Journal of Economic Behavior and Organization 239 (2025) 107251

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Searching for the external validity of social preference games: A guide of field environments based on expert perceptions☆

Daniel Navarro-Martinez [a], Sergio Pirla [b,*]

[a] Universitat Pompeu Fabra, Barcelona School of Economics & Barcelona School of Management, Spain
[b] Universidad de Zaragoza, IEDIS, Spain

ARTICLE INFO

Dataset link: Link

JEL classification:
C92
C93
D90

Keywords:
External validity
Social preference games
Field environments
Meta-analysis

ABSTRACT

The last couple of decades have witnessed a lively debate on the external validity of social preference games. Yet, scientific progress in this area has been restrained by the difficulty of delineating the field environments that social preference games should generalize to. Here we present three studies investigating the field environments and behaviors to which social preference games are expected to relate, according to specialist researchers. In Study 1, we systematically reviewed all the papers published in the top 5 economics journals that used social preference games, and we analyzed the field settings explicitly linked to the games by the authors. In Study 2, we used large language models to expand our analysis of the literature beyond the top 5. In Study 3, we conducted a survey among members of the Economic Science Association (ESA) mailing list to investigate the field environments they viewed as most closely associated with different social preference games. Overall, our results provide a rich guide to the types of field settings that are expected to relate to social preference games, according to the people who use them. This guide constitutes a useful reference to organize future research on external validity and make it more systematic.

1. Introduction

The last couple of decades have witnessed an active debate about the external validity of experimental games — that is, the extent to which behavior in these games is generalizable to other settings, mostly in the field (see, e.g., Levitt and List, 2007; Falk and Heckman, 2009; Camerer, 2011; Al-Ubaydli and List, 2013; Kessler and Vesterlund, 2015; Galizzi and Navarro-Martinez, 2019). This debate concerns a central issue in behavioral and experimental economics, given that, as noted by Lowenstein (1999): "low external validity [...] is the Achilles Heel of all laboratory experimentation" (p. F33). A lack of external validity could put economic experiments in danger of being disconnected from real-world behaviors beyond the confines of the laboratory. The issue of external validity is also largely unresolved, and research on this topic is still in its infancy. As Levitt and List (2007) pointed out: "perhaps the most fundamental question in experimental economics is whether findings from the lab are likely to provide reliable inferences outside of the laboratory" (p. 170).

☆ All the data and meta-analysis materials referenced in the paper are available in the following OSF repository: Link. Both authors contributed equally to this paper and are listed in alphabetical order. This research was funded by the BBVA Foundation, Spain (Fundacion BBVA-EI-2019-D.Navarro), the Ramon Areces Foundation (Fundacion Ramon Areces 2019-Navarro), the Spanish Ministry of Science and Innovation (PID2019-105249GB-I00, PID2022-137908NB-I00), and ICREA (ICREA Academia 2024-Daniel Navarro).

* Corresponding author.
E-mail addresses: daniel.navarro@upf.edu (D. Navarro-Martinez), spirla@unizar.es (S. Pirla).

https://doi.org/10.1016/j.jebo.2025.107251
Received 29 January 2024; Received in revised form 9 September 2025; Accepted 13 September 2025

Available online 16 October 2025

Social preference games — such as the dictator game, ultimatum game, trust game or public goods game — have been one of the main focuses of the external validity debate (see Levitt and List, 2007; Galizzi and Navarro-Martinez, 2019). These games encapsulate many of the methodological elements this debate has revolved around. They provide stylized social environments linked to game-theoretic equilibria that can be easily incentivized using monetary payoffs, yielding high degrees of internal validity. At the same time, their abstract and artificial nature raises the concern that they might not generalize to more naturalistic field environments.

One of the aspects that has constrained scientific progress on the external validity of social preference games is the lack of guidance regarding which particular field contexts these games should be expected to generalize to. That is, experimental economists are typically left to their own devices in assessing which environments constitute reasonable generalizations, and this is problematic. If behavior in a particular field setting is not found to correlate with behavior in a social preference game, it is easy to dismiss the result by arguing that this game was not intended to address this type of behavior. If, on the other hand, a correlation is found, it is easy to justify it by drawing an ad hoc analogy between the game and the field behavior. This lack of a clear correspondence between the abstract games and field behaviors leads to a disorganized and unsystematic stream of research on external validity, which is prone to issues such as selective reporting, false positives and publication bias. At the same time, it is possible that field environments that could correlate well with the games are not being systematically explored. As Galizzi and Navarro-Martinez (2019) pointed out, an indication of the unsystematic nature of this line of research is that social preference games have been associated in the literature to things "as diverse as earning or spending money, fishing shrimps, drinking beer, participating in elections for parent representatives in schools, or registering books in a library" (p. 978).

In this paper, we present a guide of field environments that are expected to be related to social preference games. In doing so, we provide the literature with a reference to organize research and discussions on the external validity of these games. To construct this guide, we relied on the view of specialist researchers in the fields of behavioral and experimental economics. Scholars in these fields know the games well and are likely to have well-developed views on the types of field environments and behaviors they expect to be related to the games. These views are also bound to permeate the discussions of results obtained using social preference games found in the scientific literature.

Based on this, we organized our research into three separate studies. In Study 1, we conducted a meta-analysis of the field behaviors that are explicitly linked to social preference games in the literature. To do so, we reviewed all the papers published in the top 5 economics journals that rely on social preference games — focusing on dictator games, ultimatum games, trust games, public goods games, and variants thereof —, and we extracted from them all the explicit references to field environments expected to be related to these games. While there are other games used in the literature that can be classified as social preference games (such as prisoner's dilemma games, gift exchange games, etc.), we decided to focus on these four types, which are context-free, cover most research on social preferences, and represent the benchmark approach to studying behaviors and measuring preferences in this area. In Study 2, we developed a large language model (LLM) approach to extend our review and meta-analysis of the literature. We validated this method using the results from Study 1, and then applied it to a broader set of economics journals. In Study 3, we run a survey among the members of the Economic Science Association (ESA) mailing list. The ESA is (arguably) the most prominent organization devoted to experimental and behavioral economics. With this survey, we investigated the field environments that experimental and behavioral economists expect to be most closely related to the different social preference games.

Our results show a high degree of convergent validity across studies and measures. In other words, the two meta-analyses and the ESA survey point to a similar set of field environments. This demonstrates that authors and other members of the experimental economics community share a fairly stable view of social preference games and what they represent. In practical terms, our work provides a cohesive guide of field settings that are expected to be related to social preference games. We believe this guide is a valuable resource to help organize research on the external validity of experimental research, making it more systematic and assessable.

The rest of the paper is organized as follows. In Sections 2–4, we present Studies 1, 2 and 3, respectively. Section 5 contains a joint analysis of the three studies and provides combined lists of field environments. Section 6 explains the main ways in which our results can be used as a guide for future research. Finally, Section 7 concludes.

2. Study 1: A meta-analysis based on the top 5 journals

2.1. Method

In this study, we conducted a meta-analysis of the existing economics literature that has employed experimental social preference games. To keep the meta-analysis manageable and avoid potential problems with the quality and reliability of the papers included, we restricted our attention to articles published in the top 5 economics journals (American Economic Review, Journal of Political Economy, Quarterly Journal of Economics and Review of Economic Studies). There is broad agreement in economics that these journals represent the highest academic standards of the profession. We considered all the papers ever published in these journals that conducted their studies using primarily social preference games. For this, we focused on four types of games: dictator games (DG), public goods games (PGG), trust games (TG) and ultimatum games (UG), including context-free modified versions or extensions of them (such as modified dictator games). We excluded papers that contain field data, as in these cases the main links with the field are established on the basis of the available data and not of the researchers' expectations of generalizability. A systematic review and meta-analysis of the literature comparing lab behavior in social preference games and actual field behavior can be found in Galizzi and Navarro-Martinez (2019).

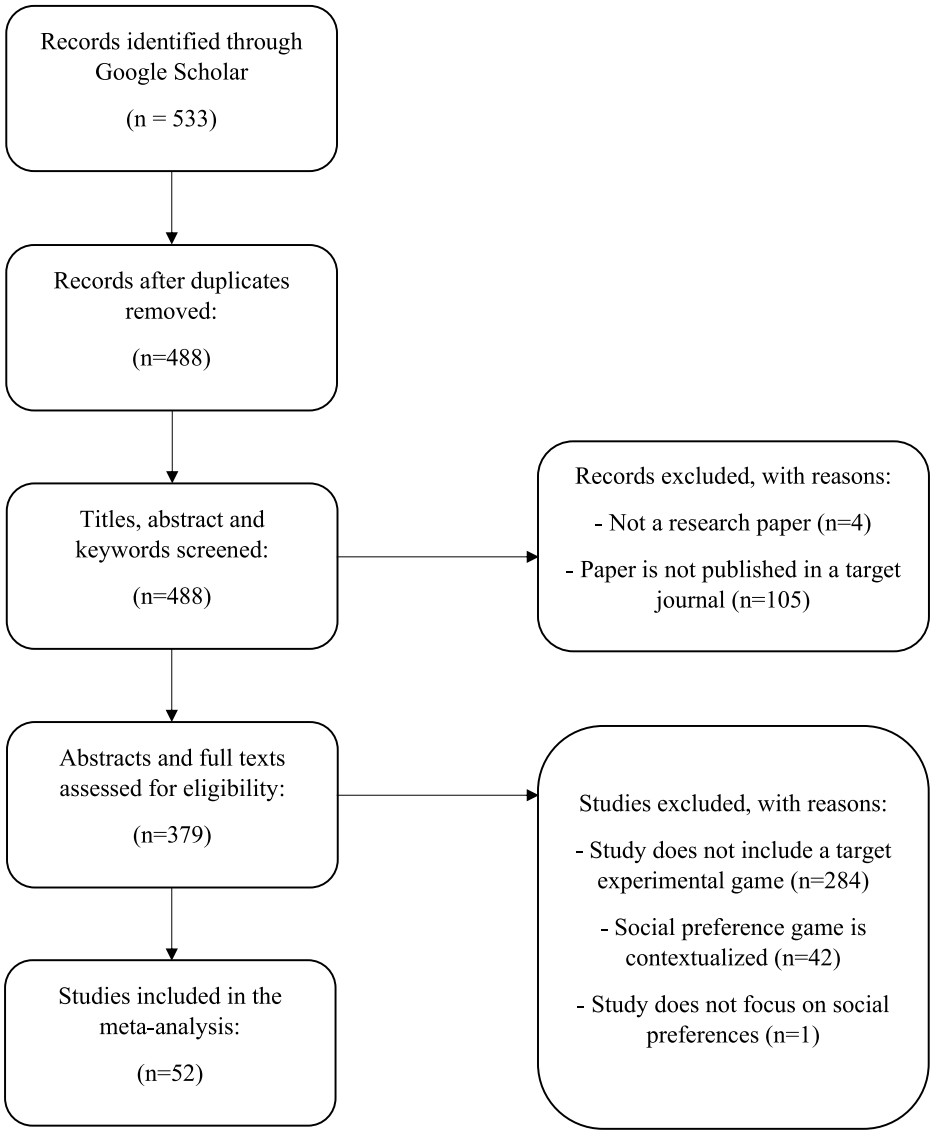

Fig. 1. PRISMA diagram of review and selection of papers in meta-analysis of the top 5.

