Online-Appendix

„Implicit Measurement of the Moral Self-Image Using the Go/No-Go Association Task (GNAT) - An Empirical Investigation of the Convergent Validity Between Explicit and Implicit Measures“

Louisa Felicitas Bläßer
Technical University of Munich

Junior Management Science 10(4) (2025) 966-984

V Appendix

V.I Pre-Study GOOD & BAD

A small-scale pre-study was conducted to identify, select, and validate stimuli words for the extended word list used in group B.

V.I.I Methodology

This pre-study was set up via a Google Forms questionnaire that was sent directly to participants. The sample consisted of 15 participants, including two native English speakers (13.33%) and thirteen people speaking German as their primary language (86.67%). Six people indicated male gender (40%), and eleven were female (60%). The average age was 28.33 years.

The participants were asked to rate English words on a nine-point Likert scale (from 1: very weak to 9: very strong) to assess how well they think the word matches the GOOD/MORAL or BAD/IMMORAL attribute, respectively. Additionally, they were instructed not to tick a box for words they find hard to understand (to identify difficult words for non-native speakers). The final score for each word was calculated as the average rating given by all participants for this word.

V.I.II Results: GOOD

The results for the GOOD stimuli words, including each word's average score and the number of missing responses (people who did not assign a number to this word), are shown in Table 6. The words chosen for the GNAT experiment are printed in bold.

GOOD	Reference	Average score	Missing responses
caring	**(Aquino & Reed, 2002)**	**7.80**	
compassionate	**(Aquino & Reed, 2002)**	**7.80**	
fair	**(Aquino & Reed, 2002; Sachdeva et al., 2009)**	**7.33**	
friendly	**(Aquino & Reed, 2002; Nosek & Banaji, 2001)**	**7.40**	
generous	**(Aquino & Reed, 2002; Sachdeva et al., 2009)**	**7.33**	
hardworking	**(Aquino & Reed, 2002)**	**7**	
helpful	**(Aquino & Reed, 2002)**	**7.53**	

honest	**(Aquino & Reed, 2002; Ferguson, 2018; Johnston et al., 2013)**	**7.40**	
kind	**(Aquino & Reed, 2002)**	**7.20**	
merciful		6.67	
altruist	**(Ferguson, 2018; Johnston et al., 2013; Perugini & Leone, 2009)**	**7.27**	
approachable		5.67	
faithful	**(Ferguson, 2018; Johnston et al., 2013)**	**7.07**	
sincere	**(Ferguson, 2018; Johnston et al., 2013; Perugini & Leone, 2009)**	**7.47**	
modest	**(Ferguson, 2018; Johnston et al., 2013; Perugini & Leone, 2009)**	**7.20**	
genuine		**7.00**	
transparent		4.40	
good	(Nosek & Banaji, 2001)	7.87	
saint		6.60	
truthful		6.53	
gentle		6.53	
joyful	**(Nosek & Banaji, 2001)**	**7.40**	
calm		5.20	
patient		**7.07**	
integrity		7.62	
respectful		**7.60**	
responsible		6.79	1
courage		6.47	
grateful		**7.00**	
loyal		**7.47**	
forgiving		**7.00**	
self-control		6.60	
perseverance		5.55	4
prudence		5.27	4
consideration		6.13	

Table 6: Stimuli words for GOOD in Pre-Study (chosen words are printed in bold)

V.I.III Results: BAD

Table 7 depicts the results for the BAD stimuli words, including each word's average score and the number of missing responses (people who did not assign a number to this word). All selected words for the GNAT experiment are printed in bold.

BAD	**Reference**	**Average score**	**Missing responses**
indifferent		4.27	
callous		6.33	5
unfair		**7.40**	
hostile		**7.80**	
stingy		7.00	5
lazy		**7.13**	
unhelpful		**7.13**	
dishonest	**(Ferguson, 2018; Johnston et al., 2013)**	**7.33**	
cruel		8.29	1
ruthless	**(Aquino & Reed, 2002)**	**7.93**	
selfish	**(Aquino & Reed, 2002; Sachdeva et al., 2009)**	**7.6**	
distant	(Aquino & Reed, 2002)	4.13	
unfaithful		6.53	
insincere		6.33	
arrogant	**(Ferguson, 2018; Johnston et al., 2013; Perugini & Leone, 2009)**	**7.07**	
cheater	**(Ferguson, 2018; Johnston et al., 2013; Perugini & Leone, 2009)**	**7.67**	
pretentious	**(Ferguson, 2018; Johnston et al., 2013)**	**7.07**	
deceptive	**(Ferguson, 2018; Johnston et al., 2013; Perugini & Leone, 2009)**	**7.73**	
bad	**(Ferguson, 2018; Nosek & Banaji, 2001)**	**7.87**	
evil	**(Nosek & Banaji, 2001)**	**7.93**	
brutal	**(Nosek & Banaji, 2001)**	**7.93**	

hateful		**7.07**	
angry	**(Nosek & Banaji, 2001)**	**7.07**	
impatient		**7.00**	
disrespectful		**7.07**	
irresponsible		6.13	
cowardice		5.29	1
ingratitude		5.8	
disloyal	**(Sachdeva et al., 2009)**	**7.33**	
vengefulness		7.73	4
impulsiveness		4.8	
quit		4.58	3
recklessness		6.93	
inconsideration		5.71	1
corrupt		8.27	
egocentric		**7.47**	

Table 7: Stimuli words for BAD in Pre-Study (chosen words are printed in bold)

V.I.IV Discussion: GOOD & BAD

The final sets of positive and negative attribute stimuli words for variation B (extended word list) were chosen based on the following two criteria:

First, no responses were missing. This implied that every participant understood the word correctly.

