

Searching for the External Validity of Social Preference Games

A Guide of Field Environments Based on Expert Perceptions

Daniel Navarro-Martinez¹ Sergio Pirla²

¹Universitat Pompeu Fabra & Barcelona GSE, Spain.

²Universidad de Zaragoza & IEDIS, Spain.

Leibniz Open Science Day 2025

- Social preference games (SPGs).
- Ongoing debate about external validity.
 - Levitt and List, 2007; Galizzi and Navarro-Martinez, 2019
- Unclear which field contexts are most relevant.

- Social preference games (SPGs).
- Ongoing debate about external validity.
 - Levitt and List, 2007; Galizzi and Navarro-Martinez, 2019
- Unclear which field contexts are most relevant.

- Social preference games (SPGs).
- Ongoing debate about external validity.
 - Levitt and List, 2007; Galizzi and Navarro-Martinez, 2019
- Unclear which field contexts are most relevant.

- A guide to the field environments to which SPGs are expected to generalize.
- Expert perceptions from behavioral and experimental economists.
- Structure:
 - Study 1: External validity claims in Top 5 economics journals.
 - Study 2: LLM-based replication and extension of Study 1.
 - Study 3: Economic Science Association (ESA) survey.

- A guide to the field environments to which SPGs are expected to generalize.
- Expert perceptions from behavioral and experimental economists.
- Structure:
 - Study 1: External validity claims in Top 5 economics journals.
 - Study 2: LLM-based replication and extension of Study 1.
 - Study 3: Economic Science Association (ESA) survey.

- A guide to the field environments to which SPGs are expected to generalize.
- Expert perceptions from behavioral and experimental economists.
- Structure:
 - Study 1: External validity claims in Top 5 economics journals.
 - Study 2: LLM-based replication and extension of Study 1.
 - Study 3: Economic Science Association (ESA) survey.

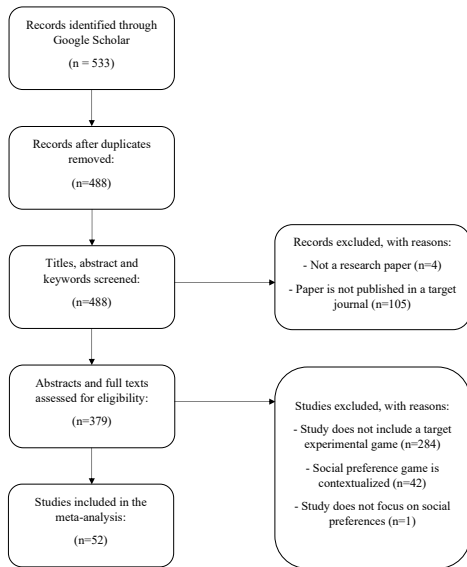
Study 1: Top 5 Analysis

- Meta-analysis of Top 5 economics journals:
 - Econometrica
 - American Economic Review
 - Journal of Political Economy
 - Quarterly Journal of Economics
 - Review of Economic Studies
- Social preference games analyzed:
 - Dictator Game (DG)
 - Public Goods Game (PGG)
 - Trust Game (TG)
 - Ultimatum Game (UG)

Study 1: Top 5 Analysis

- Meta-analysis of Top 5 economics journals:
 - Econometrica
 - American Economic Review
 - Journal of Political Economy
 - Quarterly Journal of Economics
 - Review of Economic Studies
- Social preference games analyzed:
 - Dictator Game (DG)
 - Public Goods Game (PGG)
 - Trust Game (TG)
 - Ultimatum Game (UG)

Study 1: Top 5 Analysis



Study 1: Top 5 Analysis

- Analysis of 52 papers:
 - Extracted field environments:
 - Example: “Do our results apply outside of experiments? A natural place to look for sex differences in altruism is in charitable giving.” (Andreoni & Vesterlund, 2001, *QJE*)
 - Grouped into 22 broad categories
- Summary statistics:
 - 46 distinct field settings identified
 - 65% of papers mention at least one field setting (34/52)
 - Variation across games:
 - DG: 78%, PGG: 56%, TG: 73%, UG: 42%

Study 1: Top 5 Analysis

- Analysis of 52 papers:
 - Extracted field environments:
 - Example: “Do our results apply outside of experiments? A natural place to look for sex differences in altruism is in charitable giving.” (Andreoni & Vesterlund, 2001, *QJE*)
 - Grouped into 22 broad categories
- Summary statistics:
 - 46 distinct field settings identified
 - 65% of papers mention at least one field setting (34/52)
 - Variation across games:
 - DG: 78%, PGG: 56%, TG: 73%, UG: 42%

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%

Rank	Total (N=34)	Specific settings (N=8)	Total (N=34)	DG (N=18)	PGG (N=9)	TG (N=8)	UG (N=5)	Patterns of behavior (N=14)	Games linked
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

Study 2: LLM Approach

- Evaluate the LLM-based procedure against results from Study 1.
 - ChatGPT-4o
- Use LLM to expand the meta-analysis:
 - Economic Journal, Journal of the European Economic Association, Review of Economics and Statistics
 - Experimental Economics, Games and Economic Behavior, Journal of Economic Behavior & Organization

Study 2: LLM Approach

- Evaluate the LLM-based procedure against results from Study 1.
 - ChatGPT-4o
- Use LLM to expand the meta-analysis:
 - Economic Journal, Journal of the European Economic Association, Review of Economics and Statistics
 - Experimental Economics, Games and Economic Behavior, Journal of Economic Behavior & Organization

Study 2: LLM Validation

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)?

