para estudiar los componentes emocionales presentes en las motivaciones escritas y ver cómo se correlacionan con las elecciones de los participantes. Los primeros en hacerlo con un juego económico fueron Proto et al. (2019). Ellos analizan una comunicación previa al juego en un Prisoners' Dilemma repetido. En lo mejor de mi conocimiento, este estudio es el primero en usar esta técnica en un juego de un solo golpe sin inducir un estado de ánimo particular antes de la decisión.

Los principales resultados se pueden resumir de la siguiente manera. Se encuentra que los participantes en el tratamiento de *retraso motivado* tienen menos probabilidades de colaborar (es decir, eligen *Stag*) que aquellos a los que solo se les pide que esperen 40 segundos, como en el tratamiento de *retraso en el tiempo*, o aquellos que no tienen limitaciones de tiempo para elegir una opción, como en el *control*. Además, se encuentra que fomentar la deliberación afecta el enfoque de los participantes en la estructura de pagos del juego cuando tienen que tomar una decisión. Finalmente, los hallazgos del análisis del texto sugieren que los participantes que deciden colaborar (es decir, aquellos que eligen la opción *Stag*) están en un estado de ánimo más positivo con respecto a aquellos que no colaboran (es decir, aquellos que eligen *Hare*).

Capítulo 3

El consumo sostenible de agua es relevante para la sostenibilidad general de las sociedades actuales y futuras (Wada and Bierkens, 2014; Kummu et al., 2016; Liu et al., 2017; Greve et al., 2018; Qin et al., 2019). El consumo sostenible de agua es, en muchos casos, un ejemplo de comportamiento pro-social en un dilema social (Hardin, 1968): una situación en la que existe un conflicto entre maximizar los beneficios individuales de uno y maximizar los beneficios de las generaciones presentes y futuras. Los individuos que son puramente egoístas tienen menos probabilidades de adoptar los comportamientos pro-sociales que conducen a un consumo sostenible de agua, a menos que las normas sociales ejerzan suficiente presión social para empujar a los individuos egoístas a hacer lo contrario. Dado que la adquisición de preferencias por comportamientos pro-sociales, así como la internalización de las normas sociales tienen lugar, en una parte sustancial, durante la infancia (House and Tomasello, 2018; House et al., 2020), se convierte en un objetivo fundamental crear oportunidades para que los niños pequeños desarrollen tales preferencias y internalizar normas de consumo sostenible de agua (Copple

et al., 2013; Cobo-Reyes et al., 2020). La educación de la primera infancia es el punto de partida natural para un aprendizaje permanente. Durante los últimos años, se han implementado con éxito una variedad de métodos educativos para promover la pro-socialidad en los niños. Estos métodos incluyen espacio de juego, juguetes de usos múltiples, libros dedicados, juegos en grupo y juegos organizados (Orlick, 1983). En particular, el tipo de interacciones sociales que provienen del juego en grupo y el juego organizado, así como el tiempo que los juegos pueden ocupar en la vida diaria de los niños, hacen de los programas educativos basados en juegos una herramienta candidata natural para promover comportamientos deseables. Algunos estudios, en los últimos años, han evaluado la relevancia de programas que fomentan las buenas prácticas en los referentes ambientales, como el uso del agua (Niles et al., 2013; Cuadrado et al., 2017). En un experimento de campo (Schultz et al., 2016) se estudió el papel de las normas sociales en la promoción de la conservación del agua, encontrando que las personas que recibieron información normativa sobre hogares similares en sus vecindarios consumían menos agua que el grupo de control; Además, las personas con normas personales ya sólidas se vieron menos afectadas por la información normativa que aquellas con normas personales bajas. Es importante destacar que los niños son capaces de reconocer si las normas prosociales se aplican a situaciones específicas (Blake et al., 2015), por lo que se vuelve importante que los niños comprendan qué es el consumo sostenible de agua y puedan relacionar su comportamiento con situaciones concretas y específicas como la recolección de agua o el lavado corporal.

En este capítulo aportamos evidencias sobre la efectividad de un programa educativo basado en juegos implementado durante los primeros once meses del año 2019 en el municipio de Lucca, Italia. El programa se llamó *BLUTUBE: Quién lleva el agua a casa* y tenía como objetivo promover el consumo sostenible de agua, así como la concienciación sobre el sistema municipal de agua y su uso. Los objetivos del programa eran alrededor de 1000 estudiantes de segundo a cuarto grado y sus familias. El programa se basó principalmente en la participación lúdica para los objetivos específicos de mejorar la conciencia de los estudiantes sobre el ciclo del agua en la naturaleza, el sistema de agua del municipio de Lucca y el uso diario del agua.

Nuestro enfoque de la evaluación empírica del impacto del programa se basa en la metodología de cuasi-experimento (Campbell and Stanley, 2015): no pudimos intervenir directamente en la organización del programa ¹⁰, Pero pudimos implementar un diseño simple de dos grupos (tratamiento y control) y recolectar tres medidas distintas de las variables de resultado objetivo durante un período de once meses. En particular, identificamos la conciencia de los estudiantes y sus comportamientos

¹⁰El programa basado en juegos ya estaba diseñado y organizado antes decidimos estudiar sus efectos. Entonces, las clases participantes ya estaban decididas.

sobre el consumo de agua con tres oleadas de encuestas administradas, respectivamente, inmediatamente antes de que comenzara el programa, dos días después de que finalizaran las actividades principales y después de seis meses más. Se ha demostrado que las respuestas a este tipo de cuestionarios son una fuente confiable de información sobre las perspectivas y percepciones de los niños (Danielson and Phelps, 2003; Di Riso et al., 2010; Bevans et al., 2020; Alan and Kabasakal, 2020). Nuestros hallazgos sugieren que el programa ha tenido efectos positivos, considerables y persistentes, especialmente con respecto a los hábitos y comportamientos que involucran el uso masivo o frecuente de agua (lavado de cuerpo completo, beber agua). Creemos que dicha evidencia impulsa fuertemente hacia una mayor consideración de los programas educativos basados en juegos como instrumentos de política para promover hábitos y comportamientos sostenibles, especialmente cuando los niños y sus familias pueden ser el objetivo. Este efecto se debe principalmente a un aumento en la frecuencia de comportamientos virtuosos auto informados con respecto al consumo de agua y las discusiones con los padres sobre el agua. Además, ese efecto positivo parece ser persistente: seis meses después del final de las actividades del programa, el efecto sigue siendo positivo y de magnitud apreciable.

Vale la pena enfatizar que el programa no solo brindó la oportunidad de jugar con juegos con temas de sustentabilidad. En cambio, las actividades lúdicas estructuradas se diseñaron para involucrar a los estudiantes en entornos específicos (en el hogar, en la escuela, durante el tiempo que pasan con la familia) y esto se incentivó adecuadamente en términos de las recompensas del juego que se materializaron durante un período de tiempo bastante largo (varios meses). El mensaje final resultante es que los programas basados en juegos destinados a promover comportamientos sostenibles deben diseñarse para involucrar a los participantes en su vida diaria, durante un período de tiempo considerable, y con actividades sociales que involucren a personas con las que tienen relaciones estables.

Conclusiones

Esta tesis se centra en métodos experimentales aplicados en diferentes entornos. Todas las áreas de la ciencia (incluida la economía) deben considerar todas las metodologías que se pueden aplicar. La teoría, los experimentos de laboratorio, los experimentos de campo, los experimentos en línea, la neuroeconomía, la investigación observacional y social, las encuestas y más, contribuyen a nuestra comprensión del mundo. Creo firmemente que los investigadores que utilizan estas metodologías (así como aquellos que realizan diferentes tipos de investigación económica) hacen una contribución significativa a la comprensión de los fenómenos económicos, la toma de decisiones de los individuos

en contextos económicos y el análisis de políticas.



Universitat d'Alacant
Universidad de Alicante

Introduction

The experimental approach is at the heart of some of the most interesting developments in economics.

A large number of experiments has established differences in individual decision-making (Thaler and Ganser, 2015) and the evidences from experiments on coordination, auctions, decision-making, public good is still growing (Plott and Smith, 2008; Kagel and Roth, 2016). Moreover, dedicated books on experimental methods have been published during the last 20 years (Davis and Holt, 1993; Plott and Smith, 2008), some of them tracking the methodological issues in different forms of experiments (Guala et al., 2005; Sugden, 2005; Caplin and Schotter, 2008).

Basically, experiments are used to generate controlled data. The term “controlled data” refers to the fact that the majority of the factors that influence behaviors are kept constant, with only one factor of interest (the “treatment”) changing at a time. This is the critical point for drawing causal inference. Sometimes, this generation process occurs naturally (i.e., “natural experiment”). However, most of the times, the researcher is in charge to develop and control the generation process. Economic experiments are designed to answer economic questions. Common features of economic experiments are: controlling what the decision maker can do, decide the information that he knows and the monetary incentive he can earn (Smith, 1982). As a result, an experiment can accomplish the most important features of an economic theories. But, like theoretical models, experiments are simply observations of the real world. The experimental environment is often (not always) inaccurate in the context, instructions and settings. However, thanks to this methodology, is possible to provide evidences on individuals’ preferences, test theories or understanding economic phenomena. All areas of science (including economics) need to consider all the methodologies that can be applied. Theory, lab experiments, field experiments, online experiments, neuroeconomics, observational and social research, surveys, and more, contribute to our understanding of the world.

This dissertation comprises of essays in different topics using three different experimental methodologies: an in-lab environment, an online setting and a quasi-experiment in the field. Each chapter follows the same purpose: understand individuals’ behaviors in a specific situation.

In the first chapter - a joint work with Gianluigi Albano, Angela Cipollone, Giovanni Ponti e Marco Sparro - we present the results of a laboratory experiment where subjects compete for procurement contracts which are assigned by means of a scoring rule. Given the growing importance of multi-attribute competitive mechanisms in private and public procurement markets, one might wonder how well bidders cope with the arguably more sophisticated strategic environment of scoring auctions. In the designed treatments, the buyer cares both about the financial and non-financial aspects of the submitted sellers. Specifically, the buyer asks to submit a two-dimensional bid that includes a quality offer (which affects production costs) and a financial offer, which is a rebate on the announce reserve price. The price and the quality dimensions are then converted into a one-dimensional score and the contract is awarded to the bidder with the highest score. In our procurement experiment, the quality is exogenous determined. In each period, each seller is endowed with a fixed quality level, which is independent draw (without replacement) from 11 different values (from 0 to 10). This is interesting because there are many different environments where the quality is already established before the beginning of the scoring auction.

In the lab, participants mimic a procurement auction in which a hypothetical buyer asks for a bid from 5 sellers. We run four (between-subjects) experimental sessions in which participants are randomly assigned to one of our treatments: i) the buyer prefers more the quality than the price or, ii) the price dimension matters more than quality. Each participant plays 11 rounds where observe each and every level of quality (randomized and without replacement) and submit a corresponding rebate. No feedback is given until the end of the experiment. Results show that more weight on rebate increases bids closer to the equilibrium. Moreover, more weight on quality yields on a more efficient allocation (for the buyer). This finding is mainly due to a “direct” effect of the treatment (the strategic properties of the different conditions), once it is controlled for an “indirect” effect caused by the out-of-equilibrium “trembles”, the matching groups characteristics and the individuals’ heterogeneity.

In the second chapter I explore how cognition may affect the disposition to collaborate with strangers. Before, only one study (Belloc et al., 2019) analyzed the effect of inducing less reflection on collaborative behaviors in a laboratory experiment. In contrast with them, I attempt to prompt individuals to reflect more on their decision-making. To do so, I run an online experiment with mixed design where a series of one-shot Stag-Hunt games with different payoff structures (within-subjects) are played while the participants’ cognition is manipulated (between-subject).

In the online setting, subjects are randomly assigned to three different conditions designed to

differentiate the extent of participants' reflection on their choices: in the *baseline* there is no constraint on decision-making, in the *time delay* treatment participants are forced to wait 40 seconds before picking an action, and in the *motivated delay* treatment participants are also forced to wait 40 seconds and, moreover, are required to write down a motivation for their choice before they pick an action. Recent evidence suggests that the motivated delay treatment is effective in inducing greater reflection (Bilancini et al., 2019, 2020, 2021) although it is still to be established how it compares to time delay and if asking for a motivation has additional and qualitatively different effects. So, an additional contribution of this chapter is to provide new insights on these two methodological approaches and understand if and how they promote greater reflection in an online experimental setting. Moreover, the latter methodological approach, give me the opportunity to study the emotional components present in the written motivations. The existing literature (in psychology, neuroscience and economics) reports a systematic impact of moods and emotions on individuals' behavior. Also the recent literature in economics examines the induction of positive or negative mood in one-shot economics games (Kirchsteiger et al., 2006; Capra, 2004). Then, following Proto et al. (2019), I analyzed the participants' written texts, applying the so-called "sentiment analysis", and see how individuals' emotions are correlated with their choices.

Experimental data show, consistently with previous results of (Belloc et al., 2019), that asking participants to wait 40 seconds and write a motivation for their decision before actually selecting an action (i.e., the *motivated delay* treatment) makes them less likely to collaborate than just asking to wait 40 seconds (i.e., the *time delay* treatment) or letting them choose without constraints (i.e., the *baseline*). While, no substantial difference is found between the *baseline* and the *time delay* treatment. Moreover, asking to wait 40 seconds before selecting an action has a sizeable effect on the relevance of the payoff structure for actual decisions: the effect of the expected gains from collaborating on the decision to collaborate (summarized by the basin of attraction of the action *Stag*) is about twice larger. This suggests that greater deliberation leads to give more attention to the payoff structure. Finally, the "sentiment analysis" shows that participants who chose *Stag* were more likely to motivate their choices writing a text classified as having a positive sentiment compared to those who chose *Hare*, suggesting that the choice to collaborate goes with a more positive mood. Moreover, I analyzed frequency of the most used words in the motivations. Consistently with the actual decision, words like "guarantee" and "risk" are written more frequently by those participants who deliberated in favor of the *Hare* choice, suggesting that they correctly recognized the greater safety of such action.

The third chapter is a joint work with Ennio Bilancini and Leonardo Boncinelli where we esti-

mate the impact of a game-based educational program aimed at promoting sustainable water usage among 2nd-4th grade students and their families living in the municipality of Lucca, Italy. Providing opportunities for young children to develop prosocial behaviors is a critical goal for parents and teachers (Copple et al., 2013). Educate children on prosocial water usage, make them understand the importance of the topic, and the implications to future generations is an important issue. A variety of methods for increasing prosociality in children have been successfully implemented. These methods include play space, multi-use toys, books for children, and group play and games (Orlick, 1983). In particular, the high degree of interaction that comes from group play and games for children and the time that games occupy in their lives can develop prosocial behavior in their daily activities. To this aim, the Municipality of Lucca, Italy, together with Lucca Crea¹¹ and GEAL¹², has recently launched an innovative educational program in several primary schools. The *Blutube* project relies on ludic engagement for teaching children how the water cycle works, with the aim of promoting awareness about water waste as well as efficient water usage.

In this paper, given that we cannot intervene directly on the organization of the program, we based our approach to the empirical assessment of the program's impact on the quasi-experiment methodology (Campbell and Stanley, 2015): we use a simple two-group design (treatment and control) and collect three distinct measurements of target variables from January to November 2019. Specifically, the students' awareness about water waste and efficient water usage was identified with a questionnaire collected before the program, just after and six months after the program. We ask them seven questions about their water consumption behaviors in familiar circumstances, the extent to which they talk about water with their parents, and the extent to which they eat food containing water (fruit and vegetables). We code these informations in a synthetic measure of aggregate reported behavior summing the scores of the 7 questions (each answer being in a Likert scale from 1 to 5, where 1 corresponds to the least virtuous behavior and 5 to the most virtuous one).

Our findings indicate that the game-based educational program had an impact on promoting sustainable behaviors regarding water consumption. On average, our synthetic measure of prosocial behavior in the treatment group was greater than in the control group by 2.11 points. The effect is primarily driven by two behaviors: self-reported water consumption and the extent to which students self-report talking about water with their parents. No substantial effect are found on the extent to which students self-report eating food containing water. Finally, we find that the positive effect of the program is still observed after six months from the end.

¹¹The company that organizes Lucca Comics & Games, one of the largest transmedia shows in the world.

¹²The joint-stock company that manages the city's integrated water system.

This thesis focuses on experimental methods applied to different environments. I strongly believe that researchers using these methodologies (as well as those who conduct different types of economics research) make significant contribution to the understanding of economic phenomena, individuals' decision-making in economic contexts, and policy analysis.



Universitat d'Alacant
Universidad de Alicante

Experimental Analysis of the Efficiency in Multi-Attribute Procurement Auctions

1.1 Introduction

During the last two decades, public procurement has undergone profound changes. Policy makers, academics and practitioners alike share the broad view that public procurement has evolved from a clerical signoff-ridden set of activities to a strategic tool to enhance efficiency in public organizations, to regulate markets and promote sustainable development. Thanks to a profound reformulation of public procurement regulations at a global level, promoted by forward-looking policymakers, and to the emergence of more qualified procurement workforce, as well as specialised procurement organizations, public procurement is being increasingly used to pursue objectives beyond the mere acquisition of works/products/services. Coherently with these objectives public organisations are urged to carry out competitive processes by evaluating a wide array of characteristics, comprising both financial and non-financial dimensions. For instance, the EU public procurement Directive 2014/24/EU foresees that "...contracting authorities shall base the award of public contracts on the most economically advantageous tender".¹ This implies that, under normal circumstances, public organisations shall consider both price and non-price dimensions in awarding public contracts, although the lowest-price award remains an admissible award criterion.²

Scoring (or multi-attribute) auctions are among the most widespread competitive mechanisms to evaluate heterogeneous tenders. In a scoring auction, the buyer commits to a scoring mechanism, which maps each tender's financial and non-financial attributes onto a one-dimensional score.³ In a

¹Directive 2014/24/EU, art.67(1).

²"Member States may provide that contracting authorities may not use price only or cost only as the sole award criterion or restrict their use to certain categories of contracting authorities or certain types of contracts." (Directive 2014/24/EU, art. 67(2))

³A similar mechanism is the so-called buyer-determined procurement auction, which can be considered as a multi-dimensional auction in which the scoring rule is private information. In a buyer-determined procurement auction the buyer simply sets the reserve price and a list of conditions on the quality of the good/services. Once sellers have submitted their bid, the buyer is free to assign the contract at her wish (Santamaría, 2015).

highest-score auction the tender awarded the highest score is deemed to be the winner and receives a financial payment equal to the submitted bid.⁴ In spite of the practical relevance in real procurement markets, scoring auctions have only attracted a limited theoretical investigation. Che (1993) provides the first comprehensive characterization of bidders' optimal strategies with endogenous quality choice. In his model, bidders privately observe their efficiency level (i.e., their costs for producing quality) and then, simultaneously, submit a quality-price pair. Within this framework, he is able to prove that the bidders face a price/quality decision which can be reduced to a single-dimensional problem. The reason is that rational bidders will always submit the socially efficient quality level, independently on their bidding behavior. In this reduced one-dimensional problem, bidders can be ranked according to their "productive potential" - defined as pseudo-type - that is, the highest level of social welfare they can produce. It also turns that if Che's pseudo-types are monotonic in the efficiency levels then scoring auctions can be assimilated to first-price auctions and, therefore, well-known results in price-only auctions can be applied to derive bidders' optimal behavior.⁵

Given the increasing relevance in private and public procurement markets of multi-attribute competitive mechanism, one may wonder to what extent bidders are able to cope with the arguably more sophisticated strategic environment of scoring auctions. This question becomes even more compelling as there exists a substantial experimental evidence that - even in simple price-only auctions - actual behavior may systematically differ from what theory predicts (see, for instance, Kagel and Levin, 2002, 2008).⁶ Unfortunately, testing these theoretical predictions in the field is difficult given the heterogeneity of procurement contracts in a typical database. However, such mechanisms have been studied both experimentally and in a controlled environment. Engelbrecht-Wiggans et al. (2007) compare, theoretically and experimentally, a buyer-determined and a price-based multi-attribute mechanism. Under both mechanisms, bidders are endowed with an exogenous quality level and submit only a financial bid. The authors show that the buyer-determined mechanism is able to increase the welfare of the buyer when a large number of suppliers compete for the contract. On the other hand, the buyer is better off with the use of the price-based mechanism when the number of suppliers is low and there is a low correlation between cost and quality.

Shachat and Swarthout (2010) compare the sealed-bid buyer-determined auction with an English auction with bidding credits (EBC). In the latter mechanism, the buyer endows each seller with a certain amount of bidding credits, which depend upon the quality of the goods offered. They formally

⁴This is arguably the most widely used mechanism in the family of scoring auctions.

⁵Asker and Cantillon (2008) further generalize and extend the results shown in Che (1993) by allowing for multidimensional type-space.

⁶See also (Bichler, 2000; Chen-Ritzo et al., 2005; Chang et al., 2015, 2016).

prove that the buyer-determined mechanism is less efficient than the EBC. In the experimental setting, though, they show that the actual behavior departs from the theoretical prediction. More precisely, they find that in the sealed-bid buyer-determined auction buyers and suppliers perform better due to non-equilibrium bidding and over-generous bidding credits. Strecker (2010) studies the effect of revealing information in a multi-attribute reverse English auction with one buyer and five sellers. In his setting, bids comprise one financial and two non-financial attribute. His findings suggest that efficiency is greater when the scoring rule is revealed than when only limited information is provided to sellers; however, the buyer's surplus is not significantly affected by the nature of information-revelation policy.

In this paper, we present the results of a stylized procurement auction experiment where a simulated buyer has to select the contractor out of a pool of five potential suppliers by means of a competitive mechanism. The buyer cares both about financial and non-financial aspects of the submitted tenders. More specifically, the buyer solicits two-dimensional bids comprising a quality offer (that affects production costs) and a financial offer, a rebate with respect to a publicly announced reserve (base) price. Price and quality dimensions are then mapped into a one-dimensional score and the contract is awarded to the highest-score bidder. As explained in Seshadri et al. (1991), most bidding models assume that sellers base their bids on some privately known characteristics.

In our multi-period experiment quality is exogenously determined, in that each participant, at the beginning of each period, is endowed with a fixed quality level, an independent draw (without replacement) from a finite grid. Thus our setting belongs to the class of independent private value auction models. There are several reasons for designing such an adverse-selection framework. First, there are many procurement environments where quality choices are made before - or independently of - the design of the scoring auction. This is usually the case in the procurement of medical equipment, where firms' decisions about the quality characteristics of, say, an ultrasound or Magnetic Resonance Imaging (MRI) machine are made by considering the impact on global sales rather than the competitive processes carried out by a single hospital in a specific country. This situation also applies to the procurement of IT equipment such as photocopiers or laptops. Second, a scoring auction with fixed quality levels gives rise to a less complex strategic environment for the participants in the experiment. Given that the scoring rule is known to participants before bidding, each bidder, endowed with a certain quality level, becomes immediately aware of his technical score. Hence his strategic problem boils down to computing the optimal rebate to maximize expected profits, where the event of winning coincides with the event that the same bidder has the highest score. Last, but

not least, by providing each bidder with a full range of possible qualities (without replacement) we are able to elicit a full bidding function for each participant (see Grimm et al., 2008).

The remainder of the paper is arranged as follows. Theory is presented in Section 1.2, where we model our competitive mechanism as a linear scoring auction with exogenous quality levels. Each participant's "type" (the assigned quality level) is associated with a pseudo-type, which accounts for the bidder's capacity to meet the buyer's price/quality preference, expressed by the scoring rule. Our two treatment conditions are especially designed so pseudo-types may or may not monotonically increase with quality. This depends on the relative weight of the financial attribute in the scoring rule. In one treatment the weight of the quality is sufficiently high so that the strategic environment is compatible with Che's modeling assumption and the distribution of pseudo-types is monotonically increasing in the quality level. By contrast, in the other treatment the weight of the rebate is sufficiently high so that the distribution of pseudo-types becomes a reverse U-shaped, which, in turn, implies that the seller with the highest pseudo-types lays in the interior of the support of the possible quality levels. Thus, when the scoring rule puts a relatively high weight on price, not only are bidders provided with an incentive to bid more aggressively, but also the resulting non-monotonic distribution of pseudo-types dramatically alters the strategic problem bidders face. Proposition 1.1 collects the main characteristics of these two equilibrium configurations, which depend on the relative weight of quality vs. rebate. Our theoretical analysis calls for an experimental design - described in detail in Section 1.3 - which is built upon two (between-subjects) conditions, depending on the relative weight of quality vs. price. Fixed groups of five bidders play repeatedly for 11 rounds, where each bidder is assigned to each and every quality level within the grid. Participants receive no feed-back until the end of the experiment, where a random draw selects the auction relevant for payment.

Section 1.4 reports our experimental results. We first notice that our two conditions yield a stark difference in behavior: when the relative weight on the rebate is high subjects bid more aggressively and closer to equilibrium. As expected, when the final score depends more on the rebate, subjects compete more on than when it depends on quality. We also detect a stark difference in terms of efficiency between the two treatments, where efficiency is measured by the likelihood with which the subject with the highest pseudo-type within the matching group wins the auction. Specifically, we find that, in the case in which the weight on quality is higher, the auctions are awarded, in the 95% of the cases, to the individual with the most efficient pseudo-type. This percentage drops to 43% when the rebate has a higher weight.

This striking difference in efficiency is probably due to multiple factors, which include - among others - auctions features and the impact of the latter on bidding behavior, as well as behavioral effects due to individual-specific characteristics. This suggests a more sophisticated econometric exercise whose aim is to disentangle the “direct” efficiency effect of a treatment change (i.e., the one which is only due to the difference in the strategic characteristics of the two alternative mechanisms) from the “indirect” effect (i.e., the one that depends upon the level of the deviations from equilibrium that may be also influenced by the treatment). Our “mediation analysis” (Imai et al., 2011) yields two main conclusions. First, the direct and indirect effects are both significant and point in opposite directions, favouring (hampering, respectively) efficiency in the high (low, respectively) weight on quality treatment. Second, the direct effect outweighs the indirect one, which justifies the overall difference in efficiency in favour of the high-quality treatment.

Finally, Section 1.5 concludes, followed by appendices containing the proof of Proposition 1.1 (Appendix 1.A), supplementary and statistical evidence (Appendix 1.B), a more detailed account of our econometric strategy (Appendix 1.C) and the experimental instructions (Appendix 1.D).

1.2 The Model

We consider a highest-score (procurement) auction whereby a buyer asks for bids from N firms.⁷ Each risk-neutral bidder i submits a quality-rebate pair, (q, r) , where $q_i \in [0, 1]$ is the (exogenous) privately observed quality level and r_i is the rebate offered with respect to the reserve price announced by the buyer (which is normalized to one). The bidder is then ranked according to the following linear scoring rule:

$$S(q_i, r_i) = (1 - \gamma)q_i + \gamma r_i \quad (1.1)$$

where $\gamma \in \{1/3; 2/3\}$ in our experimental implementation. Player i gets a payoff of

$$\pi(q_i, r_i) \begin{cases} \frac{1-r_i-c(q_i)}{n^*} & \text{if } S_i = \max_j(S_j(.)), \\ 0 & \text{otherwise} \end{cases} \quad (1.2)$$

where, $n^* \geq 1$ identifies the number of winners (in case of ties). By analogy with our experimental conditions, this section parametrizes the cost function as we set $c(q_i) = \frac{1}{4} + \frac{3}{4}q_i^2$.

⁷In our experimental setting N is set to 5. Some papers studied how the number of bidders influences the auction, but this is out of the scope of our paper.

A strategy for bidder i is a function $r : [0, 1] \rightarrow [0, 1]$ that maps each bidder's privately observed quality into a rebate. A symmetric Bayes-Nash equilibrium (BNE) is a vector of identical strategies, $(r(q))$, such that each bidder maximizes his expected payoff under the constraint that $0 \leq r(q_i) \leq 1 - c(q_i)$. In other words, by design, bidders can neither bid above the reserve price nor get negative profit. In a standard lowest-price auction - where bidders privately receive iid signals about their production costs and only submit a price for the procurement contract - a symmetric equilibrium can be characterized by assuming that the bidding function is strictly increasing in production costs (that is, in bidders' types). Consequently, in equilibrium, winning probabilities coincide with the probability that any bidder has drawn the lowest cost. This is not the case of our scoring auction where, to derive a BNE, we follow the approach pioneered by Che (1993), whereby the buyer derives utility from a contract that represents his true preferences and bidders are characterized by pseudo-types, which allows to rank bidders according to their winning probability. To this aim, we first introduce type- q bidder's *potential score*, $s_\gamma(q) \equiv \gamma(1 - c(q)) + (1 - \gamma)q$, which corresponds to the score when submitting a rebate $r = r_{max}(q) = 1 - c(q)$ and, by doing so, reducing to 0 the profits in case of winning. Since the scoring rule - basically - reflects the buyer's preferences with respect to the trade-off between quality and price, we can consider the bidder with the highest potential score to be the most efficient in serving the contract. For the time being, let us just assume, by analogy with Che (1993), that the higher the pseudo-type, $s_\gamma(q)$, the higher the probability for a player with type q to win the auction when the financial weight parameter in the scoring rule is γ .

As shown in Figure 1.1, depending on the value of γ , $s_\gamma(q)$ may or may not be monotonically increasing in q . More precisely, $s_\gamma(q)$ is strictly increasing in q if and only if $\gamma \leq \frac{2}{3}$, that is, when the weight associated to the financial score is sufficiently low, which is true in our experiment only when $\gamma = \frac{1}{3}$. In this case, the weight of quality evaluation in the scoring function is sufficiently high so as to make the bidder with the highest q to be the most likely winner. When $\gamma > \frac{2}{3}$, $s_\gamma(q)$ has an interior maximum, $q^* = \frac{(2(1-\gamma))}{3\gamma}$. In particular, $q^* = \frac{1}{3}$ when $\gamma = \frac{2}{3}$ (our alternative treatment).

Now, considering that a bidder observes his *potential score*, $s_\gamma(q)$, and decides to announce a score $\sigma(s_\gamma(q)) \leq s_\gamma(q)$, then:

Proposition 1.1. *If $r_\gamma^*(q)$ denotes the symmetric BNE of our scoring auction with weight equal to γ , then*

$$r_\gamma^*(q) = \max \left\{ \frac{1}{\gamma} [\sigma_\gamma^*(s_\gamma(q)) - (1 - \gamma)q], 0 \right\} \quad \text{if } \gamma = \frac{1}{3}$$

$$r_\gamma^*(q) = \frac{1}{\gamma} [\sigma_\gamma^*(s_\gamma(q)) - (1 - \gamma)q] \quad \text{if } \gamma = \frac{2}{3}$$

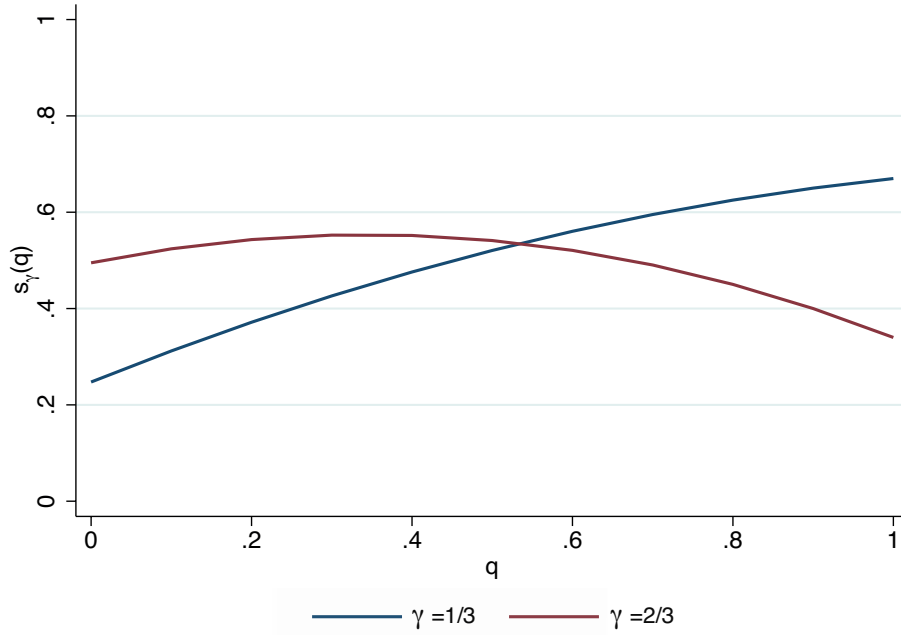


Figure 1.1: Potential score function $s_\gamma(q)$ by γ

where $\sigma^*(s)$ is the BNE of a standard first-price auction. Thus, $\sigma_\gamma^*(s_\gamma(q)) = \frac{1}{H_\gamma(s_\gamma(q))} \int_{\underline{s}_\gamma}^{s_\gamma(q)} y h_\gamma(y) dy$ with $H_\gamma(s) = G_\gamma^4(s)$, $h_\gamma = H_\gamma'(s)$ and $G_\gamma(s)$ is the c.d.f. of the random variable s and $\underline{s}_\gamma = \min_{q \in [0,1]} [s_\gamma(q)]$ is the lower bound of the potential score distribution.

While relegating the proof of Proposition 1.1 to Appendix 1.A, it may be instructive, at this point, to sketch the intuition behind our result. Following Che (1993), this is obtained by showing that our scoring auction is strategically equivalent to a first-price selling auction in which bidder i observes a signal s (his potential score) and submits a score, $\sigma_\gamma^*(s)$. At equilibrium, the submitted score $\sigma_\gamma^*(s) \leq s$ as rational bidders get positive profit by reducing the value of the rebate below its maximum level, that is, $r \leq r_{max}(q)$. The score bidding functions $\sigma_\gamma^*(s)$ associated with our treatments are reported in Figure 1.2. Notice that, coherently with the results in a “standard” first-price auction, the score bidding function $\sigma_\gamma^*(s)$ lays below the 45-degree (dashed) line, as each bidder optimally shades his bid below his value (that is, his potential score).

The explicit forms of either the score bidding function, $\sigma_\gamma^*(q)$, or its strategic equivalent rebate function, $r_\gamma^*(q)$, are complex and uninformative, but we plot them in Figure 1.3 for both values of γ (1/3 and 2/3) used in the experiment.

Given that the equilibrium bidding function $r_\gamma^*(q)$ is derived from the equilibrium of an “equivalent” first-price auction, $\sigma_\gamma^*(s_\gamma)$, it is immediate to realize that, in equilibrium, (i) bidders with the

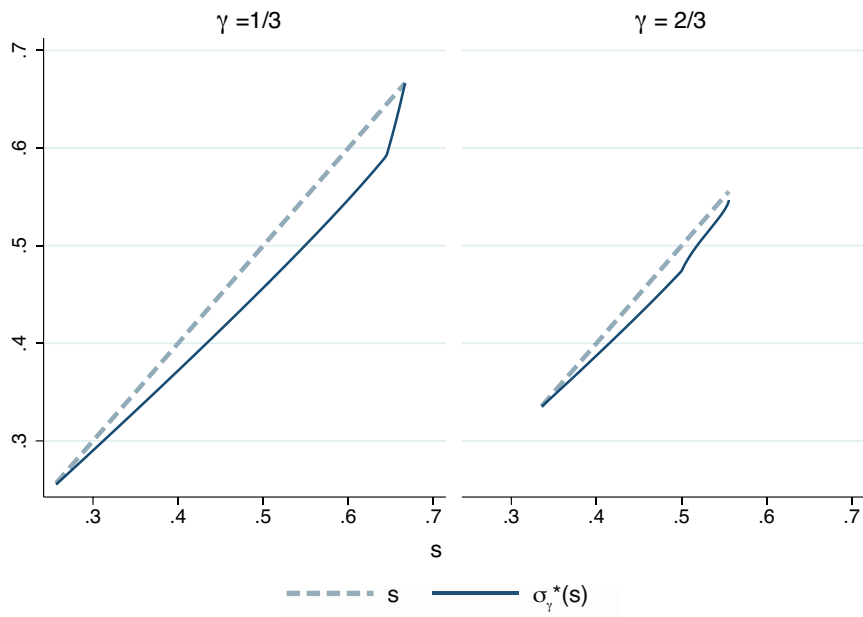


Figure 1.2: Private signals s and optimal bids $\sigma_\gamma^*(s)$ in the “modified” auctions.

