

Online Supplement to

What's moral wiggle room? A theory specification

Alina Fahrenwaldt^{*1,2}, Fiona tho Pesch^{*1,3,4}, Susann Fiedler^{1,5}, and Anna Baumert^{1,4}

*Shared first authorship (correspondence may be directed to both first authors):

alinafahrenwaldt@web.de & fiona@thopesch.de

¹Max Planck Institute for Research on Collective Goods, Bonn, Germany

²University of Cologne, Germany,

³École Normale Supérieure, Université PSL, Paris, France

⁴University of Wuppertal, Germany

⁵Vienna University of Economics and Business, Austria

Supplement A: Operationalizations & Decision Settings used in Research on the Effect of MWR

Table SA1

List of operationalizations of MWR and evaluation based on MWR-T

Name of Treatment	Short Description	References	Evaluation based on MWR-T
Original Paper on MWR (DWK)			
Hidden Information	Possibility to ignore information on outcome for recipient per distribution option (see also Outcome Ignorance)	Dana et al., 2007	Valid
Plausible Deniability	Possibility to avoid the decision by omitting any action until a specific amount of time has passed, resulting in randomness deciding (see also Hybrid Treatment)	Dana et al., 2007	Valid
Multiple Dictator	Group decision where one agent's vote suffices to implement the fair outcome, but all agents' votes are necessary to implement the unfair outcome (intransparency only in case of prosocial but not in case of selfish behavior)	Dana et al., 2007	Not valid, see Loophole 2 & Definition of MWR
MWR-T			
Additional Choice Attributes	Presentation of additional, otherwise decision-irrelevant information about each distribution option, which can be used to justify a distribution decision; may involve pitting safe agent-profiting option against risky prosocial option	Exley, 2016; Haisley & Weber, 2010; Snyder et al., 1979; tho Pesch & Dana, 2024	Valid

Decision Implementation Uncertainty	Agent's decision may be overruled by a random generator or by other agent(s); includes group decision contexts	Matthey & Regner, 2014, 2015	Valid
Default Implementation	Possibility to implement a default distribution option	Gärtner & Sandberg, 2017	Valid
Delegation	Possibility to delegate a distribution decision to chance or to another agent	Erat, 2013; Hamman et al., 2010	Valid
Exiting	Possibility to exit the decision situation	Andreoni et al., 2017; Dana et al., 2006; DellaVigna et al., 2012	Valid
Hybrid Treatment	Combination of two or more of the valid operationalizations listed in this table	Dana et al., 2007; Gärtner & Sandberg, 2017	Valid
Information Asymmetry	Only the agent knows about total size of the pie (i.e., the endowment that can be shared)	Güth & Huck, 1997; Ockenfels & Werner, 2012	Valid
Omission	Omit any action; may be combined with outcome ignorance, delegation, default implementation, or exiting, see Hybrid Treatment	Gärtner & Sandberg, 2017	Valid
Outcome Ignorance	Possibility to avoid information regarding the recipient's payoffs prior to decision	Bartling et al., 2014; Bell et al., 2017; D'Adda et al., 2018; Dana et al., 2007; Ehrich & Irwin, 2005; Grossman, 2014; Grossman & van der Weele, 2017; Matthey & Regner, 2011; Momsen & Ohndorf, 2020	Valid

Table SA2*List of decision settings in which MWR was tested and evaluation based on MWR-T*

Decision Setting	References	Evaluation based on MWR-T
Non-strategic, unilateral decision games with incentivization for the agent	Dana et al., 2007	Valid
Strategic decision games	Bolton et al., 2019	Not valid, see Auxiliary Assumption 3
Vicarious decision-making scenarios	Cerrone & Engel, 2019	Not valid, see Auxiliary Assumption 1
Reciprocation in trust games	Matthey & Regner, 2015; van der Weele et al., 2014	Not valid, see Auxiliary Assumption 3
Unethical framing & take options	Balafoutas et al., 2021; Cappelen et al., 2013; List, 2007; Zhang & Ortmann, 2014	Not valid, see Auxiliary Assumption 4

Supplement B: Specification of the original formulation of MWR effects (DWK)

Short description of the study on MWR by DWK (2007)

