Competition and moral behavior:
A meta-analysis of 45 crowd-sourced experimental designs

Christoph Huber[1], Anna Dreber[2,3], Jürgen Huber[4], Magnus Johannesson[2], Michael Kirchler[4], Utz Weitzel[5,6,7], ***« 88 more authors*[†] *»***, Felix Holzmeister[3,*]

[1]Institute for Markets and Strategy, WU Vienna University of Economics and Business, Vienna, Austria, [2]Department of Economics, Stockholm School of Economics, Stockholm, Sweden, [3]Department of Economics, University of Innsbruck, Innsbruck, Austria, [4]Department of Banking and Finance, University of Innsbruck, Innsbruck, Austria, [5]Department of Finance, School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, [6]Department of Economics and Business Economics, Nijmegen School of Management, Radboud University, Nijmegen, The Netherlands, [7]Tinbergen Institute, Amsterdam, The Netherlands. [†]**75 additional institutions in 19 different countries.**

Introduction

Morals and Markets

- **Does competition erode, promote, or not affect moral behavior?**
  - Smith (1776) argued that markets can have a civilizing effect on behavior.
  - Markets may attenuate conflict and violence (Hirschman 1977), stimulate morality, and induce trust (Henrich *et al.* 2001, 2006; Choi and Storr 2020).
  - Marx (1867) and Veblen (1899) expected markets to be innately alienating.
  - Competition may create incentives for unethical practices and undermine moral values by crowding out social norms (Shleifer 2004; Sandel 2012).

Morals and Markets

- **Does competition erode, promote, or not affect moral behavior?**
  - Smith (1776) argued that markets can have a civilizing effect on behavior.
  - Markets may attenuate conflict and violence (Hirschman 1977), stimulate morality, and induce trust (Henrich *et al.* 2001, 2006; Choi and Storr 2020).
  - Marx (1867) and Veblen (1899) expected markets to be innately alienating.
  - Competition may create incentives for unethical practices and undermine moral values by crowding out social norms (Shleifer 2004; Sandel 2012).

Morals and Markets

- **Does competition erode, promote, or not affect moral behavior?**
  - Smith (1776) argued that markets can have a civilizing effect on behavior.
  - Markets may attenuate conflict and violence (Hirschman 1977), stimulate morality, and induce trust (Henrich *et al.* 2001, 2006; Choi and Storr 2020).
  - Marx (1867) and Veblen (1899) expected markets to be innately alienating.
  - Competition may create incentives for unethical practices and undermine moral values by crowding out social norms (Shleifer 2004; Sandel 2012).

Morals and Markets

- More recently, this debate has been taken to the laboratory...
  - Falk and Szech (2013) provide evidence that subjects are less likely to forego money to prevent the death of a mouse in competitive settings.
  - Follow-up experiments question the robustness of this finding based on rather inconclusive evidence (e.g., Bartling *et al.* 2015; Kirchler *et al.* 2016; Pigors and Rockenbach 2016; Ockenfels *et al.* 2020; Bartling *et al.* 2022).

Morals and Markets

- More recently, this debate has been taken to the laboratory...
  - Falk and Szech (2013) provide evidence that subjects are less likely to forego money to prevent the death of a mouse in competitive settings.
  - Follow-up experiments question the robustness of this finding based on rather inconclusive evidence (e.g., Bartling *et al.* 2015; Kirchler *et al.* 2016; Pigors and Rockenbach 2016; Ockenfels *et al.* 2020; Bartling *et al.* 2022).

Design Heterogeneity

- **Why does empirical evidence lead to different conclusions?**
  - Sample heterogeneity: relatively small to moderate variability in effect sizes across samples (e.g., Klein *et al.* 2014, 2018; Ebersole *et al.* 2016).
  - Analytic heterogeneity: significant variance in estimates across analyses (Silberzahn *et al.* 2018; Botvinik-Nezer *et al.* 2018; Menkveld *et al.* 2021).
  - Design heterogeneity: systematic evidence is scarce (Landy *et al.* 2020).

Design Heterogeneity

- **Why does empirical evidence lead to different conclusions?**
  - Sample heterogeneity: relatively small to moderate variability in effect sizes across samples (e.g., Klein *et al.* 2014, 2018; Ebersole *et al.* 2016).
  - Analytic heterogeneity: significant variance in estimates across analyses (Silberzahn *et al.* 2018; Botvinik-Nezer *et al.* 2018; Menkveld *et al.* 2021).
  - Design heterogeneity: systematic evidence is scarce (Landy *et al.* 2020).