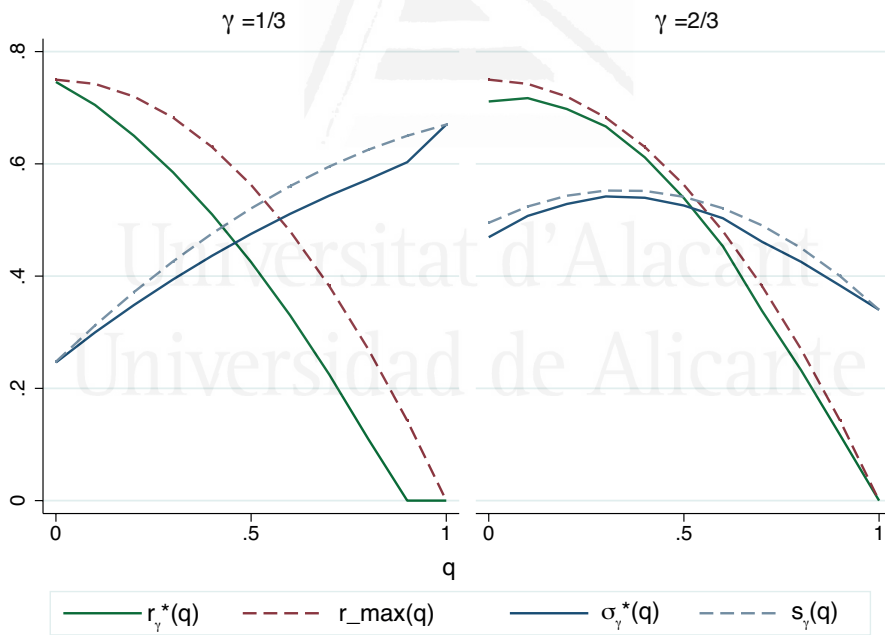


Figure 1.3: Equilibrium Analysis.

Notes: The bidding functions are plotted both in terms of submitted rebate r and obtained score σ and compared with the maximum potential rebate/score (dotted lines).

same potential score (s_γ) are expected to submit the same score $\sigma_\gamma^*(s_\gamma)$ and (ii) the winner is the bidder with the highest $s_\gamma(q)$.

Consider the graphs depicted in Figure 1.3. First, notice that the closer the equilibrium bids (solid line) to the zero-profit bids (dashed line) the lower the expected profit in case of winning. Consistently with intuition, when the weight of the rebate in the scoring rule is high ($\gamma = 2/3$), the submitted rebates are higher than in the case of $\gamma = 1/3$ for almost any q (precisely, for any $q > 0.05$). Second, when $\gamma = 1/3$, the most likely winner is the type with the highest q , because the scoring rule greatly rewards quality. It takes quite high a difference between two bidders' submitted rebates to more than compensate the score gap induced by different quality levels. Hence, in equilibrium, the types with high quality can “safely” increase their expected profit (by lowering the rebate) without considerably reduce their winning chances. In other words, the gap in the potential score among bidders with different quality levels makes it harder (relatively to the case of $\gamma = 2/3$) for less efficient bidders to overbid more efficient competitors. This also helps us to understand why $r_{1/3}^*(q)$ becomes flat above a certain threshold (approx. 0.88 with our parametrization): bidders with sufficiently high quality anticipate to be awarded with a high score for quality and would then optimally submit a discount below 0 (i.e., a price higher than the reserve price), which is not allowed by the rules of the game. The opposite is true when $\gamma = 2/3$. As shown in Figure 1.1, $s_{2/3}(q)$ is not monotonic, which shortens the length of the support of the random variable $s_{2/3}(q)$. This makes bidders closer in terms of efficiency, thus increasing their incentive to compete more aggressively and submit higher rebates. In fact, the higher weight of the rebate in the scoring rule allows bidders with lower quality to compensate their gap in quality by increasing their financial score, which is made possible by lower production costs.

1.3 Experimental Design

Our experimental sessions were conducted at the Centro di Economia Sperimentale A Roma Est (CESARE), at LUISS Guido Carli Roma. A total of 90 students were recruited among the undergraduate population of LUISS Guido Carli using the ORSEE recruiting system (Greiner et al., 2004), with no particular bias in favour of students from the Departments of Economics and Finance or Business Administration and Management. All sessions were “gender balanced”, with approximately the same number of male/female subjects.⁸ Experimental sessions were computerized. Instructions were distributed at the beginning of the experiment and were read aloud.⁹ At the end of the reading, we let subjects ask about any doubt they may have had. Moreover, given that prior experience is

⁸Descriptive statistics and their difference in means are reported in Appendix 1.B in Tables 1.B.1 and 1.B.2, respectively.

⁹The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007). Full experimental instructions can be found in Appendix 1.D.

an important dimension for procurement auctions, we let them play five warming-up rounds before each session to increase their understanding of the game.¹⁰ At the end of each session, subjects were asked to compile an extensive debriefing questionnaire (see Section 1.3.2 below), before receiving –in cash and privately– their monetary winnings.

1.3.1 Sessions & Matching

We run 4 experimental sessions in which subjects are randomly assigned to one of our treatments, $\gamma \in \{\frac{1}{3}, \frac{2}{3}\}$. In session 1 and 3, 45 subjects faced the treatment with $\gamma = \frac{1}{3}$, while in session 2 and 4, others 45 participants faced the treatment with $\gamma = \frac{2}{3}$. Each between-subjects treatment contains 5 matching groups (cohorts) of 5 players for a total of 25 participants in each session¹¹, with subjects from different cohorts never interacting with each other throughout the experiment. Matching groups remain constant throughout the experiment, with no feedback until the very end, where the period relevant for payment is publicly drawn using a random lottery incentive protocol and monetary payoffs are determined. For each treatment, subjects play 11 rounds of a procurement auction where participants act as bidders experienced, without replacement, each and every possible value of $q \in \{\frac{k}{10}\}, k = 0, 1, \dots, 10$, which is randomized across period and cohorts. Given the level of quality assigned, they have to decide the level of rebate, $r(q)$, they want to bid to obtain the final score. This permits to elicit the entire bidding function, $r(q)$, of each participant and, moreover, the lack of feedback after each round, give us the opportunity to treat each bidder decision as an independent observation (Grimm et al., 2008).

1.3.2 Debriefing

At the end of each session, subjects are asked to answer a detailed questionnaire from which we elicit proxies of their observable heterogeneity. As it turns out, one of the key variables used in Section 1.4.1 for our regression analysis is derived from the well known Cognitive Reflection Test (Frederick, 2005). The CRT is a simple test of a quantitative nature especially designed to elicit the “predominant cognitive system at work” in respondents’ reasoning:

1. A bat and a ball cost 1.10 dollars. The bat costs 1.00 dollars more than the ball. How much does the ball cost? (Correct answer: 5 cents).

¹⁰As reported in Engelbrecht-Wiggans et al. (2007), experienced bidders bid closer to theoretical predictions than inexperienced ones.

¹¹Due to the absence of some participants we lost a cohort in session 2 and 3.

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? (Correct answer: 5 minutes).
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (Correct answer: 47 days).

The CRT provides not only a measure of cognitive ability, but also of impulsiveness and, possibly, other individuals' unobservable characteristics. In this test, the "impulsive" answer (10, 100 and 24, respectively) is shown to be the modal answer (Frederick, 2005). These answers, although incorrect, may have been selected by those subjects who do not think carefully enough. Following Cueva et al. (2016), we partition individuals into three groups. *Impulsive* subjects answer the erroneous intuitive value at least in two questions, *reflective* ones answer correctly at least in two questions, and *others* are the residual group. CRT group identifiers have been used as instruments in the two-step regression analysis of Section 1.4.1.¹²

1.4 Results

The average monetary winnings were about €18 per person (including the show up fee of €10), for a 90' experiment, including debriefing and payment.

Figure 1.4 tracks average and equilibrium bidding functions by treatment, together with the treatment pseudo-types. As expected, when the scoring rule puts more weight on quality (that is, when $\gamma = 1/3$), players submit, on average, lower rebates (Mann-Whitney test: $Z = -4.144$, $p < 0.001$). This simple evidence lets us conclude that submitted bids correctly follow the incentives induced by the two treatments and, for all quality levels, players bid less aggressively when the scoring rule favours quality with respect to price. We also notice that the dispersion of bids around the average is significantly higher at low quality levels since, for higher quality levels, bids are constrained by the rule that prevents losses.¹³ In both cases, players underbid respect the optimal level; only when $\gamma = 1/3$ and for higher levels of quality, we observe an overbidding behavior.

More importantly, when $\gamma = 2/3$, players bid closer to equilibrium. Table 1.1 shows the (equilibrium) expected bid levels as well the average distance (in absolute values) from the equilibrium by treatment. Players in the higher γ treatment bid, on average, closer to the equilibrium respect to

¹²See Appendix 1.C for details.

¹³As a result, when $q = 1$, $c(q) = 1$, i.e., players are forced to bid a rebate equal to zero

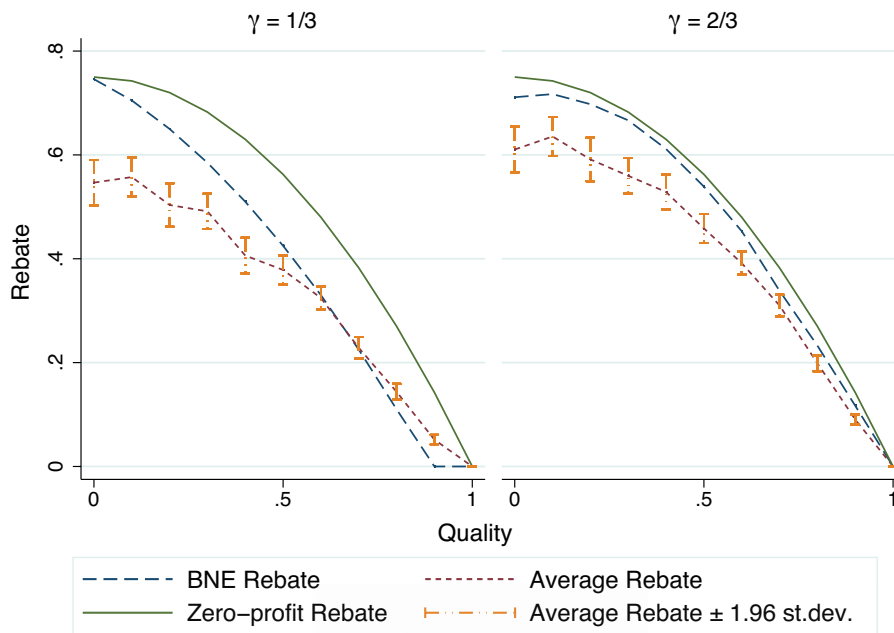


Figure 1.4: Equilibrium and empirical bidding functions by treatment. $N = 45$ per treatment.

those in the lower γ treatment. The Mann-Whitney test shows that this difference is highly statistically significant ($Z = 6.315, p < 0.001$). Moreover, the distance from the equilibrium is increasing and significantly higher for every quality level: the higher the quality level, the lower the distance from the equilibrium (this depends especially by the experimental constraint).¹⁴

Table 1.1: Average distance (in absolute values) from the equilibrium rebate by treatments.

Treatment	Subjects	Eq. Rebate	Distance (abs.)
$\gamma = 1/3$	45	0.389	.1189
$\gamma = 2/3$	45	0.462	.0734
Diff.			.0455
M-W test (<i>p-value</i>)			< .001

In other words, the level of noise is endogenous and depends on the treatment conditions. One possible explanation for this phenomenon relies on the fact that the same deviation from equilibrium, call it Δ , yields a variation of the overall score equal to $\Delta/3$ ($2\Delta/3$) if $\gamma = 1/3$ ($\gamma = 2/3$), respectively. Consistently with classic models of equilibrium with endogenous noise - take, for example, McKelvey and Palfrey (1995) Quantal Response Equilibrium - we should then expect less noise in the treatment in which the impact of the latter on the overall score is higher, as it happens when $\gamma = 2/3$.

This evidence is of extreme importance for us since bidders' noise around equilibrium may be

¹⁴See Figure 1.B.1 in Appendix 1.B.

responsible for inefficient allocations (this is what we will define as the “indirect effect” in Section 1.4.1). As a consequence, the evidence provided in Figure 1.4 - that individuals playing auctions with higher weight on the rebate play closer to equilibrium - could support the conclusion that auctions with $\gamma = 2/3$ may be characterized by higher efficiency. To this aim, in Table 1.2, for each treatment, we compute the relative frequency with which the auction has been awarded to each group member, ranked according to his relative efficiency, with RANK1 (RANK5) indicating the bidder with the highest (lowest) pseudo-type, respectively.

Table 1.2: Distribution of winners by efficiency and treatments

Auction winner	Relative Frequencies		
	<i>High weight on quality</i>	<i>Low weight on quality</i>	<i>Total</i>
$RANK_1$	94.95	43.43	69.19
$RANK_2$	5.05	42.42	23.74
$RANK_3$	0.00	11.11	5.56
$RANK_4$	0.00	3.03	1.52
Total	100.00	100.00	100.00

As Table 1.2 shows, when quality has a higher weight than price, 95% of the auctions are awarded to the most efficient player (RANK1); when the rebate has a higher weight, this percentage drops to 43%. In sum, our descriptive statistics point towards a 51.52% higher probability of getting an efficient outcome when the weight of quality in the scoring mechanism is high rather than low. This difference in efficiency is observed despite the higher noise detected in treatment which favours quality over price (see Figure 1.4 above). Section 1.4.1 aims at rationalizing this apparent contradiction.

1.4.1 Welfare analysis: a two-stage approach

In what follows, we will apply a “mediation analysis” (Imai et al., 2011), which allows us to understand which factors may have affected the differences in the efficient allocation. The idea is that the effect is not only due to a “direct” relation between the treatment condition and the efficient outcome, but also that the treatment can cause other effects that “indirectly” affect the efficiency.

As discussed in section 1.2, the effect of γ might come through the strategic properties of the treatment conditions (i.e., the shape of the potential score function). Considering that, if players always play the equilibrium, we would always observe an efficient allocation, to study to what extent γ determines the likelihood of an efficient outcome makes only sense out of equilibrium. In this sense, for any given deviation from the equilibrium, γ can have two different effects on the efficiency:

- a “direct” effect which captures the impact of the strategic characteristics of the underlying

game (e.g., the shape of the potential score function);

- an “indirect” effect which takes into account the level of noise due to the differences in the two conditions. The “trembles” around equilibrium may also depend on the auction and matching group specific characteristics (e.g., the individual heterogeneity and the realized level of quality in each group).

To this aim, we perform a “mediation analysis” to disentangle the “direct” treatment effect on efficiency from the “indirect” one. Figure 1.5 illustrates these two effects upon which we design our estimation strategy.

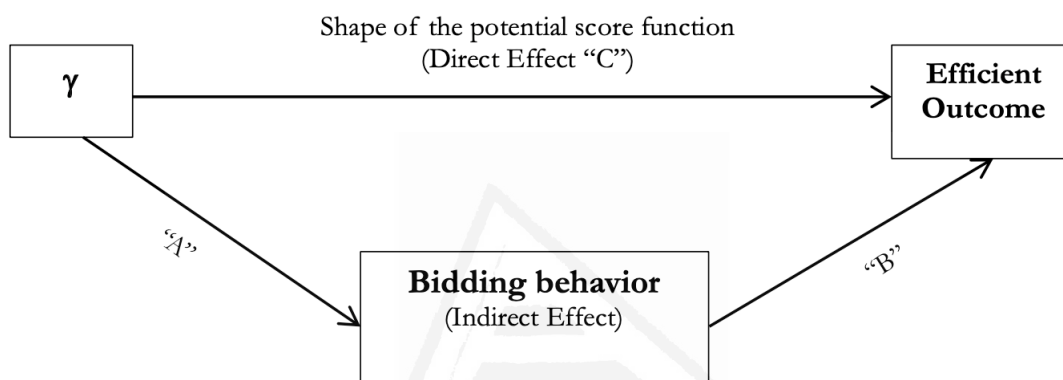


Figure 1.5: Direct and indirect effect of γ on efficiency

Looking at the Figure, the value “C” represents the “direct” effect of our treatment variable on efficiency, that is, how the potential score function characteristics would affect the probability of the efficient player to win the auction if players made identical mistakes under both treatments. The product of values “A” and “B” represents, instead, the “indirect” effect of γ on efficiency. If, say, the “direct” effect outweighs the “indirect” one, then the auction designer would be in the position to select which game is more likely to generate his preferred outcome by simply looking at the equilibrium properties of alternative game-forms, which is the standard practice of mechanism design. Conversely, if the “indirect” effect turned out to be stronger, the auction designer must also take into account behavioral and context-specific factors, which may substantially complicate his task.

With these premises, we adopt a two-stage least-squares random-effects estimator to quantify the “direct” and “indirect” effects of γ on efficiency. Our estimation strategy (see Appendix 1.C for details) relies on the following stages:

- Stage 1. We regress the difference between observed and equilibrium bids on: i) our treatment variable γ by way of a binary index, positive when $\gamma = 2/3$; ii) proxies of the auction-specific

randomized quality levels and iii) identifiers of the CRT partition (see Section 1.3.2). Stage 1 allows us to quantify the value “A” in Figure 1.5 as the marginal impact of γ on the observed “trembles” around equilibrium.

- Stage 2. We regress the likelihood of an efficient outcome on i) the predicted deviations from equilibrium estimated in Stage 1; ii) our treatment variable, γ , and iii) the same proxies in ii) used in Stage 1. Stage 2 allows us to disentangle the value “C” (as the marginal impact of γ on efficiency) from the value “B” of Figure 1.3.2 (as the marginal impact on efficiency of the predicted bidders’ trembles around equilibrium).

Detailed results from the estimation strategy are reported in Appendix 1.C. Table 1.3 reports only the estimated coefficients of the “direct”/“indirect” effects, together with their sum.

Table 1.3: Estimation of the direct/indirect effects

	<i>Marginal Impact</i>	<i>p-value</i>
“Direct” Effect	-0.708	0.003
“Indirect” Effect	+0.198	0.096
Total	-0.510	0.008

As Table 1.3 shows, we find an overall negative and significant treatment effect on efficiency in that auctions characterized by higher weight on price are 51% less likely to be awarded to the most efficient players, which is in line with the descriptive results showed in Table 1.2. Now, following the estimations obtained through the “mediation analysis”, we can shed more light on the effect of γ on this lost in efficiency. As reported in Table 1.3, the “direct” effect of γ is negative and significant, suggesting that the deviation from equilibrium, when players make the same level of mistakes, is over 70% less likely to yield efficiency when $\gamma = 2/3$. On the other hand, we find a positive and statistically significant “indirect” effect of γ when we consider the magnitude of bidders’ noise around equilibrium. Figures suggest that, accounting for the level of noise, the individuals’ heterogeneity and differences in the realized quality level for each matching group, the deviation from the equilibrium is around 20% more likely when γ is higher. This result is not enough to outweigh the “direct” effect of the treatment.

To summarize, this exercise allows us to understand if the effects of the treatment conditions (i.e., the level of γ) is principally due to the exogenous characteristics of the game or to the endogenous noise driven the choice of γ , together with the individuals’ heterogeneity and the groups compositions. The findings show that the higher efficient allocation found when the quality matters more for the

procurer, is mainly due to the estimated “direct” effect, once the “indirect” effect (i.e., the bidders’ out-of-equilibrium behavior) is considered.

1.5 Conclusion

Our experiment provides the mechanism designer with two complementary pieces of information - although confined within the very specific bounds of our parametric setting. First, more weight on rebate reduces noise, as out-of-equilibrium deviations are more costly (in terms of score) for the bidders. Second, more weight on quality yields higher efficiency, in spite of a higher level of the associated noise. It should be noticed, though, that the (quite natural, from a viewpoint of mechanism design) search for an “optimal γ ” is well beyond the scope of this paper. This is because the latter is usually influenced by contextual factors specific of each tender and by the constraints put in place by the legislators. For instance, in Italy the national Law for Public Contract makes it mandatory to use at least a weight of 0.7 on quality when public buyers wish to carry out a procurement procedure by using a scoring auction. These considerations notwithstanding, our analysis allows us to conclude that i) the level of deviation from equilibrium (the “indirect effect”) varies with the weight associated with each dimension composing the score, and that ii) in the choice of the optimal weights the designer should take into account the differences in efficiency due to both – “direct” and “indirect” - effects. The most natural extension to this paper would be to look at a procurement environment in which - by analogy with Che (1993) - participants have to decide both the level of quality and the rebate. This could be implemented by considering bidders with heterogeneous (and privately observed) productivity who have to determine - simultaneously and independently - the quality and the price of their tender.¹⁵

¹⁵Camboni et al. (2019) run a procurement experiment involving both one and two dimensional (scoring) auctions. They find that increasing the dimensionality and the size of the suppliers’ strategy space increases their tendency to make suboptimal offers, thus undermining the theoretical superiority of more complex mechanisms.

Bibliography

- ASKER, J. AND CANTILLON, E. (2008): “Properties of scoring auctions,” *The RAND Journal of Economics*, 39, 69–85.
- BICHLER, M. (2000): “An experimental analysis of multi-attribute auctions,” *Decision Support Systems*, 29, 249–268.
- CAMBONI, R., CORAZZINI, L., GALAVOTTI, S. ET AL. (2019): “Bidding on price and quality: An experiment on the complexity of scoring auctions,” Tech. rep., Dipartimento di Scienze Economiche” Marco Fanno”.
- CHANG, W.-S., CHEN, B. AND SALMON, T. C. (2015): “An investigation of the average bid mechanism for procurement auctions,” *Management Science*, 61, 1237–1254.
- CHANG, W.-S., SALMON, T. C. AND SARAL, K. J. (2016): “Procurement auctions with renegotiation and wealth constraints,” *Economic Inquiry*, 54, 1684–1704.
- CHE, Y.-K. (1993): “Design competition through multidimensional auctions,” *The RAND Journal of Economics*, 668–680.
- CHEN-RITZO, C.-H., HARRISON, T. P., KWASNICA, A. M. ET AL. (2005): “Better, faster, cheaper: An experimental analysis of a multiattribute reverse auction mechanism with restricted information feedback,” *Management Science*, 51, 1753–1762.
- CUEVA, C., ITURBE-ORMAETXE, I., MATA-PÉREZ, E. ET AL. (2016): “Cognitive (ir) reflection: New experimental evidence,” *Journal of Behavioral and Experimental Economics*, 64, 81–93.
- ENGELBRECHT-WIGGANS, R., HARUVY, E. AND KATOK, E. (2007): “A comparison of buyer-determined and price-based multiattribute mechanisms,” *Marketing Science*, 26, 629–641.
- FISCHBACHER, U. (2007): “z-Tree: Zurich toolbox for ready-made economic experiments,” *Experimental economics*, 10, 171–178.
- FREDERICK, S. (2005): “Cognitive reflection and decision making,” *Journal of Economic perspectives*, 19, 25–42.
- GREINER, B. ET AL. (2004): “The online recruitment system orsee 2.0—a guide for the organization of experiments in economics,” *University of Cologne, Working paper series in economics*, 10, 63–104.

- GRIMM, V., KOVARIK, J. AND PONTI, G. (2008): “Fixed price plus rationing: an experiment,” *Experimental Economics*, 11, 402–422.
- IMAI, K., JO, B. AND STUART, E. A. (2011): “Commentary: Using potential outcomes to understand causal mediation analysis,” *Multivariate Behavioral Research*, 46, 861–873.
- KAGEL, J. H. AND LEVIN, D. (2002): “Bidding in common-value auctions: A survey of experimental research,” *Common value auctions and the winner’s curse*, 1, 1–84.
- (2008): “Auctions: A survey of experimental research, 1995-2008,” *Handbook of experimental economics*, 2.
- MCKELVEY, R. D. AND PALFREY, T. R. (1995): “Quantal response equilibria for normal form games,” *Games and economic behavior*, 10, 6–38.
- SANTAMARÍA, N. (2015): “An analysis of scoring and buyer-determined procurement auctions,” *Production and Operations Management*, 24, 147–158.
- SESHADRI, S., CHATTERJEE, K. AND LILIEN, G. L. (1991): “Multiple source procurement competitions,” *Marketing Science*, 10, 246–263.
- SHACHAT, J. AND SWARTHOUT, J. T. (2010): “Procurement auctions for differentiated goods,” *Decision Analysis*, 7, 6–22.
- STRECKER, S. (2010): “Information revelation in multiattribute English auctions: A laboratory study,” *Decision Support Systems*, 49, 272–280.

Appendix

1.A Proof of Proposition 1

Let $s_i = s(q) \equiv 1 - c(q)$ define player i 's pseudo-type. Consider the modified game in which each bidder privately observes a "value" s and submits a bid (that is, announces a score) $\sigma(s_i) \leq s_i$. Bidder i 's expected payoff $\tilde{\pi}$ writes

$$E[\tilde{\pi}_i(s_i, \sigma(s_i))] = (s_i - \sigma(s_i)) \text{Prob}[\sigma(s_i) > \max_{j \neq i} \sigma(s_j)]$$

where $\max_{j \neq i} \sigma(s_j)$ indicates the highest score among bidder i 's competitors. This modified game is then a first-price (selling) auction where, upon observing s_i , bidder i submits a bid $\sigma_i(s_i)$. If bidder i 's bid is the highest submitted bid then bidder i gets profit equal to $(s_i - \sigma_i(s_i))$, and zero otherwise. It is easy to show that the modified game is strategically equivalent to the original one. Loosely speaking, two games, A and B, are strategically equivalent when the two games have both the same set of agents and strategies, and game B's payoff function(s) can be obtained through a transformation of game A's payoff function(s). Since $s_i = s(q_i)$ is defined as the maximum score bidder i with type q can obtain, $\sigma(s_i)$ coincides with the score obtained type q submitting a rebate $r(q)$. That is, $S(q, r(q)) = \sigma(s(q))$. Moreover, bidders' profit in the "original" game, (π_i) equal that of the modified game $(\tilde{\pi})$ except for a positive constant factor $(1/\gamma)$:

$$\sigma_i(q, r(q)) = 1 - r(q) - c(q) = \frac{1}{\gamma} [s(q) - \sigma(s(q))] = \frac{1}{\gamma} \tilde{\pi}_i(s(q), \sigma(s(q)))$$

while the winning probability of winning is exactly the same, as $\sigma s_i = \sigma(s(q)) = S(q_i, r(q))$. It results that the payoff functions in the two payoff functions differ by a multiplicative positive constant only, $\frac{1}{\gamma}$. Consequently, the two games have the same equilibria. This implies that the BNE of our original problem – that is, the equilibrium bidding function $\tilde{r}_\gamma(q)$ of the scoring auction – can be derived by deriving the equilibrium of the modified game:

$$\tilde{r}_\gamma(q) = \frac{1}{\gamma} [\sigma_\gamma^*(s(q)) - (1 - \gamma)q]$$

where $\sigma^*(s)$ is the BNE of a standard first-price auction. Thus:

$$\sigma_\gamma^*(s(q)) = \frac{1}{H_\gamma(s)} \int_{\underline{s}(\gamma)}^{s(q)} y h_\gamma(y) dy$$

with $H_\gamma(s) = G_\gamma^4(s)$, $h_\gamma = H'_\gamma(s)$ and $G_\gamma(s)$ is the distribution function of the random variable s . The only caveat is that we have not imposed any condition to ensure that, for each of the relevant parametric cases $\gamma \in \{1/3; 2/3\}$, $\tilde{r}_\gamma(q) \geq 0$. We proceed by computing first the "tentative" equilibrium rebate function $\tilde{r}_\gamma(\cdot)$, then we check that the non-negativity constraint is fulfilled. Deriving the explicit form of the equilibrium rebate function turns out to be quite cumbersome and uninformative, so it is not provided in this proof. Let $\tilde{r}_\gamma(q) = [\sigma_\gamma^*(s(q)) - (1 - \gamma)q]/\gamma$. Explicit computations show that $\tilde{r}_\gamma(q)$ assumes feasible values only for $q \in [0, 1]$ when $\gamma = 2/3$ and thus it is actually a BNE, then $r_{2/3}^*(q) = \tilde{r}_{2/3}(q)$. When $\gamma = 1/3$, instead, it becomes negative for all values of $q > q_0 \approx 0.8884$. Intuitively, this occurs because when the weight of quality in the scoring rule is sufficiently high, types with the high values of q enjoy such a large probability of winning that they would be willing to lower their rebate below 0, that is, to submit a bid above 1, which is forbidden by the rules of the game. We then conjecture that $r_{1/3}^*(q) = \max\{\tilde{r}_{1/3}(q), 0\}$.

In order to show that this is indeed an equilibrium let $q^\circ = \sup\{q : \tilde{r}_{1/3}(q) \geq 0\}$, that is, q° is the highest unconstrained type. Notice that if $r_{1/3}^*(q) = 0$ is part of an equilibrium for all $q > q^\circ$, then the probability of winning is still monotonic in the range $(q^\circ, 1]$. Consider any type $q \leq q^\circ$. The equilibrium bidding function is not affected by the constrain operating on types $q > q^\circ$. Indeed, due to the monotonicity of $s(q)$, the latter set of types would still submit a higher score than the set of types with $q \leq q^\circ$, thus leaving their probability of winning of all types $q \leq q^\circ$ unaffected. Hence, bidders with such types would have no incentives to deviate from bidding $r_{1/3}^*(q)$ if bidders with $q > q^\circ$ bid $r_{1/3}^*(q) = 0$. Consider now type $q' > q^\circ$ and suppose it envisages to submit a feasible $r_{1/3}(q') : 0 \leq r_{1/3}(q') \leq 1 - c(q')$. Then there must exist a type q'' such that

$$\begin{aligned} \sigma_{1/3}^*(q'', r_{1/3}^*(q'')) &= \sigma_{1/3}^*(q', r_{1/3}(q')) \Leftrightarrow (1 - \gamma)q'' = (1 - \gamma)q' + \gamma r_{1/3}(q') \Rightarrow \\ &\Rightarrow q'' = q' + \gamma/(1 - \gamma)r_{1/3}(q') \end{aligned}$$

Thus

$$r_{1/3}(q') = (1 - \gamma)/\gamma \Delta(q') = 2\Delta(q')$$

where $\Delta(q') = (q'' - q')$. We then need to prove that

$$[1 - c(q')](q')^4 \geq [1 - c(q') - 2\Delta(q')](q'')^4, \quad \forall q' > q^\circ, \quad \forall \Delta(q') : r_{1/3}(q') \leq 1 - c(q') \quad (1.3)$$

where the LHS of inequality 1.3 represents the expected payoff of type q' when playing the conjectured equilibrium rebate $r_{1/3}^*(q) = 0$, and the RHS of inequality (1) measures the expected payoff of type q' when playing a strictly positive rebate yielding the same score as type q'' . Inequality 1.3 can be rewritten as follows

$$\begin{aligned} [1 - c(q')](q')^4 &\geq [1 - c(q') - 2\Delta(q')](q' + \Delta(q'))^4 \Leftrightarrow \\ \Leftrightarrow [1 - c(q')][(q' + \Delta(q'))^4 - (q')^4] &\leq 2\Delta(q')(q' + \Delta(q'))^4 \end{aligned} \quad (1.4)$$

Notice that both the LHS and the RHS of inequality 1.4 are strictly increasing functions of $\Delta(q')$, and they are both equal to zero when $\Delta(q') = 0$. Call them $LHS_{(2)}(\Delta(q'))$ and $RHS_{(2)}(\Delta(q'))$, respectively. In order inequality 1.4 to hold it would then suffice to show that

$$\frac{\delta LHS_{(2)}(\Delta(q'))}{\delta \Delta(q')} \leq \frac{\delta RHS_{(2)}(\Delta(q'))}{\delta \Delta(q')}, \quad \forall \Delta(q') : r_{1/3}(q') \leq 1 - c(q'),$$

that is

$$\begin{aligned} 4[1 - c(q')](q' + \Delta(q'))^3 &\leq 2(q' + \Delta(q'))^4 + 8\Delta(q')(q' + \Delta(q'))^3 \Leftrightarrow \\ \Leftrightarrow 4[1 - c(q')](q' + \Delta(q'))^3 &\leq 2(q' + \Delta(q'))^3[(q' + \Delta(q')) + 4\Delta(q')] \Leftrightarrow \\ &\Leftrightarrow 2[1 - c(q')] \leq [q' + 5\Delta(q')] \end{aligned}$$

which is always fulfilled for every $q' > q^\circ$ and every feasible $\Delta(q')$ given the assumption on the cost function $c(\cdot)$. Hence $r_{1/3}^*(q) = \max\{\tilde{r}_{1/3}(q), 0\}$ is indeed a BNE.

1.B Supplementary Statistical Evidence

Table 1.B.1 presents the summary statistics for selected individual-level variables which we deem to represent a good proxy of the unobserved individual heterogeneity which may have an impact on bidding behavior. With respect to the Cognitive Reflection Test, we find that 38% (28%) [34%] are classified as “impulsive” (“reflective”) [“others”], respectively. From Table 1.B.2 we also notice that the sample distribution over the CRT categories has a strong gender component: while 48% of the male sample is categorized as “reflective” (and the remaining 52% is approximately equally distributed across the other categories), the same percentage of females are classified as “impulsive” and only 13% as “reflective”. This evidence is in line with previous findings in the literature (take, e.g., Frederick (2005) and Cueva et al. (2016)). With respect to the education field, the majority of our players is enrolled in an economic/business degree at a master level and expect to continue studying further. Parents’ education level is relatively high, with over 50% of the sample declaring their father/mother holds a tertiary level of education. Only 20% of our players declare to have worked during the previous week and the reported weekly cash holdings is highly disperse, ranging from 15€ to 450 €. Finally, the majority is in favour of merit-based compensation, while 67% of our players are tempered by prudent trust in others.