Fig. 1 shows a PRISMA diagram (Moher et al., 2009) summarizing the process we followed in our systematic review of the literature. To begin with, we searched Google Scholar (in September 2020) for the following terms (using the field "with the exact phrase"): "dictator game", "dictator games", "ultimatum game", "ultimatum games", "trust game", "trust games", "public good game", "public good games", "public goods game", "public goods games". Note that, while this is unlikely to happen, this procedure might potentially exclude certain games that are similar to these ones but were published in papers where these terms were never mentioned. We restricted our search to articles published in the top 5 economics journals. This resulted in 533 records that constituted the starting point of our review. All the papers are listed in an Excel file labeled as "All papers Study 1" and available in the OSF repository linked to this paper. This Excel file includes information on the authors, titles, publication years, journals, links to the papers, numbers of citations, whether the papers are included in the final meta-analysis, and the exclusion criteria applied. As shown in Fig. 1, we then sequentially applied six exclusion criteria. Specifically, we removed duplicated entries (criterion 1), records that were not research papers (criterion 2), papers that were not published in one of the target journals (criterion 3), papers that did not include one of our target experimental games (criterion 4), papers in which the games were contextualized or that included field data (criterion 5), and papers that did not focus on social preferences (criterion 6). After this process, we ended up with a final set of 52 papers to be included in our meta-analysis.

In the meta-analysis, for each of the 52 papers, we identified all the target social preference games included in the studies, and we searched for all the references to field settings expected to be related to the games. For each field setting that we found, we also

Table 1
Summary of field environments mentioned in the papers, grouped by categories (top 5 journals).

Rank	Field environments	Total (N = 34)	DG (N = 18)	PGG (N = 9)	TG (N = 8)	UG (N = 5)
1	Social and household interactions	32%	33%	22%	38%	40%
2	Political and social issues	26%	39%	0%	38%	20%
3	Compensation and sanctioning schemes design	24%	28%	33%	13%	20%
4	Charity	21%	28%	22%	0%	0%
5	Labor relations	18%	11%	0%	38%	40%
6	Health care	12%	22%	0%	0%	0%
6	Taxation	12%	17%	11%	0%	0%
6	Financial/insurance markets and investment	12%	11%	0%	25%	0%
6	Firm behavior and pricing	12%	11%	0%	0%	40%
10	Military actions and their consequences	9%	17%	0%	0%	0%
10	Tipping	9%	11%	0%	25%	20%
10	Negotiations	9%	6%	0%	13%	20%
10	Group and team dynamics	9%	0%	33%	0%	0%
14	Legal proceedings	6%	11%	0%	0%	0%
14	Environmental policy	6%	6%	11%	0%	0%
16	Consumption	3%	6%	0%	13%	20%
16	Discrimination	3%	6%	0%	13%	20%
16	Policing	3%	6%	0%	0%	0%
16	Industrial disputes	3%	0%	11%	0%	0%
16	International agreement design	3%	0%	11%	0%	0%
16	Academic writing and publishing	3%	0%	0%	13%	0%
16	Business meetings	3%	0%	0%	13%	0%

identified all the specific patterns of behavior described in relation to the field setting and all the specific game results linked to those patterns (if any). We collected all this information in an Excel file labeled as "Meta-analysis 1 A" and available in our OSF repository, which contains the following information for each paper: authors, title, publication year, journal, paper URL, number of citations on Google Scholar, games included, field settings referenced, field patterns described in relation to each field setting, game results linked to each pattern, passages where the field settings are referenced, page numbers, and a brief outline of the paper.

The meta-analysis produced a total of 46 different field settings referenced across the 52 papers. To reduce all these settings to a more manageable number, we grouped them together in 22 meaningful categories, and we conducted some basic statistical analyses on both the settings and the bigger categories. All this is contained in another Excel file, labeled as "Meta-analysis 1 B" and available in our OSF repository.

2.2. Results and discussion

Our results show that, overall, 65% of the papers included in our meta-analysis (34 out of 52) mention field settings that are expected to be related to the games. This demonstrates that most researchers using social preference games view the games as potentially having external validity and being useful to understand social behaviors in the field. This percentage shows some variation across games, with 78% for the dictator game, 56% for the public goods game, 73% for the trust game, and 42% for the ultimatum game. It is important to note, however, that given the limited sample size when subdividing by games, these differences might simply reflect random variation.

Table 1 presents a summary of the different field environments mentioned in the papers, grouped into 22 categories. The table shows the percentage of papers that mention each of the field environments (both overall and disaggregated by game) out of all the papers that mention at least one field environment. In total, "social and household interactions" is the category of field behaviors that is mentioned the most (by 32% of the papers) as expected to be related with social preference games. This category is followed by "political and social issues" (26%), "compensation and sanctioning schemes design" (24%), "charity" (21%), and "labor relations" (18%). This completes the overall top five of field environments. As shown in Table 1, the most mentioned categories are somewhat different across games, although as indicated before, the smaller number of observations for single games makes the specific ranking patterns less reliable.

Our meta-analysis also allows us to zoom in and look at the specific settings and patterns of behavior contained in the bigger categories. Table 2 does this for the highest ranked category ("social and household interactions"), which is also the one that contains the highest number of specific field settings and of associated patterns of behavior. Unpacking all the different categories into settings and patterns would be too much information for the main body of the paper. We, therefore, illustrate our analyses using the first category, and we refer the readers to the full details of settings and patterns that can be found both in the aforementioned Excel file "Meta-analysis 1B" and in another Excel file labeled as "Additional tables Study 1" (both of which can be found in our OSF repository).

As Table 2 shows, the category "social and household interactions" can be subdivided into 8 more specific field settings referenced in the papers, and these 8 settings are connected to 14 patterns of behavior mentioned by the authors. Each pattern of behavior is also linked to one or more particular games. All this information can be used in several ways. At the category and setting level, this

Table 2
Highest ranked category ("social and household interactions") unpacked into specific settings and patterns of behavior.

Rank	Total	Specific settings	Total	DG	PGG	TG	UG	Patterns of behavior	Games linked
	(N = 34)	(N = 8)	(N = 34)	(N = 18)	(N = 9)	(N = 8)	(N = 5)	(N = 14)	
1	32%	Couple dynamics	9%	17%	0%	0%	0%	Household bargaining within couples	DG
								Man distorting beliefs to justify cheating on wife	DG
								Member of couple hiring financial advisor	DG
		Social interactions	9%	6%	11%	25%	20%	Choice of tattoos, piercings	TG
								Decision to interact with someone/pursue a relationship	TG
								Expression of social acceptance	PGG
								Gossip	PGG
								Grooming before meeting someone	TG
								Peer pressure	PGG
								Secretaries performing more promptly for those who are polite or bring gifts	DG, TG, UG
								Shunning unfriendly colleagues and inviting friendly ones home	DG, TG, UG
								Social Ostracism	PGG
		Social exchange of money	3%	6%	0%	0%	0%		
		Social sharing of costs/payments	3%	6%	0%	0%	0%	Friends splitting restaurant bills equally	DG
		Evolution of social norms	3%	0%	11%	0%	0%		
		Language, discussions, agreements, and social norms in strategic interaction	3%	0%	0%	13%	0%		
		Partnerships (including husband and wife, lawyer and client, etc.)	3%	0%	0%	13%	0%		
		Household dynamics	3%	0%	0%	0%	20%	Sibling rivalry	UG

provides a ranked set of domains where researchers expect social preference games to be associated with field behaviors. In this sense, Table 1 can be thought of as a summary guide of field environments in which to look for external validity that can also be expanded by looking at the more specific settings. At the pattern of behavior level, these results provide very specific predictions about associations between the lab and the field, which could potentially be tested in appropriate lab-field studies.

3. Study 2: Extended meta-analysis using LLMs

In Study 1, we focused on papers published in the top 5 economics journals. While we believe this restriction is meaningful and it allowed us to conduct a more detailed analysis, it also implies limitations in terms of the types of outlets included and the sample sizes, especially when subdividing by games. To address these concerns, in the present study, we used an LLM approach — based on ChatGPT-4o — to extend our analysis to a broader range of journals. We first present a validation of our LLM approach (Section 3.1) and then move to the extended meta-analysis (Section 3.2).

3.1. Validation of the LLM approach

3.1.1. Method

Before analyzing a new set of journals, we validated our LLM-based approach by replicating the analysis of the top 5 journals presented in Study 1 and comparing the results with it.

Our validation exercise started from the same initial set of 533 records identified via Google Scholar search in Study 1. Like we did in Study 1, we manually removed 45 duplicate entries, 4 records that were not research papers, and 105 papers not published in one of the five target journals, which resulted in a working set of 379 papers. For all the next steps, including the implementation of the remaining exclusion criteria, we used ChatGPT-4o with a series of structured prompts that mirrored the steps followed in our Study 1 review and analysis. Table 3 summarizes the different steps of our approach, including all the prompts we used. We

Table 3
Description of steps and prompts used in LLM-based approach.