In their seminal paper, DWK found that inducing MWR by manipulating different situational features reduced the likelihood for prosociality in their participants. They argued that the employed manipulations all reduced the transparency between action and outcome:

1. In the *hidden information treatment*, subjects had the opportunity to avoid information on the consequences of their behavior for the recipient. Initially, it was hidden from the dictator which outcome each distribution option entailed for the recipient. Agents could choose to reveal this information prior to making their decision, or they could choose a distribution option without revealing. In any case, agents made their decisions (whether or not to reveal and which distribution option to implement) in private, such that only the outcome was observable for others. Accordingly, transparency was reduced by impeding the recipient's (and any other potential third-party observer's) ability to infer the agent's behavior from the observed outcomes. Moreover, transparency was also reduced for the agent in terms of a lower predictability of how their behavior affected the outcome for the recipient in case they decided not to reveal.
2. In the *plausible deniability treatment*, subjects had to make their decision in a pre-specified (and objectively sufficient) amount of time, else they were cut off from making a decision and one of the two distribution options was implemented with 50% probability instead. Knowledge as to how payoffs were determined (i.e., via the agent's behavior or via the random implementation) remained private to the agent. Accordingly, transparency was again reduced by inhibiting the inferability of the agent's behavior from the observable outcomes.
3. The *multiple dictator treatment* presented the agent with a group decision context and the recipient knew that their outcome was determined by the decision of the group of agents. Crucially, it was sufficient that one of the group members decided to implement the fair outcome for the passive recipient (versus an unfair, agent-benefitting outcome). Therefore, in this treatment, transparency was only reduced by impeding the recipient's ability to infer a specific agent's behavior from a fair outcome. In contrast, in case of an unfair outcome, the recipient could clearly infer that all agents had acted selfishly, implying full transparency. We discuss this issue in more detail in the section on the original paper's loopholes.

Table SB1*Propositions derived from the original paper on MWR (DWK)*

Proposition	Antecedence	Consequence
For all agents and situations specified in the respective auxiliary assumptions:		
1	IF MWR (instead of no MWR)	THEN higher likelihood of selfish behavior
2a	IF MWR (instead of no MWR)	THEN reduced relevance of fairness norms and constraints (i.e., feeling less compelled to give or having an excuse or justification not to give)
2b	IF reduced relevance of fairness norms and constraints (i.e., feeling less compelled to give or having an excuse or justification not to give)	THEN higher likelihood of selfish behavior

Note. The formulations in this table are abbreviations of full sentences. For example, Proposition 1 reads as “*If a person is in a situation containing moral wiggle room, then this person will be more likely to show selfish behavior relative to a situation not containing MWR.*” Furthermore, in DWK’s study designs, all proposed links between concepts appear to be viewed as probabilistic rather than deterministic and as linear relationships.

Table SB2

Concept definitions & operationalizations according to the proposition derived from original postulation of MWR (DWK)

Proposition	Concept label	Verbal Definition	Measurement/ Operationalization
1, 2a	MWR (Moral Wiggle Room)	Situational characteristics that remove the transparency (commonly known one-to-one mapping) between (selfish) behavior and the outcomes to both parties (Dana et al., 2007, p.69, p.77)	Hidden information treatment
			Plausible deniability treatment
			Multiple dictator treatment
1, 2b	Selfish behavior	Decisions that maximize one's own profit while disregarding other people's payoff (Dana et al., 2007, p.71, p.73)	Binary decision (one monetary distribution option yielding a higher payoff to the agent but a lower payoff for a passive recipient, compared to another, more egalitarian distribution option)
2a, 2b	Relevant mechanism	Not defined in the paper; <i>Fairness norms and constraints</i> interchangeably used with other terms (<i>not feeling compelled to give</i> or <i>having an excuse</i> or <i>justification</i> not to give; Dana et al., 2007, p.69, p.78)	Not operationalized in the paper

Note. DWK estimate the effect of MWR on selfish behavior at the population-level, but they seem to assume that the effects of MWR exist at the individual level as well.