Design Heterogeneity

- **Why does empirical evidence lead to different conclusions?**
  - Sample heterogeneity: relatively small to moderate variability in effect sizes across samples (e.g., Klein *et al.* 2014, 2018; Ebersole *et al.* 2016).
  - Analytic heterogeneity: significant variance in estimates across analyses (Silberzahn *et al.* 2018; Botvinik-Nezer *et al.* 2018; Menkveld *et al.* 2021).
  - Design heterogeneity: systematic evidence is scarce (Landy *et al.* 2020).

Design Heterogeneity

- **#ManyDesigns:**
  - As there are multiple valid approaches to operationalize competition and morality, we implemented a crowd-sourced project (Uhlmann *et al.* 2019).
- We eliminate sampling and analytic heterogeneity …
  - … by collecting data on various designs using a single sample
  - … by randomly assigning participants into one of the designs
  - … by standardizing the statistical analyses across designs

Design Heterogeneity

- **#ManyDesigns:**
  - As there are multiple valid approaches to operationalize competition and morality, we implemented a crowd-sourced project (Uhlmann *et al.* 2019).

- We eliminate sampling and analytic heterogeneity …
  - … by collecting data on various designs using a single sample
  - … by randomly assigning participants into one of the designs
  - … by standardizing the statistical analyses across designs

Crowd-Sourcing Research Designs

Research Teams (RTs)

- We left it to the research teams to operationalize competition and morality.
- RTs were required to design (and later program) a between-subjects study.
- RTs filed a preregistration (incl. a proposed analysis) for their experiment.

- Sample of n = 200 per treatment, i.e., n = 400 for each design/experiment.
- Envisaged sample of 50 research teams, i.e., a total of ∼20,000 participants.
- Sample of n = 400 are sufficiently large to obtain adequate statistical power to detect small to medium effect sizes (t-test: $\pi = 0.9$ for $d = 0.32$ at $\alpha = 0.05$).

- After screening applications, 102 RTs were invited to submit a research design.
- 95 RTs submitted a design, and 50 RTs were randomly selected to participate.
- 45 RTs delivered the software and were thus included in the data collection.

Research Teams (RTs)

- We left it to the research teams to operationalize competition and morality.
- RTs were required to design (and later program) a between-subjects study.
- RTs filed a preregistration (incl. a proposed analysis) for their experiment.

- Sample of n = 200 per treatment, i.e., n = 400 for each design/experiment.
- Envisaged sample of 50 research teams, i.e., a total of $\sim$20,000 participants.
- Sample of n = 400 are sufficiently large to obtain adequate statistical power to detect small to medium effect sizes (t-test: $\pi = 0.9$ for $d = 0.32$ at $\alpha = 0.05$).

- After screening applications, 102 RTs were invited to submit a research design.
- 95 RTs submitted a design, and 50 RTs were randomly selected to participate.
- 45 RTs delivered the software and were thus included in the data collection.

Research Teams (RTs)

- We left it to the research teams to operationalize competition and morality.
- RTs were required to design (and later program) a between-subjects study.
- RTs filed a preregistration (incl. a proposed analysis) for their experiment.

- Sample of n = 200 per treatment, i.e., n = 400 for each design/experiment.
- Envisaged sample of 50 research teams, i.e., a total of $\sim$20,000 participants.
- Sample of n = 400 are sufficiently large to obtain adequate statistical power to detect small to medium effect sizes (t-test: $\pi = 0.9$ for $d = 0.32$ at $\alpha = 0.05$).

- After screening applications, 102 RTs were invited to submit a research design.
- 95 RTs submitted a design, and 50 RTs were randomly selected to participate.
- 45 RTs delivered the software and were thus included in the data collection.

Design Requirements

- The design has to be eligible to obtain (fast track) IRB approval, i.e., …
  - no deception, preservation of participants' anonymity, explicit information (duration, repetitions, interactions, random processes), confidentiality, etc.
- The experiment must involve incentive compatible payments (avg. expected bonus payment of £1.70, on top of a flat participation fee of £1.30 per subject).
- The experiment must be designed such that it can be conducted via Prolific and such that it adheres to Prolific's terms and conditions for researchers.

Design Requirements

- The design has to be eligible to obtain (fast track) IRB approval, i.e., ...
  - no deception, preservation of participants' anonymity, explicit information (duration, repetitions, interactions, random processes), confidentiality, etc.