Step	Explanation	Prompt	Prompt Text
Apply exclusion criterion 1.	Identify duplicated records.	–	–
Apply exclusion criterion 2.	Identify records that are not a research paper.	–	–
Apply exclusion criterion 3.	Identify papers not published in a target journal.	–	–
Apply exclusion criterion 4.	Identify if the study does not include an experimental social preference game.	1	Each PDF contains a scientific paper. For each paper, please answer the following question: Does the paper include an experiment that implements one (or a version) of the following games: dictator game, trust game, ultimatum game, public goods game?
		2	Using the uploaded PDFs, review each paper and answer the following question: Does the paper use experimental data?
Apply exclusion criterion 5.	Identify if the study includes observational or field data or uses contextualized instructions.	3	Using the uploaded PDFs, review and answer the following question for each paper: Does the paper use observational or field data (i.e., data collected outside the laboratory)?
		4	Using the uploaded PDFs, review and answer the following question for each paper. Are the instructions read by the participants in the experimental games contextualized? Answer "Yes" if and only if there is evidence showing that the instructions read by the participants in the experiment explicitly reference real-life (i.e., outside the lab) situations.
Apply exclusion criterion 6.	Identify if the study does not focus on social preferences.	5	Using the uploaded PDFs, review each paper and answer the following question: Does the paper focus on social preferences (broadly defined)?
Identify specific games.	Identify specific games used in the studies.	6	Does the uploaded paper include an experiment that implements one (or a version) of the following games: dictator game, trust game, ultimatum game, public goods game without punishment, public goods game with punishment? If so, please specify which games are included.
Identify external validity claims.	Extract explicit references to field settings.	7	Identify and extract all explicit references to real-world (i.e., outside-the-lab) settings that the authors claim their findings relate to. For each identified setting, provide a meaningful label and the exact passage verbatim where the authors make the reference. If the paper contains no such references, write: "No claims made."
Categorize external validity claims.	Categorize explicit references to field settings into 22 categories (including an "Other" category option).	8	Categorize the following setting using the provided list of categories: Setting: [Setting included here] - Select the single most appropriate category and respond with the corresponding category number. - If no category applies, respond with: "Other."

first applied the remaining exclusion criteria by asking the model to assess, for each paper, whether it included one of our target experimental games (Prompts 1 and 2), included observational or field data or used contextualized instructions (Prompts 3 and 4), and focused on social preferences (Prompt 5). Then, for the papers passing all the criteria, we prompted the model to identify which games were included (Prompt 6), and to extract and label all explicit references to real-world settings that the authors claimed their findings related to (Prompt 7). Finally, we used the model to categorize all the extracted settings using the 22 broad categories defined in Study 1 (Prompt 8). We also included an additional category labeled as "other" for the LLM to classify the settings that were not considered to fit in any of the main categories. In our validation analyses, we focus on the set of 22 categories identified in Study 1, leaving this category aside.

3.1.2. Results and discussion

Out of the 379 papers we screened using our LLM method, the model correctly classified 90% as included vs. excluded, compared to the manual analysis presented in Study 1. The implementation of Prompts 1 to 5 resulted in a final selection of 52 papers (14% of the total) that meet all inclusion criteria, exactly the same number as in Study 1. Of these 52 papers, 33 papers (63%) overlap with the final sample obtained in Study 1 and 19 (37%) are papers that were not included in Study 1. This implies that our LLM procedure achieved a 63% probability of correctly classifying papers that were included in Study 1, and a 94% probability of correctly classifying papers that were excluded.

Focusing on the final set of 52 papers, 45 of them (87%) were identified as containing at least one reference to a real-life setting that the authors claim their results are related to. This represents a greater proportion than the one obtained in Study 1, which was 65%. As in Study 1, this percentage varies across games: 88% of the papers using the dictator game include at least one field setting,

Table 4
Summary of field environments mentioned in the papers, grouped by categories (LLM replication of top 5 analysis).

Rank	Field environments	Total (N = 45)	DG (N = 21)	PGG (N = 13)	TG (N = 9)	UG (N = 8)
1	Political and social issues	47%	57%	46%	56%	0%
2	Compensation and sanctioning schemes design	31%	33%	38%	22%	25%
3	Social and household interactions	29%	29%	31%	33%	25%
4	Group and team dynamics	20%	19%	23%	33%	0%
5	Charity	18%	33%	8%	0%	0%
6	Financial/insurance markets and investment	16%	14%	8%	33%	13%
6	Labor relations	16%	14%	8%	11%	50%
6	Negotiations	16%	14%	0%	0%	50%
9	Consumption	11%	14%	0%	22%	13%
9	Environmental policy	11%	5%	31%	0%	0%
11	Health care	9%	14%	0%	11%	0%
11	Firm behavior and pricing	9%	10%	0%	11%	25%
11	Discrimination	9%	5%	0%	22%	25%
14	Military actions and their consequences	4%	10%	0%	11%	0%
14	International agreement design	4%	5%	8%	0%	0%
14	Business meetings	4%	5%	0%	22%	0%
14	Legal proceedings	4%	5%	0%	11%	0%
14	Industrial disputes	4%	0%	8%	0%	13%
19	Taxation	2%	5%	0%	0%	0%
19	Tipping	2%	5%	0%	0%	0%
21	Academic writing and publishing	0%	0%	0%	0%	0%
21	Policing	0%	0%	0%	0%	0%

compared to 87% for the public goods game, 100% for the trust game, and 73% for the ultimatum game. Papers with at least one reference to a field setting contain an average of 4.3 settings. This number is also higher than in Study 1, where it was 2.44.

Despite these differences in the selection of papers and the number of identified field settings, the main results in terms of the classification into categories of field environments strongly support the validity of our LLM-based approach. Table 4 presents the percentage of papers (out of those containing at least one external validity claim) that are classified as referencing a field setting in each of our 22 categories. The category "other" (not included in our main analysis) was used by the model in 36% of these papers (33%, 38%, 44% and 38% for the DG, PGG, TG and UG, respectively). Overall, we found a high correlation between the distribution of field settings in Study 1 and the one obtained from our LLM approach. Specifically, we computed a Pearson correlation of $r = 0.814$ ($t = 6.264$, df = 20, $p < 0.001$) between the percentages of papers referencing each category in the original meta-analysis and the corresponding percentages produced by the LLM. We also found strong correlations when disaggregating by game. In particular, the correlation is $r = 0.820$ ($t = 6.405$, df = 20, $p < 0.001$) for papers using the dictator game, $r = 0.577$ ($t = 3.157$, df = 20, $p = 0.005$) for those using the public goods game (with or without punishment), $r = 0.565$ ($t = 3.060$, df = 20, $p = 0.006$) for the ones using the trust game, and $r = 0.735$ ($t = 4.852$, df = 20, $p < 0.001$) for those using the ultimatum game. Taken together, we believe these results provide strong support for the validity of our LLM-based approach to generate distributions of field environments based on the literature.[1]

3.2. Extended meta-analysis

3.2.1. Method

Once we validated our LLM-based approach, we used the same method to conduct a meta-analysis of external validity claims made in papers published in a broader set of journals (and excluding the top 5). On the one hand, we included three highly respected, general-interest journals outside the top 5: the Economic Journal, the Journal of the European Economic Association, and the Review of Economics and Statistics. On the other hand, we expanded to three prominent field journals in behavioral and experimental economics: Experimental Economics, Games and Economic Behavior, and the Journal of Economic Behavior & Organization. Given that these field journals have extensively published research using social preference games, this resulted in a much higher number of papers than in our meta-analysis of the top 5.

As in Study 1, we followed a systematic protocol to identify the relevant papers. Specifically, in May 2024, we searched Google Scholar using the field "with the exact phrase" and the following terms: "dictator game", "dictator games", "ultimatum game", "ultimatum games", "trust game", "trust games", "public good game", "public good games", "public goods game", and "public goods games", restricting the search to the six target journals used in this case. This process yielded 2136 records, all of which are listed (alongside the papers employed in the validation exercise) in an Excel file labeled "All papers Study 2", available in the OSF repository linked to the paper. As in Study 1, this file contains information on authors, titles, publications years, journals, links to the papers, number of citations, whether the papers are included in the final meta-analysis, and the exclusion criteria applied, this

[1] Our OSF repository includes a supplemental note with additional information on our validation exercise, including a replication of the procedure.

Table 5
Summary of field environments mentioned in the papers, grouped by categories (extended, LLM-based meta-analysis). .

Rank	Field environments	Total (N = 382)	DG (N = 116)	PGG (N = 144)	TG (N = 103)	UG (N = 75)
1	Political and social issues	59%	61%	65%	65%	49%
2	Compensation and sanctioning schemes design	31%	27%	34%	33%	23%
3	Group and team dynamics	30%	24%	43%	18%	17%
4	Social and household interactions	26%	35%	17%	35%	19%
5	Charity	15%	22%	19%	9%	4%
6	Firm behavior and pricing	15%	10%	7%	27%	21%
7	Financial/insurance markets and investment	14%	7%	9%	25%	13%
8	Labor relations	13%	10%	6%	18%	16%
9	Negotiations	10%	7%	3%	3%	35%
10	Environmental policy	9%	3%	20%	1%	0%
11	Discrimination	7%	11%	3%	11%	7%
11	Consumption	7%	5%	4%	10%	8%
13	Legal proceedings	5%	3%	4%	7%	7%
14	International agreement design	4%	0%	8%	2%	3%
15	Health care	3%	5%	5%	4%	3%
15	Taxation	3%	3%	6%	1%	1%
17	Policing	3%	1%	7%	3%	1%
18	Business meetings	3%	2%	1%	7%	1%
19	Tipping	2%	4%	0%	2%	3%
19	Academic writing and publishing	2%	3%	3%	1%	3%
21	Industrial disputes	1%	0%	3%	0%	0%
22	Military actions and their consequences	1%	1%	0%	0%	4%

time specifying also the exclusion prompts. We then manually removed duplicated entries, records that were not research papers, and papers that did not appear in one of the target journals, which resulted in a set of 1658 papers.

We applied our validated LLM-based procedure to these papers to exclude the ones that did not include one of our target games, included field data, had contextualized instructions, or did not focus on social preferences. This screening process resulted in a final set of 433 valid papers, approximately eight times more than in the Study 1 meta-analysis of the top 5. For each of these papers, our LLM identified the specific games used, extracted the passages containing external validity claims, labeled all the real-life settings referenced in those claims, and organized them into our 22 categories of field environments. As in the validation exercise, we also used the additional "other" category, which we do not include in our main analyses. This category was used for 24% of the papers (31%, 22%, 28% and 32% for the DG, PGG, TG and UG, respectively). All of the information on the extraction and classification of field environments — covering both the papers included in the extended meta-analysis and those used in our validation exercise — is compiled in an Excel file labeled "Meta-analysis Study 2", which is available in the OSF repository linked to the paper.

3.2.2. Results and discussion

Out of the final 433 papers included in our extended meta-analysis, our LLM-based approach identified external validity claims in 382 papers (88%), a percentage that is higher than the one observed in Study 1 (65%) but very similar to the one obtained in our LLM replication of the Study 1 analysis (87%). This percentage varies a bit across games but is more stable than in Study 1 and in the Study 1 replication: 87% for the dictator game, 91% for the public goods game, 90% for the trust game, and 82% for the ultimatum game. This greater stability probably comes from the larger sample of papers and the increased reliability when subdividing by games derived from it. On average, papers that reference at least one field setting mention 4.27 distinct settings, again a higher number than in Study 1 but similar to the one in our LLM replication.