Table 1.B.1: Descriptive Statistics of the sample in Study 1

Variable	Description	Obs.	Mean	St.Dev	Min	Max
CRTgroup	Cognitive Reflection Test: 1 to 3	92	1.95	0.79	1	3
	= 1 if Others	31				
	= 2 if Impulsive	25				
	= 3 if Reflective	26				
age	Age	92	22.73	2.19	19	31
woman	Gender: = 1 if woman	92	0.57	0.50		
economics	Field of education: = 1 if student of economics	92	0.61	0.49		
law	Field of education: = 1 if law	92	0.25	0.44		
political science	Field of education: = 1 if political science	92	0.11	0.31		
master_degree	Level of education: = 1 if master’s degree	92	0.64	0.48		
phd	Level of education: = 1 if Ph.D.	92	0.05	0.23		
exp_master_degree	Expected level of education: = 1 if master’s degree	92	0.46	0.50		
exp_phd	Expected level of education: = 1 if Ph.D.	92	0.50	0.50		
employed	Labour market status: = 1 if employed	92	0.20	0.40		
cash_holdings	Weekly cash holdings	88	91.08	71.62	15	450
unhappiness	Degree of unhappiness: 1 (happy) to 7 (unhappy)	90	3.79	1.72	1	6
trust	Trust in others: = 1 if yes	90	0.37	0.53	0	2
meritocracy	Preference for Meritocracy: = 1 if yes	89	0.90	0.30		
inequality	Preference for income inequality: 1 (egalitarianism) to 7 (merit)	90	4.83	1.62	1	7
RSR	Room Size Ratio	90	2.51	1.29	0.8	10

Because treatment assignment has been randomized, we should observe no statistically significant differences across characteristics between the two treatment groups. Table 1.B.2 shows that differences are indeed very moderate. In terms of individual characteristics, subjects allocated to auctions with $\gamma = 1/3$ (the control group) are comparable to those allocated to auctions with $\gamma = 2/3$ (the

treatment group).

Table 1.B.2: Difference in means between the treatment and the control group and statistical significance levels.

Variable	Description	Obs.	Difference in means T - C	p - value
CRTgroup	Cognitive Reflection Test: 1 to 3	92		
	= 1 if Others	31	0.0941	0.3453
	= 2 if Impulsive	25	0.0052	0.9596
	= 3 if Reflective	26	-0.0093	0.2955
age	Age	92	0.7296	0.1114
woman	Gender: = 1 if woman	92	0.1494	0.1517
economics	Field of education: = 1 if student of economics	92	0.0170	0.8690
law	Field of education: = 1 if law	92	-0.0761	0.4049
political science	Field of education: = 1 if political science	92	0.0388	0.5554
master_degree	Level of education: = 1 if master's degree	92	-0.0496	0.6242
phd	Level of education: = 1 if Ph.D.	92	0.0629	0.1874
exp_master_degree	Expected level of education: = 1 if master's degree	92	-0.0634	0.5471
exp_phd	Expected level of education: = 1 if Ph.D.	92	0.0652	0.5367
father_sec_educ	Father's level of education: = 1 if secondary	92	0.0293	0.7674
father_tert_educ	Father's level of education: = 1 if tertiary	92	-0.1957	0.0615
father_phd	Father's level of education: = 1 if Ph.D.	92	0.1035	0.1288
mother_sec_educ	Mother's level of education: = 1 if secondary	92	0.1811	0.0674
mother_tert_educ	Mother's level of education: = 1 if tertiary	92	-0.2189	0.0357
mother_phd	Mother's level of education: = 1 if Ph.D.	92	-0.0463	0.3738
employed	Labour market status: = 1 if employed	92	0.1220	0.1435
cash_holdings	Weekly cash holdings	88	8.2506	0.5920
unhappiness	Degree of unhappiness: 1 (happy) to 7 (unhappy)	90	-0.0128	0.9720
trust	Trust in others: = 1 if yes	90	0.1677	0.1337
meritocracy	Preference for Meritocracy: = 1 if yes	89	-0.0562	0.3855
inequality	Preference for income inequality: 1 (egalitarianism) to 7 (merit)	90	0.6195	
RSR	Room Size Ratio	90	-0.1559	0.5696

Additional Figures

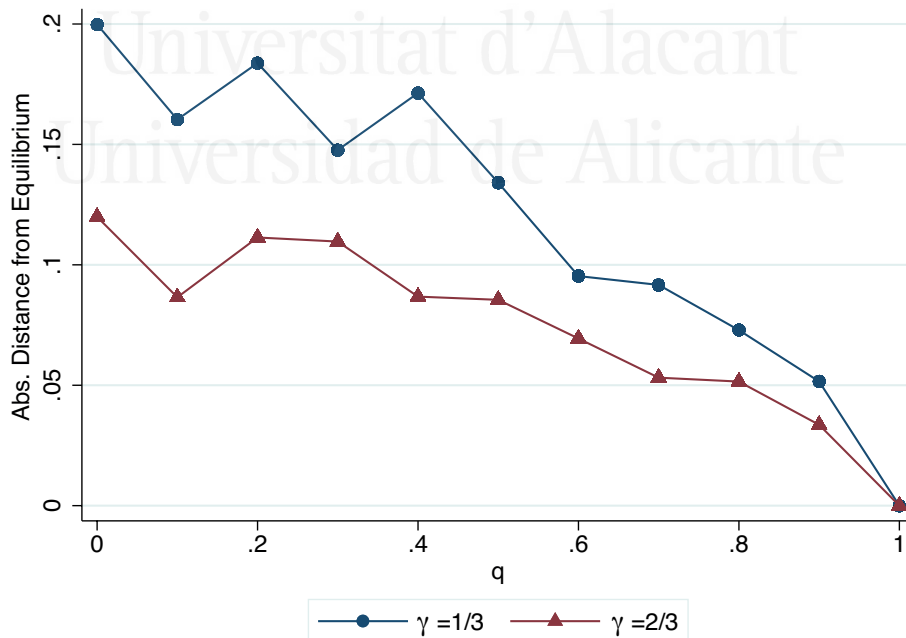


Figure 1.B.1: Average bid distance (in absolute values) from the equilibrium by treatment.

Figure 1.B.1 shows the average bid distance (in absolute values) from equilibrium for each level of quality and by treatments. When $\gamma = 2/3$, players bid closer to equilibrium for each level of quality. Those differences are statistically significant, between control and treatment, when quality is higher than 0.3 (all p-values are lower than 0.01).

1.C The Econometric Model

We estimate a system of equations using a two-stage least-squares random-effects to identify the “direct” and “indirect” effects described in section 1.4.1. The random effects approach provides consistent estimates in our context, since specific random effects are uncorrelated with the experimental design and the way subjects are randomized across periods, sessions and treatment.

First of all, we consider each matching group and round as individual observation and classify bidders according to their BNE score (that is, the score under the assumption that bidders play the BNE bids of Proposition 1.1). Doing so, we obtain an efficiency-based ranking of 5 categories, from RANK1 to RANK5, indicating the bidder with the highest (lowest) BNE score, respectively. Then, we employ the two-stages least-squares random-effect model.

1.C.1 Stage 1

As motivated in the main text, our treatment variable might have a direct influence in explaining deviations from equilibrium. This is why the latter, $(r - r^*)_{RANK1_{it}}$ and $(r - r^*)_{RANK2_{it}}$, can be treated as endogenous to the value of γ as well as to the randomized level of quality and to individual-specific characteristics (such as the results from the Cognitive Reflection Test). Thus, we run the following two regressions, where $i = 1, \dots, 18$ identifies the matching group and $t = 1, \dots, 11$ the period (i.e., the auction repetition):

$$(r - r^*)_{RANK1_{it}} = \beta_0 + \beta_1\gamma_{2/3} + \beta_2q_{RANK1_{it}} + \beta_3(q_{RANK1_{it}})^2 + \beta_4CRTgroup_{RANK1_{it}} + \epsilon_{it} \quad (1.5a)$$

$$(r - r^*)_{RANK2_{it}} = \alpha_0 + \alpha_1\gamma_{2/3} + \alpha_2q_{RANK2_{it}} + \alpha_3(q_{RANK2_{it}})^2 + \alpha_4CRTgroup_{RANK2_{it}} + \epsilon_{it} \quad (1.5b)$$

where: i) $(r - r^*)_{RANK1_{it}}$ and $(r - r^*)_{RANK2_{it}}$ are the differences between observed and equilibrium bidding function of RANK1 (RANK2) player of matching group i in period t ; ii) $\gamma_{2/3}$ is a dummy variable, positive if the weight on rebate is equal to $2/3$ (iii) $q_{RANK1_{it}}$ ($q_{RANK2_{it}}$) is the realized quality of RANK1 (RANK2) player of matching group i in period t ; (iv) $CRTgroup_{RANK1_{it}}$ ($CRTgroup_{RANK2_{it}}$) is a two-dimensional column vector of dummy variables for the group RANK1 (RANK2) player of

matching group i has been assigned to on the basis of the Cognitive Reflection Test (CRT group, see Section 1.3.2). The parameters α_1 and β_1 are the marginal effects of our treatment variable γ on bidders' "trembles" around the equilibrium bid.

1.C.2 Stage 2

To detect the "direct" effect of the treatment on the efficient allocation we run the following regression, where $i = 1, \dots, 18$ identifies the matching group and $t = 1, \dots, 11$ the period:

$$y_{it} = \delta_0 + \delta_1 \gamma_{2/3} + \delta_2 (r - r^*)_{RANK1_{it}} + \delta_3 (r - r^*)_{RANK2_{it}} + \delta_4 q_{RANK1_{it}} + \delta_5 (q_{RANK1_{it}})^2 + \delta_6 q_{RANK2_{it}} + \delta_7 (q_{RANK2_{it}})^2 + \delta_8 Period_i + c_i + \mu_i \quad (1.6)$$

where: i) y_{it} is a binary index, positive if the auction of matching group i at period t is won by the most efficient type; ii) $\gamma_{2/3}$ is a dummy variable, positive if the weight on rebate is equal to $2/3$ (remember that the treatment is randomized between groups); iii) $(r - r^*)_{RANK1_{it}}$ ($(r - r^*)_{RANK2_{it}}$) is the predicted difference between observed and equilibrium bidding function of RANK1 (RANK2) estimated in Stage 1; iv) $q_{RANK1_{it}}$ ($q_{RANK2_{it}}$) is the realized quality of RANK1 (RANK2) player of matching group i in period t ; v) $Period_i$ identifies the round number vi) c_i is a matching group specific random effect, uncorrelated with the independent variables and i.i.d. over the panel and vii) μ_{it} is an idiosyncratic error term.

The choice of variables in (iii) in equation 1.6 is motivated by the evidence that, for a given distribution of quality levels, the probability of an efficient outcome increases (decreases) when the efficient (non-efficient) bidder RANK1 (RANK2) overbids. Since RANK3 and RANK4 players never win auctions where $\gamma = 2/3$, our econometric strategy focus on the bidding behavior of RANK1 and RANK2 players. The parameter δ_1 is the marginal effect of our treatment variable γ on the probability to get an efficient outcome – holding all other variables constant at their means – and represents our "direct" effect "C" in Figure 1.5.

1.C.3 Empirical results

Table 1.C.1 reports the summary statistics of the dependent variables described above.

The system of equations in sections 1.C.1 and 1.C.2 has been estimated using a two-stage least-squares random-effect estimator. This method applies an OLS regression on both equations 1.5a and 1.5b and gets predictions for $(r - r^*)_{RANK1_{it}}$ and $(r - r^*)_{RANK2_{it}}$ (Stage 1). After substituting

Table 1.C.1: Descriptive Statistics: independent variables

Variable	Obs.	Mean	St.Dev	Min	Max
q_{RANK1}	198	.631	.293	.2	1
q_{RANK2}	198	.440	.254	0	.8
$(r - r^*)_{RANK1_{it}}$	198	-.030	.103	-.697	.160
$(r - r^*)_{RANK2_{it}}$	198	-.022	.112	-.570	.155
$Reflective_{RANK1}$	198	.323	.468	0	1
$Impulsive_{RANK1}$	198	.424	.495	0	1
$Reflective_{RANK2}$	198	.343	.476	0	1
$Impulsive_{RANK2}$	198	.459	.499	0	1

$(r - r^*)_{RANK1_{it}}$ and $(r - r^*)_{RANK2_{it}}$ with their predictions, equation 1.6 is estimated by OLS and the residuals are used to estimate the covariance matrix of equation errors (Stage 2).

As showed above, the “direct” effect reported in Table 1.3 is given by the estimated parameter δ_1 , which is the marginal effect of γ on the probability to get an efficient outcome – holding all other variables constant at their means. On the other hand, the “indirect” effect described in Section 1.4.1 is computed by the following sum of products: $(\beta_1 \times \delta_2) + (\alpha_1 \times \delta_3)$. In particular, $(\beta_1 \times \delta_2)$ measures the extent to which efficiency changes exclusively when RANK1 players’ bidding functions change by the amount they would have changed has γ moved from 1/3 to 2/3. Similarly, $(\alpha_1 \times \delta_3)$ measures the extent to which efficiency changes when RANK2 players’ bidding functions change by the amount it would have changed has γ moved from 1/3 to 2/3. The overall effect is equal to the sum of the “direct” and “indirect” effects: $\delta_1 + (\beta_1 \times \delta_2) + (\alpha_1 \times \delta_3)$.

Detailed results are presented in Table 1.C.2. Column 1 presents the results of the final Stage 2; Columns 2 and 3 present the results from Stage 1 where predictions of both the endogenous variables $(r - r^*)_{RANK1_{it}}$ and $(r - r^*)_{RANK2_{it}}$ are computed.¹⁶

Column 1 reports the Stage 2 estimation. Results show that, holding all other variables constant at their means, the probability of getting an efficient outcome when $\gamma = 2/3$ is 70% smaller than when $\gamma = 1/3$. Not surprisingly, aggressive bidding strategies of RANK2 players generate a significantly negative effect on the probability of achieving an efficient outcome. Precisely, a 10% increase in the distance of RANK2 players bids from their BNE predictions lowers the likelihood of an efficient outcome by 17%.

Looking at columns 2 and 3, results point towards a significantly role of γ on the observed differences of players’ bids from their BNE predictions. On average, the marginal impact of γ is negative

¹⁶Since the variables $r_{RANK1} - r^*_{RANK1}$ and $r_{RANK2} - r^*_{RANK2}$ range from negative to positive values, we choose to add a constant value to the data (that is the minimum value of $r_{RANK1} - r^*_{RANK1}$ to $r_{RANK1} - r^*_{RANK1}$ and the minimum value of $r_{RANK2} - r^*_{RANK2}$ to $r_{RANK2} - r^*_{RANK2}$) in order to get a clear intuition of the estimated impacts without loss of generality.

Table 1.C.2: Determinants of the probability of the most efficient type to win the auction, Marginal Effect Values (MEs).

Variable	<i>Prob. Efficient Winner</i>	$r_{RANK1} - r_{RANK1}^*$	$r_{RANK2} - r_{RANK2}^*$
$\gamma_{2/3}$	-0.708*** (0.234)	-0.172*** (0.039)	-0.032 (0.020)
q_{RANK1}	-0.582 (0.779)	0.083 (0.149)	
q_{RANK1}^2	0.386 (0.650)	-0.149 (0.123)	
q_{RANK2}	-0.947** (0.462)		-0.331*** (0.117)
q_{RANK2}^2	1.294** (0.582)		0.481*** (0.144)
<i>Period_i</i>	-0.020* (0.011)	-0.003 (0.002)	-0.000 (0.002)
$r_{RANK1} - r_{RANK1}^*$	-0.827 (0.618)		
$r_{RANK2} - r_{RANK2}^*$	-1.759*** (0.464)		
<i>Impulsive</i> _{RANK1}		-0.015 (0.013)	
<i>Reflective</i> _{RANK1}		0.012 (0.014)	
<i>Impulsive</i> _{RANK2}			-0.024 (0.018)
<i>Reflective</i> _{RANK2}			0.033* (0.019)
Constant	2.946*** (0.688)	0.792*** (0.065)	0.580*** (0.033)
Observations	198	198	198
R^2	0.29		
<i>Underidentification test:</i>			
Kleibergen-Paap rk LM statistic	13.16		
<i>p</i> – value	0.0105		
<i>Overidentification test:</i>			
Hansen J statistic	3.08		
<i>p</i> – value	0.5445		

Standard errors (in parentheses) are clustered at the group level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

and more relevant for RANK1 than for RANK2 player. This means that, *ceteris paribus*, the observed bidding functions are closer to the equilibrium bidding function when $\gamma = 2/3$. Interestingly, bids from non-efficient reflective players are significantly higher than their BNE predictions compared to bids from other CRT-based groups. Figure 1.C.1 shows that, on average, the bidding functions of reflective players are closer to the equilibrium bidding functions compared to those of non-reflective players. This means that they are, on average, more likely to make a correct guess about the level of quality associated with the highest pseudo-type. Hence, when endowed with a sub-optimal (second-best) level of quality, reflective players probably realize that they need to overbid if they want to have a chance of winning the auction.

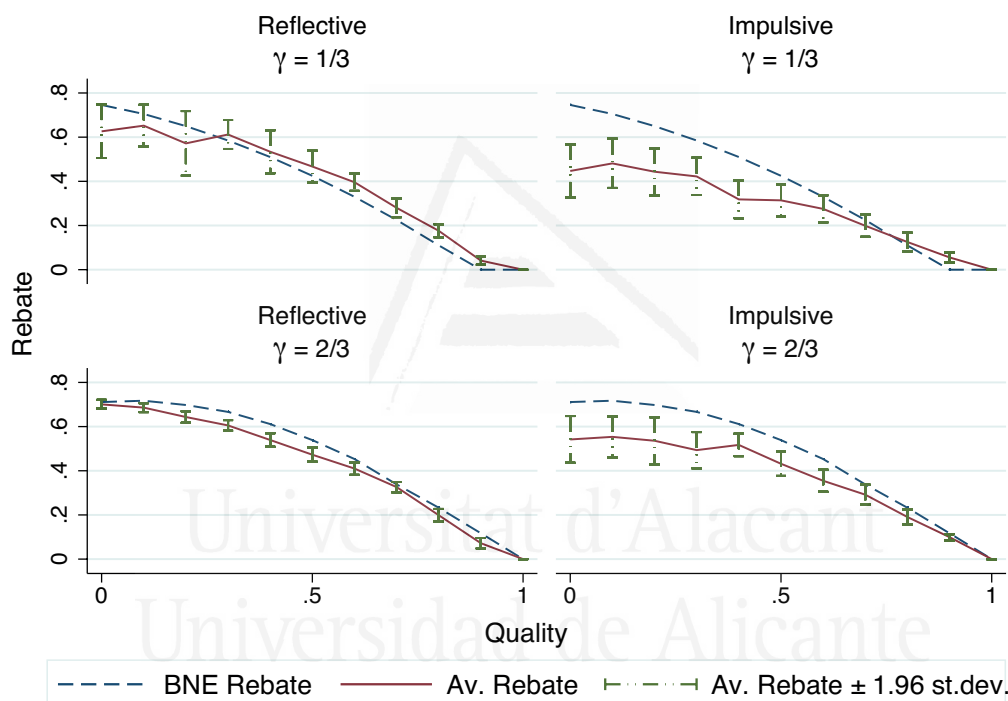


Figure 1.C.1: Equilibrium and empirical bidding functions by treatment and CRT groups

Our goodness-of-fit measures indicate that the model fits the data well. In particular, our under-identification tests - based on the Kleibergen-Paap rk LM statistics for estimates with heteroskedasticity-robust and clustered standard errors - reject the null hypothesis that Stage 1 and Stage 2 equations are underidentified (that is, the matrices of reduced form coefficients on the excluded instruments are full column rank). The Hansen statistics do not reject the null hypothesis that the instruments are valid instruments (that is, they are uncorrelated with the error term), and, hence, the excluded instruments are correctly excluded from the estimated equations.

1.D Experimental Instructions

WELCOME TO THE EXPERIMENT!

- This is an economic experiment on individual decision-making. We are only interested in your choices, not in who make them. Pay attention to your decisions because your behavior will affect your final reward.
- In this experiment, you will play for 11 periods in which you must take a decision. Each decision and its result is independent from any other; namely, every decision that you take in a specific period does not have any effect on the results of other experiment's periods.
- At the beginning of the experiment, the computer will match everyone anonymously and randomly in groups of 5 players. This matching will be the same during all the experiment.

How you can gain a reward during the experiment?

- First of all, you will receive €5 as a “show-up fee”, to acknowledge your availability. Moreover, at the end of the experiment, one period will be drawn randomly and your winnings in that specific period will be summed up to the show-up fee and it will be privately paid to you in cash at the end of the experiment.
- In what follows we will explain which decisions you have to take in each situation and how to deal with the user interface of the computer to implement them.
- Please do not disturb other participants during the course of the experiment. If you need help, raise your hand up and wait in silence. One of the proctors will come to help you as soon as possible.

Good Luck!

The Experiment

- In each period of the experiment, you will participate, together with the other 4 members of your group, in an auction in which everyone has to make an offer to win a “contract”, the object of which is the service that you produce.
- In case you win the auction, your profit will be the difference between your price, paid to buy your service, and the cost incurred to achieve it. In case you lose the auction, you will not

receive anything from the buyer and, moreover, you don't have any cost; namely, your profit will be zero. At the beginning of each period, you will receive the most relevant information about your bid, that is:

1. Your service quality, Q , for that period. This parameter (which value goes from 0, 10, 20 ... to 100) is not chosen by you (neither by other group members) but is assigned randomly and independently by the computer, with equal probability for each value.
2. The cost, C , associated to the assigned quality, Q , that you have to pay in case you win the auction.

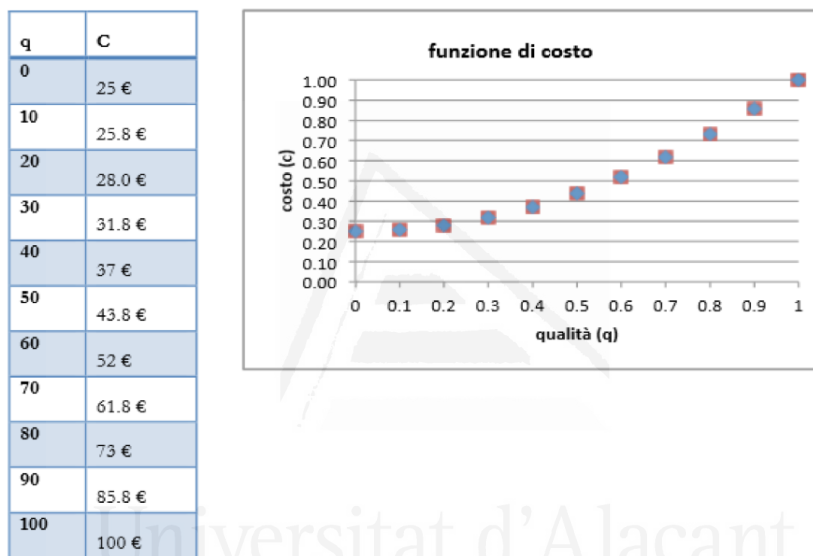


Figure 1.D.1

- The cost parameter C depends on the quality: a higher quality is associated with a higher cost. In the following graph (Figure 1.D.1) you can see the cost function, which is the same for all the players, in each phase and period. The graph and the table below report the level of quality assigned to each player, Q , and associated the cost, C , that each player has to pay in case he wins the auction: e.g., a quality level equal to 40 will always correspond to a cost equal to 37.
- In each period, after having observed your quality and the associated cost, you can make your bid as follows:
- Define your rebate, R , with respect to the baseline price (fixed at 100 € for all the players). The offered price (which correspond to $(100-R)$ €) cannot be lower than your production cost, C , in

that period. In this sense, the computer will prevent you from fixing a rebate that is too high so as to yield a monetary loss in case you win the auction (given that the price would cover your production costs for the service).

- Choose your rebate, R , and put it in the corresponding box of the user interface. The relationship between your rebate and the price you offer is given by $Price = 100 - R$
- As we said before, the computer does not allow any rebate R greater than $\text{€}(100 - C)$.

How is the winner determined?

- In each period, after that all the players have chosen their value R , they receive a Total Score (TS), which is a weighted average between the quality, Q , and a score associated to each bid, namely FINANCIAL SCORE (FS).
- The winner will be the player in your group who obtains the highest TS. In case two or more players obtain the highest TS, all of them will be considered as winners and their payoff will be shared across the winners. The Total Score, TS, is calculated with the following formula, which will be the same during the entire experiment:

$$TS = \gamma Q + (1 - \gamma)R$$

where Q is the level of quality determined by computer and R is the rebate, that is, a rebate of 10 € corresponds to 10 points, a rebate of 20 € corresponds to 20 points, and so on.

- The graph below (Figure 1.D.2) shows how varies the score associated with the rebate as a function of R .
- In the example of Table 1.D.1 (where $\gamma = 2/3$), the winner is Player 5 because he obtained the best TS. His profit (assuming a cost $C = 43.8$) would be: $100 - R - C = 100 - 37 - 43.8 = 19.2e$

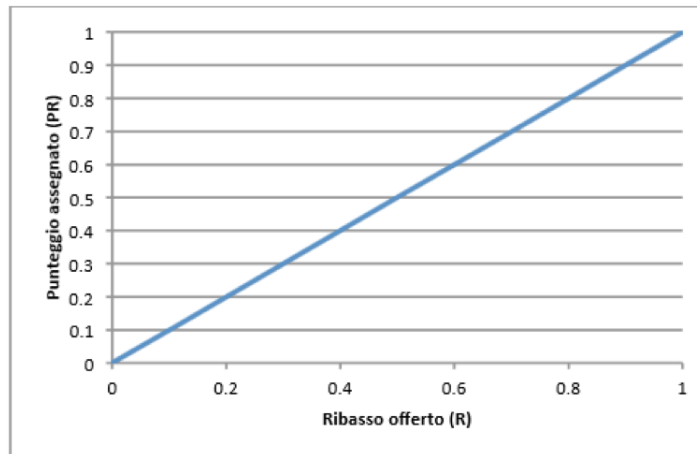


Figure 1.D.2: The rebate score as a function of R

Table 1.D.1: An example of one period: when $\gamma = 2/3$

Player	1	2	3	4	5
Quality (Q)	60	40	90	30	50
Cost (C)	52	37	85.8	31.8	43.8
Rebate (R)	25	39	1	30	37
$(PT = Q \times \frac{1}{3}) + (PR \times \frac{2}{3})$	37	39	31	30	41.3
Profit	0	0	0	0	19.2

To summarize:

- You are competing with other 4 players for the award of a service. The group will be the same during the entire experiment.
- In each period, the computer will assign randomly a value between 0 and 100, which is the quality of your service, Q. A cost C corresponds to each level of quality, Q, where C is an increasing function of Q, as specified in the table above. In each period, you must decide your rebate with respect to the baseline price (fixed for all players in each period to €100), R.
- The service will be assigned to your group member who obtains the highest TOTAL SCORE, TS, according to this formula:

$$TS = 1/3 \times Q + 2/3 \times FS$$

- • The winner's profit is equal to $(100 - R - C)$ €.
- In the case of multiple winners, the profits are equally divided between them.

- The losers in each group obtain a profit equal to zero.
- Before starting the experiment, you will participate in 5 dry periods, to better understand the rules of the game described above. The computer will simulate the behavior of the other 4 group members of your group, and it will provide you with a feedback at the end of each period.
- After these 5 periods, you will start playing for good. During 11 periods of the experiment, you will not receive any feedback (your score, your opponents' score, your profit, your opponent's profit, etc.). You will just receive all result of this experiment at the end of it.



Universitat d'Alacant
Universidad de Alicante

Promoting greater deliberation in the one-shot Stag-Hunt game played online

2.1 Introduction

The fact that behavior in social dilemmas can be affected by the extent of reflection has been the object of a lively debate in recent years (see Capraro, 2019, for a recent survey). Economists showed a great interest in exploring the uses of cognitive processes in social decision making (Rubinstein, 2007, 2013, 2016; Harrison et al., 2008; Piovesan and Wengström, 2009; Alós-Ferrer and Strack, 2014). One interesting way is to use the “dual-process” approach, which relies on contrasting a more automatic and “intuitive” process with a more considered and “deliberative” one (Kahneman, 2003; Frederick, 2005; Achtziger and Alós-Ferrer, 2014; Brocas and Carrillo, 2014). The amount of time that an individual takes to make a decision, better known as Response Time (RT), is one of the simplest measures to understand the process that led to that choice (Fehr and Rangel, 2011; Caplin and Dean, 2015), although it was recently noticed that response time might be affected by the heterogeneity of subjects’ types and by different contexts (Krajbich et al., 2015; Evans et al., 2015; Chen and Fischbacher, 2020). One way to mitigate this problem is to exogenously impose a constraint on the response time. Given that subjects under time constraint might rely more on intuition and that subjects who must wait before make a choice might rely more on deliberative process; place them under time constraint or time delay is commonly used to understanding the impact on choices (Rand et al., 2012, 2014; Rand, 2019).

While the ongoing debate has been focusing mostly on cooperative behavior (Rand, 2016, 2017; Bouwmeester et al., 2017; Kvarven et al., 2020; Alós-Ferrer and Garagnani, 2020) and, to a lesser extent, on the disposition to donate (Achtziger et al., 2015; Rand et al., 2016; Merkel and Lohse, 2019; Bago et al., 2020; Fromell et al., 2020),¹ only one paper (Belloc et al., 2019) has explored the

¹See Hallsson et al. (2018) for survey focusing behaviors in the Ultimatum Game.

effect of little reflection on collaborative behavior. Belloc et al. (2019) study the effect of inducing less reflection on collaborative behavior in a series of Stag-Hunt games. In the Stag-Hunt game an individual has to choose between a more efficient but risky action, i.e., collaborating to hunt a *Stag*, and a safer action with a lower maximum reward, i.e., going alone to hunt a *Hare* (Skyrms, 2004). Differently from the Prisoner's Dilemma game – where each player can incur a personal cost to generate a larger benefit to the other – the Stag-Hunt is a coordination game where an individual faces the trade-off between risky collaboration that can provide the largest reward and safer non-collaborative behavior which provides a smaller reward with certainty. The trade-off between efficiency and safety has been investigated in different ways. Schmidt et al. (2003) find that changes in risk dominance significantly affect subjects' behaviors, whereas changes in the level of payoff dominance do not. Capraro et al. (2020) have explored the motives for collaborating in the Stag-Hunt game finding that it is primarily driven by preferences for efficiency, rather than moral preferences. However, to the best of my knowledge, only Belloc et al. (2019) have attempted to see how manipulating cognition can affect collaboration, and they did so in a laboratory experiment where intuitive decisions were prompted by imposing a time constraint of 10 seconds (the so-called time pressure condition, see Spiliopoulos and Ortmann, 2018) to pick an action in the Stag-Hunt game. Their experimental data show that the probability to be more collaborative is higher for participants in the time pressure treatment.

I contribute to this latter line of research with a pre-registered online experiment where, contrary to Belloc et al. (2019), I focus on the role of deliberation applying two distinct methods aimed at inducing greater reflection. This is potentially important because there are some criticisms on the use of time constraints. First, time constraint must be set in the decision screen and this imply that subjects have all the time to think about the decision in the instruction screen. Secondly, the compliance with the time pressure condition is not always possible. If participants do not comply to respond within time, the researcher can allow to respond after the limit or not. In the latter case there will be a selection bias, while if they are allowed to submit their choice also after the constraint, then those subjects might be different from the others in some unobservable variable. Moreover, is also interesting to see if and to what extent previous findings obtained in the laboratory by (Belloc et al., 2019) carry on to the online setting, also if not directly comparable.

In the experiment I attempt to prompt individuals to reflect more on their decision-making. In particular, I rely on two methods to induce greater reflection in participants. The first one is a *time delay* treatment, which is the opposite of a time pressure treatment: participants are required to

wait at least 40 seconds before they can pick an action. The second method is a *motivated delay* treatment (Bilancini et al., 2017): participants are required to wait at least 40 seconds (as in the time delay treatment) and to write down a motivation for their decision before they can pick an action². Recent evidence suggests that the motivated delay treatment is effective in inducing greater reflection (Bilancini et al., 2019, 2020, 2021) although it is still to be established how it compares to time delay and if asking for a motivation has additional and qualitatively different effects. In order to keep the two treatments comparable I require participants to wait the same number of seconds. So, a further contribution of this work is to provide new insights on these two methodological approaches and how they might promote greater reflection in an online experimental setting. To be in line with Belloc et al. (2019), I also elicited risk aversion and trust. These two measures are directly related with the Stag-Hunt game: on one side, choosing *Stag*, is a more risky and trusting while, on the other hand, choosing *Hare* is safer.

Finally, according with the existing literature (psychology, neuroscience, economics and management), moods and emotions may systematically affect individual's behavior. A lot of works recognize the important links between emotion and decision-making (Loewenstein and Lerner, 2003; Rick and Loewenstein, 2008) and emotion and social interaction (Heilman et al., 2010). Moreover, the recent experimental literature, also investigate the induction of positive or negative mood in the one-shot economics interaction (Kirchsteiger et al., 2006; Capra, 2004). They find that a more positive mood induces a more altruistic and trusting behaviors. To fully exploit the *motivated delay* treatment, I perform a text analysis aimed at compare the characteristics of the participants' written texts. The purpose is to use the sentiment analysis to study the emotional components present in the written motivations and see how them are correlated with participants' choices. The firsts to do so with an economic game were Proto et al. (2019). They analyze a pre-play communication in a repeated Prisoners' Dilemma. To the best of my knowledge, the present study is the first to use this technique in a one-shot game without inducing a particular mood before the decision.

The main results can be summarized as follow. I find that participants in the *motivated delay* treatment are less likely to collaborate (i.e., choose *Stag*) than those who are asked only to wait 40 seconds, as in the *time delay* treatment, or those who have no time constraints to pick a choice, as in the *baseline*. Moreover, I find that encouraging deliberation affects the focus of participants on the payoffs of the game when they have to make a decision. Finally, the findings on the text analysis suggest that participants who decide to collaborate (i.e., those who choose the *Stag* option) are in a

²Participants can start typing during the 40 seconds and do not have any time constraint.

more positive mood with respect to those who do not collaborate (i.e., those who choose *Hare*).

Section 2.2 describes the experimental design and the procedures used in the experiment. Section 2.3 shows the main results while Section 2.4 reports the text analysis on the written motivation. Finally, in Section 2.5 I discuss about the experimental evidences and the implications in the recent literature.

2.2 Experimental Design

I conducted a pre-registered³ online experiment on the platform Prolific (Palan and Schitter, 2018). The experimental interface was entirely developed using Qualtrics. I recruited 902 subjects based in the US, aged between 18 and 40 years, who completed at least two studies on Prolific and with an approval rate of at least 50%. At the beginning of the experiment participants received detailed instructions about the tasks and the payments and they provided their informed consent.

The experimental design is mixed, consisting of a between-subject part concerning the manipulation to encourage deliberation and a within-subject part concerning the variation of game payoffs in the Stag-Hunt game.

To foster deliberation, participants were randomly assigned to three different treatment groups. In the *baseline*, participants played a series of four different Stag-Hunt games without any further requirement or constraint. In the *time delay* treatment, participants were required to wait at least 40 seconds before they could pick an action in each Stag-Hunt game; in this time period, the game payoffs (associated with the actions) remained displayed on the screen. In the *motivated delay* treatment, participants were required to wait at least 40 seconds (as in the time delay treatment) and to write down a motivation for their choice (of at least 40 characters), before they can pick an action. The basic difference between the two treatments is that, in the latter, during the 40 seconds and even after, participants have to write a motivation for their choice, with any time constraint.

