Problem: Current research paradigms use narrow sets of stimuli that do not fully probe the conceptual robustness of the effect.

➢ How much variation is there in the study designs and in the estimated results when you let several teams conceptually replicate the same hypothesis?
➢ Can experts predict which claims are more likely to replicate? Within the same claim, can experts predict which research designs are more likely to lead to a successful replication?

Goals of our work:
➢ Make transparent how design choices affect research results
➢ Study by means of a forecasting survey whether the effects of researcher design choices on estimated effect sizes are predictable

Research components
➢ 5 original hypotheses in psychology regarding implicit cognition, negotiation and moral judgments
➢ Conceptual replications - 15 research teams independently designed studies to test the original hypotheses (main study)
➢ Direct replications:
  • Of each original design
  • Of the main study (N > 15,000)
➢ Forecasting study - For each set of materials, independent researchers predicted:
  • Likelihood that a significant result would be found
  • Standardized effect size of the replication

Results of direct and conceptual replications

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Cons. p &lt; .05</th>
<th>Cons. p &gt; .05</th>
<th>Incons. p &lt; .05</th>
<th>Incons. p &gt; .05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Awareness of prejudice</td>
<td>54% (7)</td>
<td>0% (0)</td>
<td>8% (1)</td>
<td>38% (5)</td>
</tr>
<tr>
<td>Negotiation and trust</td>
<td>92% (12)</td>
<td>8% (1)</td>
<td>0% (0)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>Praise for needless work</td>
<td>77% (10)</td>
<td>8% (1)</td>
<td>8% (1)</td>
<td>8% (1)</td>
</tr>
<tr>
<td>Proximal authorities drive legitimacy</td>
<td>42% (5)</td>
<td>25% (3)</td>
<td>25% (3)</td>
<td>8% (1)</td>
</tr>
<tr>
<td>Morality and happiness</td>
<td>23% (3)</td>
<td>62% (8)</td>
<td>15% (2)</td>
<td>0% (0)</td>
</tr>
</tbody>
</table>

Direct replications:
➢ All 5 original designs replicated: the hypotheses were supported when the original materials were used, both in the Main study and in the Replication study

Conceptual replications:
➢ Fairly consistent support for 2 hypotheses out of 5
➢ Substantial variability in results across and within hypotheses

Forecasting study: significance and effect size

Fig 1: Histograms of average predictions regarding whether the result is statistically significant (p < .05) or not at study-design level.

Fig 2: Correlation between average predicted effect size and observed effect size, separately for each of the five hypotheses.