Table 5 presents the main results of our LLM meta-analysis in relation to the classification of the identified field settings into our 22 categories. Specifically, for each category, the table reports the percentage of papers (out of those containing at least one reference to a field setting) that include at least one setting classified under that category. The top five most frequently referenced categories are: "political and social issues" (59%), "compensation and sanctioning schemes design" (31%), "group and team dynamics" (30%), "social and household interactions" (26%), and "charity" (15%). These top categories strongly overlap with the ones identified in Study 1 (four out of five), with the only exception that here "group and team dynamics" appears among the top five in place of "labor relations". There is also a 100% overlap in these categories between this extended meta-analysis and the LLM replication of Study 1.

Another aspect that stands out is that, due to the substantially larger sample, the percentages obtained when subdividing by games are more reliable. This makes them much more fine-grained and meaningful. The percentages are broadly similar across games but with some important differences. For the dictator game, the top five categories are exactly the same as the total ones, and the ranking looks generally very similar. For the public goods game, the ranking is also very similar, with the exception that the category "environmental policy" becomes part of the top five. This is a meaningful pattern, given that environmental issues are a classic example of public goods. For the trust game, there are two notable differences. First, the category "charity" drops down and moves out of the top five categories. This makes sense because the trust game is typically not linked to altruism and charitable giving. Second, the categories "firm behavior and pricing", "financial/insurance markets and investment" and "labor relations"

increase and move up to the top five. This is also a reasonable pattern, given that behavior in firms, investment, and labor relations have been used as typical trust game settings. In fact, the trust game is sometimes referred to as the investment game, and the gift exchange game can be understood as a contextualized derivation of the trust game in the domain of labor relations. Finally, for the ultimatum game, there are also two differences (with respect to the total) that stand out. First, like in the case of the trust game, the category "charity" decreases. This is again a meaningful pattern that captures the fact that the ultimatum game is typically not linked to altruism but to more strategic considerations. Second, there is a substantial increase in the category "negotiation", which is also logical because the ultimatum game has been extensively used to study negotiation behavior. Taken together, these patterns across games demonstrate that our extended meta-analysis has reliably captured differences in the way the games are understood and interpreted in the literature.

Importantly, the distribution of field settings across categories in this LLM meta-analysis shows a strong correlation with the results of Study 1. When including all papers, the correlation between the percentage of papers referencing each category in Study 1 and in this extension is $r = 0.734$ ($t = 4.827$, df = 20, $p < 0.001$). The correlations are also fairly high and significant when disaggregating by game: for the dictator game, $r = 0.746$ ($t = 5.003$, df = 20, $p < 0.001$); for the public goods game, $r = 0.487$ ($t = 2.492$, df = 20, $p = 0.021$); for the trust game, $r = 0.632$ ($t = 3.647$, df = 20, $p = 0.002$); and for the ultimatum game, $r = 0.544$ ($t = 2.902$, df = 20, $p = 0.009$). The correlations are even higher between this extended meta-analysis and the LLM replication of Study 1: all papers, $r = 0.952$ ($t = 13.901$, df = 20, $p < 0.001$); dictator game, $r = 0.932$ ($t = 11.463$, df = 20, $p < 0.001$); public goods game, $r = 0.872$ ($t = 7.972$, df = 20, $p < 0.001$); trust game, $r = 0.844$ ($t = 7.048$, df = 20, $p < 0.001$); and ultimatum game, $r = 0.464$ ($t = 2.340$, df = 20, $p = 0.030$).

Overall, this study developed and validated an LLM-based approach to identify and classify external validity claims in the experimental economics literature. We used this method to meta-analyze a set of 433 papers published in six general interest and field journals, substantially expanding the scope of analysis beyond the top 5 journals examined in Study 1. This broader analysis allows us to derive a second set of independently constructed measures of the field environments that social preference games are expected to relate to, based on a wider and more diverse body of literature. This broader analysis shows a high degree of convergent validity in relation to the results of Study 1.

4. Study 3: ESA survey

4.1. Method

In Study 3, we conducted an online survey among the members of the ESA mailing list to investigate the field environments that they expected to be most correlated with social preference games. The ESA is (arguably) the most important international association of experimental and behavioral economists, and by publicizing our survey in its mailing list, we got access to this population. Specifically, we announced our study by asking for volunteers to participate in a "survey to investigate the perceptions of behavioral and experimental economists regarding the external validity of social preference games". In the announcement, we explained that 10% of the participants would be randomly selected to be paid $50 (which was implemented via PayPal). The survey was completed by a total of 89 people (out of a total of 191 who clicked on the link) and had an approximate median completion time of 14 minutes.

In the survey, we asked participants two main types of questions. First, they were presented with open-ended questions where we asked them to "briefly describe at least 1 and up to 3 field situations, contexts or behaviors" that they thought were closely related to each of our target games (in this case, dictator game, ultimatum game, trust game, public goods game without punishment, and public goods game with punishment). Second, we asked them to provide an overall rating (on a scale from 0 to 10) for each of the 22 categories of field environments identified in Study 1 in terms of how closely linked behavior in each domain was expected to be to behavior in our target social preference games.[2] In addition to the 22 field environments from the meta-analysis, we asked them to rate two other environments that we believed were not closely associated to social preference games: saving for retirement and financial information avoidance. We expected these two field environments to be rated particularly low in terms of their connection to social preference games, providing a further test of the quality and meaningfulness of our responses.

Participants also reported additional information on their level of seniority, fields of specialization, whether they had used social preference games as part of their research in the past, the extent to which they saw social preference games as a good tool to study social behavior, whether they were aware of the literature on the external validity of social preference games, and the extent to which they thought that research on this topic is important. Note that it is in principle possible that the same person is an author of one of the papers included in our meta-analyses and a respondent of our ESA survey (approximately 86% of our participants reported having used social preference games in the past). In any case, such a coincidence would simply reflect the fact that we are analyzing in different ways the perceptions of experts in the same area, and it would not undermine our results in any obvious way. Appendix A contains our verbatim experimental instructions; Appendix B provides summary statistics and regression analyses involving the additional information provided by the participants.

[2] Initially, our list of categories included 24 items. However, after further consideration, we decided to merge "evolution of social norms" into the broader category "social and household interactions", and to merge "group dynamics (including those related to the Kyoto Protocol, UN Security Council, work teams, and sports teams)" under the more general category "group and team dynamics", and we consistently used this classification throughout the whole paper. As a result, participants rated a total of 24 categories, but for our analyses we treated these two pairs as single merged categories. For these cases, we estimated their scores as the average of the initial ratings. The average initial rating was 6.76 for "social and household interactions", 7.16 for "group and team dynamics", and 6.82 for both "evolution of social norms" and "group dynamics (including those related to the Kyoto Protocol, UN Security Council, work teams, and sports teams)".

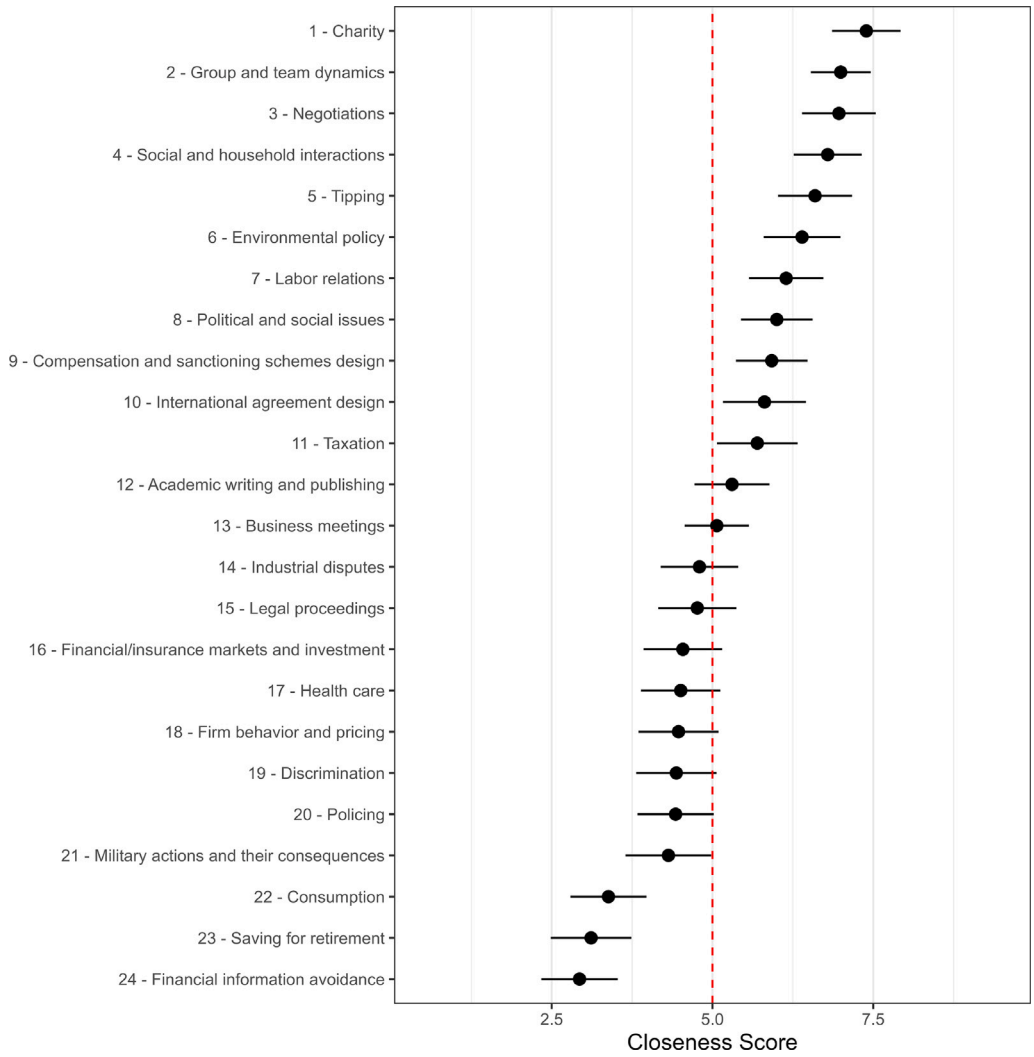

Fig. 2. Average rating for the 22 field environments, with 95% confidence intervals.

4.2. Results and discussion

4.2.1. The ratings of the field environments

We start by presenting the results of the ratings for the 22 field environments originating from the meta-analysis (plus the two additional environments we included) in terms of how closely associated to social preference games they are perceived to be by our sample of ESA respondents. Fig. 2 shows the average rating for each environment and 95% confidence intervals.