Study 2: LLM Validation

- 379 entries retained after applying exclusion criteria 1–3.
- Of these papers:
 - 90% correctly classified
- Final sample:
 - 52 papers
 - 33 out of 52 also included in the final sample of Study 1 (66%)
 - 19 out of 52 not included (37%)

Study 2: LLM Validation

- 379 entries retained after applying exclusion criteria 1–3.
- Of these papers:
 - 90% correctly classified
- Final sample:
 - 52 papers
 - 33 out of 52 also included in the final sample of Study 1 (66%)
 - 19 out of 52 not included (37%)

Study 2: LLM Validation

- 379 entries retained after applying exclusion criteria 1–3.
- Of these papers:
 - 90% correctly classified
- Final sample:
 - 52 papers
 - 33 out of 52 also included in the final sample of Study 1 (66%)
 - 19 out of 52 not included (37%)

Study 2: LLM Validation

Step	Explanation	Prompt	Prompt Text
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."

- Summary results:
 - 45 out of 52 papers (87%) contained at least one reference to a real-life setting (vs. 65% in Study 1)
 - Average number of field settings per paper: 4.3 (vs. 2.4 in Study 1)
- Correlation of percentage of papers referencing each category (compared with Study 1):
 - All games: $r = 0.814$, $p < 0.001$
 - DG: $r = 0.820$, $p < 0.001$
 - PGG: $r = 0.577$, $p = 0.005$
 - TG: $r = 0.565$, $p = 0.006$
 - UG: $r = 0.735$, $p < 0.001$

- Summary results:
 - 45 out of 52 papers (87%) contained at least one reference to a real-life setting (vs. 65% in Study 1)
 - Average number of field settings per paper: 4.3 (vs. 2.4 in Study 1)
- Correlation of percentage of papers referencing each category (compared with Study 1):
 - All games: $r = 0.814$, $p < 0.001$
 - DG: $r = 0.820$, $p < 0.001$
 - PGG: $r = 0.577$, $p = 0.005$
 - TG: $r = 0.565$, $p = 0.006$
 - UG: $r = 0.735$, $p < 0.001$

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%

- Extension to additional journals:
 - Economic Journal, Journal of the European Economic Association, Review of Economics and Statistics
 - Experimental Economics, Games and Economic Behavior, Journal of Economic Behavior & Organization
- Summary statistics:
 - Initial records: 2,136
 - Papers retained after applying inclusion criteria 1–3: 1,658
 - Final sample: 433 papers

- Extension to additional journals:
 - Economic Journal, Journal of the European Economic Association, Review of Economics and Statistics
 - Experimental Economics, Games and Economic Behavior, Journal of Economic Behavior & Organization
- Summary statistics:
 - Initial records: 2,136
 - Papers retained after applying inclusion criteria 1–3: 1,658
 - Final sample: 433 papers

- Summary results:

- 382 out of 433 papers (88%) contained at least one reference to a real-life setting (vs. 65% in Study 1, 87% in the replication)
- Average number of field settings per paper: 4.3 (vs. 2.4 in Study 1, 4.3 in the replication)

- Correlation of percentage of papers referencing each category (compared with Study 1):

- All games: $r = 0.734, p < 0.001$
- DG: $r = 0.746, p < 0.001$
- PGG: $r = 0.487, p = 0.021$
- TG: $r = 0.632, p = 0.002$
- UG: $r = 0.544, p = 0.009$

- Summary results:
 - 382 out of 433 papers (88%) contained at least one reference to a real-life setting (vs. 65% in Study 1, 87% in the replication)
 - Average number of field settings per paper: 4.3 (vs. 2.4 in Study 1, 4.3 in the replication)
- Correlation of percentage of papers referencing each category (compared with Study 1):
 - All games: $r = 0.734$, $p < 0.001$
 - DG: $r = 0.746$, $p < 0.001$
 - PGG: $r = 0.487$, $p = 0.021$
 - TG: $r = 0.632$, $p = 0.002$
 - UG: $r = 0.544$, $p = 0.009$

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%

- Studies 1 and 2:
 - Searched for mentions of real-life contexts in nearly 500 papers
 - Classified these mentions into 22 categories
 - Analyzed the prevalence of each category in the literature

Study 3: ESA Survey

- Distributed via ESA mailing list; 10% chance of receiving \$50.
- $N = 90$; median completion time: 14 minutes.
- Two main tasks:
 - Closed ratings:
 - Overall rating (0–10) for each of the 22 categories
 - Open questions:
 - Describe 1–3 field contexts most closely related to each game
- Also collected respondent information:
 - Level of seniority, fields of specialization, use of social preference games, etc.