To explore the role of payoffs, I asked participants in each treatment group to play four Stag-Hunt games, which differ between them for the size of the basin of attraction of the action *Stag*. All games' payoffs are displayed in Table 2.2.1. Each number in the payoff matrix corresponds to the Points gained by the row player in the four games, depending on the choices of both players. Following Belloc et al. (2019), I use the same four (symmetric) games with variable size of the basin of attraction of the action *Stag*. The basin of attraction of *Stag* is defined as 1 minus the minimum probability of the other player playing *Stag* that is sufficient to make playing *Stag* a best reply, i.e.,

³The experiment was pre-registered on AsPredicted.org. The pre-registration is available at the following link.

yielding an expected payoff which is higher than or equal to the one given by playing *Hare*. In other words, the basin of attraction of *Stag* is a measure of how little must be the minimum probability that one player plays *Stag* in order to induce the other player to play *Stag* too. The basin of attraction of Game 1 is equal to $\frac{1}{4}$, i.e., *Stag* is best reply if and only if the probability of the opponent playing *Stag* lies in the interval $[\frac{3}{4}, 1]$. Game 2 has the largest basin of attraction of *Stag*, i.e., $\frac{3}{8}$, Game 3 has the smallest one, i.e., $\frac{1}{8}$, while Game 4 is a transformation of Game 1 (by adding 1 to every payoff entry) and so it has the same basin of attraction. Game 1 was always played first, while the order of the remaining three Games was randomized across players and, hence, across treatments too. This allows to restrict the study of the treatment effects to just Game 1 (with a pure between-subject design). No feedback information was provided on any aspect of the game outcomes.

The choice pages for each treatment are reported in Appendix 2.A. The only differences between the baseline and the treatments are the following: in the *time delay* treatment I remember to the participants that they have to wait 40 second before making a decision while, in the *motivated delay* treatment, they find a “text box” where is asked to write down a motivation for their choice (40 characters at least). In both cases the choices’ buttons do not appear before 40 seconds. It is important to notice some points. The words *Stag* and *Hare* are never used during the experiment and I opted to call them *Option A* and *Option B* to be more neutral. Moreover, although the Options order in the instruction screen is fixed (in the first rows is displayed *Option A* and in the second rows *Option B*), the choice buttons are showed horizontally and in random order.

Table 2.2.1: The four Stag-Hunt games. Reported payoffs refer to the row player.

	Game 1		Game 2	
	<i>Stag</i>	<i>Hare</i>	<i>Stag</i>	<i>Hare</i>
<i>Stag</i>	4	0	4	0
<i>Hare</i>	3	3	2.5	2.5
	Game 3		Game 4	
	<i>Stag</i>	<i>Hare</i>	<i>Stag</i>	<i>Hare</i>
<i>Stag</i>	4	0	5	1
<i>Hare</i>	3.5	3.5	4	4

After each Stag-Hunt game, participants’ belief about the behavior of other participants are elicited (Manski, 2004). To this aim participants were asked to guess the percentage of people playing *Stag* in that game. Participants earned 1 Point if their guess was within 5% of the actual average behavior in that game for their treatment group.

After the four games were played and the beliefs elicited, participants were asked to do the “Bomb

Risk Elicitation Task” (Crosetto and Filippin, 2013) aimed at measuring their risk taking attitude⁴. This task was incentivized by giving to participants 0.01 Point for each box opened, provided the bomb was not found. Subsequently, participants were asked their trust in other people using the item proposed by Falk et al. (2018). It corresponds to a self-assessment question, where is asked whether participants assume that other people only have best intentions (Likert scale, 0-10). The choice to ask individuals the trust question after the games was driven by two reasons: (i) asking them their trust in other people before the actual decisions may have influenced the main variable of interest; (ii) given that no feedback information is given after each game, trust might not be influenced by the results of the game, although we cannot exclude that reported trust is affected by the reasoning during the previous decision-making process. Lastly, subjects were asked two comprehension questions about the qualitative characteristics of the actions involved in the games: (i) “Which option is more collaborative?” and (ii) “Which option is the safer one?”. This was done to understand which representation of the game participants had.

At the end of the experimental session, pairs of participants were formed randomly and all payoffs computed for the Stag-Hunt games and the other tasks. At the beginning of the experiment, participants were informed that a random lottery incentive protocol would be used for payments regarding games and guesses; namely, one game would be drawn at random for each pair and those two participants would be paid according to the payoffs earned in such game.⁵ Points earned in all tasks were converted into money at the rate of 0.4 GBP for 10 Points earned.

Finally, in order to control that the treatment samples are well-balanced, I collected some individual characteristics as control variables (age, gender, student status, employment status). Full experimental instructions can be found in Appendix 2.A.

2.3 Results

I recruited 902 participants, of which 298 were randomly assigned to the *baseline* treatment, 302 to the *time delay* treatment and 302 to the *motivated delay* treatment. The average earning was 0.56 GBP (including the show up fee of 0.30 GBP) and the average length of the experiment was about

⁴As opposed to Belloc et al. (2019) I decided to use an incentivized measure of risk elicitation. While they use a series of not incentivized questions, I opted to choose the Bomb Risk Elicitation Task, which is an incentivized measure of risk aversion

⁵I opted to pay both the game outcome and the belief elicitation task although this increases the room for potential biases due to hedging motives. I do not believe this is a problem in the setup since the hedging problem is not transparent for subjects (it is hard to say what guess should help in hedging risk) and the gains from hedging are very little (about 1/4 of the game payoff) (see Blanco et al., 2010, for a detailed discussion of the issue).

7 minutes. The average reward per hour was 5.86 GBP, which is above the minimum required by Prolific (i.e., 5.00 GBP).

Random assignment to treatment groups produced balanced samples. Table 2.3.1 reports descriptive statistics for the participants across the three treatment groups, comparing groups pairwise. Mann-Whitney tests confirm good balance across treatment groups: all participants' characteristics (gender, age, student status, risk preferences, self assessment of trusting behavior) are similarly distributed across treatment groups and there are no statistically significant differences in the pairwise comparisons of means.

Table 2.3.1: Descriptive Statistics

Variable	Mean			MW test, p -values diff=0		
	B	TD	MD	B-TD	B-MD	TD-MD
Female	0.49	0.52	0.55	0.5423	0.1909	0.4810
Age	27.44	27.54	26.96	0.9271	0.3044	0.3263
Student Status	0.38	0.33	0.37	0.2755	0.9234	0.3186
Risk Preferences	36.18	35.45	37.96	0.9540	0.3618	0.3780
Trust	5.74	5.54	5.81	0.2322	0.7662	0.1382
Previous Experience	0.39	0.41	0.37	0.2830	0.5123	0.6811
N	298	302	302			

Note: B = Baseline; TD = Time Delay and MD = Motivated Delay. Female=1 if the individual is a female; Age is the individual's age; Student Status=1 if the individual is a student at the moment of the experiment; Risk Preferences is the average number of "boxes" opened in the BRET task; Trust assume values from 0 to 10, where 10 is whether the individual assumes that other people only have the best intentions; Previous Experience=1 if individuals have already had experience the Stag-Hunt game.

The *time delay* and *motivated delay* treatments were aimed at encouraging participants to spend more time reflecting on their decisions. A first test of the effectiveness of these treatments can be done by looking at their impact on response times. Figure 2.3.1 shows, on the left chart, the effect of treatments on response times in the Stag-Hunt games. On average, participants in the *motivated delay* treatment spent 60 seconds more in picking an action than those in the *baseline* treatment, and 22 seconds more than those in the *time delay* treatment. In particular, the average response time in the *baseline* treatment was 16.36 seconds, in the *time delay* treatment was 55.70 seconds, and in the *motivated delay* treatment was 77.74 seconds. The Kruskal-Wallis test rejects the hypothesis that response times in the *baseline*, the *time delay* and the *motivated delay* treatments are equally distributed ($\chi^2_{(2)} = 3206.176$, $p < 0.001$). Mann-Whitney tests return that this also applies to a pairwise comparisons of the means (B-TD: $z = -48.980$, $p < 0.001$; B-MD: $z = -48.980$, $p < 0.001$; TD-MD: $z = -49.143$, $p = 0.006$). Similar results hold when I test for response times for each game separately. It is interesting to notice that subjects in the *time delay* treatment spent 16 seconds

more than the 40 required. This is similar to the time spent by participants in the *baseline* to decide their action. It may suggest that individuals zoned out for 40 seconds and then engaged once they can effectively choose an option. Although, the analysis performed in Table 2.3.2 below, show significant differences in the choice of *Stag* between the *baseline* and the *time delay* treatment when it is interacted with the basin of attraction, suggesting that participants put more attention on the payoffs during that time.

The right chart of Figure 2.3.1 shows the percentage of times it is chosen *Stag* by treatment group. In the *baseline* treatment, the percentage times of playing *Stag* is 70.30%, against 71.93% in the *time delay* treatment. The Fisher's exact test cannot reject the null hypothesis of equal percentages ($p=0.201$). The percentage times of playing *Stag* in the *motivated delay* treatment is 66.72%, lower than in the other two treatment groups. Fisher's exact test confirms that this percentage is statistically different from the one obtained in the *baseline* treatment ($p = 0.065$) and the one obtained in the *time delay* treatment ($p < 0.01$).

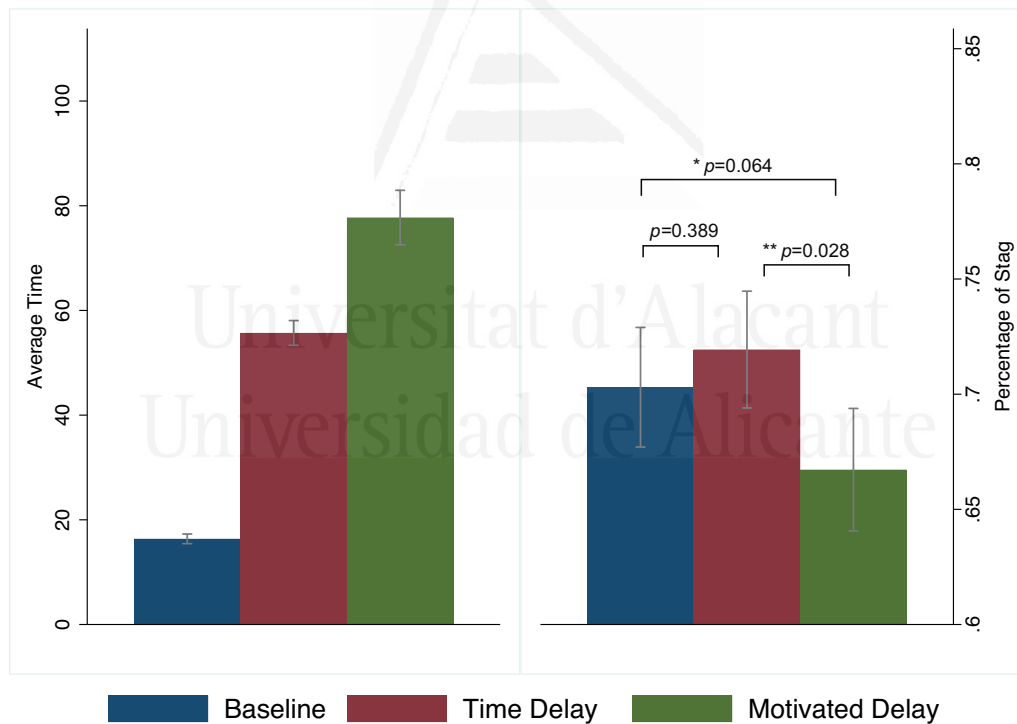


Figure 2.3.1: **Average time in seconds for decision (left chart) and average play of *Stag* (right chart) in the baseline treatment, the time delay and the motivated delay treatments.** Behaviors differ significantly across treatments (Chi-squared test, $\chi^2_{(2)} = 8.12$, $p=0.017$). While the percentage of *Stag* choices in the *baseline* and *time delay* treatments are not significantly different (Fisher's exact test, $p=0.201$), they are different in the *time delay* and *motivated delay* treatments (Fisher's exact test, $p=0.065$) and in the *time delay* and *motivated delay* treatments (Fisher's exact test, $p = 0.006$).

To exploit the entire potential of the mixed (between-within) design I ran a number of logit regressions where I could jointly study the impact of the treatments (between-subject variation), the basin of attraction (within-subject variation) and their interaction (further details are reported in Appendix 2.B.1) on the likelihood to choose *Stag*. Three different models were estimated, as shown in Table 2.3.2. In all cases I evaluated statistical significance with standard errors (in brackets) clustered at individual level.

In Model (1) the choice of *Stag* is regressed on a dummy variable for the *time delay* treatment, a dummy variable for the *motivated delay* treatment – the omitted category being the *baseline* treatment – and the variable *basin* – being equal to the size of the basin of attraction of *Stag* in the game considered normalized at the minimum level (i.e., $basin = 1/8$)– which captures the impact of payoff variations. The estimated treatment effect is positive for the *time delay* treatment and negative for the *motivated delay* treatment, but neither is statistically significant. The estimated coefficient of *basin* is positive, large and statistically significant, suggesting that the payoff structure had a primary role in shaping decisions: since $\exp(2.286) = 9.835$, I can say that for an increase of 0.01 in the basin of attraction, I expect to see the 8.83% increase in the odds of choosing *Stag*, keeping all the other variables at a fixed value. As one would expect, a greater basin of attraction of *Stag* implies a greater likelihood to play *Stag*.

In Model (2) I add the interaction between treatment variables and *basin*, normalized at the minimum level in the experiment. Estimated treatment effects turn out to be in line with the results of the non-parametric analysis reported in Figure 2.3.1. In particular, when the basin of attraction is at its lower level, the estimated effect of the *motivated delay* treatment is negative and statistically significant, while the estimated effect of the *time delay* treatment is positive but not statistically significant. The estimated coefficient of *basin* is positive and statistically significant. In this case, since $\exp(1.288) = 3.62$, an increase of 0.01 in the basin of attraction yields an increase of 2.62% in the odds of choosing *Stag* in the *baseline*. The estimated coefficients of the interaction terms are positive, statistically significant and similar in magnitude, suggesting an increase of more than 3% in the odds of choosing *Stag* in both *time delay* and *motivated delay* treatment when there is an increase of 0.01 in *basin* ($\exp(1.562) = 4.76$ and $\exp(1.42) = 4.14$, respectively). These estimates suggest that the payoff structure plays an important role in all treatments, but it does much more so, and in a similar way, in the *time delay* and *motivated delay* treatments. On top of this, the *motivated delay* treatment seems to have a negative effect on the likelihood to choose *Stag*.

In Model (3) I add a number of controls: risk aversion as measured by BRET (Crosetto and

Table 2.3.2: Logit regression

Variable	Model (1)	Model (2)	Model (3)
Time Delay	0.080 (0.141)	-0.104 (0.168)	-0.077 (0.170)
Motivated Delay	-0.167 (0.141)	-0.336** (0.162)	-0.344** (0.164)
Basin	2.286*** (0.297)	1.288** (0.531)	1.392*** (0.541)
Time Delay \times Basin		1.562** (0.771)	1.532* (0.790)
Motivated Delay \times Basin		1.421** (0.703)	1.336* (0.718)
BRET			-0.001 (0.002)
Trust			0.149*** (0.027)
Cons	0.584** (0.104)	0.703*** (0.120)	-0.304 (0.428)
Individual Characteristics	No	No	Yes
No. of Obs.	3608	3608	3584
N	902	902	896
χ^2	62.86***	71.17***	96.81***
Pseudo R^2	0.0088	0.0094	0.0283

Note: The dependent variable is binary and equal to 1 if the choice is *Stag*, 0 otherwise. The error terms are clustered at individual level and reported in parenthesis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Filippin, 2013), reported trust in others (Falk et al., 2018), and other individual characteristics (See Table 2.B.1 in Appendix 2.B.1 for a reported regression with all controls). Estimates confirm the findings of Model (2). Moreover, the estimated coefficients of BRET and reported trust are of some interest. The estimated coefficient of BRET is almost zero suggesting that, contrary to what found by Belloc et al. (2019), risk-taking behavior does not go with a greater likelihood to choose *Stag*. The estimated coefficient of reported trust is instead positive and statistically significant, again in contrast to what found by Belloc et al. (2019). One possible explanation for this difference is the fact that I adopted different measures for both risk taking and trust and the sample composition is different (i.e., older participants, less students). Another possibility is that the online setting reduces the role of risk taking and increases that of trust.

I also studied participants' beliefs about the behaviour of other participants in the Stag-Hunt games, with the aim of better understanding the channel through which treatments affect behavior. I found a positive and highly significant correlation between the one's belief that others play *Stag*

and one's own choice of *Stag* ($r(902) = 0.615, p < 0.001$), suggesting consistency between individuals choices and beliefs. Given such strong correlation, it seems natural to see if the treatments affect beliefs in the same way in which they affect the choice of *Stag*. Figure 2.3.2 shows how beliefs vary across treatments. Treatments seem to have an effect which is qualitatively similar to that on behavior, as shown in the right panel of Figure 2.3.1. In particular, participants in the *motivated delay* treatment are less likely to believe that the other players will choose *Stag* with respect to participants assigned to the *baseline* and *time delay* treatments. According with this result seems that individuals realize that when other participants have to motivate their choices they will cooperate less and, consequently, they behave consistently. This is consistent with the fact that treatments affect behavior by shaping beliefs about others' behavior.

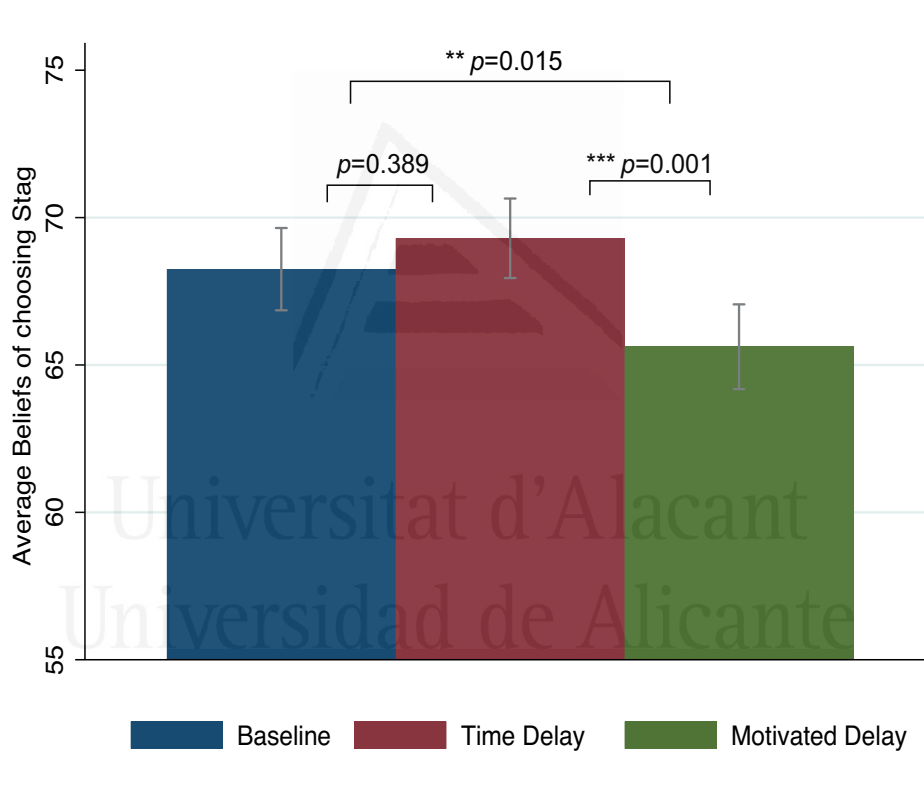


Figure 2.3.2: **Average beliefs about the behavior of other participants in the baseline, the time delay and the motivated delay conditions.** Beliefs about others' behavior in the Stag-Hunt games differ significantly across treatments (Kruskal-Wallis test, $\chi^2_{(2)} = 11.71, p < 0.01$). While beliefs in the *baseline* and *time delay* treatments are not significantly different (Mann-Whitney test, $z = -0.861, p = .389$), they are in the *time delay* and *motivated delay* treatments (Mann-Whitney test, $z = 2.426, p = .0153$) and in the *time delay* and *motivated delay* treatments (Mann-Whitney test, $z = 3.307, p < 0.01$).

2.4 Text Analysis

In the *motivated delay* treatment participants were asked to write down a motivation for their decisions in each of the four Stag-Hunt games. In particular, they had to write a motivation (of at least 40 characters) after looking at the payoffs of a game and before actually selecting an action. Besides aiming at inducing greater reflection, this treatment returns written texts that allow to perform an interesting quantitative text analysis in one-shot game⁶.

More specifically, following the recent experimental literature of emotions which highlights the effect of emotion on social preferences (Capra, 2004; Kirchsteiger et al., 2006), I try to compare the characteristics of the texts written by the participants who chose to collaborate (*Stag*) and those written by the participants who chose the safer option (*Hare*). To this purpose I use the Matlab Text Analytic ToolboxTM, and in particular the VADER algorithm, to carry out sentiment analysis on the written motivations.⁷ Sentiment analysis aims at detecting the polarity (negative or positive) of emotions within a written text. To do so VADER relies on a dictionary that maps lexical features to emotion intensities, known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text. The VADER dictionary uses a lexicon with words annotated with a score from -1 to 1, where -1 indicates words with strong negative sentiment, 1 indicates words with strong positive sentiment and scores close to 0 indicate neutral sentiment. For example, words like “love”, “enjoy”, and “happy” are representative of a positive sentiment, while words like “hate”, “ugly”, and “pain” are representative of a negative sentiment. Moreover, VADER is trained enough to identify the basic context of these words, such that it recognizes “did not love” as a negative statement.

In Figure 2.4.1 I report the sentiment scores relative to the motivations written by those who chose *Stag* and those who chose *Hare*. In the left panel, which reports the sentiment scores over all four games, I can observe that the text written by participants choosing *Stag* appears to have, on average, a more positive sentiment score than the text written by participants who have chosen *Hare*. This suggests that participants who chose *Stag* were in a more positive mood with respect to those who chose *Hare*. In the right panel of Figure 2.4.1, which reports the sentiment scores for each Stag-Hunt game separately, I can observe a similar pattern. Looking specifically at the differences between

⁶To the best of my knowledge, only one paper conducted a text analysis using a repeated Prisoners’ Dilemma.

⁷The toolbox provides algorithms and visualizations for analyzing and modeling text data. Currently, many tools, such as Linguistic Inquiry and Word Count (Tausczik and Pennebaker, 2010), are used to extract advanced features from texts. The Matlab Text Analytics ToolboxTM uses the specific tool called VADER (Gilbert and Hutto, 2014). It was developed by Gilbert and compared with 11 typical state-of-the-practice benchmarks, including Linguistic Inquiry and Word Count (LIWC), Senti WordNet, Affective Norms for English Words (ANEW), the General Inquirer, and machine learning-oriented techniques.

the sentiment scores in each game, I can observe that the largest difference is obtained in Game 2, where the basin of attraction of *Stag* is the largest, while the smallest difference is obtained in Game 3, where the basin of attraction of *Stag* is the smallest. In short, the difference in sentiment scores grows in the relative advantage, in term of risk-dominance, of choosing *Stag*. This may suggest that, while collaboration always go with more positive sentiments, the extent of its riskiness – with respect to the safer action – tends to re-balance sentiments. This is conforming with Kim and Kanfer (2009), which show that when individuals have to made a risk judgement for a more cognitively demanding task, their mood states are attenuated. Moreover, Drichoutis and Nayga Jr (2013) find that when participants are in a more negative mood, they increase their risk aversion. This is in line with our results, where a moer negative mood is related with the choice of Hare. Another possible explanation is that the difference in sentiment between *Stag* and *Hare* positively depends on the difference in payoffs between the two equilibria. The experimental data cannot help to decide between these two explanations, since a larger basin of attraction of Stag goes with a larger payoff difference between equilibria.

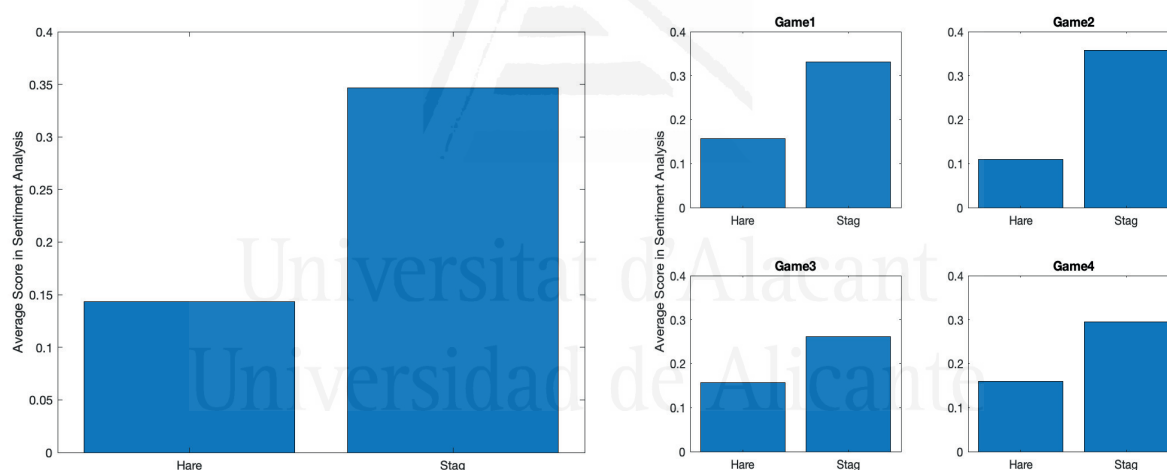


Figure 2.4.1: **Emotions in the Text Analysis of the written motivations in the motivated delay treatment.** The left-side chart reports the average sentiment score, which takes values from -1 (negative sentiment) to 1 (positive sentiment), for those participants who have chosen *Hare* and those who have chosen *Stag*. The right-side chart reports same results divided by Games.

Finally, it is interesting to notice that the words most frequently used were consistent with the actual decisions. As shown in Figure 2.B.2 in Appendix 2.B.3, the word “A”, which is referred to the *Stag* choice, appears more frequently in the motivations written by those who have chosen *Stag* while the word “B”, which is associated to the *Hare* choice, appears more frequently in the motivations written by those who have chosen *Hare*. Moreover, words like “guarantee” and “risk” appears only in

the written motivation by those who decided to choose *Hare*, suggesting that they correctly recognized the greater safety of such action.

2.5 Discussion

Many recent studies have explored if and how the manipulation of cognition can affect behavior in social dilemmas. Belloc et al. (2019) explore how behavior is affected in the one-shot Stag-Hunt game played in the lab, showing that participants are more likely to collaborate (i.e., to choose the action *Stag*) when they are forced to decide within 10 seconds (i.e., when they are put under *time pressure*). This suggests that collaborative behavior may be more likely when participants deliberate less, at least when collaboration is risky as opposed to safe non-collaboration – as it is the case in the Stag-Hunt game.

I contribute to this line of research with an online experiment where: first, I can study the role of deliberation in greater detail since I apply two distinct methods, both aimed at inducing greater reflection, secondly, I can see if and to what extent the insights of Belloc et al. (2019) carry on to the online setting.

I find that asking participants to wait 40 seconds and write a motivation for their decision before selecting an action (i.e., the *motivated delay* treatment in the experiment) makes participants less likely to collaborate than just asking them to wait 40 seconds (i.e., the *time delay* treatment) or letting them choose without constraints (i.e., the *baseline* treatment). Specifically, the size of the treatment effect is about 5 percent points. Moreover, I find that asking to wait 40 seconds before selecting an action has a sizeable effect on the relevance of the payoffs for actual decisions: the effect of the expected gains from collaborating on the decision to collaborate (summarized by the basin of attraction of the action *Stag*) is about twice larger. This suggests that greater deliberation leads to give more weight to the payoffs.

Findings are consistent with those of Belloc et al. (2019), overall suggesting similar conclusions. However, the different experimental setting requires some further qualifications regarding the results obtained which, I think, can provide additional insights. While Belloc et al. (2019) found that collaboration is more likely when participants have to decide in 10 seconds, I find that: (i) the basin of attraction of action *Stag* is more likely to affect the decision when I ask to wait at least 40 seconds before selecting an action, and (ii) collaboration is less likely to occur when I also ask to write down a motivation for the decision. Since both the *time delay* treatment and *motivated delay* treatment are designed to promote deliberation, findings strongly suggest that the payoffs acquire greater relevance

in more deliberative decisions, with no appreciable difference between the two treatments. So, in this regard writing a motivation *per se* seems to have no effect. However, the *motivated delay* treatment does seem to have an additional effect in reducing collaboration, and this asks for an explanation. While it is not possible to use the experimental data to give a certain answer, it seems fair to say that there are at least two possibilities to be considered: while waiting 40 seconds have not affected deliberation, but individuals have zoned out during that time, the additional effect observed in the *motivated delay* treatment may be done by the fact that it is more effective in promoting deliberation, or it could be due to the fact that it affects how participants deliberate.

In order to see if these results have to do with the comprehension of the qualitative feature of the payoff structure of a Stag-Hunt game (not specifically the size of the basin of attraction of the action *Stag*), at the end of the experiment I asked the following two simple questions: “Which option is more collaborative?” and “Which option is the safer one?”. As shown in Figure 2.5.1 most participants (about 85%) answered that choosing *Stag* is more collaborative, instead the remaining 15% said that it is playing *Hare*. The latter may have thought that payoffs are less different under that choice and thus it is more collaborative. In this case no substantial difference is found across treatments. On the contrary, for the second question responses show a greater rate of participants in the *motivated delay* treatment (about 73%) answering that *Hare* is the safer option, with respect to both the *time delay* treatment (about 60%) and the *baseline* treatment (about 67%). This suggests that prompting participants to write down a motivation for their decision may improve the understanding of what is the safe action (more details can be found in Appendix 2.B.3, Table 2.B.3). It is also important to notice that, the lower percentage of participants in the *time delay* treatment respect to the *baseline*, although not statistically significant, may be found because they were less focus on the task.

The *motivated delay* treatment has been recently proposed by Bilancini et al. (2017) as an experimental condition aimed at promoting deliberation in an effective way without imposing longer waiting times with respect to the more standard *time delay* treatment. While it has still to be established if, and to what extent, the *motivated delay* treatment is more effective in promoting deliberation, the richness of the information embedded in the written motivations gives us the opportunity to take advantage of the tools that have been developed for the quantitative analysis of written texts (see, e.g., Proto et al., 2019). I carried out such an investigation by applying the so-called sentiment analysis which allowed us to impute “negative” as opposed to “positive” sentiment scores to the written motivations. According to the results the participants who chose *Stag* were more likely to write a motivation classified as having a positive sentiment compared to those who chose *Hare*. Moreover, I

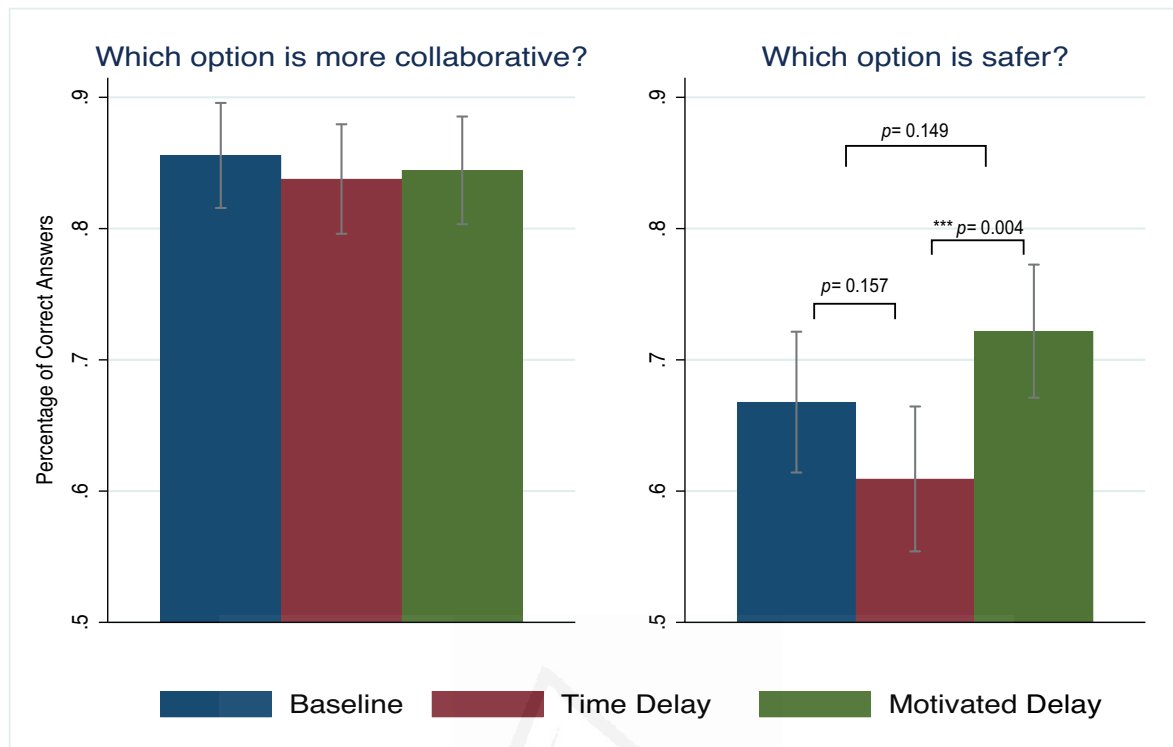


Figure 2.5.1: **Percentage of correct answers to the comprehension questions in the baseline, the time delay and the motivated delay conditions.** On the right chart, participants are asked which option is more collaborative. Most of them (85%) answered correctly without any difference between treatment. On the left chart, participants are asked which option is safer. The percentage of correct answers differ significantly across treatments (Chi-squared test, $\chi^2 = 8.61$, $p = 0.013$). While the percentage of correct answers is significantly higher in the *motivated delay* and the *time delay* treatments (Fisher’s exact test, $p = 0.004$), there are no significant differences in the *baseline* and *time delay* treatments (Fisher’s exact test, $p = 0.157$) and in the *baseline* and the *motivated delay* treatments (Fisher’s exact test, $p = 0.149$).

could see that the words most frequently used were consistent with the actual decisions⁸. For example, the word “A”⁹ appears more frequently in the motivations written by those who have chosen *Stag* while the word “B” appears more frequently in the motivations written by those who have chosen *Hare*. Also, the words “guarantee” and “risk” are more frequently written by those participants who deliberated in favor of the *Hare* choice, suggesting that they correctly recognized the greater safety of such action.

A few comments regarding the choice of a mixed design are worth doing. Mixing between-subject and within-subject has both advantages and disadvantages. The between-subject design encourage

⁸See Figure 2.B.2 in Appendix 2.B.3 for the word frequencies by choice of action and Figure 2.B.3 for the frequencies by the games played.

⁹In the experiment, to avoid the nouns *Stag* and *Hare*, I chose to write down the two alternatives in a different and neutral way: *Option A* for the *Stag* choice and *Option B* for the *Hare* choice.

deliberation was a natural choice because I did not want to mix the effects of manipulations over time, which is a dimension that is rather hard to control, especially online. While I could randomize treatments in order to minimize the issue in a within-subject design, I could not see any substantial advantage for adopting it for such treatments. Instead, for the analysis of the payoffs I did opt for a within-subject design, and this is a less straightforward choice. Indeed, I could have adopted a between-subject design also in this case but, as a result, I would have been forced to increase the number of participants substantially as I were interested in assessing also interaction effects. The drawback is that some learning effect could be at work, biasing the results. In short, the choice of a within-subject design is the result of a greater relevance given to power with respect to potential biases due to learning. At any rate, in order to see if the learning effect is sizeable I did not randomize the position of Game 1, having it played first by every participant, so that I could run the analysis of the treatment effects only on this game. Such restricted analysis confirms what shown in Figure 2.3.1: asking participants to write a motivation for their choice makes them less likely to collaborate than just asking to wait time or letting them choose without constraints (see Figure 2.B.4 in the Appendix 2.B.3).