Apart from the specific ranking of the field environments, Fig. 2 contains two important pieces of information. First, there is substantial variation in the average scores given to the different environments, which indicates that they were perceived quite differently in terms of their connection to the games. Second, the two additional environments that we introduced are ranked as being the most disconnected from social preference games. This confirms that our participants evaluated the relationship between the environments and the games in a meaningful way.

Comparing the results with the ones obtained in the Study 1 and 2 meta-analyses, we can see that the ranking shows some clear similarities but also some differences. Focusing on the categories ranked in the top five, there is an overlap in two of them in relation to Study 1 and three of them in relation to the Study 2 extended meta-analysis. The three categories included in the top five in Studies 1 and 2 that are not there this time are “labor relations”, “political and social issues” and “compensation and sanctioning schemes design”. All these categories are still ranked relatively high in this analysis, being positioned at 7, 8 and 9. The two categories that are in the top five this time and not in Studies 1 and 2 are “negotiations” and “tipping”. “Negotiations” is ranked in positions 10 and 9 in the Study 1 and 2 meta-analyses, respectively. So, it is still consistently positioned in the first half of the ranking. “Tipping”, on the other hand, might constitute the biggest difference, being positioned at 10 and 19 in the previous

analyses. Note, however, that "tipping" is in position 17 (not counting the option "other category") in our analysis of the open responses (explained in the next section), which suggests that this is a peculiarity of the results obtained in the rating task. Beyond establishing these comparisons, it is hard to know what produced the differences. Given that in this case the question was a general one about social preference games, it might have to do with the specific games that participants had in mind when responding. For instance, thinking about the ultimatum game might result in a higher score for "negotiations"; thinking about player 2 in the trust game may lead to a higher rating for "tipping".

Apart from these comparisons, we can analyze how well the rating results correlate with the meta-analytical results presented in Studies 1 and 2. This will tell us more precisely the extent to which the most referenced field environments in the literature are also the ones that our ESA respondents perceived to be most associated to social preference games. The correlation between the proportion of papers that reference a given category in Study 1 and its score in the ESA survey ratings is $r = 0.579$ ($t = 3.332$, df = 22, $p = 0.003$). Excluding the two additional environments that were not present in the meta-analysis, the correlation is $r = 0.494$ ($t = 2.546$, df = 20, $p = 0.019$). So, the correlations are relatively high and statistically significant. Similarly, the correlation between the proportion of papers that reference a given category in the extended meta-analysis of Study 2 and the score of that category in the ESA survey ratings is $r = 0.447$ ($t = 2.346$, df = 22, $p = 0.028$). Excluding the two additional categories that are not present in the extended meta-analysis, the correlation is $r = 0.385$ ($t = 1.866$, df = 20, $p = 0.077$). Overall, this confirms that the results of the Study 1 and Study 2 meta-analyses of the literature and the ratings of the field environments by experimental and behavioral economists have a significant degree of convergent validity.

4.2.2. The open-ended questions

After removing nonsensical entries, in our open-ended questions, we obtained a total of 743 proposed field situations, contexts or behaviors, 150 associated to the dictator game, 131 to the ultimatum game, 149 to the trust game, and 313 to the public goods game (with and without punishment). We then further removed responses that were too generic, meaning that they did not clearly specify a field situation, context or behavior. Examples of this are entries that simply read "cooperation" or "conflict". This resulted in a total of 664 valid responses across games.

To analyze these responses and be able to compare them to our meta-analyses and to the ratings discussed in the previous section, we classified them into our 22 categories of field environments. To do this, we recruited 300 participants on Prolific and gave each of them the task of classifying 20 randomly selected responses into 23 categories (the 22 categories used in the meta-analyses plus an additional "other category").[3] Verbatim instructions are included in Appendix C. To present the responses to the Prolific participants, we also removed the ones that were repeated, reducing the entries to be classified to 527. On average, every response was rated by 11.4 people, with a minimum number of ratings per response of 4. Based on this methodology, we constructed three different indexes of the prevalence of the 23 categories in the open-ended questions.

In Index 1, we assigned each open response to the category most frequently chosen for that particular response by our Prolific raters. In the case of a tie, the response was categorized under all the equally selected top categories. In Index 2, we assigned each open response to all categories chosen for that particular response by the raters. In Index 3, we assigned each open response to all categories chosen for that particular response, weighted by the frequency with which the response was assigned to each category. As an illustration, consider a hypothetical open response classified by our Prolific raters as corresponding to the category of charity (by 6 raters), health care (by 3 raters), and policing (by 1 rater). Using Index 1, this response would only be classified as part of the charity category. With Index 2, the response would be equally classified as belonging to charity, health care and policing. With Index 3, the response would be classified as charity (with a 0.6 weight), health care (0.3 weight), and policing (0.1 weight). The three indexes are calculated in terms of the percentage probability that a given random response is classified in each category, as defined by the corresponding index (and accounting for the presence of repeated responses). Hence, Index 1 can be understood as representing the probability that a random open response is maximally assigned to a given category; Index 2 as the probability that at least 1 Prolific rater assigns a random open response to a given category; and Index 3 as the overall prevalence of a given category in the classification performed by our raters (i.e., the likelihood that a category is selected by a random rater when classifying a random open response).

Table 6 shows the prevalence of each category of field environments in the open responses according to the three indexes. The correlations between all three indexes are very high and significant ($r_{12} = 0.746$ $t = 5.130$, df = 21, $p < 0.001$; $r_{23} = 0.782$ $t = 5.750$, df = 21, $p < 0.001$; $r_{13} = 0.992$ $t = 36.329$, df = 21, $p < 0.001$), which confirms that our results are robust to different ways of measuring the prevalence of field environments. Index 3 is arguably the most meaningful measure, as it captures the prevalence of the different field environments in the open responses in a more comprehensive way. For this reason, we focus on it in our subsequent analyses, but the results are very similar using the other two indexes.

Looking at the categories ranked in the top five (not counting the option "other category"), there is an overlap of four out of five with respect to the results of the ratings. As explained in Section 4.2.1, the category "tipping" is in the top five based on the ratings but not based on the open responses. This is consistent with the Study 1 and 2 meta-analyses, where "tipping" is also not in the top five. In the results from the open responses, "tipping" is substituted in the top five by "financial/insurance markets and investment", a category that is in position 6 in the Study 1 analysis and 7 in the extended meta-analysis of Study 2.

[3] Participants in this exercise were presented with 24 distinct categories, treating "social and household interactions" and "evolution of social norms" as separate items. We then decided to merge "evolution of social norms" into the bigger category "social and household interactions". So in our analyses, these two items are treated as a single unified category, which is consistent with the classification used throughout the paper.

Table 6
Prevalence of each category of field environments in the open responses.

Rank	Field environments (22+1)	Index 1	Index 2	Index 3
1	Charity	15.36%	23.19%	14.10%
2	Social and household interactions	16.27%	37.80%	13.91%
3	Group and team dynamics	11.30%	27.26%	9.57%
4	Other category	7.83%	42.47%	7.94%
5	Financial/insurance markets and investment	7.08%	17.32%	6.02%
6	Negotiations	7.23%	23.04%	5.75%
7	Taxation	6.02%	9.49%	5.58%
8	Environmental policy	6.17%	9.94%	5.24%
9	Political and social issues	4.67%	24.7%	5.04%
10	Labor relations	4.97%	15.96%	3.87%
11	Business meetings	3.16%	20.33%	3.53%
12	Health care	4.37%	6.33%	3.43%
13	Consumption	2.71%	12.50%	2.56%
14	Firm behavior and pricing	1.96%	17.47%	2.51%
15	Legal proceedings	2.11%	10.84%	1.84%
16	Compensation and sanctioning schemes design	1.20%	13.25%	1.74%
17	Academic writing and publishing	1.51%	5.87%	1.46%
18	Tipping	1.05%	6.02%	1.41%
19	Policing	1.81%	7.38%	1.34%
20	Industrial disputes	1.66%	7.08%	1.26%
21	International agreement design	1.05%	4.52%	0.85%
22	Discrimination	0.15%	4.22%	0.58%
23	Military actions and their consequences	0.45%	1.05%	0.46%

Table 7
Prevalence of each category of field environments in the open responses by game (Index 3).

Rank	Field environments (22+1)	Total (N = 664)	DG (N = 148)	PGG (N = 260)	TG (N = 141)	UG (N = 115)
1	Charity	14.10%	47.99%	7.04%	2.30%	0.92%
2	Social and household interactions	13.91%	15.18%	13.56%	19.19%	7.32%
3	Group and team dynamics	9.57%	5.01%	11.23%	11.26%	9.40%
4	Other category	7.94%	6.47%	7.46%	10.99%	7.19%
5	Financial/insurance markets and investment	6.02%	2.79%	1.38%	19.24%	4.47%
6	Negotiations	5.75%	2.20%	0.59%	4.61%	23.41%
7	Taxation	5.58%	0.56%	11.81%	3.71%	0.25%
8	Environmental policy	5.24%	0.65%	12.57%	0.09%	0.86%
9	Political and social issues	5.04%	3.04%	7.74%	3.69%	3.20%
10	Labor relations	3.87%	1.68%	1.86%	4.11%	10.66%
11	Business meetings	3.53%	1.67%	2.30%	4.87%	6.80%
12	Health care	3.43%	1.36%	7.98%	0.07%	0.00%
13	Consumption	2.56%	1.92%	0.97%	4.30%	4.88%
14	Firm behavior and pricing	2.51%	1.05%	0.81%	4.33%	5.96%
15	Legal proceedings	1.84%	0.77%	1.48%	1.51%	4.40%
16	Compensation and sanctioning schemes design	1.74%	1.60%	1.82%	1.21%	2.41%
17	Academic writing and publishing	1.46%	0.18%	2.78%	0.69%	1.09%
18	Tipping	1.41%	5.22%	0.42%	0.19%	0.20%
19	Policing	1.34%	0.39%	2.83%	0.54%	0.18%
20	Industrial disputes	1.26%	0.00%	0.56%	1.86%	3.75%
21	International agreement design	0.85%	0.17%	0.91%	1.03%	1.35%
22	Discrimination	0.58%	0.06%	1.22%	0.15%	0.33%
23	Military actions and their consequences	0.46%	0.05%	0.68%	0.06%	0.98%

Crucially, the results obtained from the open responses (Index 3) correlate quite highly and significantly with the ratings presented in the previous section ($r_{3\text{-rate}} = 0.649$, $t = 3.821$, df = 20, $p = 0.001$), with the outcomes of the meta-analysis in Study 1 ($r_{3\text{-meta S1}} = 0.622$, $t = 3.549$, df = 20, $p = 0.002$), and with the expanded meta-analysis of Study 2 ($r_{3\text{–meta S2}} = 0.417$, $t = 2.049$, df = 20, $p = 0.054$), indicating a high degree of convergent validity. It is also important to note that, as captured by Index 3, the option "other category" only corresponded to approximately 8% of the responses. This suggests that our categories of field environments were perceived as a meaningful way to classify the field situations, contexts and behaviors by our sample of Prolific participants.