- The experiment must involve incentive compatible payments (avg. expected bonus payment of £1.70, on top of a flat participation fee of £1.30 per subject).
- The experiment must be designed such that it can be conducted via Prolific and such that it adheres to Prolific's terms and conditions for researchers.

Data Collection

- All data was collected in a single Prolific study, set up by the coordinators.
- Participants were directed to a common welcome screen, signed a captcha, provided informed consent, and completed a common attention check item.
- After that, participants were redirected to one of 45 × 2 = 90 treatments in batches of four (to mitigate attrition for designs using real-time interaction).

- We collected the data in ten time slots during the two weeks from January 17 to January 28, 2022, with one slot per day, from Monday to Friday in each week.
- Eventually, we reached a sample of 18,123 completed (and valid) observations.

Data Collection

- All data was collected in a single Prolific study, set up by the coordinators.
- Participants were directed to a common welcome screen, signed a captcha, provided informed consent, and completed a common attention check item.
- After that, participants were redirected to one of 45 × 2 = 90 treatments in batches of four (to mitigate attrition for designs using real-time interaction).

- We collected the data in ten time slots during the two weeks from January 17 to January 28, 2022, with one slot per day, from Monday to Friday in each week.
- Eventually, we reached a sample of 18,123 completed (and valid) observations.

Peer Evaluations

- Participating RTs were asked to assess each others' designs anonymously.
- RTs involving two members were required to submit one rating per design.

- In particular, each RT was asked to assess ten other randomly selected designs (based on the pre-registration template submitted by each RT):

> *To what extent does this design [..] provide an informative test of the research question: "Does competition affect moral behavior?"*
>
> *→ 0 (not informative at all) to 10 (extremely informative)*

- To account for RT fixed effects, in all analyses, we demean the RTs' quality ratings before estimating the mean peer evaluation score for each design.

Peer Evaluations

- Participating RTs were asked to assess each others' designs anonymously.
- RTs involving two members were required to submit one rating per design.

- In particular, each RT was asked to assess ten other randomly selected designs (based on the pre-registration template submitted by each RT):

> *To what extent does this design [..] provide an informative test of the research question: "Does competition affect moral behavior?"*
>
> → *0 (not informative at all) to 10 (extremely informative)*

- To account for RT fixed effects, in all analyses, we demean the RTs' quality ratings before estimating the mean peer evaluation score for each design.

Peer Evaluations

- Participating RTs were asked to assess each others' designs anonymously.
- RTs involving two members were required to submit one rating per design.

- In particular, each RT was asked to assess ten other randomly selected designs (based on the pre-registration template submitted by each RT):

> *To what extent does this design [..] provide an informative test of the research question: "Does competition affect moral behavior?"*
>
> → *0 (not informative at all) to 10 (extremely informative)*

- To account for RT fixed effects, in all analyses, we demean the RTs' quality ratings before estimating the mean peer evaluation score for each design.

Two Analytic Approaches ...

A. For each research design, we estimate the effect size and standard error according to the analytic specification that has been proposed by the RT. *(Requirement: ordinary least squares regression on a treatment indicator.)*

B. To remove as much of the analytical variation across RTs as possible, we employ a standardized analytic specification for all 45 research designs. *(No controls, no exclusions, individual level, robust standard errors.)*

Two Analytic Approaches ...

A. For each research design, we estimate the effect size and standard error according to the analytic specification that has been proposed by the RT.

(Requirement: ordinary least squares regression on a treatment indicator.)

B. To remove as much of the analytical variation across RTs as possible, we employ a standardized analytic specification for all 45 research designs.

(No controls, no exclusions, individual level, robust standard errors.)

Hypotheses

- **Primary hypotheses:**

  1A/1B Competition affects moral behavior.

  2A/2B Estimated effect size are heterogeneous.

- **Secondary hypotheses:**

  1A/1B Effect size estimates vary systematically with mean peer ratings.

  2A/2B Effect sizes are heterogeneous after controlling for mean ratings.

- *Pre-registered exploratory analyses and robustness tests:*
  - Analytic approach B with the exclusion criteria as used in approach A.
  - Analytic approach B with standard errors clustered on the batch variable.
  - Primary hypothesis tests for the 50% with the highest/lowest peer rating.

Hypotheses

- **Primary hypotheses:**

  1A/1B Competition affects moral behavior.

  2A/2B Estimated effect size are heterogeneous.

- **Secondary hypotheses:**

  1A/1B Effect size estimates vary systematically with mean peer ratings.

  2A/2B Effect sizes are heterogeneous after controlling for mean ratings.