Bibliography

- ACHTZIGER, A. AND ALÓS-FERRER, C. (2014): “Fast or rational? A response-times study of Bayesian updating,” *Management Science*, 60, 923–938.
- ACHTZIGER, A., ALÓS-FERRER, C. AND WAGNER, A. K. (2015): “Money, depletion, and prosociality in the dictator game,” *Journal of Neuroscience, Psychology, and Economics*, 8, 1.
- ALÓS-FERRER, C. AND GARAGNANI, M. (2020): “The cognitive foundations of cooperation,” *Journal of Economic Behavior & Organization*, 175, 71–85.
- ALÓS-FERRER, C. AND STRACK, F. (2014): “From dual processes to multiple selves: Implications for economic behavior,” *Journal of Economic Psychology*, 41, 1–11.
- BAGO, B., BONNEFON, J.-F. AND DE NEYS, W. (2020): “Intuition rather than deliberation determines selfish and prosocial choices,” *Journal of Experimental Psychology: General*.
- BELLOC, M., BILANCINI, E., BONCINELLI, L. ET AL. (2019): “intuition and Deliberation in the Stag Hunt Game,” *Scientific Reports*, 9, 1–7.
- BILANCINI, E., BONCINELLI, L., CAPRARO, V. ET AL. (2020): “The effect of time pressure and motivated delay on cooperation and social norms in the online one-shot public goods game,” *Working Papers*.
- BILANCINI, E., BONCINELLI, L. AND CELADIN, T. (2021): “Social Value Orientation and Conditional Cooperation in the Online One-Shot Public Goods Game,” *Working Papers*.
- BILANCINI, E., BONCINELLI, L., LUINI, L. ET AL. (2017): “Does focality depend on the mode of cognition? Experimental evidence on pure coordination games,” Tech. rep.
- BILANCINI, E., BONCINELLI, L. AND SPADONI, L. (2019): “Motivating Risky Choices Increases Risk Taking,” *Working Papers*.
- BLANCO, M., ENGELMANN, D., KOCH, A. K. ET AL. (2010): “Belief elicitation in experiments: is there a hedging problem?” *Experimental Economics*, 13, 412–438.
- BOUWMEESTER, S., VERKOEIJEN, P. P., ACZEL, B. ET AL. (2017): “Registered replication report: Rand, greene, and nowak (2012),” *Perspectives on Psychological Science*, 12, 527–542.

- BROCAS, I. AND CARRILLO, J. D. (2014): “Dual-process theories of decision-making: A selective survey,” *Journal of economic psychology*, 41, 45–54.
- CAPLIN, A. AND DEAN, M. (2015): “Revealed preference, rational inattention, and costly information acquisition,” *American Economic Review*, 105, 2183–2203.
- CAPRA, M. C. (2004): “Mood-driven behavior in strategic interactions,” *American Economic Review*, 94, 367–372.
- CAPRARO, V. (2019): “The dual-process approach to human sociality: A review,” *Available at SSRN 3409146*.
- CAPRARO, V., RODRIGUEZ-LARA, I. AND RUIZ-MARTOS, M. J. (2020): “Preferences for efficiency, rather than preferences for morality, drive cooperation in the one-shot Stag-Hunt Game,” *Journal of Behavioral and Experimental Economics*, 101535.
- CHEN, F. AND FISCHBACHER, U. (2020): “Cognitive processes underlying distributional preferences: a response time study,” *Experimental Economics*, 23, 421–446.
- CROSETTO, P. AND FILIPPIN, A. (2013): “The “bomb” risk elicitation task,” *Journal of Risk and Uncertainty*, 47, 31–65.
- DRICHOUTIS, A. C. AND NAYGA JR, R. M. (2013): “Eliciting risk and time preferences under induced mood states,” *The Journal of Socio-Economics*, 45, 18–27.
- EVANS, A. M., DILLON, K. D. AND RAND, D. G. (2015): “Fast but not intuitive, slow but not reflective: Decision conflict drives reaction times in social dilemmas.” *Journal of Experimental Psychology: General*, 144, 951.
- FALK, A., BECKER, A., DOHMEN, T. ET AL. (2018): “Global evidence on economic preferences,” *The Quarterly Journal of Economics*, 133, 1645–1692.
- FEHR, E. AND RANGEL, A. (2011): “Neuroeconomic Foundations of Economic Choice—Recent Advances,” *Journal of Economic Perspectives*, 25, 3–30.
- FREDERICK, S. (2005): “Cognitive reflection and decision making,” *Journal of Economic perspectives*, 19, 25–42.
- FROMELL, H., NOSENZO, D. AND OWENS, T. (2020): “Altruism, fast and slow? Evidence from a meta-analysis and a new experiment,” *Experimental Economics*, 1–23.

- GILBERT, C. AND HUTTO, E. (2014): “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Available at (20/04/16) [http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf](http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf), vol. 81, 82.
- HALLSSON, B. G., SIEBNER, H. R. AND HULME, O. J. (2018): “Fairness, fast and slow: A review of dual process models of fairness,” *Neuroscience & Biobehavioral Reviews*, 89, 49–60.
- HARRISON, G. W. ET AL. (2008): “Neuroeconomics: A critical reconsideration,” *Economics and Philosophy*, 24, 303.
- HEILMAN, R. M., CRIŞAN, L. G., HOUSER, D. ET AL. (2010): “Emotion regulation and decision making under risk and uncertainty,” *Emotion*, 10, 257.
- KAHNEMAN, D. (2003): “Maps of bounded rationality: Psychology for behavioral economics,” *American Economic Review*, 93, 1449–1475.
- KIM, M. Y. AND KANFER, R. (2009): “The joint influence of mood and a cognitively demanding task on risk-taking,” *Motivation and Emotion*, 33, 362–372.
- KIRCHSTEIGER, G., RIGOTTI, L. AND RUSTICHINI, A. (2006): “Your morals might be your moods,” *Journal of Economic Behavior & Organization*, 59, 155–172.
- KRAJBICH, I., BARTLING, B., HARE, T. ET AL. (2015): “Rethinking fast and slow based on a critique of reaction-time reverse inference,” *Nature communications*, 6, 1–9.
- KVARVEN, A., STRØMLAND, E., WOLLBRANT, C. ET AL. (2020): “The intuitive cooperation hypothesis revisited: a meta-analytic examination of effect size and between-study heterogeneity,” *Journal of the Economic Science Association*, 1–16.
- LOEWENSTEIN, G. AND LERNER, J. S. (2003): “The role of affect in decision making.” .
- MANSKI, C. F. (2004): “Measuring expectations,” *Econometrica*, 72, 1329–1376.
- MERKEL, A. L. AND LOHSE, J. (2019): “Is fairness intuitive? An experiment accounting for subjective utility differences under time pressure,” *Experimental Economics*, 22, 24–50.
- PALAN, S. AND SCHITTER, C. (2018): “Prolific. A subject pool for online experiments,” *Journal of Behavioral and Experimental Finance*, 17, 22–27.

- PIOVESAN, M. AND WENGSTRÖM, E. (2009): “Fast or fair? A study of response times,” *Economics Letters*, 105, 193–196.
- PROTO, E., SGROI, D. AND NAZNEEN, M. (2019): “Happiness, cooperation and language,” *Journal of Economic Behavior & Organization*, 168, 209–228.
- RAND, D. G. (2016): “Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation,” *Psychological science*, 27, 1192–1206.
- (2017): “Social dilemma cooperation (unlike Dictator Game giving) is intuitive for men as well as women,” *Journal of experimental social psychology*, 73, 164–168.
- (2019): “Intuition, deliberation, and cooperation: Further meta-analytic evidence from 91 experiments on pure cooperation,” *Available at SSRN 3390018*.
- RAND, D. G., BRESKOLL, V. L., EVERETT, J. A. ET AL. (2016): “Social heuristics and social roles: Intuition favors altruism for women but not for men.” *Journal of Experimental Psychology: General*, 145, 389.
- RAND, D. G., GREENE, J. D. AND NOWAK, M. A. (2012): “Spontaneous giving and calculated greed,” *Nature*, 489, 427.
- RAND, D. G., PEYSAKHOVICH, A., KRAFT-TODD, G. T. ET AL. (2014): “Social heuristics shape intuitive cooperation,” *Nature Communications*, 5, 3677.
- RICK, S. AND LOEWENSTEIN, G. (2008): “The role of emotion in economic behavior,” *Handbook of emotions*, 3, 138–158.
- RUBINSTEIN, A. (2007): “Instinctive and cognitive reasoning: A study of response times,” *The Economic Journal*, 117, 1243–1259.
- (2013): “Response time and decision making: An experimental study.” *Judgment & Decision Making*, 8.
- (2016): “A typology of players: Between instinctive and contemplative,” *The Quarterly Journal of Economics*, 131, 859–890.
- SCHMIDT, D., SHUPP, R., WALKER, J. M. ET AL. (2003): “Playing safe in coordination games: the roles of risk dominance, payoff dominance, and history of play,” *Games and Economic Behavior*, 42, 281–299.

SKYRMS, B. (2004): *The stag hunt and the evolution of social structure*, Cambridge University Press.

SPILIOPOULOS, L. AND ORTMANN, A. (2018): “The BCD of response time analysis in experimental economics,” *Experimental economics*, 21, 383–433.

TAUSCZIK, Y. R. AND PENNEBAKER, J. W. (2010): “The psychological meaning of words: LIWC and computerized text analysis methods,” *Journal of language and social psychology*, 29, 24–54.



Universitat d'Alacant
Universidad de Alicante

Appendix

2.A Experimental Instructions

Welcome

This study aims at understanding how people make choices in different scenarios.

You will face **4 different scenarios**, in which you have to make a decision, and a short questionnaire.

We ask you to focus on the tasks and try to avoid any distraction; please silence your mobile phone and turn off the television/music.

Your privacy

No personal or identifying information will be collected during the study.

Your data will be anonymous and confidential, meaning that any information you provide cannot be traced back to you.

The results of this study may be published on journal articles and/or presented at conferences.

The raw data (from which you cannot be identified) will be kept for a minimum period of five years after the publication process is complete.

Please note you have the right to withdraw consent at any time.

Contact

You can reach out to the researcher (Roberto Di Paolo, email: roberto.dipaolo@imtlucca.it) if you have any questions related to this study.

Clicking on the **I Agree** button below indicates that:

- You have read the above information;
- You voluntarily agree to participate;
- You are at least 18 years old.

If you do not wish to participate in the research study, please decline participation by clicking on the **I Disagree** button and you will be redirected to Prolific main page.

Please, copy below your Prolific ID

Baseline - Instructions Page

Instructions

In this experiment you will face **4 different scenarios**.

In each scenario you have the chance to obtain a certain amount of money, but such amount **differs** across scenarios.

Choice

In each scenario you will be randomly matched with another participant and both you and the other participant will be asked to choose simultaneously between two options, Option A and Option B. Both you and the other participant will gain Points depending on the combination of the options chosen.

Guess

At the end of each scenario you will be asked to guess how frequently Option A is chosen by other participants.

You will receive **1 extra Point** if your guess is correct.

A guess is considered correct if it is within a range of +/- 5% from the real value.

Payment

Finally, at the end of the experiment, one of the four scenarios will be randomly selected and you will be paid according to the Points you obtained in that particular scenario.

10 Points correspond to £0.4 (around 0.5\$).

Time Delay - Instructions Page

Instructions

In this experiment you will face **4 different scenarios**.

In each scenario you have the chance to obtain a certain amount of money, but such amount **differs** across scenarios.

Choice

In each scenario you will be randomly matched with another participant and both you and the other participant will be asked to choose simultaneously between two options, Option A and Option B. Both you and the other participant will gain Points depending on the combination of the options chosen.

Time Delay

In each scenario, before you can choose an option, you have to **wait 40 second**.

Guess

At the end of each scenario you will be asked to guess how frequently Option A is chosen by other participants.

You will receive **1 extra Point** if your guess is correct.

A guess is considered correct if it is within a range of +/- 5% from the real value.

Payment

Finally, at the end of the experiment, one of the four scenarios will be randomly selected and you will be paid according to the Points you obtained in that particular scenario.

10 Points correspond to £0.4 (around 0.5\$).

Motivated Delay - Instructions Page

Instructions

In this experiment you will face **4 different scenarios**.

In each scenario you have the chance to obtain a certain amount of money, but such amount **differs** across scenarios.

Choice

In each scenario you will be randomly matched with another participant and both you and the other participant will be asked to choose simultaneously between two options, Option A and Option B. Both you and the other participant will gain Points depending on the combination of the options chosen.

Motivation

In each scenario, before you can choose an option, you have **to write a motivation of at least 40 characters** (and wait 40 second before you can actually choose your option).

Guess

At the end of each scenario you will be asked to guess how frequently Option A is chosen by other participants.

You will receive **1 extra Point** if your guess is correct.

A guess is considered correct if it is within a range of +/- 5% from the real value.

Payment

Finally, at the end of the experiment, one of the four scenarios will be randomly selected and you will be paid according to the Points you obtained in that particular scenario.

10 Points correspond to £0.4 (around 0.5\$)

Baseline - GAME 1 Page

Scenario

If you choose **Option A** :

- you obtain **4 Points** and the other obtains **4 Points**, if the other chooses **Option A**;
- you obtain **0 Points** and the other obtains **3 Points**, if the other chooses **Option B**.

If you choose **Option B**:

- you obtain **3 Points** and the other obtains **0 Points**, if the other chooses **Option A**;
- you obtain **3 Points** and the other obtains **3 Points**, if the other chooses **Option B**.

Choose between

Option A

Option B

Time Delay - GAME 1 Page**Scenario**

If you choose **Option A** :

- you obtain **4 Points** and the other obtains **4 Points**, if the other chooses **Option A**;
- you obtain **0 Points** and the other obtains **3 Points**, if the other chooses **Option B**.

If you choose **Option B**:

- you obtain **3 Points** and the other obtains **0 Points**, if the other chooses **Option A**;
- you obtain **3 Points** and the other obtains **3 Points**, if the other chooses **Option B**.

Remember you have to wait **40 seconds** before making your decision.

Choose between

Option A	Option B
-----------------	-----------------

Motivated Delay - GAME 1 Page**Scenario**

Before making a decision, motivate your choice (40 characters at least):

If you choose **Option A** :

- you obtain **4 Points** and the other obtains **4 Points**, if the other chooses **Option A**;
- you obtain **0 Points** and the other obtains **3 Points**, if the other chooses **Option B**.

If you choose **Option B**:

- you obtain **3 Points** and the other obtains **0 Points**, if the other chooses **Option A**;
- you obtain **3 Points** and the other obtains **3 Points**, if the other chooses **Option B**.

Choose between

Option A	Option B
-----------------	-----------------

Your guess

Guess the **percentage** of participants choosing **Option A**.

Below we remind you how Points are allocated depending on the options chosen.

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%



If **Option A** is chosen:

- the decision maker obtains **4 Points** and the other participant obtains **4 Points**, if the other chooses **Option A**;
- the decision maker obtains **0 Points** and the other participant obtains **3 Points**, if the other chooses **Option B**.

If **Option B** is chosen:

- the decision maker obtains **3 Points** and the other participant obtains **0 Points**, if the other chooses **Option A**;
- the decision maker obtains **3 Points** and the other participant obtains **3 Points**, if the other chooses **Option B**.

Final questions (page 1)

Your task is to decide on the number of boxes to collect out of 100 such boxes numbered 1 through 100.

You collect the boxes continuing until the box whose number is equal to the number of boxes you decide to collect.

Exactly one of these 100 boxes contains a bomb.

You do not know the bomb's location.

You only know that the bomb is equally likely to be in any of the 100 boxes.

You have to choose a number between 1 and 100 in order to collect the boxes from 1 to the number chosen.

At the end of the experiment, the number of the box containing the bomb will be randomly determined by the computer.

If you happen to have collected the box where the bomb is located you will earn **0 extra Points**.

If the bomb is located in a box that you did not collect you will earn **0.01 extra Points** for each collected box.

How many boxes do you want to open?

Final questions (page 2)

How well do the following statement describe you as a person?

Please indicate your answer on a scale from 0 to 10, where 0 means “does not describe me at all” and a 10 means “describes me perfectly”.

	0	1	3	4	5	6	7	8	9	10
I assume that people have only the best intentions.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Final question (page 3)

To what extent have you participated in studies like this one before?

Nothing like this scenario

Somewhat like this scenario

Exactly this scenario

Recall the following Scenario:

If you choose **Option A** :

- you obtain **4** Points and the other obtains **4** Points, if the other chooses **Option A**;
- you obtain **0** Points and the other obtains **3** Points, if the other chooses **Option B**.

If you choose **Option B**:

- you obtain **3** Points and the other obtains **0** Points, if the other chooses **Option A**;
- you obtain **3** Points and the other obtains **3** Points, if the other chooses **Option B**.

	Option A	Option B
Which option is more collaborative?	<input type="radio"/>	<input type="radio"/>
Which option is safer?	<input type="radio"/>	<input type="radio"/>

THANK YOU!

Thank you for participating.

You will see your earnings on your Prolific Profile after the experimenter has confirmed your payment.

Please, click on the button below to complete the survey.

2.B Experimental Data & Analysis

2.B.1 Regression analysis

I estimate the following logistic regression model:

$$\ln \left(\frac{\text{prob}(Y_i = 1)}{1 - \text{prob}(Y_i = 1)} \right) = \beta_0 + \beta_1 TDtreat_i + \beta_2 MDtreat_i + \beta_3 basin_g + \beta_4 TDtreat_i * basin_g + \beta_5 MDtreat_i * basin_g + \mathbf{X}_i + \epsilon_i, \quad (2.1)$$

where $Y_i = 1$ if individual i chose *Stag* and 0 otherwise; $TDtreat_i = 1$ if i is in the *time delay* treatment and = 0 otherwise while $MDtreat_i = 1$ if i is in the *motivated delay* treatment and = 0 otherwise (omitted category is the *baseline* treatment), $basin_g$ is the value of the basin of attraction of *Stag* in game $g = 1, \dots, 4$; \mathbf{X}_i is a vector controls variables (see below), and ϵ_i is the error term. The vector \mathbf{X}_i includes the following controls: $female_i = 1$ if i is female and = 0 otherwise; age_i is i 's age; $student_i = 1$ if i is a student and = 0 otherwise; $previous_studies_i = 1$ if i already saw some game similar to the Stag-Hunt and = 0 otherwise; $BRET_i$ is the number of boxes opened in the BRET task (a measure of i 's risk aversion); $trust_i$ is i 's self-assessed trust in others.

Table 2.B.1 reports the estimates of all regressors for the Model (1)-(3) of Table 2.3.2 described in the main text, including the omitted controls.

2.B.2 Tobit Regression on Beliefs

I estimate a Tobit model with the same regressors as in (2.1) using as dependent variable the participants' beliefs about others' choices of *Stag* – a variable ranging between 0 (nobody plays *Stag*) and 100 (everybody plays *Stag*) and taking all integer values. Figure 2.B.1 shows the distribution of beliefs. Indeed, data seem censored, with a substantial fraction of beliefs being equal to 100.

Table 2.B.2 shows the estimates obtained in the Tobit regressions. In sum, I found that the *motivated delay* treatment reduces the beliefs that other participants' choice will be *Stag*, while, as expected, the effect of the basin of attraction increases them. Moreover, both the *time delay* and the *motivated delay* treatments have a sizeable effect on beliefs. These results are in line with the ones presented in Table 2.3.2, suggesting that both treatments and payoffs affect behavior, at least in part, through their impact on beliefs.

Table 2.B.1: Logit regression

Variable	Model (1)	Model (2)	Model (3)
Time Delay	0.080 (0.141)	-0.104 (0.168)	-0.077 (0.179)
Motivated Delay	-0.167 (0.141)	-0.336** (0.162)	-0.344** (0.164)
Basin	2.286 (0.297)	1.288** (0.531)	1.392*** (0.541)
Time Delay \times Basin		1.562** (0.771)	1.532* (0.790)
Motivated Delay \times Basin		1.421** (0.703)	1.336* (0.718)
BRET			-0.001 (0.002)
Trust			0.149*** (0.027)
Female			0.085 (0.114)
Age			0.004* (0.011)
Student			-0.013 (0.149)
Previous Studies			0.151 (0.117)
Cons	0.584** (0.104)	0.703*** (0.120)	-0.304 (0.428)
Individual Characteristics	No	No	Yes
No. of Obs.	3608	3608	3584
N	902	902	896
χ^2	62.86***	71.17***	96.81***
Pseudo R^2	0.0088	0.0094	0.0283

Note: The dependent variable is binary and equal to 1 if the choice is *Stag*, 0 otherwise. The error terms are clustered at individual level and reported in parenthesis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

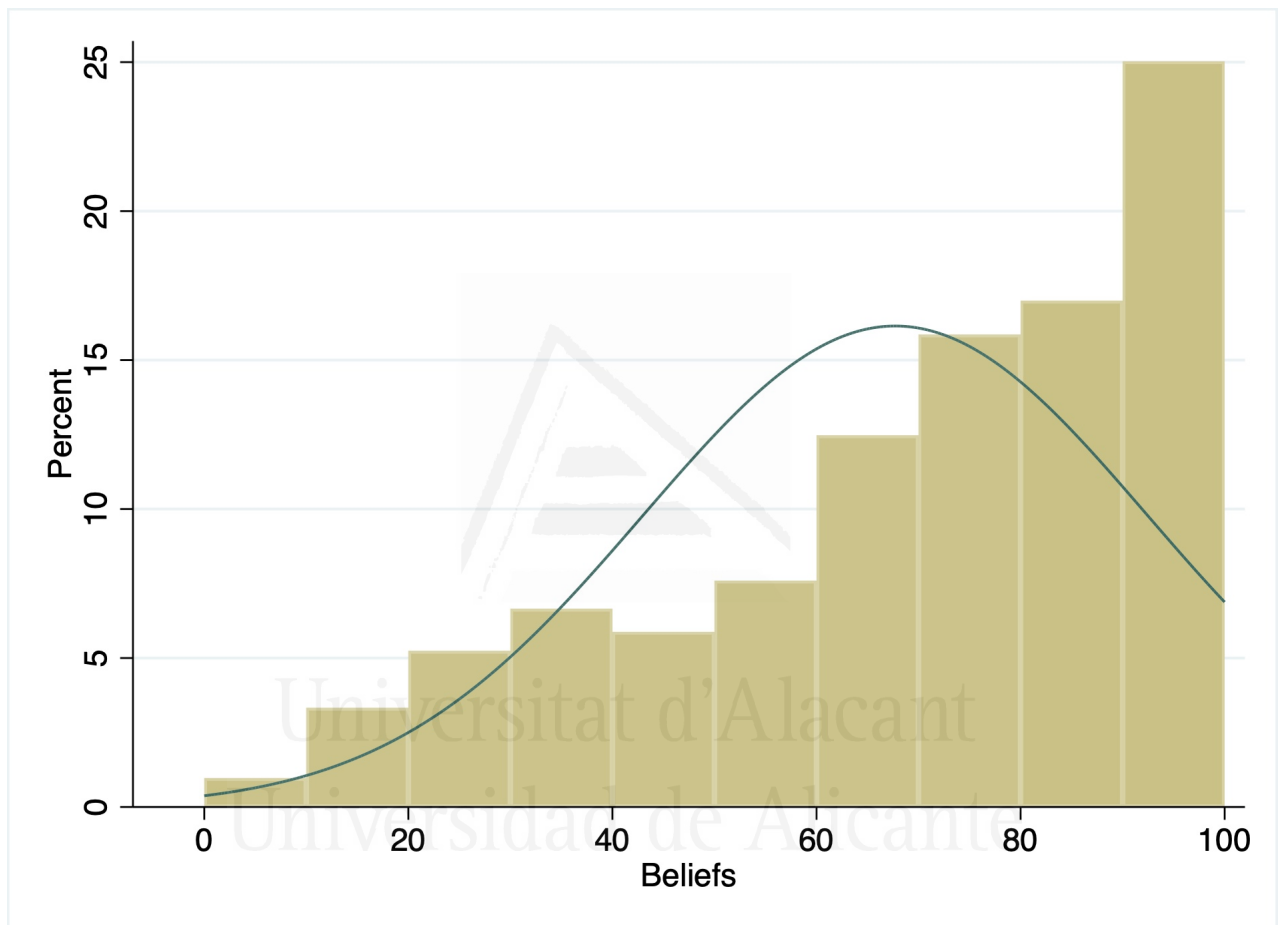


Figure 2.B.1: Histogram distribution of beliefs about other participants' choices of *Stag*. After each Stag-Hunt game, participants are asked: "Guess the percentage of participants choosing Option A" with a slider, from 0 to 100, to answer the question.

Table 2.B.2: Tobit regression (right-censored)

Variable	Model (1)	Model (2)	Model (3)
Time Delay	0.819 (1.885)	-1.940 (2.155)	-1.481 (2.110)
Motivated Delay	-2.771 (1.955)	-4.241* (2.211)	-4.095* (2.147)
Basin	23.366*** (2.541)	12.007*** (4.310)	12.332*** (4.331)
Time Delay \times Basin		22.105*** (6.112)	21.749*** (6.146)
Motivated Delay \times Basin		11.773* (6.189)	11.552* (6.233)
BRET			-0.001 (0.033)
Trust			2.562*** (0.358)
Female			-1.124 (1.445)
Age			0.242 (0.146)
Student			-0.933 (1.852)
Previous Studies			1.785 (1.441)
Cons	63.274*** (1.433)	67.689*** (1.533)	46.454*** (5.488)
No. of Obs.	3608	3608	3584
left-censored	0	0	0
uncensored	3,374	3374	3352
right-censored	234	234	232
N	902	902	896
χ^2	29.55***	20.95***	14.62***
Pseudo R^2	0.0011	0.0012	0.0072

Note: The dependent variable is continuous and it is equal to participants' belief about the other participants' choices of *Stag*. The error terms are clustered at individual level and reported in parenthesis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

2.B.3 Figures and Tables

Table 2.B.3: Percentage of correct answers in the comprehension questions divided by treatment.

	More Collaborative Option			Safer Option		
	<i>Baseline</i>	<i>Time Delay</i>	<i>Motivated Delay</i>	<i>Baseline</i>	<i>Time Delay</i>	<i>Motivated Delay</i>
<i>Wrong</i>	14.43%	16.23%	15.56%	33.22%	39.07%	27.81%
<i>Correct</i>	85.57%	83.77%	84.44%	66.78%	60.93%	72.19%

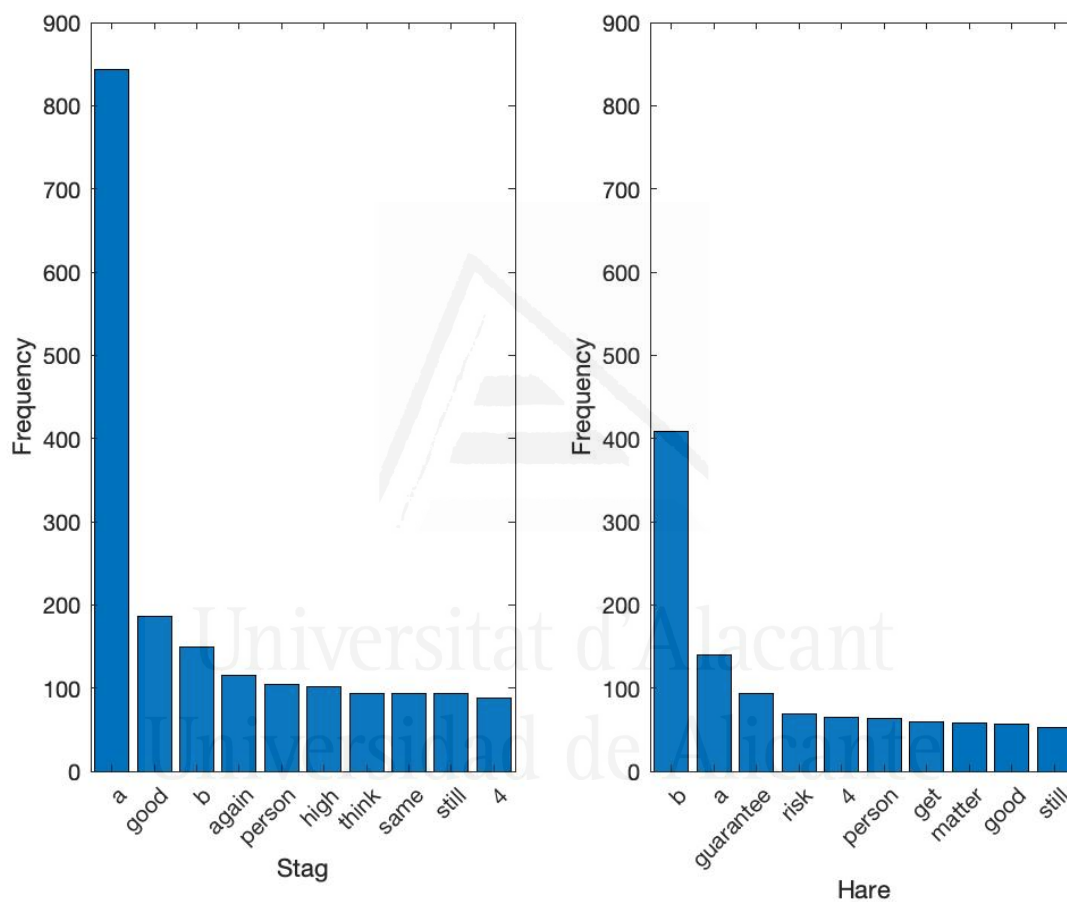


Figure 2.B.2: Bar plot of the most frequent words written to motivate choices in the *motivated delay* treatment, by chosen strategy (*Stag* or *Hare*). The words distribution seems consistent with the participants' decision. Words as "a" and "b" refer to "Option A" (*Stag*) and "Option B" (*Hare*).

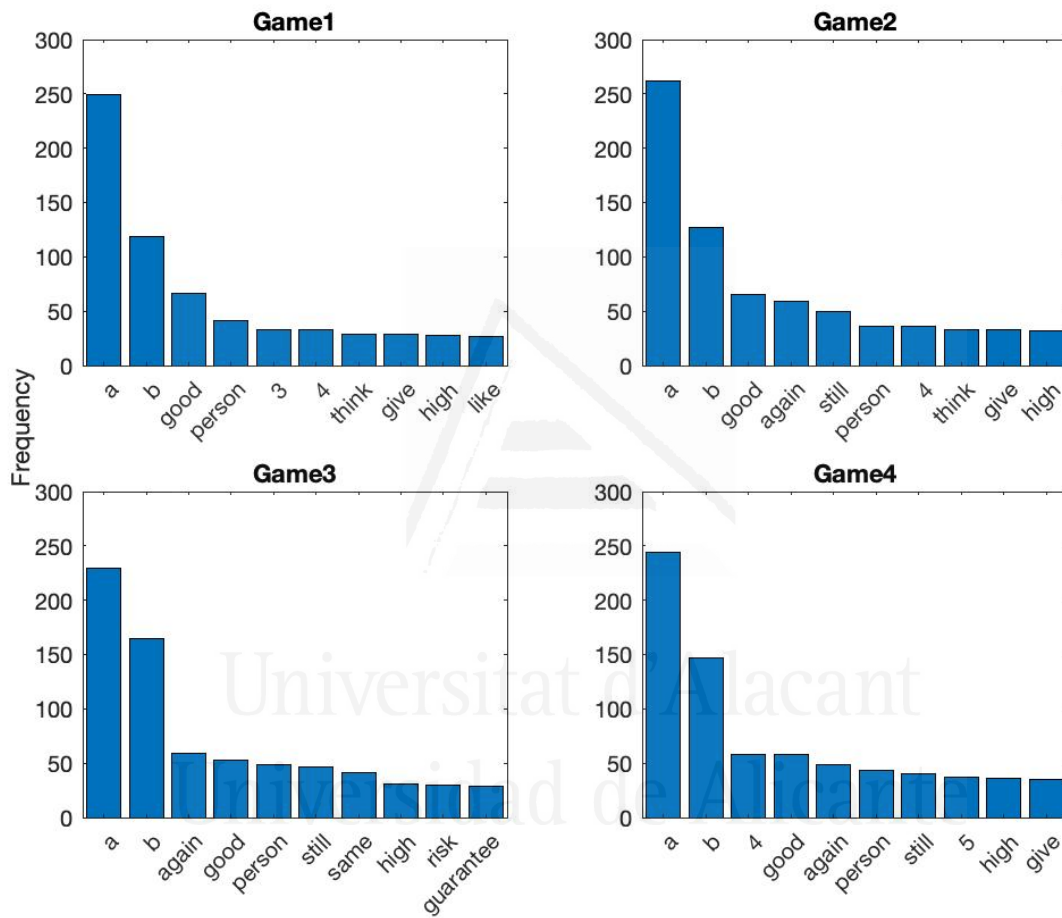


Figure 2.B.3: Bar plot of the most frequent words written to motivate choices in the *motivated delay* treatment, by Stag-Hunt game. Words as “a” and “b” refer to “Option A” (*Stag*) and “Option B” (*Hare*).

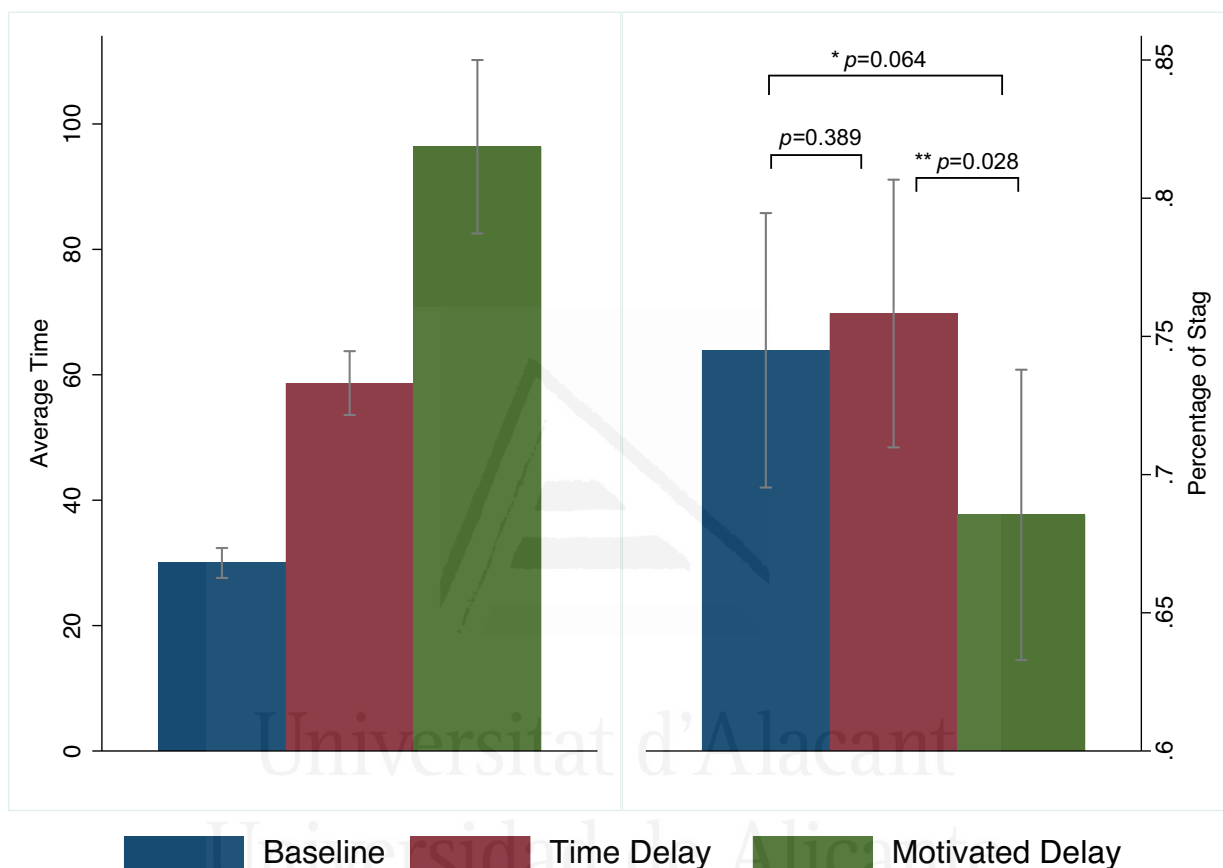


Figure 2.B.4: Average play of stag (right chart) and time in seconds for decision (left chart) in the baseline, the time delay and the motivated delay conditions for Game 1. Behaviors differ weakly across treatments (Chi-squared test, $\chi^2=4.60$, $p=0.100$). While the percentage of *Stag* choices in the *baseline* and *time delay* treatments are not significantly different (Fisher’s exact test, $p=0.389$), they are in the *time delay* and *motivated delay* treatments (Fisher’s exact test, $p=0.064$) and in the *time delay* and *motivated delay* treatments (Fisher’s exact test, $p=0.028$).