Table 7 shows the prevalence of the different categories of field environments in the open responses divided into the different games, based on Index 3. In these results by games, we can observe similar patterns to the ones described in Section 3.2.2 in relation to the extended meta-analysis of Study 2. The dictator game shows a ranking that is fairly similar to the total one, with the difference that the category "charity" (the highest ranked in both cases) has a considerably larger percentage for the dictator game. This is reasonable, given that both dictator game contributions and charitable giving are typically linked to altruism. For the public goods

game, we see again that "environmental policy" shows a higher percentage, this time together with "taxation". For the trust game, again "charity" goes down and "financial/insurance markets and investment" goes up. Finally, for the ultimatum game, we see one more time that the percentage for "charity" decreases and the one for "negotiations" substantially increases (see the comments in relation to these patterns in Section 3.2.2).

The correlations between Index 3 and the results of the Study 1 meta-analysis subdivided by games are also relatively high and mostly statistically significant ($r_{DG} = 0.430$, $t = 2.130$, df = 20, $p = 0.045$; $r_{PGG} = 0.476$, $t = 2.423$, df = 20, $p = 0.025$; $r_{TG} = 0.466$, $t = 2.358$, df = 20, $p = 0.028$; $r_{UG} = 0.384$, $t = 1.862$, df = 20, $p = 0.077$). Similarly, we find mostly sizable and significant correlations when examining the relationship between Index 3 and the results of the Study 2 extended meta-analysis disaggregated by games ($r_{DG} = 0.312$, $t = 1.467$, df = 20, $p = 0.158$; $r_{PGG} = 0.478$, $t = 2.434$, df = 20, $p = 0.024$; $r_{TG} = 0.442$, $t = 2.206$, df = 20, $p = 0.039$; $r_{UG} = 0.550$, $t = 2.944$, df = 20, $p = 0.008$). Again, this indicates a notable degree of convergent validity for specific games.

5. Combining all studies

After separately analyzing the literature (Studies 1 and 2) and the ratings and open questions included in the ESA survey (Study 3), this section combines these different elements to provide a joint summary and ranking of the different categories of field environments. To conduct our combined evaluation, we used principal component analysis. Specifically, we first standardized (z-normalized) the percentages corresponding to each category in the meta-analysis of Study 1, the extended meta-analysis of Study 2, and the ratings and open responses (using Index 3 as presented in Table 6) from the ESA survey. Then, we performed a principal component analysis using these four standardized measures, which showed a relatively high degree of internal consistency (Cronbach's $\alpha = 0.830$). The first component explains most of the variance in the data (approximately 67%) and is positively associated with the four measures. We therefore used the loadings of our four variables on this component to derive a combined measure. Table 8 shows the standardized resulting measure ("PC combined"), together with the standardized variables from the two meta-analyses, the ratings in the ESA survey, and the open responses in the ESA survey. We also included an additional variable combining the two meta-analyses by averaging them. The ranking of the 22 field environments based on the PC combined measure constitutes an ordered list of the environments that social preference games are expected to be related to, based both on the literature and the informed opinion of experimental and behavioral economists.

In Tables 9 to 12, we expand on these analyses by presenting combined results by game. To do so, we focus on the results of the two meta-analyses and the ESA open responses for each game separately (as in Tables 2, 5 and 7). These results do not include the ESA ratings, as these are a general measure and cannot be linked to specific games. The scores presented in the tables were obtained by replicating our principal component procedure with our game-specific measures. This exercise yielded similar results in terms of internal consistency and variance explained (DG: Cronbach's $\alpha = 0.747$, variance explained = 67.2%; PGG: Cronbach's $\alpha = 0.735$, variance explained = 65.4%; TG: Cronbach's $\alpha = 0.760$, variance explained = 67.7%; UG: Cronbach's $\alpha = 0.745$, variance explained = 66.3%). The rankings by games in Tables 9 to 12 largely reproduce the meaningful differences between games discussed in Sections 3.2.2 and 4.2.2.

6. How to use our results as a guide

Our results provide a guide of field environments and behaviors related to social preference games that can be used in several ways. To help potential users, this section explains three main ways in which it can be utilized.

First, when studies report significant correlations between social preference games and field behaviors, we can use this guide to evaluate whether those correlation are expected or not. If they are expected, this lends further support to the correlations and confirms the expectations. If they are not expected, this raises the flag of potential spurious correlations or false positives that would have to be confirmed, and at the same time suggests a potential new association that researchers have not been considering. In any case, our guide will provide additional insights when evaluating this kind of research. The same holds for papers that report non-significant correlations. If they are expected, this lends further support for the lack of a relationship and confirms the expectations. If they are not expected, this can lead to a reconsideration of the methods used, while it challenges an established belief.

This kind of comparison of results with expectations can be done at a general level using Table 8, which shows the scores and ranks obtained by our 22 categories of field environments in the two meta-analyses, in the ESA ratings, in the ESA open responses and in all sources combined. If we want to do comparisons focusing on specific games, we can use Tables 9 to 12, which contain scores and ranks based on a combination of results from the two meta-analyses and the ESA open responses for specific games. Potentially, we can dig even deeper by looking at more specific field settings and behaviors uncovered in the detailed meta-analysis of Study 1. We can do this by checking the meta-analysis materials, as illustrated in Table 2. The most relevant material for this would be our Excel file "Additional tables Study 1" included in our OSF repository.

Second, if a researcher plans to run a study related to a particular field domain of social behavior, this guide can be used to evaluate if this domain is likely to be associated with social preference games, and to which particular games. This can be useful, among other things, to evaluate the possibility of using social preference games to address specific behaviors of interest. As explained in the previous paragraph, the guide we provide can be used for this at a general level using Table 8, at the level of specific games using Tables 9 to 12, or at a more detailed level using our meta-analysis materials.

Third, if someone wants to conduct research on the external validity of social preference games, this guide points to the specific field environments in which external validity is expected to be high. Starting by focusing on these environments seems an optimal

Table 8
Standardized scores for each category of field environments in the two meta-analyses, ESA ratings, ESA open responses, and a combined measure from principal component analysis.

Rank	Field environments	Meta-Analysis 1	Meta-Analysis 2	Average Meta-Analyses	ESA Closed	ESA Open	PC Combined
1	Social and household interactions	2.55	1.02	1.79	1.20	2.49	2.62
2	Charity	1.16	0.25	0.71	1.75	2.54	2.09
3	Political and social issues	1.86	3.40	2.63	0.47	0.22	1.32
4	Group and team dynamics	−0.22	1.29	0.53	1.38	1.38	1.07
5	Financial/insurance markets and investment	0.13	0.12	0.12	−0.87	0.47	0.35
6	Compensation and sanctioning schemes design	1.51	1.38	1.45	0.40	−0.63	0.32
7	Labor relations	0.82	0.04	0.43	0.60	−0.08	0.23
8	Negotiations	−0.22	−0.17	−0.19	1.36	0.40	0.19
9	Taxation	0.13	−0.62	−0.25	0.19	0.36	0.17
10	Environmental policy	−0.57	−0.24	−0.40	0.83	0.27	−0.03
11	Health care	0.13	−0.62	−0.25	−0.90	−0.19	−0.21
12	Firm behavior and pricing	0.13	0.19	0.16	−0.93	−0.43	−0.22
13	Business meetings	−0.91	−0.66	−0.78	−0.38	−0.17	−0.52
14	Tipping	−0.22	−0.73	−0.48	1.02	−0.71	−0.63
15	Consumption	−0.91	−0.35	−0.63	−1.93	−0.42	−0.66
16	Legal proceedings	−0.57	−0.47	−0.52	−0.66	−0.60	−0.66
17	Military actions and their consequences	−0.22	−0.79	−0.50	−1.07	−0.96	−0.84
18	Academic writing and publishing	−0.91	−0.73	−0.82	−0.17	−0.70	−0.87
19	Policing	−0.91	−0.64	−0.77	−0.97	−0.73	−0.89
20	Industrial disputes	−0.91	−0.77	−0.84	−0.63	−0.75	−0.92
21	International agreement design	−0.91	−0.56	−0.74	0.29	−0.86	−0.93
22	Discrimination	−0.91	−0.35	−0.63	−0.96	−0.92	−0.97

Table 9
Dictator Game: Standardized scores for each category of field environments in the two meta-analyses, ESA ratings, ESA open responses, and a combined measure from principal component analysis.

Rank	Field environments	Meta-Analysis 1	Meta-Analysis 2	Average Meta-Analyses	ESA Open	PC Combined
1	Charity	1.37	0.71	1.04	4.25	2.53
2	Political and social issues	2.34	3.39	2.86	−0.12	2.28
3	Social and household interactions	1.85	1.64	1.75	1.06	1.88
4	Compensation and sanctioning schemes design	1.37	1.06	1.21	−0.26	0.93
5	Health care	0.88	−0.40	0.24	−0.28	0.18
6	Taxation	0.40	−0.57	−0.09	−0.36	−0.15
7	Labor relations	−0.09	−0.05	−0.07	−0.25	−0.15
8	Tipping	−0.09	−0.45	−0.27	0.09	−0.17
9	Group and team dynamics	−1.06	0.88	−0.09	0.07	−0.17
10	Firm behavior and pricing	−0.09	−0.05	−0.07	−0.31	−0.18
11	Financial/insurance markets and investment	−0.09	−0.28	−0.18	−0.14	−0.20
12	Military actions and their consequences	0.40	−0.69	−0.15	−0.41	−0.21
13	Legal proceedings	−0.09	−0.57	−0.33	−0.34	−0.37
14	Discrimination	−0.57	0.01	−0.28	−0.41	−0.43
15	Negotiations	−0.57	−0.28	−0.43	−0.20	−0.45
16	Consumption	−0.57	−0.40	−0.48	−0.23	−0.50
17	Environmental policy	−0.57	−0.57	−0.57	−0.35	−0.61
18	Policing	−0.57	−0.69	−0.63	−0.37	−0.66
19	Business meetings	−1.06	−0.63	−0.84	−0.25	−0.83
20	Academic writing and publishing	−1.06	−0.57	−0.81	−0.40	−0.86
21	International agreement design	−1.06	−0.75	−0.90	−0.40	−0.93
22	Industrial disputes	−1.06	−0.75	−0.90	−0.41	−0.93

strategy to uncover the domains to which the games generalize. In this respect, we believe our results can lead to a more organized and systematic investigation of external validity. Again, as explained in the previous paragraphs, the target environments to be tested can be obtained at a general level using Table 8, in terms of specific games in Tables 9 to 12, and in terms of more specific settings and behaviors by using our meta-analysis materials.