- *Pre-registered exploratory analyses and robustness tests:*
  - Analytic approach B with the exclusion criteria as used in approach A.
  - Analytic approach B with standard errors clustered on the batch variable.
  - Primary hypothesis tests for the 50% with the highest/lowest peer rating.

Hypotheses

- **Primary hypotheses:**
  - 1A/1B Competition affects moral behavior.
  - 2A/2B Estimated effect size are heterogeneous.

- **Secondary hypotheses:**
  - 1A/1B Effect size estimates vary systematically with mean peer ratings.
  - 2A/2B Effect sizes are heterogeneous after controlling for mean ratings.

- *Pre-registered exploratory analyses and robustness tests:*
  - Analytic approach B with the exclusion criteria as used in approach A.
  - Analytic approach B with standard errors clustered on the batch variable.
  - Primary hypothesis tests for the 50% with the highest/lowest peer rating.

Results

Meta-Analytic Effect & Heterogeneity

- **Primary hypotheses:**

  1A/1B Competition affects moral behavior.

  2A/2B Estimated effect size are heterogeneous.

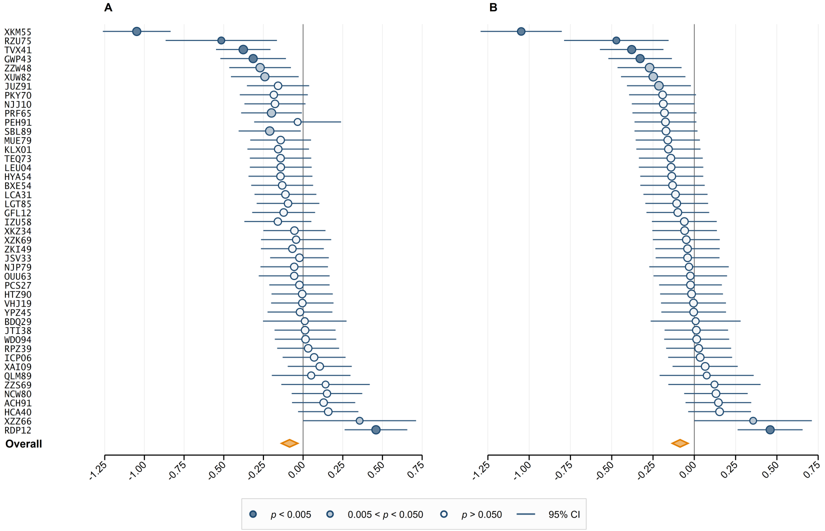
- **Primary hypothesis tests:**
  - Random effects meta-analysis (DerSimonian and Laird 1986)
  - z-test based on the overall effect size and its standard error (1A/1B).
  - Cochran's Q-test (χ^2-test); heterogeneity measures τ and I^2 (2A/2B).

Meta-Analytic Effect & Heterogeneity

- **Primary hypotheses:**
  - 1A/1B Competition affects moral behavior.
  - 2A/2B Estimated effect size are heterogeneous.

- **Primary hypothesis tests:**
  - Random effects meta-analysis (DerSimonian and Laird 1986)
  - z-test based on the overall effect size and its standard error (1A/1B).
  - Cochran's Q-test (χ^2-test); heterogeneity measures τ and I^2 (2A/2B).

Meta-Analytic Effect & Heterogeneity



Meta-Analytic Effect & Heterogeneity

	Approach A	**Approach B**
Meta-analytic effect	$d = -0.085^{*}$ ($p = 0.008$)	$d = -0.086^{**}$ ($p = 0.004$)
# $d < 0$, $p < 0.05$	8 (17.8%)	7 (15.7%)
# $d > 0$, $p < 0.05$	2 (4.4%)	2 (4.4%)
Cochran's Q	$Q(44) = 181.1^{**}$ ($p < 0.001$)	$Q(44) = 161.5^{**}$ ($p < 0.001$)
I^2	72.8%	75.7%
τ	0.185	0.169
τ/σ	1.69	1.57

Moderating Effects of Design Quality?

- **Secondary hypotheses:**

  1A/1B Effect size estimates vary systematically with mean peer ratings.

  2A/2B Effect sizes are heterogeneous after controlling for mean ratings.

- **Secondary hypothesis tests:**
  - Meta-regression on the peers' average (demeaned) quality ratings (1A/1B).
  - Q, τ, and I^2 for the residual heterogeneity, i.e., for the heterogeneity that remains after adjusting for the effect of the moderator variable (2A/2B).