Game-based educational program promotes sustainable water use

3.1 Introduction

Sustainable water consumption is relevant for the general sustainability of current and future societies (Wada and Bierkens, 2014; Kummur et al., 2016; Liu et al., 2017; Greve et al., 2018; Qin et al., 2019). Sustainable water consumption is, in many cases, an instance of prosocial behaviour in a social dilemma (Hardin, 1968): a situation in which a conflict exists between maximizing one's individual benefits and maximizing the benefits of the present and future generations. Individuals who are purely self-interested are less likely to adopt the prosocial behaviors that lead to sustainable water consumption, unless social norms exert sufficient social pressure to push self-interested individuals to do otherwise. Since the acquisition of preferences for prosocial behaviors as well as the internalization of social norms take place, in a substantial part, during childhood (House and Tomasello, 2018; House et al., 2020), it becomes a critical goal to create opportunities for young children to develop such preferences and internalize norms of sustainable water consumption (Copple et al., 2013; Cobo-Reyes et al., 2020).

Early childhood education is the natural starting point for a life-long learning. During the past years, a variety of educational methods to promote prosociality in children have been successfully implemented. These methods include play space, multi-use toys, dedicated books, group play, and organized gaming (Orlick, 1983). In particular, the kind of social interactions that come from group play and organized gaming, as well as the time that gaming can occupy in children's daily lives, make game-based educational programs a natural candidate tool for promoting desirable behaviors. Some studies, in recent years, evaluated the relevance of programs which encourage good practices in environmental benchmarks, such as the use of water (Niles et al., 2013; Cuadrado et al., 2017). In a field experiment (Schultz et al., 2016) the role of social norms in promoting water conservation

was studied, finding that people who received normative information about similar household in their neighborhoods consumed less water than the control group; moreover, people with already strong personal norms were less affected by the normative information than those with low personal norms. Importantly, children are able to recognize if prosocial norms apply to specific situations (Blake et al., 2015), so that it becomes important that children understand what is sustainable water consumption and can relate their behavior to concrete and specific situations such as water collection or body washing.

In this paper we provide evidence regarding the effectiveness of a game-based educational program implemented during the first eleven months of the year 2019 in the municipality of Lucca, Italy. The program was named *BLUTUBE: Who brings the water home* and was aimed at promoting sustainable water consumption as well as awareness about the municipal water system and its usage. The targets of the program were around 1000 students from 2nd-4th grades and their families. The program relied primarily on ludic engagement for the specific objectives of improving students' awareness about the water cycle in nature, the water system of the municipality of Lucca, and the daily usage of water.

Our approach to the empirical assessment of the program's impact is based on the quasi-experiment methodology (Campbell and Stanley, 2015): we were unable to intervene directly on the organization of the program¹, but we were able to implement a simple two-group design (treatment and control) and collect three distinct measurements of target outcome variables over a period of eleven months. In particular, we identify the students' awareness and their behaviors about water consumption with three waves of surveys administered, respectively, immediately before the program started, two days after the main activities were over, and after six further months. Responses to this kind of questionnaires have been shown to be a reliable source of information on children's perspectives and perceptions (Danielson and Phelps, 2003; Di Riso et al., 2010; Bevans et al., 2020; Alan and Kabasakal, 2020).

Our main finding is that the program had a positive impact on the awareness of water usage. This effect is primarily driven by an increase in the frequency of self-reported virtuous behaviors regarding water consumption and discussions with parents about water. Moreover, such positive effect appears to be persistent: six months after the end of the main activities of the program the effect is still positive and of appreciable size.

¹The game-based program was already design and organized before we decided to study its effects. Then, the participating classes were already decided.

3.2 Material and Methods

3.2.1 The game-based educational program

The program was designed and implemented by the Provincial Education Office of Lucca (Provveditorato agli Studi), Lucca Crea s.r.l. (a company 100% owned by the municipality of Lucca which is in charge of organizing and managing cultural events)², and GEAL s.p.a. (the water utility company of the municipality of Lucca).

The program was titled “*BLUTUBE - Chi porta l’acqua a casa*” (*BLUTUBE - Who brings the water home*) and had its main engine made of gaming activities, for which an urban and a board game were developed ad hoc by Lucca Crea and its collaborators, also in partnership with GEAL and the municipality of Lucca. The gaming activities were tuned to fit 2nd, 3rd, and 4th grades students from the primary schools in Lucca. The main aim of the program was to bring about greater awareness of the daily use of water resources and their sustainable consumption together with knowledge of the integrated water system of the municipality of Lucca and the water cycle in general. Games and gaming activities were specifically designed for this purpose, although the board game (also named BLUTUBE) was designed to be playable, and enjoyable, as a stand alone game too (more details on the games can be found in Appendix 3.A).

The program was divided in three distinct phases. The first phase was titled *How not to drown in a glass of water*. During February, a group of educators, specifically selected for the program, went to each class participating in the program to give a short talk on the importance of water resources and their consumption as well as to explain the working of the gaming activities (program phases, allocation of game points, publication of rankings) and, in particular, to teach students how to play the board game BLUTUBE. Moreover, each student got its own box of the board game (for playing at home) and each class was also endowed with a copy of the board game (for playing in class).

The second phase was titled *Bring the water to your mill* and starts just after the educators talk in each class. It lasted 6 weeks during which the students participating in the program had the chance to play as much as they wanted, and accumulate points accordingly, for two distinct rankings: the individual ranking and the class ranking. There were four different ways to obtain points:

- *playing the board game BLUTUBE at school*: each student can play during school time. The teacher records each time a student plays on a scoreboard and each week a picture is sent to

²Lucca Crea s.r.l. primary job is to organize Lucca Comics & Games, one of the largest transmedia shows in the world focusing on comics, games and pop culture.

the program organizers. For each recorded play a student earns 10 points, up to a total of 2500 for the whole phase also considering the points earned for playing at home (see below);

- *playing the board game BLUTUBE at home*: each student can play at home with their family or friends and gain points every time they send a picture of their playing with parents, familiars or friends, to the program organizers, also indicating the name, the surname, the school and the class. For each appropriate picture sent a student earns 10 points, up to 2500 points in total also considering the points earned for playing at school (see above);
- *visiting the “hidden water places” in Lucca*: each student can visit, together with parents or other family members, a number of specific places labelled as “water places” in municipality of Lucca. Such places are reported in the map describing the program and distributed at the beginning with the board game. A student can send to the program organizers a picture proving a visit in one distinct water place indicated in the map, also indicating the student’s name, the surname, the school and the class. For each appropriate picture sent the student earns 150 points, up to 2500 points in total.
- *providing evidence of sustainable behavior*: each student can send to the program organizers a picture where the student is making a sustainable use of water, e.g., eating vegetables, filling the can at the fountain, turning the faucet off when they are brushing their teeth. The picture has also to indicate the student’s name, the surname, the school and the class. A student gains between 10 to 200 points for each appropriate picture, depending on the actual behavior (e.g., sending a photo eating a vegetable is awarded by 50 points or make a video in where they teach others how not to waste water is awarded with 200 points), up to 5000 points in total.

Starting from the second week of the second phase both individual and class scores were published in a dedicated website and in local newspapers. In this way, the participating students, their parents, and others in their schools could see their weekly progress and compare their scores with those of other participants.

The last phase of the program was titled *BLUTUBE Tournaments* and consisted in a tournament with restricted participation where the only way to accumulate points was playing with the board game BLUTUBE. Specifically, the 16 classes with the highest total score in the second phase (among the 53 classes participating) were selected to participate in four distinct group stage tournaments (each comprising 4 of the 16 classes). The winner of each group stage tournament qualified to participate in the final stage tournament which took place during the Lucca Comics and Games

festival held in 2019. The final stage tournament allowed to win a full paid holiday trip themed “Environment”, where students could learn methods to create electricity through the use of heat while respecting the environment.

All activities related to the game-based educational program had been carried out between January and November, 2019. The participation protocol was as follows. Most primary schools in the municipality of Lucca were involved. Actual participation in the program was determined at the class level, under consent by the school head teacher. Lucca Crea, which was in charge of promoting the program across the schools, talked to the head teacher of each school asking for classes who were available to participate in the program. In most cases, the decision about whether to participate or not was taken by the head teacher of each class, and in no case there was a possibility for the students of the class to affect such decision, which was made on the basis of the overall workload of the class in terms of extra-curricular activities. A few remarks are worth doing. First, the participation protocol led to a situation where in the same schools there were classes which participated and classes which did not participate. Second, participation was exogenous to the students’ desire to participate. This because the extra activities that classes will follow during the year are decided before the beginning of the school by the school council. Third, actual participation was often exogenous to the teachers’ desire to participate too. This is because the teachers’ decision was often constrained by the fact that their class was already involved in a number of extra-curricular activities, and hence could not actually participate, or by the fact that it had to add extra-curricular activities and they are forced by the school principal to add the program in their activities.

This participation protocol allows the applicability and effectiveness of our method of analysis, in that the assignment to the program, although not fully randomized, might be considered exogenous to schools, students’ and teachers’ preferences.

3.2.2 Method and data

The program described in Subsection 3.2.1 qualifies as a nonequivalent group design (a type of quasi-experiment (Cook et al., 1979)) for which we designed a pre/post control-treatment study that we implemented using a questionnaire (designed ad hoc) administered three times: just before the program, immediately after the end of phase two (three month later than the beginning), and then again at the end of the program (after six months).

The study includes 28 primary schools. From those schools, 53 classes (around the 45% of the total classes from the 2nd-4th grades) were directly involved in the program, forming the treatment

group. For the control group we selected other 53 classes that were not directly involved in the program. We tried to build the best possible counterfactual between the 63 classes remained out of the program, controlling for the grade and the number of students in the classes. This was not an easy task because the total of 106 classes covers about the 90% of the entire population of 2nd-4th grades students in the municipality of Lucca (the overall number of classes being 116). So, together the treatment and control groups represent almost the entire student's population.

Students' awareness about the efficient use of water was identified by means of a paper-based survey regarding students' behaviors and habits related to water use and consumption (the original and the English-translated questionnaires can be found in Appendices 3.A.3 and 3.B, respectively). Specifically, the survey contained seven distinct questions about water consumption in familiar circumstances, the extent to which students talk about water with their parents, and the extent to which students eat food containing water (fruit and vegetables). These questions are: *Teeth*: "How much do you keep the faucet turned on when you brush your teeth?"; *Shower*: "Are you having more often a bath or a shower?"; *Fountain*: "Do you drink water more from plastic bottles or from fountains/faucets?"; *Vegetables*: "Are you eating fruit or vegetables during your meals?"; *Hands*: "When you wash your hands, do you turn the faucet off while you soap your hands?"; *Parents*: "Do you talk with your parents on how the water gets to your house?"; *Waste*: "Do you talk with your parents on how not to waste water?". Each question was chosen to measure the main purposes of the program. The main targets, that Lucca Crea and GEAL had in mind while were developing the educational program, are the children's habits. Thus, we construct the 7 questions with the purpose to measure the particular habits (and not the frequency, which can be assumed randomly distributed between families) involved in the urban and board game.

Answers were recorded using a 1-to-5 Likert scale which was proposed in three cases with categories going from the least virtuous to the most virtuous and in the remaining four cases in the reverse order. For the analysis presented in Section 3.3 we recoded all answers such that category 1 is always the least virtuous and category 5 is always the most virtuous (the original scale for each question can be found in Appendix 3.B).

The survey also contained questions related to relational activities, ludic habits and ludic preferences, that we do not exploit in the following analysis as they were meant for different research purposes. In addition, we measure cognitive skills using logical and mathematical questions taken from the tests produced by the INVALSI (Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione) and the ones developed by TIMSS (Trends in Mathematics and Science

Study).

The first survey was collected during February 2019, before the beginning of the program. The parents of students involved signed an informed consent form, with the specific consent for the possibility to link students' answers to their scores in the program. Teachers received only general information about the research project, and specifically no details about what we were trying to elicit. The second survey was administered at the end of the second phase, during the month of May 2019. The survey was identical to the previous one but for the questions aiming at eliciting cognitive skills which we opted to substitute with new ones of comparable difficulty. This was done because students might have learnt the answer to a particular mathematical/logical question (e.g., comparing with classmates) and not having increased their cognitive skills. The second survey was administered to the classes involved following the same procedures as in the first wave. Lastly, a third survey was administered six months after, when the program was officially over. This last survey was identical to the previous two but for the questions aiming at eliciting cognitive skills. Also in this case the survey was administered to the classes involved following the same procedures as in the first two waves. Figure 3.2.1 reports the timing of the program and the three survey waves.

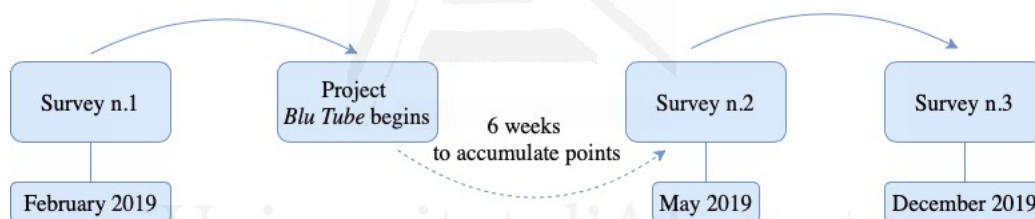


Figure 3.2.1: Timeline of quasi-experimental study of the intervention program.

3.3 Results

Our final sample consists of 52 classes in the treatment group (one class envelope was lost during the collection process) and 53 in the control group, for a total of 105 classes and 5273 questionnaires (up to three per student). Table 3.3.1 reports the number of questionnaires by group and survey wave, showing that the sizes are balanced within groups and across groups.

Table 3.3.2 reports the summary statistics by treatment and control groups for the pre-program survey. Figures show that the two groups are not well balanced: while the difference in the number of students per class is only marginally not statistically significant ($Z = -1.95, p = .051$), the difference in the measured students' cognitive skills is statistically significant ($Z = -2.17, p = .031$) as well as the distribution of grades ($Z = 4.99, p < .0001$). These differences are mainly due to the fact that

Table 3.3.1: Sample composition by conditions and periods

	Classes	Students			
	<i>Assignment</i>	<i>Pre Program</i>	<i>Post Program</i>	<i>Post 6 Months</i>	<i>Total</i>
Treatment	52	869	895	908	2601
Control	53	869	860	872	2672
Total	105	1738	1755	1780	5273

The pre-program wave involved 869 students in both treatment and control group (Column 3). The post-program wave involved 895 students in the treatment group and 860 students in the control group (Column 4). The post-6 months wave (six months after the end of the second phase) involved 908 students from the treatment group and 872 from the control group (Column 5). The final sample consists of 5273 questionnaires (up to three per student).

the distribution of students across grades is quite different between the treatment and the control group³ (for the 2nd grade there are 818 students in the treatment group and 855 in the control group; for the 3rd grade there are 621 students in the treatment group and 1123 in the control group; for the 4th grade there are 1162 students in the treatment group and 694 in the control group). In the light of this, we checked whether there is any difference in the reported behavior in the pre-program survey. Importantly, there is no statistically significant difference in the aggregate reported behavior between the control and the treatment group ($Z = -1.30, p = .193$). Aggregate reported behavior is constructed summing up the answers to all 7 questions of relevance here, so that (with a Likert scale going from 1 to 5) the aggregate variable ranges from a minimum of 7 (least sustainable reported behavior) to a maximum of 35 (most sustainable reported behavior).

We also looked at the distribution of answers in the pre-program survey for each of the 7 questions, testing for statistically significant differences. In four cases we found that the distribution of answers are not statistically different between the treatment and the control group, namely: *Shower* ($Z = -0.18, p = .849$); *Fountain* ($Z = 0.84, p = .397$); *Vegetables* ($Z = -0.69, p = .488$); *Waste* ($Z = 1.69, p = .091$), while in 3 cases we found higher statistically significant differences in the treatment: *Teeth* ($Z = -3.05, p = .002$), *Hands* ($Z = -2.36, p = .018$) and *Parents* ($Z = -2.27, p = .023$).

In the light of these results we adopt a two-step strategy. First, we carry out a non-parametric analysis of the treatment effect on the aggregate reported behavior. This is possible because, although the treatment and control groups are not perfectly balanced, the aggregate variable comes with similar levels in the two groups for the pre-program survey. We then check the robustness of non-parametric results by running regressions for each wave, including controls for the sample characteristics in order to correct for the lack of sample balancedness.

³Cognitive skills are higher for students in higher grades respect those in lower grades (2nd-3th: $Z = -4.487, p < 0.001$, 2nd-4th: $Z = -9.143, p < 0.001$; 3rd-4th: $Z = -4.589, p < 0.001$)

Table 3.3.2: Mean difference of independent samples in the pre-program sample

Variable	Control	Treatment	Min.	Max.	p-value
Grade	3.14	2.94	2	4	<.001
Students	17.08	17.33	9	25	.051
Cognitive skills	0.50	0.56	0	1	.031
Aggregate reported behavior	22.04	22.32	9	34	.193
• <i>Teeth</i>	4.49	4.60	1	5	.002
• <i>Shower</i>	3.91	3.86	1	5	.849
• <i>Fountain</i>	2.41	2.36	1	5	.397
• <i>Vegetables</i>	3.17	3.21	1	5	.488
• <i>Hands</i>	3.46	3.65	1	5	.018
• <i>Parents</i>	1.99	2.13	1	5	.023
• <i>Waste</i>	2.61	2.51	1	5	.091

Descriptive Statistics. *Grade* is the students' year group. *Students* is the number of students in each class. *Cognitive Skills* is equal to 1 if the result obtained in the logical and mathematical questions are higher than the median, 0 otherwise. *Aggregate reported behavior* is the average sum of the first seven questions of the questionnaire. *Teeth*: "How much do you keep the faucet turned on when you brush your teeth?". *Shower*: "Are you having more often a bath or a shower?". *Fountain*: "Do you drink water more from the plastic bottles or from the fountain/faucet?". *Vegetables*: "Are you eating fruit or vegetables during your meals?". *Hands*: "When you wash your hands, do you turn the faucet off while you soap your hands?". *Parents*: "Do you talk with your parents on how the water gets to your house?". *Waste*: "Do you you talk with your parents on how not to waste water?". The Mann-Whitney test of the equality of the means is reported in the last column.

Second, we study the treatment effect on the reported behavior for each of the 7 questions using ordered logit regressions where we pool all data and we control for sample characteristics, the 3-survey structure, and their interaction with the treatment. This allows us to obtain indications about the source of the treatment effects estimated at the aggregate level, taking into account the fact that some reported behaviors do not come with similar levels in the pre-program survey. Also, we previously carry out a non-parametric analysis of the treatment effect for each of the 7 questions in order to give a complete picture about the differences in reported behavior across both the three surveys and the treatment and control groups.

Finally, one might wonder if the answers to the 7 questions can be accounted for by a few common factors. Correlation analysis and principal component analysis suggest that this is not quite the case (see Appendix 3.C.2).

3.3.1 Aggregated reported behavior

Figure 3.3.1 reports the cumulative distribution function of the aggregated reported behavior in the three waves (pre-program, post-program, and post6-program, i.e., 6 months after post-program) for both control and treatment groups. While the distributions of treatment and control groups in the

pre-program do not appear to be different, in the post-program and post6-program the distributions of the treatment group are shifted to the right; in particular, the distribution of the treatment group appears to first order stochastically dominate the distribution of the control group. Epps-Singleton test of the equality of the distributions confirms this: we reject the hypothesis that the distributions of treatment and control groups are the same in both the post-program survey and the post6-program survey ($W2 = 62.243, p < .001$ and $W2 = 30.943, p < .001$, respectively), while we cannot reject the hypothesis that the distributions of treatment and control groups are the same in pre-program survey ($W2 = 2.331, p = .675$).

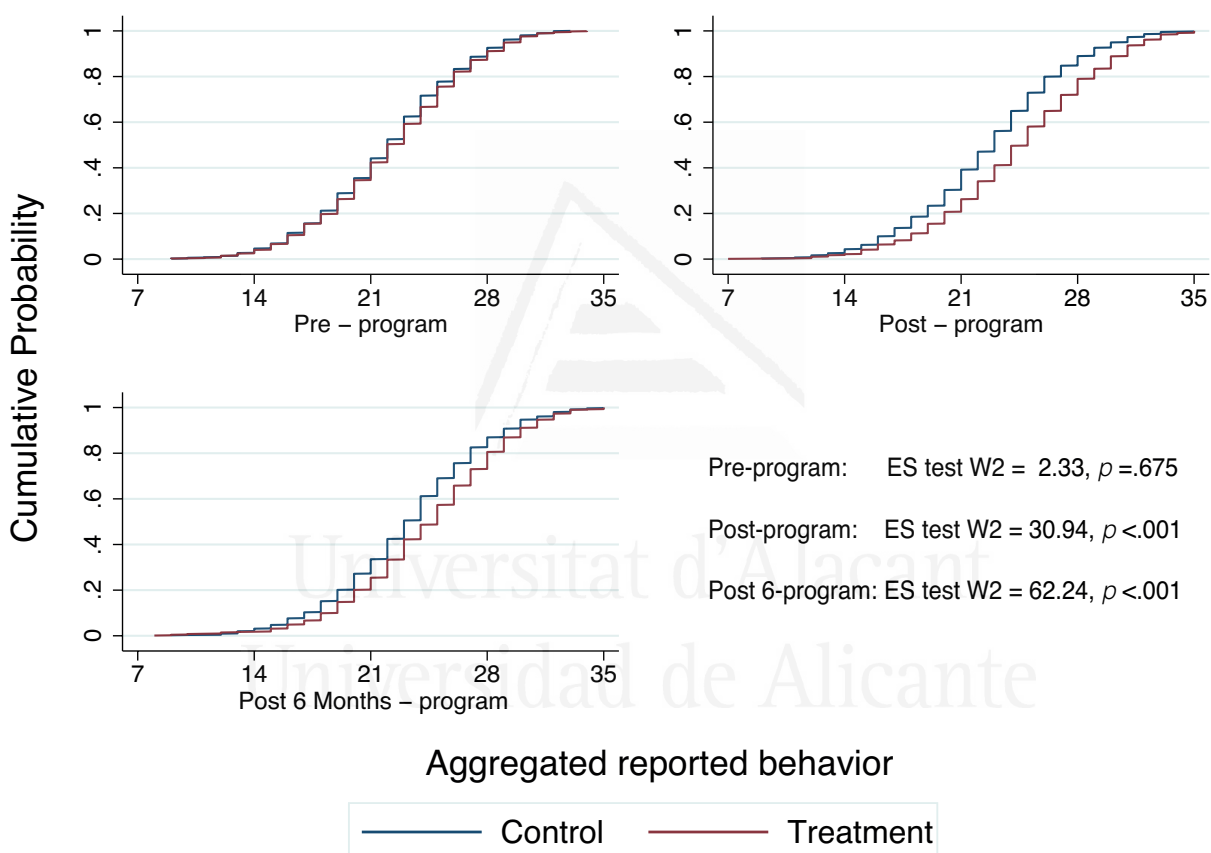


Figure 3.3.1: Cumulative distribution function of the aggregated reported behavior by conditions and waves. Distributions in the post-program and 6 months after are shifted to the right in the treatment group, with a statistically significant differences between conditions. ES stands for Epps-Singleton test.

In Figure 3.3.2 the means of the aggregated reported behavior are compared between treatment and control groups, by survey wave. No statistically significant difference is found for the pre-program survey ($Z = -1.300, p = 0.193$). In contrast, for the post-program we find that the treatment group has a statistically higher average of about 2.11 with respect to the pre-program treatment group

($Z = -9.055, p < 0.001$) and a statistically significant higher average of about 1.72 with respect to the post-treatment control group ($Z = -7.479, p < 0.001$). These numbers range from 1.32% to 7.04% of the pre-program average, suggesting that the treatment has had an impact between the pre-program and the post-program surveys.

Furthermore, Figure 3.3.2 shows that there is no appreciable difference between the aggregated behavior in the treatment group between the post-program survey and the post6-program survey ($Z = 0.165, p = 0.869$). Also, although the the average aggregated behavior of the control group increases of about 0.56 points between the post-program and the post6-program surveys, we still find a statistically significant difference between the treatment and the control groups in the post6-treatment survey ($Z = 5.271, p < 0.001$). Together, these findings suggest that the effect of the treatment is persistent, at least until the official end of the program (about 9 months after its start).

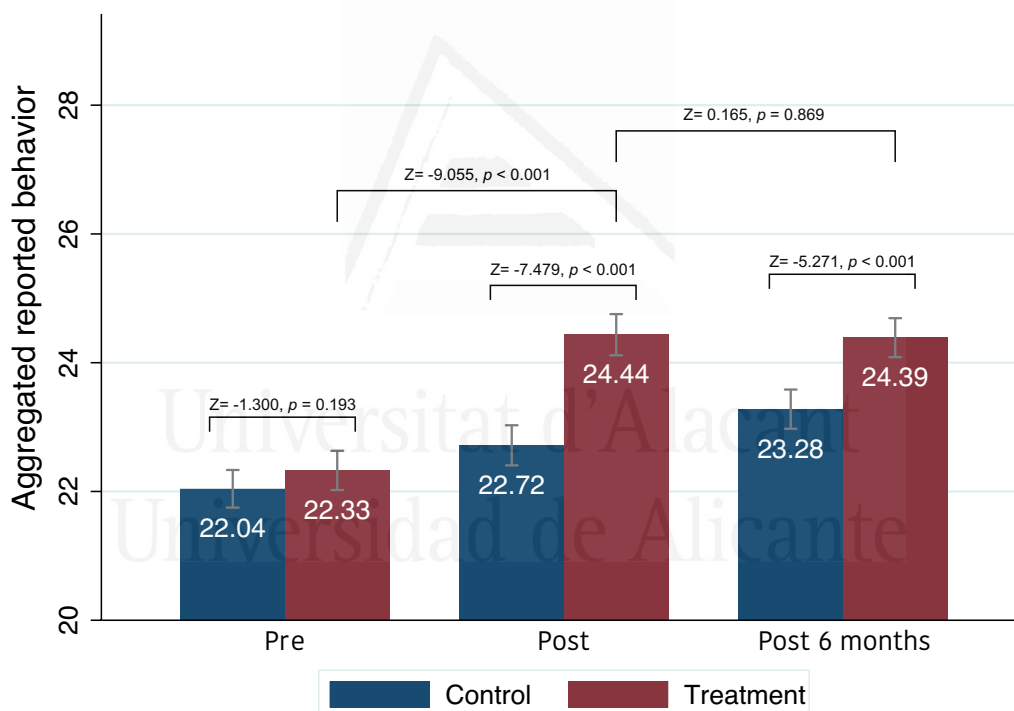


Figure 3.3.2: Average of the aggregated reported behavior by conditions and waves. In the pre-program period, the aggregated reported behavior in the treatment group is not significantly different from the control group (Mann-Whitney test, $Z = -0.638, p = 0.523$). In the post-program period and after 6 months, the aggregated reported behavior in the treatment group is significantly higher respect to the control group (Mann-Whitney test, $Z = -7.479, p < 0.001$ and $Z = -5.271, p < 0.001$, respectively). The treatment effect is stable after 6 month (Mann-Whitney test, $Z = 0.165, p = 0.869$). Error bars represents the 95% confidence interval.

The findings described above rely on the assumption that the lack of balance between treatment and control groups did not bias our estimates. In order to control for such potential problem we

run linear regression models where aggregated reported behavior is predicted by the treatment and a number of controls. Importantly, since students came from different schools and classes, and that in one school there is the possibility to have more than one class treated, we are able to control for schools including school fixed effects. In addition, besides a dummy variable for the treatment (which is equal to 1 if the student belongs to the treatment group), we include a dummy for the grade (omitted category is 2nd grade), an index of cognitive skills (fraction of correct answers in logical/mathematical questions), and the number of students in the class. We run similar regressions for the pre-program, the post-program, and the post6-program surveys. Results are reported in Table 3.3.3.

Table 3.3.3: Linear Fixed Effect Regression

	Pre		Post		Post 6 Months	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	.534 (.379)	1.13 (.593)	1.89*** (.394)	2.17*** (.550)	1.12** (.384)	1.27* (.594)
3rd Grade		1.07 (.610)		.927 (.574)		1.21 (.703)
4th Grade		2.05*** (.418)		1.65*** (.443)		1.10* (.492)
TR × 3rd		-.811 (.802)		-.481 (.782)		-.645 (.809)
TR × 4th		-.088 (.770)		.182 (.736)		.018 (.776)
Cognitive Skills		.187 (.227)		-.011 (.233)		.647* (.261)
Students		-.002 (.045)		.055 (.038)		.092 (.048)
Constant	21.9*** (.262)	20.6*** (.830)	22.6*** (.251)	20.7*** (.727)	23.3*** (.268)	20.5*** (.793)
N	1685	1685	1732	1732	1765	1765

The dependent variable is the aggregated reported behavior on good/bad practices of water usage. *Treatment* is equal to 1 if the students are in the treatment group, 0 otherwise. *Grade* is the students' year group. *2nd Grade* is the reference category. *Cognitive Skills* is equal to 1 if the result obtained in the logical and mathematical questions are higher than the median, 0 otherwise. *Students* is the number of students in each class. In all cases, we control for school fixed effects. Standard errors (in parenthesis) are clustered at class level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Column (1) reports the results of the school-fixed effect regression in the pre-program survey. No statistical significant effect of the treatment is found in this case. The result is confirmed by the estimates reported in Column (2) where the regressors include controls for the grade of the students, their interaction with the treatment, an index of cognitive skills and the number of students in each class. Among these, the only regressor with statistically significant coefficient is *4th Grade*, suggesting

that it may be the source of potential pre-program differences in reported behavior.

Column (3) reports the results of the school-fixed effect regression in the post-program survey. The coefficient of the treatment variable is positive (1.89) and statistically significant, confirming the results of the non-parametric test. Similar results are found in column (4) where the regressors include the controls used for the regression in column (2). In particular, the coefficient of the treatment variable is positive (2.17) and statistically significant. Again, the only regressor with statistically significant coefficient is *4th Grade*, in line with the idea that it may be the source of potential pre-program differences in reported behavior.

Column (5) reports the results of the school-fixed effect regression in the post6-program survey. The coefficient of the treatment variable is positive (1.12) and statistically significant, somewhat lower than in column (3). This confirms the result showed in the non-parametric test that the effect of the program is persistent after 6 months, although it is reduced magnitude. Similar results are found in column (6) where the regressors include the controls used for the regression in column (2) and (4). Specifically, the coefficient of the treatment variable is positive (1.27) and statistically significant. Again, the coefficient of *4th Grade* is positive and statistically significant, but in this case it is not the only one: also the coefficient of *Cognitive Skills* is positive and statistically significant. The sum of these two coefficients is about of the same magnitude that the one of *4th Grade* in column (4), suggesting that in the longer run cognitive skills might be a substitute for grade seniority.

3.3.2 Disaggregated reported behaviors

Figure 3.3.3 reports the means of reported behaviors for each of the 7 questions comparing treatment and control groups, by survey wave. As already noted in Table 3.3.2, 3 out of 7 reported behaviors (*Teeth*, *Hands*, and *Parents*) appear to be statistically different in the pre-program survey, with the treatment group coming with a higher mean.

Looking at the differences between treatment and control groups in the post-program survey, we find that 4 out of 7 variables show a statistically significant difference, with a higher mean for the treatment group: *Teeth* ($Z = -4.248, p < 0.001$); *Fountain* ($Z = -3.149, p = 0.0016$); *Hands* ($Z = -5.429, p < 0.001$); *Parents* ($Z = -6.115, p < 0.001$) and *Waste* ($Z = -5.284, p < 0.001$). Moreover, 3 of these 4 variables appear to be statistically different also in the post6-program survey: *Teeth* ($Z = -2.587, p = 0.009$); *Hands* ($Z = -5.020, p < 0.001$) and *Parents* ($Z = -3.881, p = 0.001$); in addition, we also find a statistically significant difference for the variable *Shower*, again with a higher mean in the treatment group ($Z = -5.125, p < 0.001$).