Potentially, the different elements of the guide we provide and the specific ways in which we have classified games and field environments could also serve as a starting point for alternative classification approaches. Future analyses could also look in more detail into the specific game characteristics and experimental protocols and how they affect perceptions of external validity. In this sense, we believe that our materials (tables, Excel files, data, etc.) allow for much experimentation and may lead to interesting future studies.

Table 10
Public Goods Game: Standardized scores for each category of field environments in the two meta-analyses, ESA ratings, ESA open responses, and a combined measure from principal component analysis.

Rank	Field environments	Meta-Analysis 1	Meta-Analysis 2	Average Meta-Analyses	ESA Open	PC Combined
1	Group and team dynamics	2.33	1.91	2.12	1.55	2.15
2	Social and household interactions	1.33	0.27	0.80	2.07	1.89
3	Environmental policy	0.32	0.49	0.40	1.85	1.49
4	Taxation	0.32	−0.37	−0.03	1.68	1.20
5	Political and social issues	−0.69	3.29	1.30	0.78	1.01
6	Charity	1.33	0.44	0.89	0.63	0.91
7	Compensation and sanctioning schemes design	2.33	1.35	1.84	−0.53	0.57
8	Health care	−0.69	−0.46	−0.57	0.84	0.30
9	Labor relations	0.32	−0.37	−0.03	−0.52	−0.35
10	International agreement design	0.32	−0.24	0.04	−0.73	−0.47
11	Policing	−0.69	−0.33	−0.51	−0.30	−0.48
12	Academic writing and publishing	−0.69	−0.55	−0.62	−0.32	−0.53
13	Industrial disputes	0.32	−0.55	−0.11	−0.81	−0.58
14	Business meetings	−0.69	−0.68	−0.68	−0.42	−0.63
15	Financial/insurance markets and investment	−0.69	−0.20	−0.44	−0.63	−0.68
16	Legal proceedings	−0.69	−0.50	−0.59	−0.60	−0.72
17	Discrimination	−0.69	−0.59	−0.64	−0.66	−0.78
18	Firm behavior and pricing	−0.69	−0.33	−0.51	−0.75	−0.79
19	Consumption	−0.69	−0.50	−0.59	−0.72	−0.80
20	Negotiations	−0.69	−0.55	−0.62	−0.80	−0.87
21	Military actions and their consequences	−0.69	−0.76	−0.72	−0.78	−0.90
22	Tipping	−0.69	−0.76	−0.72	−0.84	−0.94

Table 11
Trust Game: Standardized scores for each category of field environments in the two meta-analyses, ESA ratings, ESA open responses, and a combined measure from principal component analysis.

Rank	Field environments	Meta-Analysis 1	Meta-Analysis 2	Average Meta-Analyses	ESA Open	PC Combined
1	Social and household interactions	1.97	1.38	1.68	2.73	2.72
2	Financial/insurance markets and investment	1.05	0.77	0.91	2.74	2.29
3	Political and social issues	1.97	3.25	2.61	−0.06	1.38
4	Labor relations	1.97	0.35	1.16	0.01	0.70
5	Group and team dynamics	−0.80	0.35	−0.22	1.30	0.69
6	Business meetings	0.13	−0.37	−0.12	0.15	0.04
7	Compensation and sanctioning schemes design	0.13	1.26	0.69	−0.51	0.03
8	Consumption	0.13	−0.19	−0.03	0.05	0.02
9	Firm behavior and pricing	−0.80	0.90	0.05	0.05	0.02
10	Negotiations	0.13	−0.62	−0.24	0.10	−0.05
11	Tipping	1.05	−0.68	0.19	−0.70	−0.30
12	Discrimination	0.13	−0.13	0.00	−0.70	−0.45
13	Taxation	−0.80	−0.74	−0.77	−0.06	−0.47
14	Charity	−0.80	−0.25	−0.53	−0.31	−0.51
15	Academic writing and publishing	0.13	−0.74	−0.31	−0.61	−0.54
16	Legal proceedings	−0.80	−0.37	−0.59	−0.46	−0.63
17	Industrial disputes	−0.80	−0.80	−0.80	−0.39	−0.70
18	International agreement design	−0.80	−0.68	−0.74	−0.54	−0.77
19	Policing	−0.80	−0.62	−0.71	−0.63	−0.81
20	Health care	−0.80	−0.56	−0.68	−0.72	−0.85
21	Environmental policy	−0.80	−0.74	−0.77	−0.71	−0.89
22	Military actions and their consequences	−0.80	−0.80	−0.80	−0.72	−0.91

7. Conclusion

We have investigated the field environments that are expected to be associated to social preference games according to the people who use them. In Studies 1 and 2, we have done this by conducting a systematic review and meta-analysis of the literature, and in Study 3 by asking specialist researchers in the fields of experimental and behavioral economics. Our results show a high degree of convergent validity among our different measures, which has allowed us to put together a cohesive guide of field environments that can be used at different levels of detail.

An important methodological contribution of our paper is the development and validation of an LLM–based method to identify and extract external validity claims from the literature. This approach allows for a systematic and scalable analysis of much broader sets of papers than would be feasible using manual coding alone. LLM approaches have the potential limitation of relying on the model's interpretation of the texts, which may not always align with the intended meaning of the original authors or the

Table 12
Ultimatum Game: Standardized scores for each category of field environments in the two meta-analyses, ESA ratings, ESA open responses, and a combined measure from principal component analysis.

Rank	Field environments	Meta-Analysis 1	Meta-Analysis 2	Average Meta-Analyses	ESA Open	PC Combined
1	Negotiations	0.62	1.89	1.25	3.61	3.10
2	Labor relations	1.97	0.41	1.19	1.21	1.39
3	Social and household interactions	1.97	0.62	1.30	0.58	1.04
4	Political and social issues	0.62	3.05	1.83	−0.19	0.95
5	Firm behavior and pricing	1.97	0.84	1.40	0.33	0.94
6	Group and team dynamics	−0.74	0.52	−0.11	0.97	0.63
7	Compensation and sanctioning schemes design	0.62	0.94	0.78	−0.34	0.21
8	Consumption	0.62	−0.22	0.20	0.12	0.16
9	Financial/insurance markets and investment	−0.74	0.20	−0.27	0.05	−0.08
10	Business meetings	−0.74	−0.75	−0.74	0.49	−0.08
11	Legal proceedings	−0.74	−0.33	−0.53	0.03	−0.25
12	Discrimination	0.62	−0.33	0.14	−0.73	−0.44
13	Industrial disputes	−0.74	−0.85	−0.80	−0.09	−0.49
14	Tipping	0.62	−0.64	−0.01	−0.76	−0.55
15	International agreement design	−0.74	−0.64	−0.69	−0.54	−0.73
16	Military actions and their consequences	−0.74	−0.54	−0.64	−0.61	−0.74
17	Charity	−0.74	−0.54	−0.64	−0.62	−0.75
18	Academic writing and publishing	−0.74	−0.64	−0.69	−0.59	−0.76
19	Environmental policy	−0.74	−0.85	−0.80	−0.63	−0.85
20	Health care	−0.74	−0.64	−0.69	−0.79	−0.89
21	Taxation	−0.74	−0.75	−0.74	−0.75	−0.89
22	Policing	−0.74	−0.75	−0.74	−0.76	−0.90

interpretation of human raters. In this sense, it is reassuring that our validation exercise shows a high degree of agreement with manual coding. More generally, given that human raters are also prone to mistakes and subjectivities in their assessments, further research is needed to determine which of the two approaches is more reliable.

Going back to our opening paragraph, external validity has been the Achilles Heel of laboratory experimentation on social preferences (and in economics more generally), and more systematic research is needed to evaluate the extent of the problem and devise effective solutions for it. Some recent developments have already shown that there is potential to use social preference games in ways that are externally valid by using approaches such as contextualizing them (Wang and Navarro-Martinez, 2023a) or aggregating measures to reduce measurement error (Wang and Navarro-Martinez, 2023b). We hope that the guide we present in this paper provides a valuable resource to help advance research on external validity by making it more systematic and assessable.

Funding

This research was funded by the BBVA Foundation, Spain (Fundacion BBVA-EI-2019-D.Navarro), the Ramon Areces Foundation (Fundacion Ramon Areces 2019-Navarro), the Spanish Ministry of Science and Innovation (PID2019-105249GB-I00, PID2022-137908NB-I00), and ICREA (ICREA Academia 2024-Daniel Navarro).

Declaration of competing interest

The authors declare no competing interests.

Appendix A. Instructions ESA survey

Thank you for participating in this study.

This survey is designed to investigate the perceptions of behavioral and experimental economists regarding the external validity of social preference games (specifically, the dictator game, trust game, ultimatum game and public goods game).

To advance research on the explanatory power of these games in relation to field behavior, it is important to understand which are the types of field situations that are most closely associated with the games. To this end, we will ask you to provide up to 3 field situations, contexts or behaviors that you think are closely related to particular social preference games. We will also ask you to rate how informative the behavior of participants in social preference games is to understand behavior in various specific contexts.

If you complete the survey, there is a 10% chance that you are selected to receive a $50 prize (paid through a Paypal transfer). If you are selected to receive the prize, at the end of the survey you will be redirected to a separate (optional) questionnaire where you can provide your contact information to receive the payment. Note that by using a separate questionnaire, we ensure that your responses to the main survey are fully anonymous, as there will be no way of matching them to your contact information.

The anonymous responses collected in this study may be made public in a research data repository, and the resulting findings may be presented at scientific meetings or published in scientific journals.

By clicking on the button below, you are giving your consent to participate in this study.

Field Situations/Contexts/Behaviors Related to Social Preference Games

In the following boxes, please briefly describe at least 1 and up to 3 field situations, contexts or behaviors that you think are closely related to the specified games:

Dictator game.

[]

Ultimatum game.

[]

Trust game.

[]

Public goods game, without punishment.

[]

Public goods game, with punishment.