Moderating Effects of Design Quality?

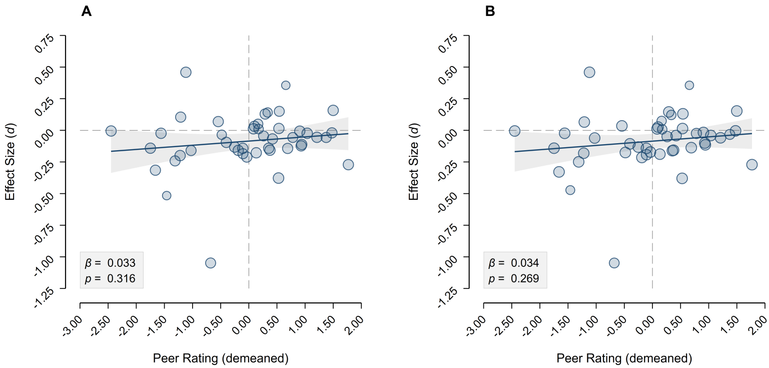
- **Secondary hypotheses:**

  1A/1B Effect size estimates vary systematically with mean peer ratings.

  2A/2B Effect sizes are heterogeneous after controlling for mean ratings.

- **Secondary hypothesis tests:**
  - Meta-regression on the peers' average (demeaned) quality ratings (1A/1B).
  - Q, τ, and I^2 for the residual heterogeneity, i.e., for the heterogeneity that remains after adjusting for the effect of the moderator variable (2A/2B).

Moderating Effects of Design Quality?



Residual heterogeneity remains significant (p < 0.001) for both analytic approaches; and the heterogeneity measures τ and I^2 are virtually unaffected by the moderator.

Summary and Conclusion

Summary and Conclusion

- We find evidence of an **adverse effect of competition on moral behavior**, yet the estimated negative effect size is quite small with a Cohen's *d* of about 0.1.
- We find strong evidence of **substantial design heterogeneity**, i.e., systematic variation in effect sizes across designs, above and beyond sampling variance.

For Illustration Purposes...

- The substantial design heterogeneity identified in our study suggests that the informativeness and generalizability of a single study protocol can be limited.

- Consider randomly implementing one of the 45 designs ...
  - The average sample standard error for our 45 designs is $\sigma = 0.108$.
  - The estimated standard deviation of the true effect size is $\tau = 0.169$.

  - Considering the uncertainty due to design choice ...

    → the standard error doubles ($\sqrt{\sigma^2 + \tau^2} = 0.200$)

    → results in a very wide 95% CI of $[-0.477, 0.308]$

For Illustration Purposes...

- The substantial design heterogeneity identified in our study suggests that the informativeness and generalizability of a single study protocol can be limited.

- Consider randomly implementing one of the 45 designs ...
    - The average sample standard error for our 45 designs is $\sigma = 0.108$.
    - The estimated standard deviation of the true effect size is $\tau = 0.169$.

    - Considering the uncertainty due to design choice ...

      → the standard error doubles ($\sqrt{\sigma^2 + \tau^2} = 0.200$)

      → results in a very wide 95% CI of $[-0.477, 0.308]$

For Illustration Purposes...

- The substantial design heterogeneity identified in our study suggests that the informativeness and generalizability of a single study protocol can be limited.

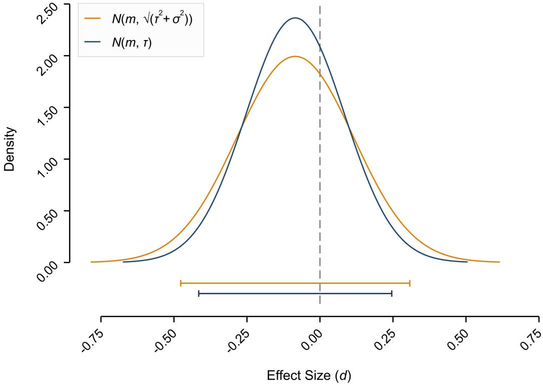
- Consider randomly implementing one of the 45 designs ...
  - The average sample standard error for our 45 designs is $\sigma = 0.108$.
  - The estimated standard deviation of the true effect size is $\tau = 0.169$.

  - Considering the uncertainty due to design choice ...

    → the standard error doubles ($\sqrt{\sigma^2 + \tau^2} = 0.200$)

    → results in a very wide 95% CI of [−0.477, 0.308]

For Illustration Purposes...