Table SB3*Auxiliary assumptions (per proposition derived from original postulation of MWR)*

Proposition	Auxiliary Assumption
1-2b	(1) The decision must have consequences for oneself and others, and the interests of these parties must conflict. (Dana et al., 2007, p. 75, footnote 10)
1, 2a	(2) The presence of MWR in the social decision situation must not restrict the agent's choice (i.e., their ability to implement any of the outcomes available without MWR). (Dana et al., 2007, p. 69)
2a, 2b	(3) The interaction between the agent and the recipient is non-strategic and unilateral. (Dana et al., 2007, p. 70)
1-2b	(4) There are two independent and sometimes conflicting motives active in agents in the population: an agent's preferences over payoff distributions AND an agent's self- or social image concerns. (Dana et al., 2007, pp.68-69, pp.70-78)

Reasons for the auxiliary assumptions: If Auxiliary Assumption 1 is not met, selfish behavior is indistinguishable from prosocial behavior. If Auxiliary Assumption 2 is not met, selfish behavior may be an artifact of not being able to choose freely. If Auxiliary Assumption 3 is not met, the measured effects may be confounded by other factors (e.g., expectation of reciprocity). Moreover, if Auxiliary Assumption 4 is not met, the behavioral effect of MWR (and its underlying mechanisms) may not be observable.

Table SB4*Overview of loopholes and their solutions*

Loophole No.	Content	Solution
1	Inconsistency between the definition of MWR and the proposed explanatory mechanism	More consistent definition of MWR as “situational characteristics that obfuscate the signal which the outcome of an own-payoff-maximizing (i.e., potentially selfish) behavior sends to others about one’s intention to behave selfishly” (Appendix, Table A2)
2	Unsuitable operationalizations of the concept of MWR	List with explanations for suitable operationalizations (Appendix, Table A2) and exclusion of unsuitable operationalizations (e.g., multiple dictator treatment) from final theory (Online Supplement A, Table SA1)
3	Lack of differentiation of psychological mechanisms	Differentiation into three interrelated psychological mechanisms (see Figure 1 & Table 2, Propositions 2a-4b): anticipated social image damage, perceived social norms, anticipatory guilt
4	Lack of clear definition and operationalization of mechanism concepts	Refined definitions & operationalizations (for an overview see Appendix, Table A2)
5	Lack of clear definitions and operationalizations of agents’ motivations	Refined definitions & operationalizations (Appendix, Table A2) and focus on social image (consistently in motives & proposed mechanisms, Appendix, Tables A1 & A3)
6	No specification of individual differences	Provision of additional propositions accounting for heterogeneity (Propositions 6 & 7, Appendix, Table A1)
7	No explicit specification of the response format & action space available to agents	Extension of the auxiliary assumptions (Appendix, Table A3, Auxiliary Assumption 4)

Supplement C: Different degrees of MWR (Extension of MWR-T)

In MWR-T, we only considered comparisons between settings with and without MWR. However, we argue that different treatments may actually introduce different *degrees* of MWR (with lower degrees implying less transparent situations). This idea is especially interesting for the differential effects propositions (Props 6 & 7), since it allows for more fine-grained predictions. Crucially, we posit that the effects of other-regarding preferences, social image concerns and degree of MWR interact to shape behavior (for a visualization, see Figure C1). The lower the transparency, the more MWR there is to be exploited relative to the baseline condition without MWR, and the greater is the expected increase in selfish behavior as moderated by social image concerns and other-regarding preferences.

As already stated in Auxiliary Assumption 5, for people who score very low on social image concerns (i.e., $y = 0$), the introduction of any degree of MWR does not play a role and the distribution decision is purely determined by other-regarding preferences. The higher the social image concerns, the more a person is influenced in their behavior by the degree of MWR. Similarly, for people who are truly prosocial, social image concerns and MWR do not influence their decision, as they are already very likely to behave prosocially and we can see a ceiling effect (see Figure C1).