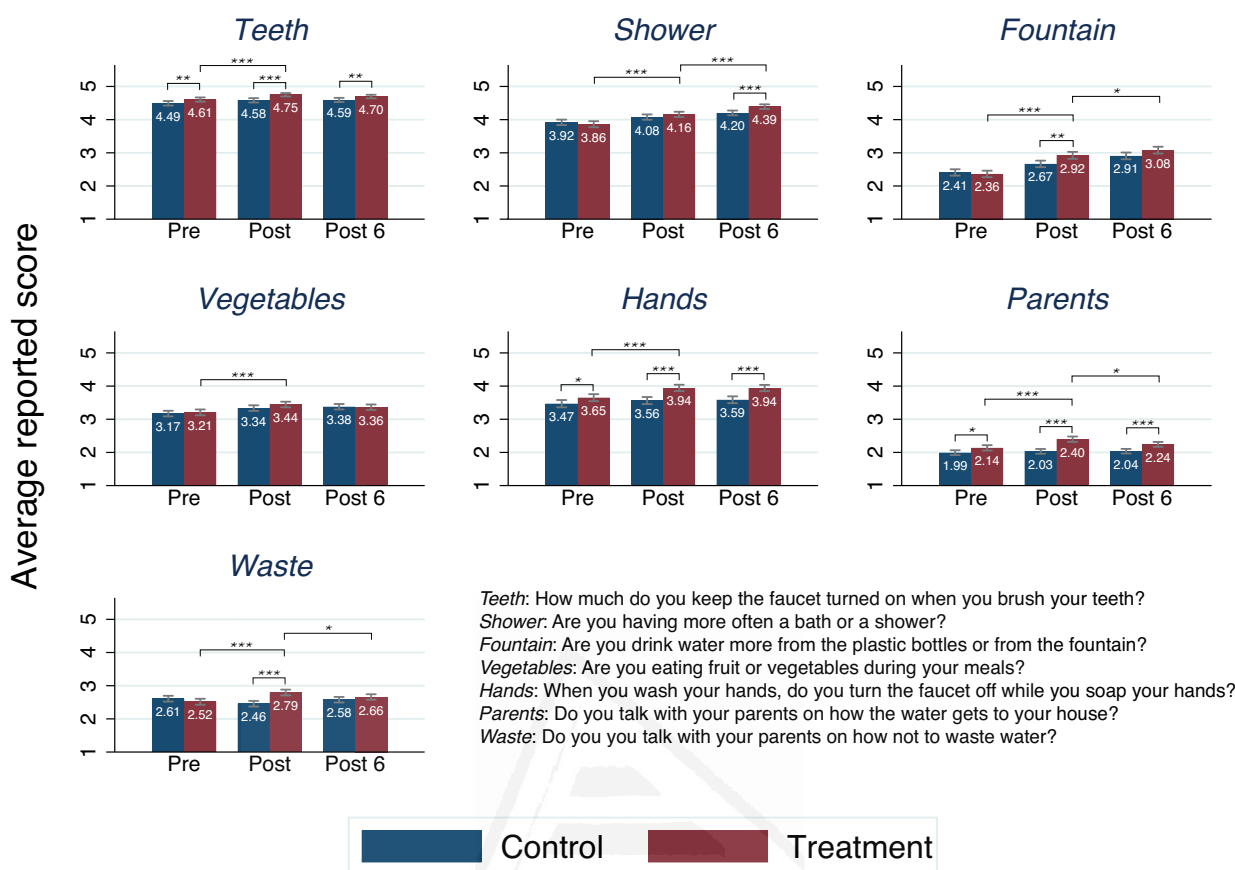


Figure 3.3.3: Average reported behavior by questions, conditions and survey wave. Each answer assume values from 1 to 5. Questions are reported in the figure. Statistically significant difference between conditions are reported above columns (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Error bars represent the 95% confidence interval.

In order to control for potential confounding factors that potentially persisted across the three waves – and which could explain the differences described above – we pool data of the three survey waves and we run ordered logit regressions for each of the 7 variables, also adding the control variables used in the analysis of aggregated reported behavior. In this case we prefer not to use a liner regression models because of the 5-tier ordinal structure of answers.

Figure 3.3.4 reports the estimates of the relevant coefficients of the ordered logit regressions (detailed estimates can be found in Table 3.C.4 in Appendix 3.C). Specifically, the coefficients of interests are those of the interactions between *Treatment* and *Post* (the treatment effect just after the end of the program) and between *Treatment* and *Post 6* (the treatment effect 6 months after the end of the program), whereas the base of reference is the control group in the pre-program survey. According to this analysis the program has had a positive effect on *Fountain*, *Hands*, *Parents* and *Waste*. These effects are still detectable after six months for *Fountain* and *Waste*, when also a

positive treatment effect on *Shower* is found.

These results suggest that the program has had a positive effect especially on two dimensions, namely the habits and behaviors that involve massive or frequent use of water (full body washing, hands washing, drinking) and the discussions with parents about water (from where it comes, how not to waste it), while other dimensions involving more indirect or limited use of water (eating products requiring water to be produced, teeth brushing) seem to have been less affected. Moreover, while the effect on the discussions with parents seems to have faded away towards the end of the program, the effect on the habits and behaviors that involve massive or frequent use of water seems to have persisted beyond the end of the program.

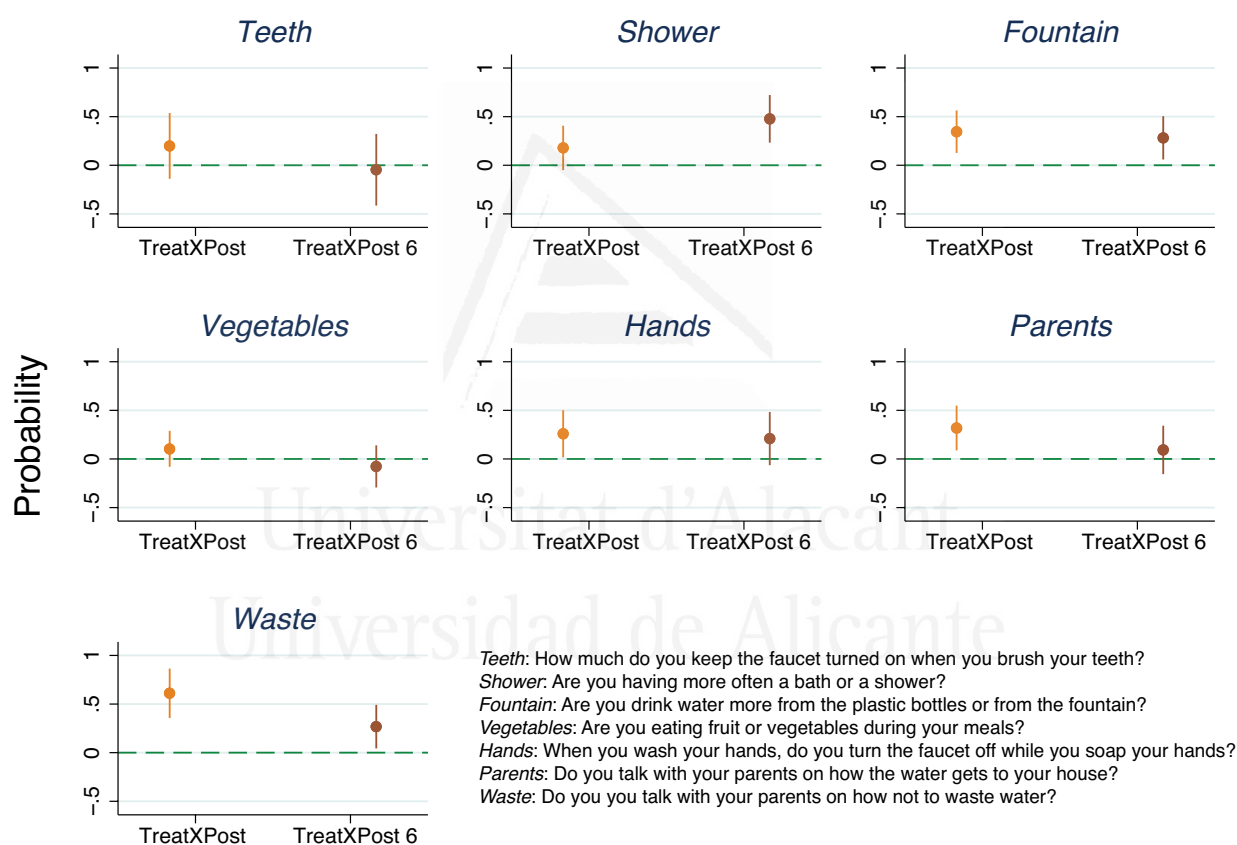


Figure 3.3.4: Estimated coefficient of the Ordered Logit regression in Table 3.C.4 in Appendix 3.C.3. The dependent variables are the 7 questions, which assume values from 1 to 5. Questions are reported in the figure. Error bars represent the 95% confidence interval.

3.4 Discussion

Our results provide field evidence about the effectiveness of promoting sustainable behaviors regarding water consumption by means of game-based educational programs. Our analysis exploited a unique dataset built from a quasi-experiment involving about two thousand Italian students of 2nd-4th grades, all from the same municipality (Lucca, Italy). Specifically, our findings suggest that the program has had positive, sizeable and persistent effects, especially with regard to habits and behaviors that involve massive or frequent use of water (full body washing, water drinking). We believe that such evidence strongly pushes towards a greater consideration of game-based education programs as policy instruments to promote sustainable habits and behaviors, especially when children and their families can be targeted.

It is worth emphasizing that the program had not just provided a chance to play with sustainability-themed games. Instead, the structured ludic activities were designed to engage students in specific settings (at home, at school, during time spent with the family) and this was properly incentivized in terms of the game rewards that materialized over a rather long period of time (several months). The resulting take-home message is that game-based programs aiming at promoting sustainable behaviors should be designed to engage participants in their daily life, for a substantial length of time, and with social activities involving people with whom they have stable relationships.

One important aspect of our results which deserves to be highlighted is that the decline in the treatment effect over the last part of the program is entirely due to an improvement of the reported behavior in the control group, and not to a progressive deterioration of the reported behavior in the treatment group – which actually does not decline. This dynamic could have at least two different sources. One is independent learning by students over the nine months of the program, which might have led students to improve their behaviors over time just through standard channels which have nothing to do with the program and that are common to all classes and schools. Some evidence of this is found in the positive correlation between the 4th grade and virtuous behaviors. If this is the correct explanation, then the program has had in part the effect of accelerating such learning in the first months, implying a deceleration in the last months. Another source of explanation is the presence of peer effects beyond students' own classes, that is, students in the control group might have been exposed indirectly to the program through their social connections outside their own classes. This latter explanation would imply that the treatment effect is far larger than our estimates indicate. With our data we cannot establish which explanation works better. Additional specific data have to be collected for this purpose.

A standard limitation of quasi-experiment is that, since the randomization protocol cannot be managed directly, one cannot conclude about the causal effect of the treatment. We think that such limitation, although not absent, is less severe in our study because the assignment procedure was largely exogenous to students' and teachers' desires, with constraints for eligibility and required participation that left little room for self-selection. Moreover, we could control for systematic differences in the characteristics of control and treatment groups, such as grade, cognitive abilities, class size, and school.

Another limitation of this study is that we could only use self-reported behavior and not directly observe relevant behaviors. It is possible that the observed treatment effect is due by the fact that students learn what is better for the society and then answers accordingly, without having change their behavior. Unfortunately, it turned out that the observation of direct water consumption by the families involved in the program was unfeasible, mostly due to the absence of a reliable way to collect these data, either from the local water utility or from the families themselves.

Perhaps the most important limitation of this study is the fact that we were not allowed to connect individual response in the three surveys for the control group, where we are forced to take their surveys anonymous (while we could do so for the treatment group since the names of the students were public). This has forced us to rely on class averages to get a longitudinal structure of the data, greatly reducing the statistical power and necessarily limiting the scope of our analysis (e.g., we could not properly exploit individual characteristics). We cannot do much in this regard if not stressing that such information should be made a priority in future studies.

Starting from the results of this study there are at least three avenues of future research that seem promising. Firstly, one may dig into the collected data regarding ludic habits and preferences to see whether these modulate the effects of the program, and whether they are affected by the participation in the program. Ludic habits and preferences are important for students' wellbeing and life-long learning. Secondly, one may want to run follow-up field experiments with the aim of observing actual behavior regarding water use. This can only be done with a substantial smaller number of students, but full randomization is likely to be more easily implementable in such a case. Lastly, one may want to run similar studies employing game-based educational programs aimed at promoting different sustainable behaviors and habits, such as waste production, recycling, and energy consumption, in order to check to what extent our results can be generalized.

Bibliography

- ALAN, Ü. AND KABASAKAL, K. A. (2020): “Effect of number of response options on the psychometric properties of Likert-type scales used with children,” *Studies in Educational Evaluation*, 66, 100895.
- BEVANS, K. B., AHUVIA, I. L., HALLOCK, T. M. ET AL. (2020): “Investigating child self-report capacity: a systematic review and utility analysis,” *Quality of Life Research*, 29, 1147–1158.
- BLAKE, P. R., PIOVESAN, M., MONTINARI, N. ET AL. (2015): “Prosocial norms in the classroom: The role of self-regulation in following norms of giving,” *Journal of Economic Behavior & Organization*, 115, 18–29.
- BUCCIOL, A. AND PIOVESAN, M. (2011): “Luck or cheating? A field experiment on honesty with children,” *Journal of Economic Psychology*, 32, 73 – 78.
- CAMPBELL, D. T. AND STANLEY, J. C. (2015): *Experimental and quasi-experimental designs for research*, Ravenio Books.
- CAPRARA, G. V., BARBARANELLI, C., PASTORELLI, C. ET AL. (2000): “Prosocial foundations of children’s academic achievement,” *Psychological science*, 11, 302–306.
- COBO-REYES, R., DOMINGUEZ, J. J., GARCÍA-QUERO, F. ET AL. (2020): “The development of social preferences,” *Journal of Economic Behavior & Organization*, 179, 653–666.
- COHN, A. AND MARÉCHAL, M. A. (2016): “Priming in economics,” *Current Opinion in Psychology*, 12, 17–21.
- COOK, T. D., CAMPBELL, D. T. AND DAY, A. (1979): *Quasi-experimentation: Design & analysis issues for field settings*, vol. 351, Houghton Mifflin Boston.
- COPPLE, C., SIGEL, I. E. AND SAUNDERS, R. (2013): *Educating the young thinker: Classroom strategies for cognitive growth*, Routledge.
- CUADRADO, E., TABERNERO, C., GARCÍA, R. ET AL. (2017): “The role of prosocialness and trust in the consumption of water as a limited resource,” *Frontiers in Psychology*, 8, 694.
- DANIELSON, C. K. AND PHELPS, C. R. (2003): “The assessment of children’s social skills through self-report: A potential screening instrument for classroom use,” *Measurement and Evaluation in Counseling and Development*, 35, 218–229.

- DI RISO, D., SALCUNI, S., CHESSA, D. ET AL. (2010): “The Strengths and Difficulties Questionnaire (SDQ). Early evidence of its reliability and validity in a community sample of Italian children,” *Personality and Individual Differences*, 49, 570–575.
- FUNTOWICZ, S. O. AND RAVETZ, J. R. (1995): “Science for the post normal age,” in *Perspectives on ecological integrity*, Springer, 146–161.
- GILL, D. AND PROWSE, V. (2016): “Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis,” *Journal of Political Economy*, 124, 1619–1676.
- GREVE, P., KAHIL, T., MOCHIZUKI, J. ET AL. (2018): “Global assessment of water challenges under uncertainty in water scarcity projections,” *Nature Sustainability*, 1, 486–494.
- HARDIN, G. (1968): “The tragedy of the commons,” *science*, 162, 1243–1248.
- HOUSE, B. R., KANNGIESSER, P., BARRETT, H. C. ET AL. (2020): “Universal norm psychology leads to societal diversity in prosocial behaviour and development,” *Nature Human Behaviour*, 4, 36–44.
- HOUSE, B. R. AND TOMASELLO, M. (2018): “Modeling social norms increasingly influences costly sharing in middle childhood,” *Journal of experimental child psychology*, 171, 84–98.
- IMBENS, G. W. AND WOOLDRIDGE, J. M. (2009): “Recent developments in the econometrics of program evaluation,” *Journal of economic literature*, 47, 5–86.
- KUMMU, M., GUILLAUME, J. H., DE MOEL, H. ET AL. (2016): “The world’s road to water scarcity: shortage and stress in the 20th century and pathways towards sustainability,” *Scientific reports*, 6, 1–16.
- LIU, J., YANG, H., GOSLING, S. N. ET AL. (2017): “Water scarcity assessments in the past, present, and future,” *Earth’s future*, 5, 545–559.
- MEYER, B. D. (1995): “Natural and quasi-experiments in economics,” *Journal of business & economic statistics*, 13, 151–161.
- MUSSEN, P. AND EISENBERG-BERG, N. (1977): *Roots of caring, sharing, and helping: The development of pro-social behavior in children.*, WH Freeman.
- NILES, M. T., LUBELL, M. AND HADEN, V. R. (2013): “Perceptions and responses to climate policy risks among California farmers,” *Global Environmental Change*, 23, 1752–1760.

- ORLICK, T. (1983): “Enhancing love and life mostly through play and games.” *Journal of Humanistic Counseling, Education & Development*.
- QIN, Y., MUELLER, N. D., SIEBERT, S. ET AL. (2019): “Flexibility and intensity of global water use,” *Nature Sustainability*, 2, 515–523.
- SCHULTZ, P. W., MESSINA, A., TRONU, G. ET AL. (2016): “Personalized normative feedback and the moderating role of personal norms: A field experiment to reduce residential water consumption,” *Environment and Behavior*, 48, 686–710.
- WADA, Y. AND BIERKENS, M. F. (2014): “Sustainability of global water use: past reconstruction and future projections,” *Environmental Research Letters*, 9, 104003.



Universitat d'Alacant
Universidad de Alicante

Appendix

3.A Material

The material used and the activities in the quasi-natural experiment are summarized by the pictures below.

3.A.1 Informed consent

The informed consent form is shown below. It explains the purpose of the program and the possibility to collect information about the children. Parents have to sign the authorization to participate.

GEAL
NUOVA VITA PER L'ACQUA.

LIBERATORIA ALUNNI MINORENNI

I sottoscritti _____ e _____, genitori dell'alunno
_____ della classe _____ dell'istituto scolastico _____ con la presente:

AUTORIZZANO

Nell'ambito delle attività del progetto di Blu Tube di GEAL SpA a riprendere l'alunno/a con fotocamere e/o videocamere nei vari momenti dell'attività scolastica in occasione degli interventi in classe nonché delle attività previste di tornei extra scolastiche e in occasione di gite e visite d'istruzione, da solo, con i compagni, con insegnanti ed operatori ai soli fini di:

- Formazione, ricerca e documentazione dell'attività didattica
- Divulgazione della ricerca didattica e delle esperienze didattiche effettuate sotto forma di documento in convegni, sito internet e altri ambiti di studio.

AUTORIZZANO

A pubblicare eventuali foto, video e produzioni personali dell'alunno derivanti dallo svolgimento delle attività didattiche del progetto, sul sito internet del progetto e degli enti organizzatori e sui quotidiani. Tale autorizzazione deve ritenersi valida per l'intera durata del progetto.

Lucca, li _____ Firma di entrambi i genitori _____

In applicazione del GDPR UE 679/16 e del D.Lgs. 196/2003 così come modificato dal D.Lgs. 101/18, i dati personali sono trattati in modo lecito, secondo correttezza e con adozione di idonee misure di protezione relativamente: all'ambiente in cui vengono custoditi, al sistema adottato per elaborarli, ai soggetti incaricati del trattamento. Titolare del Trattamento dei dati è GEAL SpA, nella figura del suo rappresentante legale. I dati in nessun caso vengono comunicati a soggetti privati senza il preventivo consenso scritto dell'interessato. Al soggetto interessato sono riconosciuti il diritto di accesso ai dati personali e gli altri diritti definiti dalla normativa. Il sottoscritto, ricevuta l'informativa di cui all'art. 13 del GDPR UE 679/16, esprime il proprio consenso affinché i dati personali forniti con la presente richiesta possano essere trattati nel rispetto della normativa vigente per gli adempimenti connessi alla presente procedura.

Firma di entrambi i genitori _____

Figure 3.A.1: Informed consent signed by parents

English translation:

“The undersigned parents _____ of the student _____ from the class _____ of the school _____, hereby: They authorize, as part of the activities of the Blu Tube program of GEAL SpA to record the student with cameras and/or video-cameras in different moments of the school activities, during the attendance in the classroom as well as during the activities planned for extracurricular tournaments and in the occasion of trips and educational visits, alone, in company, with teachers and with the educators for the sole purpose of:

- Training, research and documentation of the teaching activities;
- Dissemination of didactic research and didactic experiences carried out in the form of a document in conferences, website and other fields of study.

They authorize to publish any photos, videos and personal productions of the student deriving from the carrying out of the project’s educational activities on the website of the project and of the organizing bodies and in newspapers. This authorization must be considered valid for the entire duration of the project.”

3.A.2 The board game *Blutube*

The *Blutube* board game box was delivered to all children in the classes participating actively to the program. Each box contains: 58 water network cards, 28 city cards, 20 pawns, a scoring board, 8 water loss cards and the game instructions.



Figure 3.A.2: The *Blutube* board game

The material also includes a flier in which are described all the activities and the relative scores that children can obtain during the phase two of the *BLUTUBE* program.



(a) Front



(b) Back

Figure 3.A.3: Descriptive flier of the *BLUTUBE* activities and relative scores

3.A.3 The Survey

Educators received the surveys in a closed envelope marked with the class number. The first page in each envelope includes the instruction for the educators to ensure the same protocol in each class. The content of the envelope is showed below.

ISTRUZIONI PER LA SOMMINISTRAZIONE DEL QUESTIONARIO

**Leggere questo questa pagina per intero
prima di distribuire il questionario agli studenti**

Procedura da seguire per la somministrazione:

- (1) Consegnare 1 questionario a ciascuno studente (in qualunque ordine), chiarendo che si dovrà scrivere solo sul foglio consegnato
- (2) Verificare che ogni studente scriva con la penna, NON con la matita
- (3) Quando tutti gli studenti sono pronti a compilare il proprio questionario, leggere a voce alta il testo del questionario allegato a queste istruzioni e chiedere di dare una risposta subito dopo ogni domanda
- (4) Tra la lettura di una domanda e la lettura della domanda seguente, lasciare il tempo agli studenti di rispondere, controllando che non comunichino tra di loro e che non guardino le risposte degli altri
- (5) Per le domande della prima parte (tutte tranne gli indovinelli) cercare di non superare un'attesa di circa 30 secondi per risposta
- (6) Per le domande della seconda parte (gli indovinelli logico-matematici) cercare di non superare un'attesa di 3-4 minuti per risposta
- (7) Dare spiegazioni solo se esplicitamente richiesto, EVITANDO SEMPRE di indicare una risposta come corretta o sbagliata; eventualmente dire: "Rispondi nel modo che pare giusto a te", "Rispondi per come hai capito tu", "Rispondi secondo la tua interpretazione delle parole"
- (8) **IMPORTANTE:** gli indovinelli della seconda parte sono tarati per avere solo pochi studenti che rispondono correttamente e molti altri studenti (anche più del 50%) che non rispondano correttamente, soprattutto delle classi II e III; è importante non dare tempo aggiuntivo e non aiutare gli studenti che non sanno rispondere, altrimenti i dati potrebbero risultare inutilizzabili
- (9) Finita la compilazione, ritirare i questionari e metterli nella busta da cui sono stati presi, assieme alle presenti istruzioni
- (10) EVITARE che si facciano foto del questionario o che se ne copino delle parti altrove

GRAZIE MILLE PER LA COLLABORAZIONE !

Figure 3.A.4: Instructions given to the educators

NOME: INIZIALE COGNOME:

Leggi le domande. Rispondi mettendo una crocetta con la penna sul simbolo che si trova vicino alla risposta che vuoi dare. Quando trovi i puntini scrivi direttamente la tua risposta sui puntini.

Per quanto tempo fai scorrere l'acqua dal rubinetto quando ti lavi i denti?
 TUTTO IL TEMPO
 PIU' DELLA META' DEL TEMPO
 CIRCA META' DEL TEMPO
 MENO DELLA META' DEL TEMPO
 SOLO IL TEMPO NECESSARIO

Fai più spesso il bagno o la doccia?
 SEMPRE IL BAGNO
 PIU' VOLTE IL BAGNO CHE LA DOCCIA
 BAGNO E DOCCIA UGUALMENTE SPESSO
 PIU' LA DOCCIA CHE IL BAGNO
 SEMPRE LA DOCCIA

Bevi più spesso l'acqua dalle bottiglie acquistate o l'acqua presa dal rubinetto o fontana?
 SEMPRE IN BOTTIGLIA
 PIU' IN BOTTIGLIA CHE DAL RUBINETTO O FONTANA
 IN BOTTIGLIA E DAL RUBINETTO O FONTANA UGUALMENTE SPESSO
 PIU' DAL RUBINETTO O FONTANA CHE IN BOTTIGLIA
 SEMPRE DAL RUBINETTO O FONTANA

Mangi frutta o verdura durante i tuoi pasti?
 SI, SEMPRE
 SI, IL PIU' DELLE VOLTE
 CIRCA LA META' DELLE VOLTE
 SI MA POCHE VOLTE
 NO, MAI

Quando lavi le mani col sapone chiudi l'acqua del rubinetto mentre ti insaponi?
 SI, SEMPRE
 SI, IL PIU' DELLE VOLTE
 CIRCA LA META' DELLE VOLTE
 SI MA POCHE VOLTE
 NO, MAI

Parli con i tuoi genitori dell'acqua e di come fa ad arrivare a casa tua?
 SI, CONTINUAMENTE
 SI, SPESSO
 SI, QUALCHE VOLTA
 MI E' CAPITATO DI FARLO
 NO, NON L'HO MAI FATTO

Parli con i tuoi genitori di come non sprecare l'acqua?
 SI, CONTINUAMENTE
 SI, SPESSO
 SI, QUALCHE VOLTA
 MI E' CAPITATO DI FARLO
 NO, NON L'HO MAI FATTO

Quanti amici del cuore hai?
 0 1 2 3 4 5 più di 5

Quanti dei tuoi amici del cuore sono in classe con te?
 0 1 2 3 4 5 più di 5

I tuoi amici sono più maschi o più femmine?
 MASCHI FEMMINE META' E META'

Quanto spesso giochi con i tuoi amici dopo la scuola?
 TUTTI I GIORNI
 ALMENO UNA VOLTA ALLA SETTIMANA
 ALMENO UNA VOLTA AL MESE
 QUASI MAI
 MAI

Quanto spesso giochi con i tuoi amici nei giorni in cui non vai a scuola?
 TUTTI I GIORNI
 ALMENO UNA VOLTA ALLA SETTIMANA
 ALMENO UNA VOLTA AL MESE
 QUASI MAI
 MAI

Quanto ti serve aiuto per fare qualcosa chiedi ai tuoi amici di aiutarti?
 SEMPRE
 LA MAGGIOR PARTE DELLE VOLTE
 QUALCHE VOLTA
 RARAMENTE
 MAI

Quanto spesso giochi su cellulare, tablet, computer o console?
 TUTTI I GIORNI
 ALMENO UNA VOLTA ALLA SETTIMANA
 ALMENO UNA VOLTA AL MESE
 QUASI MAI
 MAI

Come giochi su cellulare, tablet, computer o console?
 SOPRATTUTTO DA SOLO SPESSO INSIEME AD ALTRI

Quanto spesso giochi a giochi da tavolo o ai giochi di carte?
 TUTTI I GIORNI
 ALMENO UNA VOLTA ALLA SETTIMANA
 ALMENO UNA VOLTA AL MESE
 QUASI MAI
 MAI

Giochi più spesso a:
 GIOCHI IN CUI DEVO PENSARE GIOCHI IN CUI DEVO MUOVERMI

Preferisci:
 GIOCHI IN CUI DEVO PENSARE GIOCHI IN CUI DEVO MUOVERMI


Giochi con più piacere a giochi in cui:
 DEVO SCONFIGGERE GLI ALTRI
 DEVO COLLEZIONARE OBIETTIVI
 DEVO AIUTARE GLI ALTRI
 DEVO SCOPRIRE COSE

Quanto spesso fai sport?
 TUTTI I GIORNI
 ALMENO UNA VOLTA ALLA SETTIMANA
 ALMENO UNA VOLTA AL MESE
 QUASI MAI
 MAI

Scrivi i nomi dei 3 giochi a cui giochi più spesso:
 -
 -
 -


Quanto è lungo il filo? 5 6 7 8 9

Un coniglio e una rana cominciano a saltare nello stesso momento. Il salto del coniglio è lungo il doppio di quello della rana. Ogni volta che il coniglio fa un salto, anche la rana fa un salto.



In quale casella si incontrano? F G H I L

Osserva la linea dei numeri.





Quale numero va scritto nella casella vuota?


Nel negozio di vernici vendono dei contenitori da 5 litri. Marco ha bisogno di 37 litri di vernice. Quanti ne deve comprare? 5 6 7 8

INDOVINELLI
 Leggi le domande e cerca di rispondere all'indovinello. Non ti preoccupare se non riesci a trovare la soluzione corretta, prova comunque a trovare una soluzione.

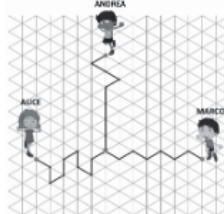
Lisa ha due tipi di caramelle. Quante caramelle ha Lisa di ogni tipo?



 NUMERO:

 NUMERO:

Alice, Marco e Andrea giocano alla caccia al tesoro. Chi fa il percorso più lungo per arrivare al tesoro (indicato con una X)?



ALICE MARCO ANDREA

Figure 3.A.5: Original questionnaire

3.A.4 Participants and ranking

Students from the 2nd-4th grades of 53 classes in the primary schools in Lucca participated to the *BLUTUBE* program. Figures below show some students with their personal board game.



Figure 3.A.6: Primary students participating to the *BLUTUBE* program

During the second phase, the participating students accumulate points with the activities described in Section 3.2. Each week, both individual and class scores were published in a dedicated website and in the local newspapers. Figures below show an example of the published ranking.

Nome	Classe	Scuola	Istituto	Gioco Casa	Gioco Scuola	L'acqua nascosta	Luoghi	Punteggio
Clarissa G.	3B	Donatelli S. Vito	Lucca 6	2500	870	5000	2500	10870
Noemi P.	2	Montuolo	Lucca 7	730	150	5000	2500	8380
Giorgio M.	4	Nave	Lucca 7	50	770	5000	2500	8320
Elisa P.	2	Montuolo	Lucca 7	660	120	5000	2500	8280
Yara A.	2	Montuolo	Lucca 7	190	120	5000	2500	7810
Nathan F.	2B	Pascoli	Lucca 1	200	40	5000	2500	7740
Andrea P.	4	S. Maria del giudice	Lucca 2	0	800	5000	1800	7600
Gloria D.	3C	Pascoli	Lucca 1	40	280	5000	1500	6820
Agnese B.	4	Nave	Lucca 7	10	310	3400	2500	6220
Giulio N.	3	Montuolo	Lucca 7	60	50	3050	2500	5660

(a) Online

CLASSIFICA INDIVIDUALE			CLASSIFICA PER CLASSE			
Nome	Classe	Scuola	Punteggio	Classe	Scuola	Punteggio
Clarissa G.	3B	Donatelli S. Vito	8130	4	Nave	25710
Noemi P.	2	Montuolo	7860	2.	Montuolo	20820
Elisa P.	2	Montuolo	7850	4	S. Maria del Giudice	19180
Giorgio M.	4	Nave	7450	4	Sant'Alessio	16970
Andrea P.	4	S. Maria del giudice	6000	3A	San Vito	14570
Agnese B.	4	Nave	5210	3	Montuolo	14080
Aurora P.	3	Montuolo	5200	3B	San Vito	13890
Giulio N.	3	Montuolo	4310	2B	Ponte a Moriano	8280
Martina M.	4	Nave	3950	3	Antraccoli	8040
Yara A.	2	Montuolo	3900	3A	S. Angelo	7880

(b) Local newspaper

Figure 3.A.7: Public ranking

3.B Translated Survey

Name: _____ Beginning Last Name: _____

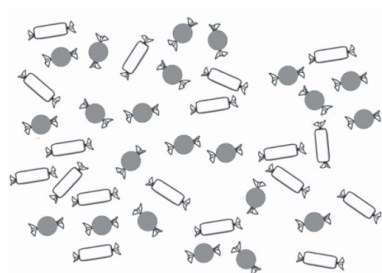
Read the following questions. Answer to them putting a cross with your pen on the symbol which is close to the answer. When you find the . . . , write directly you answer.

- How much time do you keep the faucet turned on when you brush your teeth?
 - All the time – More than half of the time – Half of the time – Less than half of the time
 - Only the time needed
- Are you having more often a bath or a shower?
 - Always a bath – More often a bath than a shower – A bath and a shower equally – More often a shower than a bath – Always a shower
- Do you drink more from the plastic bottles or from the fountain/faucet?
 - Always from the plastic bottles – More from the plastic bottles than from the fountain/faucet – From the plastic bottles and from the fountain/faucet equally – More from the fountain/faucet than from the plastic bottle – Always from the fountain/faucet
- Are you eating fruit or vegetables during your meals?
 - Yes, always – Yes, most of the times – Half of the times – Yes, but few times – No, never
- When you wash your hands, do you turn the faucet off while you soap your hands?
 - Yes, always – Yes, most of the times – Half of the times – Yes, but few times – No, never
- Do you talk with your parents on how the water gets to your house?
 - Yes, constantly – Yes, often – Yes, sometimes – I happened to do it – No, I never do it
- Do you you talk with your parents on how not to waste water?
 - Yes, constantly – Yes, often – Yes, sometimes – I happened to do it – No, I never do it
- How many best friends do you have?
 - 0 – 1 – 2 – 3 – 4 – 5 – More than 5
- How many best friends do you have in your class?
 - 0 – 1 – 2 – 3 – 4 – 5 – More than 5

- Your best friends are more males or females?
 - Male – Female – Half and Half
- How many time do you play with your friends after school?
 - Every day – At least once a week – At leas once a month – Hardly ever – Never
- How many time do you play with your friends when you do not go to school?
 - Every day – At least once a week – At leas once a month – Hardly ever – Never
- Do you ask your friends' help when you need it?
 - Always – Most of the times – Sometimes – Rarely – Never
- How much do you play at mobile phone, tablet, computer or console?
 - Every day – At least once a week – At leas once a month – Hardly ever – Never
- How do play on mobile phone, tablet, computer or console?
 - Alone – With others
- How much do you play at board games?
 - Every day – At least once a week – At leas once a month – Hardly ever – Never
- I often play:
 - games where I have to think – games where I have to move
- I prefer:
 - games where I have to think – games where I have to move
- I play with more pleasure in games in which:
 - I must defeat others – I must collect goals – I must help others – I must discover things
- How many times do you play sport?
 - Every day – At least once a week – At leas once a month – Hardly ever – Never
- Write the names of three games you play often.

RIDDLES: Read the questions and answer the riddle. Don't worry if you cannot find the correct answer, in all cases, try to find a solution.

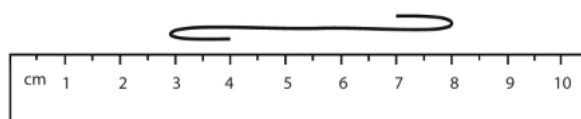
- Lisa has 2 types of candies. How many candies of each type has Lisa?



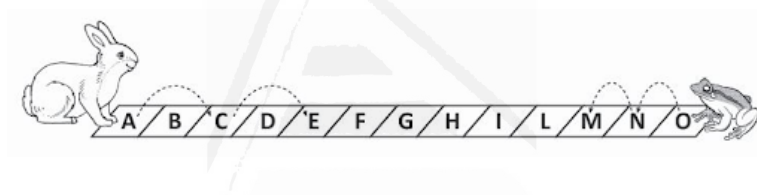
– How many are white? ____

– How many are black? ____

- How long is the thread? – 5 – 6 – 7 – 8 – 9

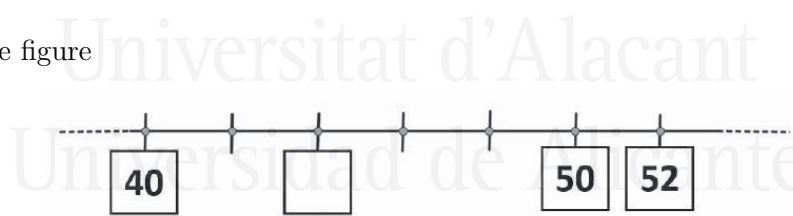


- A rabbit and a frog start to jump at the same moment. The rabbit's jump is twice as long as that of the frog. Each time that the rabbit jumps, the frog jumps too.