Summary and Conclusion

- To obtain more reliable scientific evidence, researchers should conduct studies based on multiple conceivable designs pooled in a meta-analysis.
- Moving towards much larger data collections and more team science could improve the informativeness and generalizability of experimental research.

Thank you!

Christoph Huber

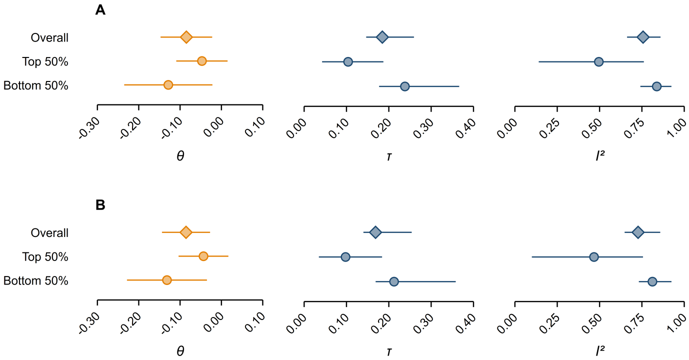
WU Vienna University of Economics and Business

✉ christoph.huber@wu.ac.at

🌐 chr-huber.com

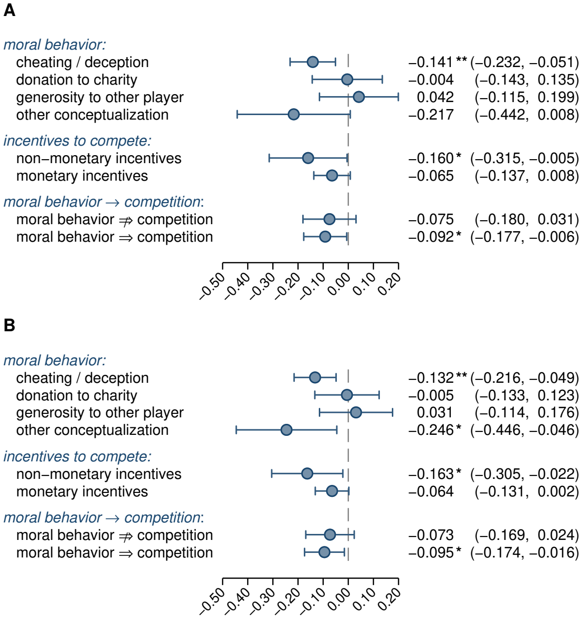
Appendix

Moderating Effects of Design Quality?



Moderating Effects of Design Quality?

Analytic Approach B	**Top 50%**	**Bottom 50%**
Meta-analytic effect	$d = -0.043$	$d = -0.132^{*}$
	($p = 0.159$)	($p = 0.008$)
# $d < 0$, $p < 0.05$	2 (9.1%)	5 (21.7%)
# $d > 0$, $p < 0.05$	1 (4.5%)	1 (4.5%)
Cochran's Q	$Q(44) = 39.4^{*}$	$Q(44) = 117.0^{**}$
	($p = 0.009$)	($p < 0.001$)
I^2	46.7%	81.2%
τ	0.098	0.212
τ/σ	0.89	2.01


A
moral behavior:
cheating / deception
−0.141 ** (−0.232, −0.051)
donation to charity
−0.004 (−0.143, 0.135)
generosity to other player
0.042 (−0.115, 0.199)
other conceptualization
−0.217 (−0.442, 0.008)
incentives to compete:
non−monetary incentives
−0.160 * (−0.315, −0.005)
monetary incentives
−0.065 (−0.137, 0.008)
moral behavior → competition:
moral behavior ⇏ competition
−0.075 (−0.180, 0.031)
moral behavior ⇒ competition
−0.092 * (−0.177, −0.006)
−0.50
−0.40
−0.30
−0.20
−0.10
0.00
0.10
0.20
B
moral behavior:
cheating / deception
−0.132 ** (−0.216, −0.049)
donation to charity
−0.005 (−0.133, 0.123)
generosity to other player
0.031 (−0.114, 0.176)
other conceptualization
−0.246 * (−0.446, −0.046)
incentives to compete:
non−monetary incentives
−0.163 * (−0.305, −0.022)
monetary incentives
−0.064 (−0.131, 0.002)
moral behavior → competition:
moral behavior ⇏ competition
−0.073 (−0.169, 0.024)
moral behavior ⇒ competition
−0.095 * (−0.174, −0.016)
−0.50
−0.40
−0.30
−0.20
−0.10
0.00
0.10
0.20