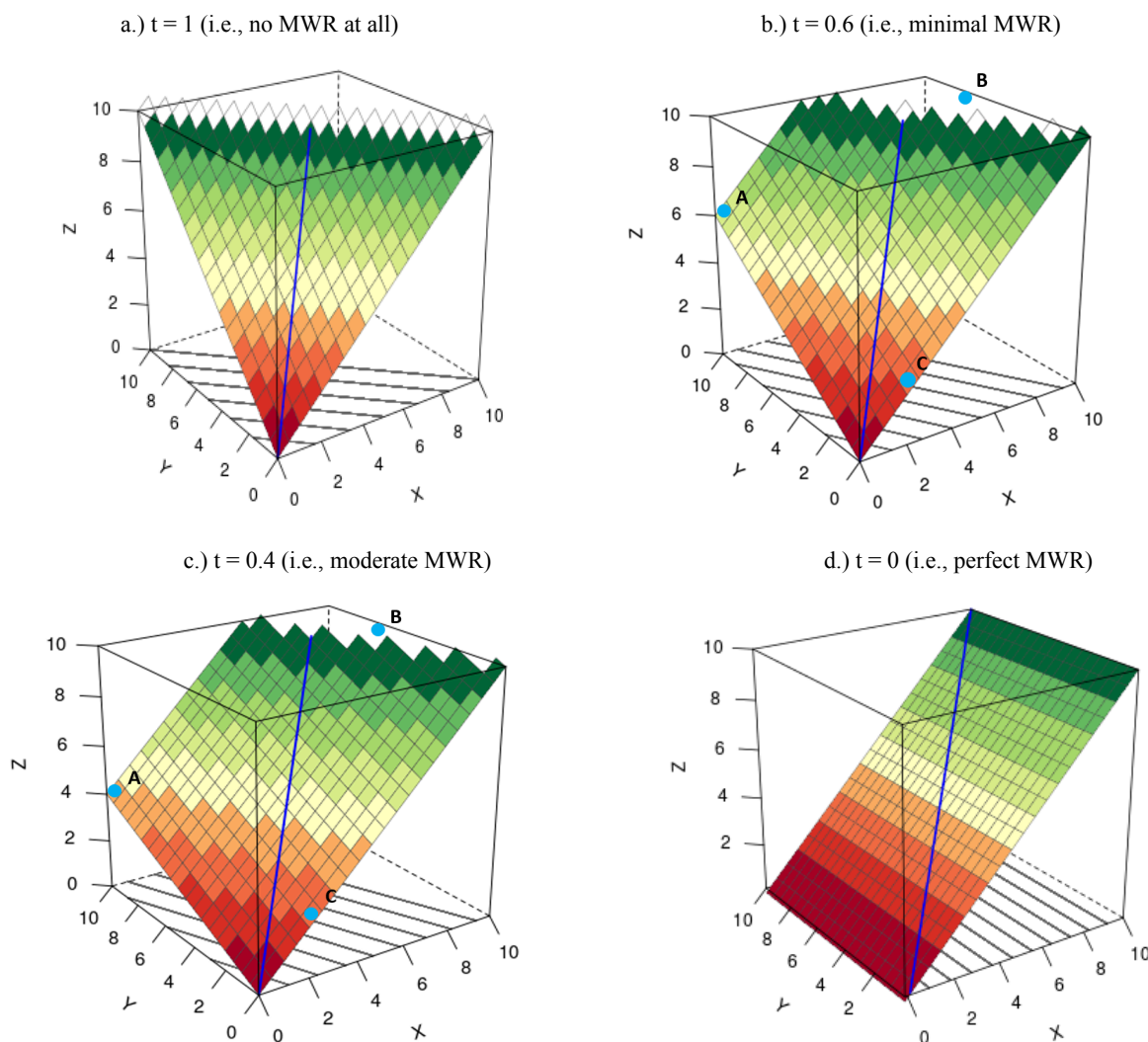
Our graphs (Figure C1 a-b) depict social image concerns on the y-axis ($y = 0$ would have no social image concerns), other-regarding preferences on the x-axis ($x = 0$ would be purely selfish), and the probability for prosocial behavior on the z-axis. For simplicity, social behavior is assumed to be measured in a binary fashion. Note that a 50% probability to implement the prosocial decision (i.e., $z = 5$) represents the point of indifference between selfish and prosocial motives. In a completely transparent situation ($t = 1$, i.e., no MWR at all) social image concerns are just as (or even more) important for behavior as other-regarding preferences (see Figure C1a, slopes for y and x are equally steep), which means that observed behavior is not very helpful in understanding an agent's intention and latent other-regarding preferences. In a completely intransparent situation ($t = 0$, i.e., perfect MWR) social image concerns do not play a role at all (see Figure C1d, only slope for x), allowing for reliable inference of an agent's true other-regarding preferences from their observed behavior. It is rather unrealistic to create a fully transparent or a fully intransparent situation, so that all possible situations should fall somewhere in between. In the following examples, we will therefore focus on Figures C1b and C1c.