In which box they will meet? – F – G – H – I – L

- Look at the figure

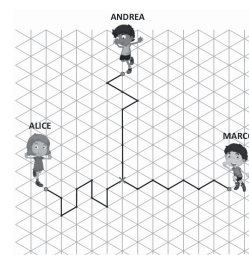


Which number is going into the empty box? ____

- In a paint shop they sell 5-liter containers. Matteo needs 37 liters of paint. How many containers he has to buy? – 5 – 6 – 7 – 8

- Alice, Marco and Andrea are playing treasure hunt. Who is going to do the longer path to arrive at the treasure?

– Alice – Marco – Andrea



3.C Supplementary Analysis

3.C.1 Robustness check

We next conduct some robustness checks in which we explore if our findings are robust using different samples of our data.

First of all, we resample the treatment and control group to ensure that their are balanced by grade, number of students and cognitive skills. Table 3.C.1 reports the summary statistics by treatment and control groups for the balance randomly chosen sample in the pre-program survey. In this case, figures show that the sample is balanced between treatment and control group, specifically: no differences are found in the distribution of grades ($Z = 1.369, p = .171$), in the number of students per class ($Z = -0.551, p = .581$) and in the measured students' cognitive skills ($Z = -0.948, p = .343$). Importantly, we do not find any difference in the aggregated reported behavior between the control and treatment groups ($Z = -0.638, p = .523$).

Table 3.C.1: Mean difference of independent samples in the pre-program sample

Variable	Control	Treatment	Min.	Max.	p-value
Grade	3.07	2.99	2	4	.171
Students	16.56	16.70	9	25	.581
Cognitive skills	0.516	0.567	0	1	.343
Aggregate reported behavior	22.03	22.32	9	34	.523
• <i>Teeth</i>	4.49	4.59	1	5	.056
• <i>Shower</i>	3.92	3.85	1	5	.609
• <i>Fountain</i>	2.38	2.36	1	5	.855
• <i>Vegetables</i>	3.18	3.21	1	5	.809
• <i>Hands</i>	3.47	3.63	1	5	.314
• <i>Parents</i>	1.97	2.13	1	5	.069
• <i>Waste</i>	2.59	2.52	1	5	.508

Descriptive Statistics. *Grade* is the students' year group. *Students* is the number of students in each class. *Cognitive Skills* is equal to 1 if the result obtained in the logical and mathematical questions are higher than the median, 0 otherwise. *Aggregate reported behavior* is the average sum of the first seven questions of the questionnaire. *Teeth*: "How much do you keep the faucet turned on when you brush your teeth?". *Shower*: "Are you having more often a bath or a shower?". *Fountain*: "Do you drink water more from the plastic bottles or from the fountain/faucet?". *Vegetables* "Are you eating fruit or vegetables during your meals?". *Hands*: "When you wash your hands, do you turn the faucet off while you soap your hands?". *Parents*: "Do you talk with your parents on how the water gets to your house?". *Waste*: "Do you talk with your parents on how not to waste water?". The Mann-Whitney test of the equality of the means is reported in the last column.

Figure 3.3.2 shows the means of the aggregated reported behavior by conditions and survey waves for the balance randomly chosen sample. The findings support our results. Specifically, no statistically significant difference is found for the pre-program survey ($Z = -0.638, p = 0.523$). On the other hand, for the post-program we find that the treatment group has a statistically higher

average of about 2.13 with respect to the pre-program treatment group ($Z = -5.590, p < 0.001$) and a statistically significant higher average of about 1.71 with respect to the post-treatment control group ($Z = -4.342, p < 0.001$). Furthermore, we find a statistically significant increase in the aggregated reported behavior in the post6-program for the treatment group ($Z = -3.229, p < 0.001$), while no appreciable difference is found in the treatment group between the post-program survey and the post6-program survey ($Z = 0.156, p = 0.870$).

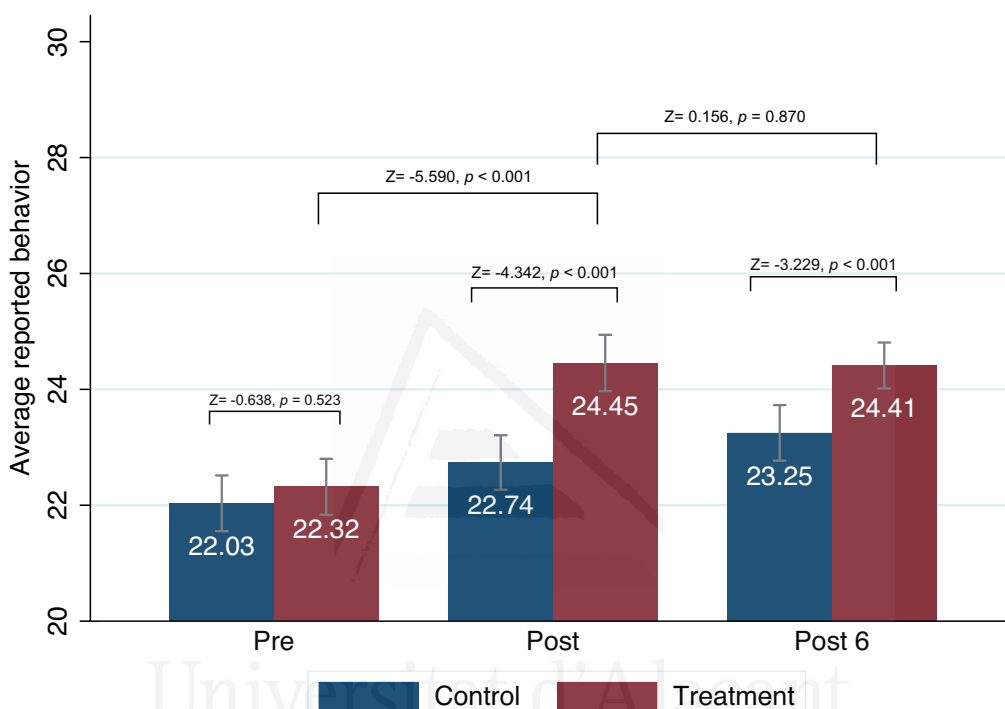


Figure 3.C.1: Average of the aggregated reported behavior by conditions and waves for a balance subsample. In the pre-program period, the aggregated reported behavior in the treatment group is not significantly different from the control group (Mann-Whitney test, $Z = -0.638, p = 0.523$). In the post-program period and after 6 months, the aggregated reported behavior in the treatment group is significantly higher respect to the control group (Mann-Whitney test, $Z = -4.342, p < 0.001$ and $Z = -3.229, p < 0.001$, respectively). The treatment effect is stable after 6 month (Mann-Whitney test, $Z = 0.156, p = 0.870$). Error bars represents the 95% confidence interval.

Secondly, given that 2nd grade students are fairly-balanced, we run the same non-parametric analysis to see the effects across control and treatment groups. Our sample is reduced to 1673 observations (818 students in the control group and 855 students in the treatment group).

Figure 3.C.2 shows the means of the average reported behavior by conditions and survey waves 2nd grade students. In this case, the findings weakly support our results. Specifically, a weakly statistically significant difference is found for the pre-program survey ($Z = -2.365, p = 0.018$). However, for the post-program we find a higher statistical difference between the treatment group

and the control group ($Z = -4.509, p < 0.001$). This difference remains statistically significant higher after 6 months from the end of the program ($Z = -3.138, p < 0.001$). Furthermore, we find a statistically significant increase between the aggregated reported behavior in the treatment groups between the pre-program and post-program surveys ($Z = -4.631, p < 0.001$), while no appreciable difference is found between the post-program survey and the post6-program survey ($Z = 0.112, p = 0.910$).

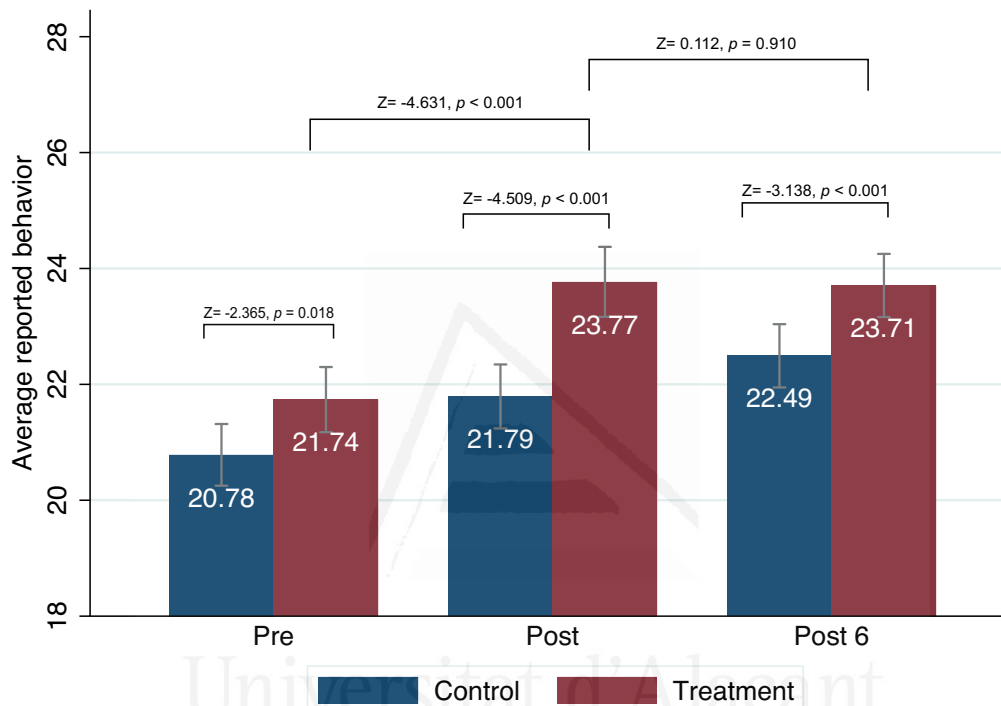


Figure 3.C.2: Average of the aggregated reported behavior by conditions and waves for 2nd grade students. In the pre-program period, the aggregated reported behavior in the treatment group is weakly significantly different from the control group (Mann-Whitney test, $Z = -2.365, p = 0.018$). In the post-program period and after 6 months, the aggregated reported behavior in the treatment group is significantly higher respect to the control group (Mann-Whitney test, $Z = -4.509, p < 0.001$ and $Z = -3.138, p < 0.001$, respectively). The treatment effect is stable after 6 month (Mann-Whitney test, $Z = 0.112, p = 0.910$). Error bars represents the 95% confidence interval.

3.C.2 Principal Component Analysis

Table 3.C.2 reports the pairwise correlation matrix of the 7 questions in the pre-program survey wave. Looking at the correlation coefficient, most of the pairwise result lowly correlated. Only in the case of *Parents* with *Waste* the correlation is significantly higher ($\rho = 0.439; p < 0.001$).

According with the principal component analysis, Table 3.C.3 reports the eigenvalues of the 7 components, the percentage of the variance explained by each component and the cumulative

Table 3.C.2: Correlation coefficient between the 7 questions in the pre-program survey wave

Variables	Teeth	Shower	Fountain	Vegetables	Hands	Parents	Waste
Teeth	1.000						
Shower	0.091***	1.000					
Fountain	0.044	0.050*	1.000				
Vegetables	0.004	0.041	0.063**	1.000			
Hands	0.127***	0.008	0.045	0.165***	1.000		
Parents	0.043	-0.042	0.043	0.166***	0.215***	1.000	
Waste	0.087***	0.030	0.031	0.172***	0.288***	0.439***	1.000

Pairwise correlation matrix between the 7 questions in the pre-program survey wave. *Teeth*: “How much do you keep the faucet turned on when you brush your teeth?”. *Shower*: “Are you having more often a bath or a shower?”. *Fountain*: “Do you drink water more from the plastic bottles or from the fountain/faucet?”. *Vegetables*: “Are you eating fruit or vegetables during your meals?”. *Hands*: “When you wash your hands, do you turn the faucet off while you soap your hands?”. *Parents*: “Do you talk with your parents on how the water gets to your house?”. *Waste*: “Do you talk with your parents on how not to waste water?”. *** Significant correlation at 0.001 (both sides); ** Significant correlation at 0.01 (both sides). * Significant correlation at 0.05 (both sides).

contribution rate. The cumulative percentage of the variance of the first 5 principal component is 81%.

Table 3.C.3: Principal Component analysis

Component	Eigenvalue	% of Variance	Cum. % of Var.
<i>Comp1</i>	1.7863	0.2552	0.2552
<i>Comp2</i>	1.1159	0.1594	0.4146
<i>Comp3</i>	.99145	0.1416	0.5562
<i>Comp4</i>	.93436	0.1335	0.6897
<i>Comp5</i>	.85916	0.1227	0.8125
<i>Comp6</i>	.76678	0.1095	0.9220
<i>Comp7</i>	.54596	0.0780	1.0000
N.Obs			1685

Principal component analysis of the data in the pre-program survey wave.

Figure 3.C.3 reports the cumulative percentage of the explained variance for all the principal components. As shown in the figure, the first 5 components explain about the 81% of the total variance of the data in the pre-program survey wave.

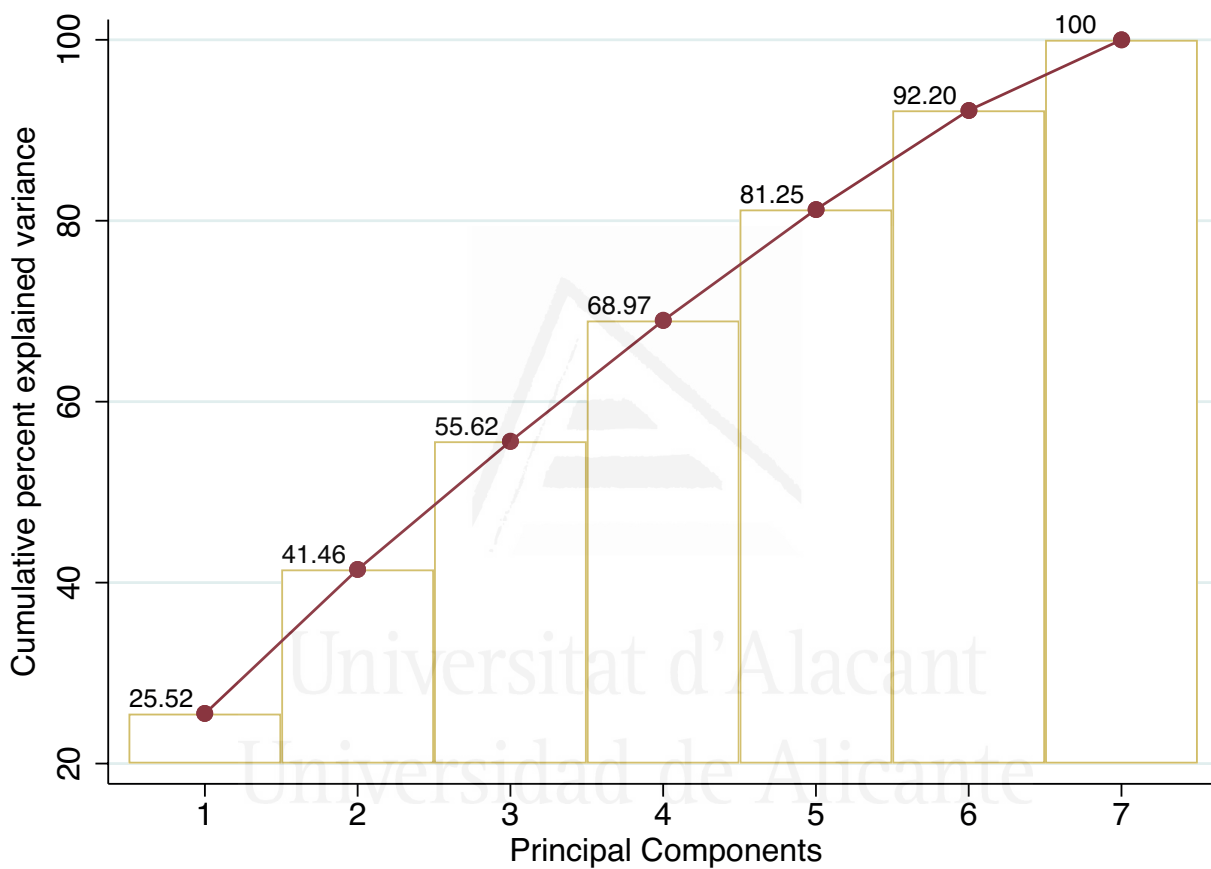


Figure 3.C.3: Cumulative explained variance in the principal component analysis. The first 5 components explain about the 81% of the variance of the data in the pre-program survey wave.

3.C.3 Ordered Logit Regression Analysis

Table 3.C.4 reports the ordered logit regression results for each of the 7 questions which form the aggregated reported behavior analyzed in Section 3.3.1. The interaction between *Treatment* and *Post* shows the effect of the treatment in the post-program survey wave respect to the pre-program. Results show that the treatment had a positive significant effect on *Fountain*, *Hands*, *Parents* and *Waste*. These effects are still visible after 6 months from the end of the project (the interaction between *Treatment* and *Post 6*) for *Fountain* and *Waste*. Moreover, there is a positive and significant effect of the treatment on *Shower* in the long run period.

Table 3.C.4: Ordered Logit regression

	Teeth		Shower		Fountain		Vegetables		Hands		Parents		Waste	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Treatment	0.363** (0.137)	0.156 (0.203)	0.019 (0.114)	-0.009 (0.142)	-0.073 (0.115)	-0.014 (0.175)	0.060 (0.113)	0.174 (0.172)	0.217 (0.143)	0.360 (0.230)	0.212 (0.130)	0.581** (0.221)	-0.156 (0.126)	-0.125 (0.204)
Post	0.264** (0.096)	0.270** (0.095)	0.272*** (0.078)	0.285*** (0.080)	0.317*** (0.075)	0.343*** (0.073)	0.235*** (0.060)	0.248*** (0.060)	0.086 (0.079)	0.082 (0.079)	0.106 (0.078)	0.105 (0.080)	-0.218* (0.088)	-0.224* (0.088)
Post 6	0.220 (0.139)	0.207 (0.139)	0.427*** (0.078)	0.415*** (0.082)	0.575*** (0.082)	0.543*** (0.087)	0.287*** (0.070)	0.282*** (0.069)	0.111 (0.094)	0.126 (0.097)	0.135 (0.091)	0.151 (0.092)	-0.037 (0.075)	-0.023 (0.077)
Treatment × Post	0.196 (0.173)	0.199 (0.173)	0.158 (0.114)	0.179 (0.116)	0.343** (0.108)	0.345** (0.111)	0.099 (0.095)	0.104 (0.094)	0.269* (0.122)	0.259* (0.124)	0.331** (0.115)	0.319** (0.117)	0.617*** (0.130)	0.611*** (0.130)
Treatment × Post 6	-0.053 (0.188)	-0.046 (0.188)	0.457*** (0.123)	0.477*** (0.125)	0.276* (0.108)	0.282* (0.114)	-0.074 (0.111)	-0.076 (0.111)	0.219 (0.138)	0.210 (0.140)	0.095 (0.125)	0.093 (0.127)	0.262* (0.114)	0.268* (0.114)
3rd Grade		-0.060 (0.159)		0.059 (0.134)		0.334 (0.186)		0.295 (0.155)		0.174 (0.247)		0.445** (0.166)		0.243 (0.193)
4th Grade		-0.143 (0.130)		0.502*** (0.123)		0.368** (0.125)		0.375** (0.123)		0.389* (0.177)		0.574*** (0.141)		0.285* (0.128)
Treatment × 3rd Grade		0.195 (0.225)		0.122 (0.174)		-0.134 (0.237)		-0.264 (0.224)		-0.035 (0.304)		-0.518* (0.245)		-0.132 (0.257)
Treatment × 4th Grade		0.386 (0.217)		0.163 (0.191)		-0.081 (0.202)		0.038 (0.184)		-0.291 (0.263)		-0.412 (0.250)		0.122 (0.240)
Cognitive Skills		0.077 (0.078)		0.207*** (0.061)		0.337*** (0.058)		0.171** (0.054)		-0.111 (0.069)		-0.140* (0.067)		-0.133* (0.059)
Students		0.012 (0.012)		0.007 (0.012)		0.030** (0.011)		-0.004 (0.010)		0.009 (0.015)		0.019 (0.013)		0.012 (0.013)
cut1	-2.985*** (0.122)	-2.819*** (0.266)	-2.509*** (0.095)	-2.068*** (0.256)	-0.416*** (0.086)	0.520* (0.231)	-2.452*** (0.101)	-2.198*** (0.203)	-1.417*** (0.121)	-1.108*** (0.264)	-0.247* (0.101)	0.376 (0.259)	-1.033*** (0.090)	-0.712** (0.229)
cut2	-2.528*** (0.119)	-2.362*** (0.264)	-1.524*** (0.084)	-1.074*** (0.252)	0.403*** (0.082)	1.354*** (0.234)	-0.506*** (0.069)	-0.236 (0.202)	-0.680*** (0.107)	-0.367 (0.257)	0.906*** (0.085)	1.541*** (0.261)	-0.036 (0.079)	0.291 (0.227)
cut3	-1.803*** (0.108)	-1.636*** (0.254)	-0.802*** (0.082)	-0.340 (0.250)	1.035*** (0.088)	1.996*** (0.239)	0.335*** (0.068)	0.617** (0.204)	-0.293** (0.104)	0.023 (0.257)	2.130*** (0.088)	2.769*** (0.266)	1.026*** (0.084)	1.357*** (0.230)
cut4	-1.129*** (0.096)	-0.961*** (0.255)	0.177* (0.081)	0.656** (0.251)	1.694*** (0.091)	2.661*** (0.241)	1.350*** (0.068)	1.639*** (0.207)	0.239* (0.098)	0.557* (0.253)	3.481*** (0.112)	4.120*** (0.284)	2.166*** (0.092)	2.499*** (0.233)
Observations	5269	5269	5269	5269	5262	5262	5258	5258	5260	5260	5245	5245	5243	5243
Pseudo R ²	0.007	0.008	0.010	0.018	0.010	0.016	0.002	0.006	0.005	0.007	0.005	0.009	0.002	0.004

The dependent variables are the reported behavior for each question of the aggregated reported behavior and are defined in the top row. *Treatment* is equal to 1 if the students are in the treatment group, 0 otherwise. *Post* and *Post 6* are equal to 1 if the survey is taken in the post-program or post6-program wave, respectively. *Pre* is the reference category and refers to the survey taken in the pre-program wave. *Grade* is the students' year group. *2nd Grade* is the reference category. *Cognitive Skills* is equal to 1 if the result obtained in the logical and mathematical questions are higher than the median, 0 otherwise. *Students* is the number of students in each class. Standard errors (in parenthesis) are clustered at class level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figures below show the effects of the treatment on the estimated probability for each possible answer level (from 1 to 5) to each reported behavior, by survey wave. In most of the cases, there are positive effects in the estimated probability for higher answer levels, while the effects on the estimated probability for lower answer are negative.

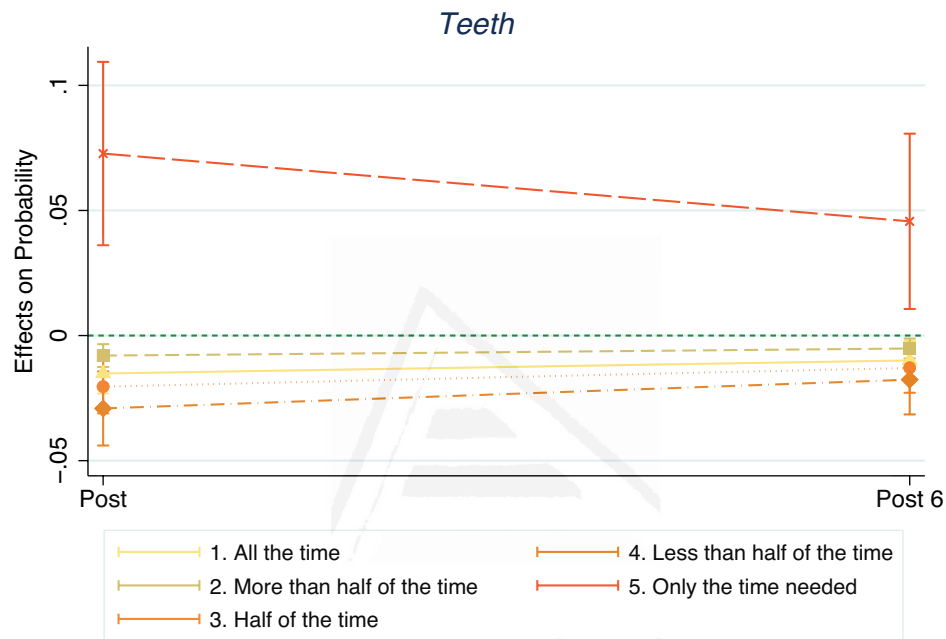


Figure 3.C.4: *Teeth* - “How much do you keep open the tap when you brush your teeth?”. Effects on the estimated probability for each possible answer, by survey way. Error bars represents the 95% confidence interval.

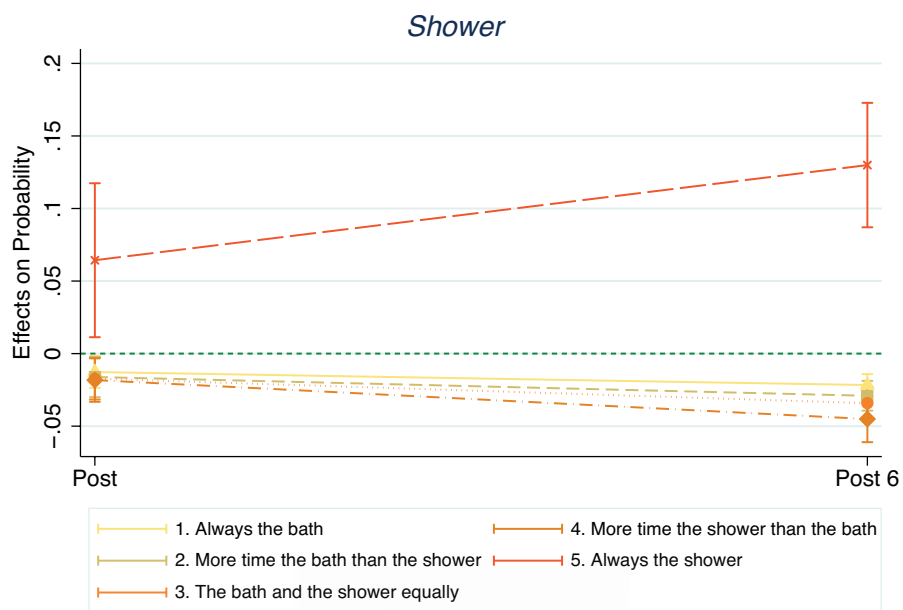


Figure 3.C.5: *Shower* - “Are you doing more often the bath or the shower?”. Effects on the estimated probability for each possible answer, by survey way. Error bars represents the 95% confidence interval.

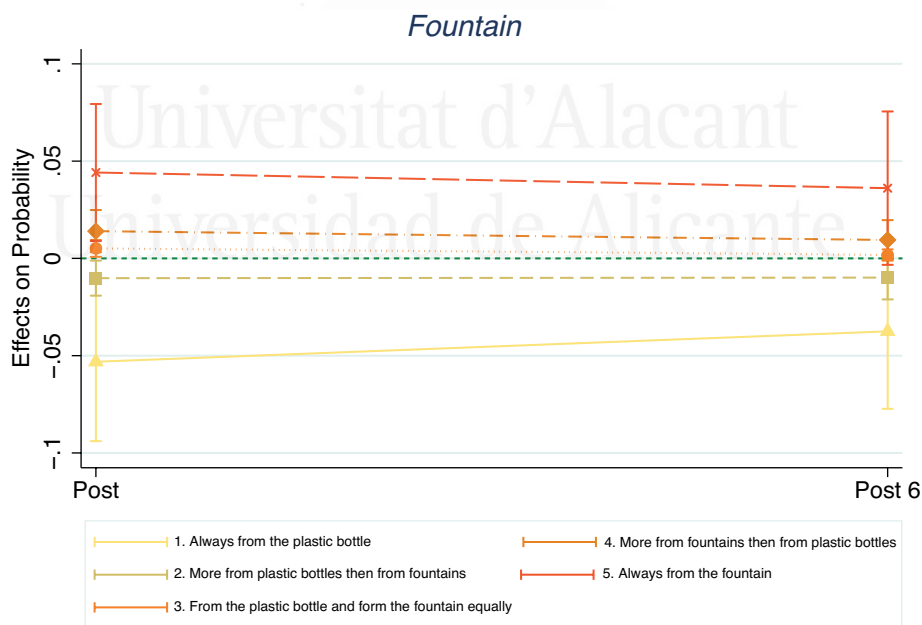


Figure 3.C.6: *Fountain* - “Are you drink more the water from the plastic bottles or from the fountain/faucet?”. Effects on the estimated probability for each possible answer, by survey way. Error bars represents the 95% confidence interval.

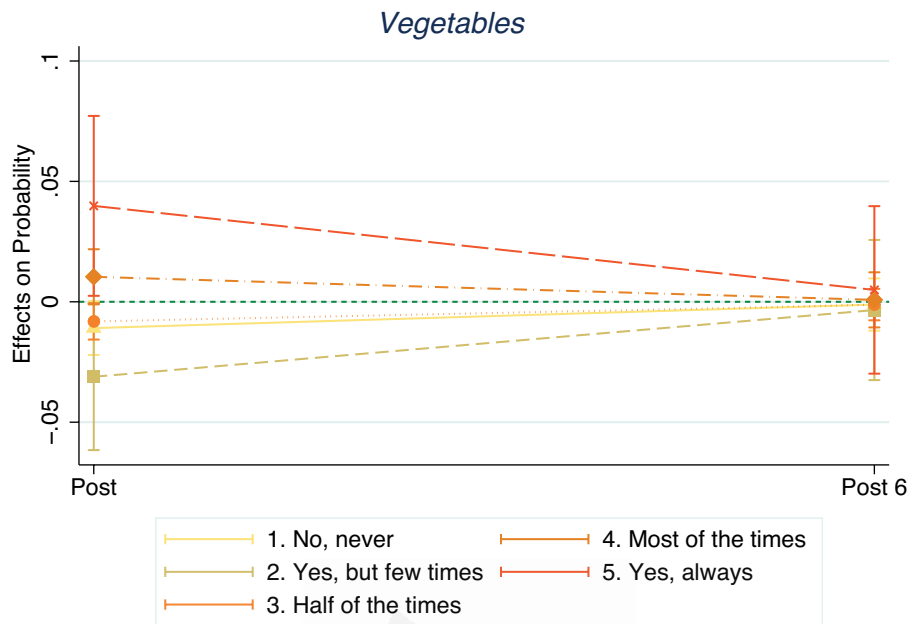


Figure 3.C.7: *Vegetables* - “Are you eating fruit or vegetables during your meals?”. Effects on the estimated probability for each possible answer, by survey way. Error bars represents the 95% confidence interval.



Figure 3.C.8: *Hands* - “When you wash your hands, do you close the tap while you soap your hands?”. Effects on the estimated probability for each possible answer, by survey way. Error bars represents the 95% confidence interval.

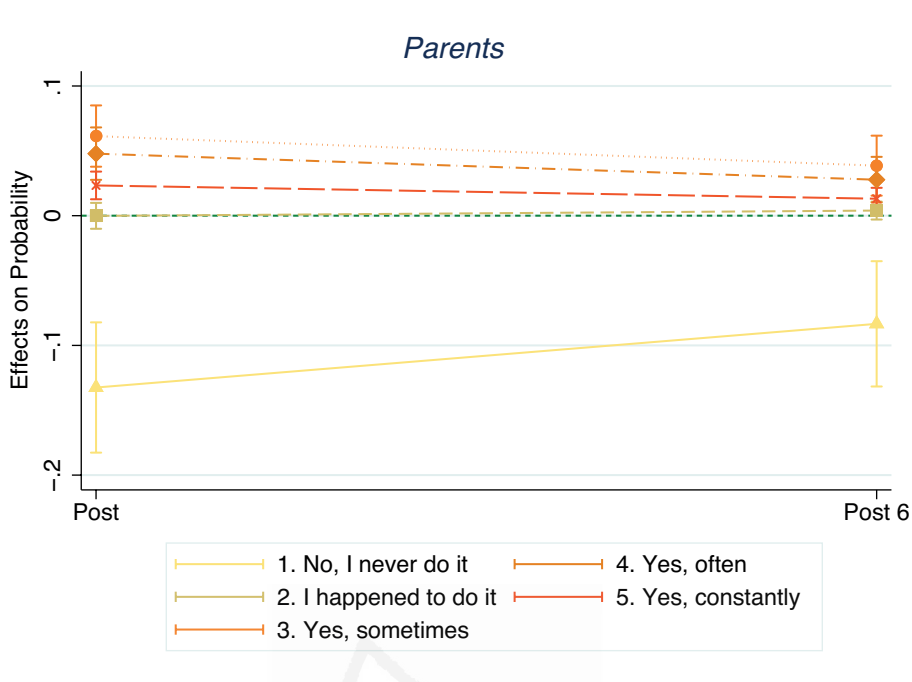


Figure 3.C.9: *Parents* - “Do you talk with your parents on how the water gets to your house?”. Effects on the estimated probability for each possible answer, by survey way. Error bars represents the 95% confidence interval.

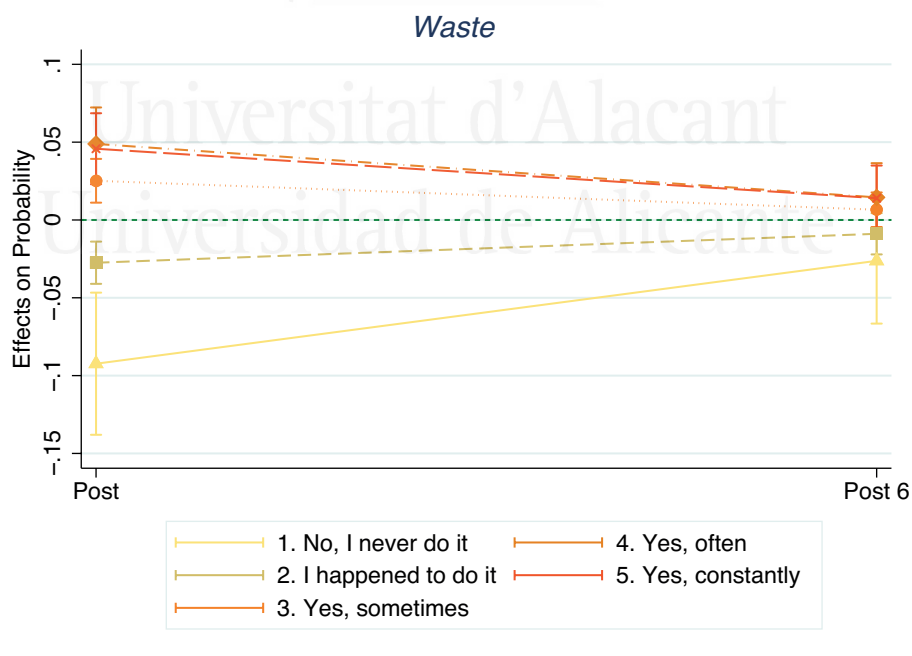


Figure 3.C.10: *Waste* - “Do you talk with your parents on how don’t waste water?”. Effects on the estimated probability for each possible answer, by survey way. Error bars represents the 95% confidence interval.