[]

Rating of Specific Field Contexts

In general, how closely related do you think is behavior in any of the social preference games we mentioned before (dictator game, ultimatum game, trust game, public goods game with or without punishment) to behavior in the following field contexts? [Slider from 0 – Not related at all to 10 – Very closely related]

– Political and social issues.
– Compensation and sanctioning schemes design.
– Health care.
– Financial/insurance markets and investment.
– Social and household interactions.
– Labor relations.
– Group and team dynamics.
– Firm behavior and pricing.
– Environmental policy.
– Military actions and their consequences.
– Business meetings.
– Industrial disputes.
– Charity.
– Legal proceedings.
– Negotiations.
– Tipping.
– Taxation.
– Policing.
– International agreement design.
– Academic writing and publishing.
– Consumption.
– Evolution of social norms.
– Group dynamics (including those related to the Kyoto Protocol, UN security council, work teams in companies and sports teams).
– Discrimination.
– Saving for retirement.
– Financial information avoidance.

Personal Information:

– Seniority (years since PhD)
o No PhD
o Current PhD Student
o Less than 10 years since PhD defense.
o Between 10 and 20 years since PhD defense.
o More than 20 years since PhD defense.
– Field(s) of specialization: I have worked/published in the following fields (select all that apply):
o Applied Economics.
o Behavioral Economics.
o Econometrics/Statistics.
o Experimental Economics.
o Judgement and Decision Making.
o Macroeconomics.
o Microeconomics.
o Psychology.

– Have you ever used social preference games (the ones mentioned in this study or others) as part of your research?
o Yes
o No
– In general, to what extent do you think social preference games are a good tool to understand naturally occurring social behavior in the field?
[Slider from 0 – Not good at all to 10 – Very good]
– Are you aware of the literature on the external validity of social preference games?
o Not aware
o Somewhat aware
o Very aware
– To what extent do you think research on the external validity of social preference games is important?
[Slider from 0 – Not important at all to 10 – Very important]
– If you have any comments, please write them in the following box:
[]

End of Survey Message [Those Who Don't Receive Money]

Thank you for completing this survey! Your responses have been saved.
Unfortunately, you were not selected to receive the $50 prize.
Once again, we thank you for your time.

End of Survey Message [Those Who Receive Money]

Thank you for completing this survey! Your responses have been saved.
Congratulations, you were selected to receive the $50 prize.
Please click on the following link to access a questionnaire where we will collect your contact information to send you the payment. Note that by using a separate questionnaire, we ensure that your responses to the main survey cannot be matched to your contact information.
Link
Once again, we thank you for your time.

Appendix B. Analysis of additional information provided by the participants

Apart from the summary statistics (see Table B.1), it is interesting to investigate which personal characteristics affect the perceptions of external validity of social preference games. To this end, we calculated an overall score for each individual by averaging the ratings given to the 22 environments. This provides an indicator of how externally valid social preference games are perceived to be. We then regressed this measure (using OLS) on the additional variables we collected, including level of seniority, fields of specialization, whether participants had used social preference games as part of their research, whether they were aware of

Table B.1
Summary statistics ESA sample.

Variable	N	Mean	Std. Dev.
Seniority	88		
... No PhD.	5	5.7%	
... Current PhD student.	17	19.3%	
... Less than 10 years since PhD defense.	39	44.3%	
... Between 10 and 20 years since PhD defense.	20	22.7%	
... More than 20 years since PhD defense.	7	8%	
Field(s) of specialization	88		
... Applied economics	27	30.3%	
... Behavioral economics	71	79.8%	
... Experimental economics	72	80.9%	
... Judgement and decision making	20	22.5%	
... Macroeconomics	3	3.4%	
... Microeconomics	32	36%	
... Psychology	18	20.2%	
... Econometrics/Statistics	8	9%	
Used SPG in the past	88		
... No	12	13.6%	
... Yes	76	86.4%	
Awareness external validity literature	88		
... Not aware	14	15.9%	
... Somewhat aware	54	61.4%	
... Very aware	20	22.7%	
Usefulness SPG	88	6.716	2.269
Importance external validity literature	88	7.295	2.6

Note: Demographic information was missing for one participant, so our N here is 88.

Table B.2
Regression analysis of additional variables.

	Dependent variable:			
	EV (1)	Usefulness (2)	Awareness lit EV (3)	Importance lit EV (4)
Seniority	0.287	0.535*	−0.080	−0.576*
	(0.237)	(0.282)	(0.079)	(0.336)
Applied econ	0.451	−0.206	0.065	−0.399
	(0.426)	(0.508)	(0.145)	(0.616)
Behavioral econ	−0.492	−0.146	0.031	−0.546
	(0.476)	(0.568)	(0.162)	(0.688)
Experimental econ	0.762	1.427**	0.177	2.125***
	(0.513)	(0.612)	(0.165)	(0.702)
JDM	−0.633	0.152	0.350**	−0.423
	(0.482)	(0.574)	(0.159)	(0.675)
Macroeconomics	0.410	−0.593	0.780**	−1.548
	(1.115)	(1.329)	(0.367)	(1.558)
Microeconomics	0.041	0.696	−0.032	0.012
	(0.407)	(0.485)	(0.139)	(0.591)
Psychology	−0.041	−0.261	0.240	1.940***
	(0.492)	(0.586)	(0.158)	(0.672)
Stats and econometrics	−0.411	0.476	0.224	0.404
	(0.692)	(0.825)	(0.235)	(0.998)
Used SPG	−0.807	0.263	0.506**	−0.011
	(0.599)	(0.715)	(0.197)	(0.834)
Awareness lit EV	−0.433	−0.450		
	(0.335)	(0.400)		
Importance lit EV	0.188**	0.332***		
	(0.079)	(0.094)		
Constant	4.606***	2.693**	0.369	6.495***
	(0.862)	(1.028)	(0.237)	(1.005)
Observations	88	88	88	88
R^2	0.206	0.316	0.229	0.208
Adjusted R^2	0.079	0.207	0.129	0.105
Residual std. error	1.695	2.021	0.580	2.460
F statistic	1.618	2.891***	2.292**	2.024**

Notes: Seniority was transformed into a numeric variable by assigning a value of 0 to the categories "No PhD" and "Current PhD student", a value of 1 to "Less than 10 years since PhD defense", a value of 2 to "Between 10 and 20 years since PhD defense", and a value of 3 to "More than 20 years since PhD defense". Awareness was transformed into a numeric variable by assigning a value of 0 to the category "Not aware", a value of 1 to "Somewhat aware" and a value of 2 to "Very aware". Demographic information was missing for one participant, so our models are estimated on the data provided by 88 respondents.

* Stands for statistical significance at the 10% level.

** Stands for statistical significance at the 5% level.

*** Stands for statistical significance at the 1% level.

the literature on the external validity of social preference games, and the extent to which they thought that research on this topic is important. Table B.2 shows our regression results (Model 1). Only the importance given to research on this topic is significantly associated to perceived external validity. The relationship is positive, which means that people who report giving more importance to research on external validity also rate higher the overall external validity of the games. To further explore our data, Table B.2 contains three other regressions. Model 2 uses the extent to which participants believed social preference games are a good tool to study social behavior as a dependent variable, Model 3 uses awareness of the external validity literature, and Model 4 the importance given to it.

Appendix C. Instructions classification of open responses (Prolific participants)

Instructions:

In a previous survey, a number of individuals had to provide examples of certain types of real-life behaviors, contexts, and situations (i.e., giving to charity). We now need to classify these responses into 23 predefined categories that we have.

In the next pages, we will show you a total of 20 of these past responses (one per page). Please, **assign each response to the prespecified category that you think is most closely associated with it.**

IMPORTANT NOTES:
1 - Some responses are longer, and others are very short and might be ambiguous. Please do your best to classify each response in one of the given categories based on what the individual wrote. If you really think it does not fit in any of the categories, there is an "Other category" option.
2 - Some responses might fit into various categories. In these cases, simply select the category that you think is most closely associated with the given response.
3 - Some responses clearly correspond to one specific category. **If you fail to appropriately categorize more than one of these clear responses, you might not be allowed to continue with the task and receive the payment.** So, please pay attention to every response.

Click "next" to begin with the classification task:

Before starting with the classification task, please take a minute to familiarize yourself with the 23 prespecified categories:

Academic writing and publishing
Business meetings
Charity
Compensation and sanctioning schemes design
Consumption
Discrimination
Environmental policy
Evolution of social norms
Financial/insurance markets and investment
Firm behavior and pricing
Group and team dynamics
Health care
Industrial disputes
International agreement design
Labor relations
Legal proceedings
Military actions and their consequences
Negotiations
Policing
Political and social issues
Social and household interactions
Taxation
Tipping

Please classify the following response:

"XXX"

o Academic writing and publishing
o Business meetings
o Charity
o Compensation and sanctioning schemes design
o Consumption
o Discrimination
o Environmental policy
o Evolution of social norms
o Financial/insurance markets and investment
o Firm behavior and pricing
o Group and team dynamics
o Health care
o Industrial disputes
o International agreement design
o Labor relations
o Legal proceedings
o Military actions and their consequences
o Negotiations
o Policing
o Political and social issues
o Social and household interactions
o Taxation
o Tipping
o Other category

Data availability

All the data and meta-analysis materials referenced in the paper are available in the following OSF repository:Link.

References

Al-Ubaydli, O., List, J.A., 2013. On the Generalizability of Experimental Results in Economics: With a Response To Camerer. Working Paper 19666, National Bureau of Economic Research.
Camerer, C., 2011. The promise and success of lab-field generalizability in experimental economics: A critical reply to levitt and list. SSRN. 1977749.
Falk, A., Heckman, J.J., 2009. Lab experiments are a major source of knowledge in the social sciences. Sci. 326 (5952), 535–538.
Galizzi, M.M., Navarro-Martinez, D., 2019. On the external validity of social preference games: A systematic lab-field study. Manag. Sci. 65 (3), 976–1002.
Kessler, J.B., Vesterlund, L., The external validity of laboratory experiments: The misleading emphasis on quantitative effects. 18, 392–405.
Levitt, S.D., List, J.A., 2007. What do laboratory experiments measuring social preferences reveal about the real world? J. Econ. Perspect. 21 (2), 153–174.
Lowenstein, G., 1999. Experimental economics from the vantage-point of behavioral economics. Econ. J. 109, 25–34.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P., 2009. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. Ann. Intern. Med. 151 (4), 264–269.
Wang, X., Navarro-Martinez, D., 2023a. Bridging the gap between the economics lab and the field: Dictator games and donations. Judgm. Decis. Mak. 18, e18.
Wang, X., Navarro-Martinez, D., 2023b. Increasing the external validity of social preference games by reducing measurement error. Games Econom. Behav. 141, 261–285.