In Figure C1b, you see three different potential individuals. Person A is purely selfish ($x = 0$), while having very high social image concerns ($y = 10$). As the situation is relatively transparent ($t = 0.6$), this person is more likely to behave prosocially, i.e., overcome the point of indifference between selfish and prosocial motives. In a situation with less transparency ($t = 0.4$, Figure C1c), however, this person becomes more likely to behave selfishly. Person B is someone who is extremely prosocial ($x = 10$), and also has relatively high social image concerns ($y = 7$). This person is very likely to behave prosocially, both with high ($t = 0.6$, Figure C1b) and with low transparency ($t = 0.4$, Figure C1c). You can even see that this high level of prosociality makes it irrelevant how much a person cares about their social image, as that person truly wants to behave prosocially, so that neither social image concerns, nor the degree of transparency can change their decision (matching Proposition 7 in MWR-T). Person C is relatively selfish ($x = 2$), and has no social image concerns ($y = 0$). This person is

likely to behave selfishly, both in the rather transparent ($t = 0.6$) and the rather intransparent ($t = 0.4$) situation (i.e., this person will never overcome the point of indifference).

Figure SC1a-d

A graphic depiction of the effects of individual characteristics influencing the likelihood of showing prosocial behavior



Note. X-axis = other-regarding preferences, Y-axis = social image concerns, Z-axis = probability for prosocial behavior. The color gradient indicates the degree of selfishness (dark red: probability of choosing the prosocial outcome = 0, dark green: probability of choosing the prosocial outcome = 1).

Note that, in a binary decision task (as assumed here), any behavioral effect of MWR will be driven mainly by people with a medium score on other-regarding preferences (see also Table 2, Prop. 7), as they will be the most likely to switch from a prosocial to a selfish option when MWR is introduced. In continuous dependent variables, however, we expect to see changes in all participants with high social image concerns. This hypothesis is supported by results of Dana et al. (2006) who found that all kinds of participants would exit a dictator game at a cost, not only the ones who initially share a lot or little.

References (Online Supplement)

- Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, 125(3), 625–653. <https://doi.org/10.1086/691703>
- Balafoutas, L., Sandakov, F., & Zhuravleva, T. (2021). No moral wiggle room in an experimental corruption game. *Frontiers in Psychology*, 12, 3509. <https://doi.org/10.3389/fpsyg.2021.701294>
- Bartling, B., Engl, F., & Weber, R. A. (2014). Does willful ignorance deflect punishment? An experimental study. *European Economic Review*, 70, 512–524. <https://doi.org/10.1016/j.eurocorev.2014.06.016>
- Bell, E., Norwood, F. B., & Lusk, J. L. (2017). Are consumers willfully ignorant about animal welfare? *Animal Welfare*, 26(4), 399–402. <https://doi.org/10.7120/09627286.26.4.399>
- Bolton, G. E., Kusterer, D. J., & Mans, J. (2019). Inflated reputations: Uncertainty, leniency, and moral wiggle room in trader feedback systems. *Management Science*, 65(11), 5371–5391. <https://doi.org/10.1287/mnsc.2018.3191>
- Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., & Tyran, J.-R. (2013). Give and take in dictator games. *Economics Letters*, 118(2), 280–283. <https://doi.org/10.1016/j.econlet.2012.10.030>
- Cerrone, C., & Engel, C. (2019). Deciding on behalf of others does not mitigate selfishness: An experiment. *Economics Letters*, 183, 108616. <https://doi.org/10.1016/j.econlet.2019.108616>
- D’Adda, G., Gao, Y., Golman, R., & Tavoni, M. (2018). *It’s so hot in here: Information avoidance, moral wiggle room, and high air conditioning usage* (Working Paper ID 3149330). Social Science Research Network.