Study 3: ESA Survey

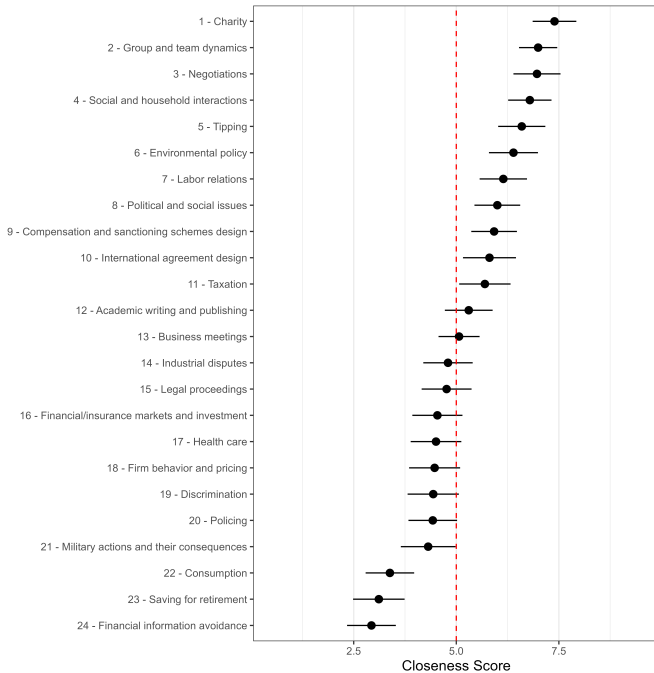
- Distributed via ESA mailing list; 10% chance of receiving \$50.
- $N = 90$; median completion time: 14 minutes.
- Two main tasks:
 - Closed ratings:
 - Overall rating (0–10) for each of the 22 categories
 - Open questions:
 - Describe 1–3 field contexts most closely related to each game
- Also collected respondent information:
 - Level of seniority, fields of specialization, use of social preference games, etc.

Study 3: ESA Survey

- Distributed via ESA mailing list; 10% chance of receiving \$50.
- $N = 90$; median completion time: 14 minutes.
- Two main tasks:
 - Closed ratings:
 - Overall rating (0–10) for each of the 22 categories
 - Open questions:
 - Describe 1–3 field contexts most closely related to each game
- Also collected respondent information:
 - Level of seniority, fields of specialization, use of social preference games, etc.

Study 3: ESA Survey

- Distributed via ESA mailing list; 10% chance of receiving \$50.
- $N = 90$; median completion time: 14 minutes.
- Two main tasks:
 - Closed ratings:
 - Overall rating (0–10) for each of the 22 categories
 - Open questions:
 - Describe 1–3 field contexts most closely related to each game
- Also collected respondent information:
 - Level of seniority, fields of specialization, use of social preference games, etc.



- Open-ended questions:

- A total of 743 field settings were mentioned.
- Prolific raters classified them into the 22 predefined categories (plus an “Other” category).
- Each response was classified by an average of 11.4 raters.

- Convergent validity:

- $r_{\text{ESA ratings} - \text{meta S1}} = 0.494, p = 0.019$
- $r_{\text{ESA ratings} - \text{meta S2}} = 0.385, p = 0.077$
- $r_{\text{open} - \text{meta S1}} = 0.622, p = 0.002$
- $r_{\text{open} - \text{meta S2}} = 0.417, p = 0.054$
- $r_{\text{ESA ratings} - \text{open}} = 0.649, p = 0.001$

- Open-ended questions:
 - A total of 743 field settings were mentioned.
 - Prolific raters classified them into the 22 predefined categories (plus an “Other” category).
 - Each response was classified by an average of 11.4 raters.
- Convergent validity:
 - $r_{\text{ESA ratings} - \text{meta S1}} = 0.494, p = 0.019$
 - $r_{\text{ESA ratings} - \text{meta S2}} = 0.385, p = 0.077$
 - $r_{\text{open} - \text{meta S1}} = 0.622, p = 0.002$
 - $r_{\text{open} - \text{meta S2}} = 0.417, p = 0.054$
 - $r_{\text{ESA ratings} - \text{open}} = 0.649, p = 0.001$

- Aggregation of results across studies:
 - Principal Component Analysis (PCA)
 - Inputs: Study 1 (meta-analysis of Top 5), Study 2 (LLM extension), open-ended responses, and closed ratings
 - Cronbach's $\alpha = 0.830$
 - Results computed for:
 - All games combined
 - Each game independently

- Aggregation of results across studies:
 - Principal Component Analysis (PCA)
 - Inputs: Study 1 (meta-analysis of Top 5), Study 2 (LLM extension), open-ended responses, and closed ratings
 - Cronbach's $\alpha = 0.830$
 - Results computed for:
 - All games combined
 - Each game independently

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

- A guide to field environments based on expert perceptions
- All materials are available online (OSF repository)
- Just published at JEBO

- A guide to field environments based on expert perceptions
- All materials are available online (OSF repository)
- Just published at JEBO

- A guide to field environments based on expert perceptions
- All materials are available online (OSF repository)
- Just published at JEBO