<https://papers.ssrn.com/abstract=3149330>

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100(2), 193–201. <https://doi.org/10.1016/j.obhdp.2005.10.001>

Processes, 100(2), 193–201. <https://doi.org/10.1016/j.obhdp.2005.10.001>

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67–80.

<https://doi.org/10.1007/s00199-006-0153-z>

DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, 127(1), 1–56.

<https://doi.org/10.1093/qje/qjr050>

Ehrich, K. R., & Irwin, J. R. (2005). Willful ignorance in the request for product attribute information. *Journal of Marketing Research*, 42(3), 266–277.

<https://doi.org/10.1509/jmkr.2005.42.3.266>

Erat, S. (2013). Avoiding lying: The case of delegated deception. *Journal of Economic Behavior & Organization*, 93, 273–278. <https://doi.org/10.1016/j.jebo.2013.03.035>

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2), 587–628. <https://doi.org/10.1093/restud/rdv051>

Gärtner, M., & Sandberg, A. (2017). Is there an omission effect in prosocial behavior? A laboratory experiment on passive vs. active generosity. *PLOS ONE*, 12(3), 1–21.

<https://doi.org/10.1371/journal.pone.0172496>

Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences.

Management Science, 60(11), 2659–2665. <https://doi.org/10.1287/mnsc.2014.1989>

Grossman, Z., & van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1), 173–217.

<https://doi.org/10.1093/jeea/jvw001>

- Güth, W., & Huck, S. (1997). From ultimatum bargaining to dictatorship—An experimental study of four games varying in veto power. *Metroeconomica*, 48(3), 262–299.
<https://doi.org/10.1111/1467-999X.00033>
- Haisley, E. C., & Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, 68(2), 614–625.
<https://doi.org/10.1016/j.geb.2009.08.002>
- Hamman, J. R., Loewenstein, G., & Weber, R. A. (2010). Self-interest through delegation: An additional rationale for the principal-agent relationship. *American Economic Review*, 100(4), 1826–1846. <https://doi.org/10.1257/aer.100.4.1826>
- List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3), 482–493. <https://doi.org/10.1086/519249>
- Matthey, A., & Regner, T. (2011). Do I really want to know? A cognitive dissonance-based explanation of other-regarding behavior. *Games*, 2(1), 114–135.
<https://doi.org/10.3390/g2010114>
- Matthey, A., & Regner, T. (2014). *More than outcomes: The role of self-image in other-regarding behavior* (Working Paper 2014–036). Jena Economic Research Papers. <https://www.econstor.eu/handle/10419/108543>
- Matthey, A., & Regner, T. (2015). *Do reciprocators exploit or resist moral wiggle room? An experimental analysis* [Working Paper].
<https://www.econstor.eu/handle/10419/144897>
- Momsen, K., & Ohndorf, M. (2020). When do people exploit moral wiggle room? An experimental analysis of information avoidance in a market setup. *Ecological Economics*, 169, 106479. <https://doi.org/10.1016/j.ecolecon.2019.106479>
- Ockenfels, A., & Werner, P. (2012). ‘Hiding behind a small cake’ in a newspaper dictator game. *Journal of Economic Behavior & Organization*, 82(1), 82–85.

<https://doi.org/10.1016/j.jebo.2011.12.008>

- Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology*, 37(12), 2297–2306. <https://doi.org/10.1037/0022-3514.37.12.2297>
- tho Pesch, F., & Dana, J. (2024). Attributional ambiguity reduces charitable giving by relaxing social norms. *Journal of Experimental Social Psychology*, 110, 104530.
- van der Weele, J. J., Kulisa, J., Kosfeld, M., & Friebe, G. (2014). Resisting moral wiggle room: How robust is reciprocal behavior? *American Economic Journal: Microeconomics*, 6(3), 256–264. <https://doi.org/10.1257/mic.6.3.256>
- Zhang, L., & Ortmann, A. (2014). The effects of the take-option in dictator-game experiments: A comment on Engel's (2011) meta-study. *Experimental Economics*, 17(3), 414–420. <https://doi.org/10.1007/s10683-013-9375-7>