Second, the average score resides within a range of 7 and 7.8 for the positive word list and within a range of 7 and 7.93 for the negative word list, which assures that the evaluative intensity of the chosen words is similar to each other, as suggested by Nosek and Banaji (2001). Further, a score of 7 or higher provides additional evidence that the words selected are mutually exclusive and clearly belong to one category (GOOD/MORAL or BAD/IMMORAL) with sufficient intensity.

V.II Explicit Moral-Self-Image Questionnaire

Please respond to the following statements as they apply to you.

Compared to the caring person I want to be, I am:								
1	2	3	4	5	6	7	8	9
Much less caring than the person I want to be				Exactly as caring as the person I want to be				Much more caring than the person I want to be
○	○	○	○	○	○	○	○	○

Compared to the compassionate person I want to be, I am:								
1	2	3	4	5	6	7	8	9
Much less compassionate than the person I want to be				Exactly as compassionate as the person I want to be				Much more compassionate than the person I want to be
○	○	○	○	○	○	○	○	○

Compared to the fair person I want to be, I am:								
1	2	3	4	5	6	7	8	9
Much less fair than the person I want to be				Exactly as fair as the person I want to be				Much more fair than the person I want to be
○	○	○	○	○	○	○	○	○

Compared to the friendly person I want to be, I am:								
1	2	3	4	5	6	7	8	9
Much less friendly than the person I want to be				Exactly as friendly as the person I want to be				Much more friendly than the person I want to be
○	○	○	○	○	○	○	○	○

Compared to the generous person I want to be, I am:								
1	2	3	4	5	6	7	8	9
Much less generous than the person I want to be				Exactly as generous as the person I want to be				Much more generous than the person I want to be
○	○	○	○	○	○	○	○	○

Compared to the hard-working person I want to be, I am:								
1	2	3	4	5	6	7	8	9
Much less hard-working than the person I want to be				Exactly as hard-working as the person I want to be				Much more hard-working than the person I want to be
○	○	○	○	○	○	○	○	○

Compared to the helpful person I want to be, I am:								
1	2	3	4	5	6	7	8	9
Much less helpful than the person I want to be				Exactly as helpful as the person I want to be				Much more helpful than the person I want to be
○	○	○	○	○	○	○	○	○

Compared to the honest person I want to be, I am:								
1	2	3	4	5	6	7	8	9
Much less honest than the person I want to be				Exactly as honest as the person I want to be				Much more honest than the person I want to be
○	○	○	○	○	○	○	○	○

Compared to the kind person I want to be, I am:								
1	2	3	4	5	6	7	8	9
Much less kind than the person I want to be				Exactly as kind as the person I want to be				Much more kind than the person I want to be
○	○	○	○	○	○	○	○	○

Next

Table 8: Explicit Moral Self-Image Questionnaire, as used in the experiment

V.III Formula of the Pearson Correlation Coefficient

The Pearson correlation coefficient is determined by the following formula, as cited in Fahrmeir et al. (2016):

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

with x_i being a value in the x dataset and $\overline{x}$ representing the mean (analogously for y)

It investigates the direction and strength of a linear relationship between two variables, x and y, in a dataset. A positive linear connection between the two variables is suspected when the values are distributed that with an increasing x, y is also increasing (Asuero et al., 2006; Fahrmeir et al., 2016).

V.IV Implicit Moral Self-Image Score Based on Reaction Times

V.IV.I Methodology

Statistical Analysis of Sensitivity (d') based on Reaction Times

Analogously to the signal detection theory, as cited in Macmillan (2002), the formula used to determine the sensitivity based on Hit (H) rate and False Alarm (FA) rate was adapted to define a d' based on reaction times (RT):

$$d'(reaction\ times\ (RT)) = Z\left(\frac{Average\ RT\ (Hits)}{Response\ deadline}\right) - Z\left(\frac{Average\ RT\ (False\ Alarms)}{Response\ deadline}\right)$$

It is important to note that a negative d' in this context indicates a higher sensitivity and ability to discriminate words. This is because it is generally assumed that people react faster if it is easier to discriminate signals from noise, indicating that they are more sensitive (Greenwald et al., 1998; Nosek & Banaji, 2001). Therefore, faster responses for Hits (lower ratio) led to a lower Z-score. Simultaneously, slower responses for False Alarms (higher ratio) resulted in a higher Z-score, further decreasing d' after the subtraction.

To highlight this relationship with a fictional example:

Given that the average reaction time for Hits was 480ms, the average reaction time for False Alarms was 600ms, and the response deadline was 700ms for all trials. Inserting into the formula, this yielded: $d'(RT)(ME + BAD) = Z\left(\frac{480ms}{700ms}\right) - Z\left(\frac{600ms}{700ms}\right) = -0.5838.$

When the average reaction time changed to 430ms for Hits and stayed constant at 600ms for False Alarms, this yielded: $d'(RT)(ME + BAD) = Z\left(\frac{430ms}{700ms}\right) - Z\left(\frac{600ms}{700ms}\right) = -0.7771.$

With the average reaction time for Hits being held constant at 480ms and for False Alarms changed to 650ms: $d'\,(RT)(ME + BAD) = Z\left(\frac{480ms}{700ms}\right) - Z\left(\frac{650ms}{700ms}\right) = -0.9815.$

Both adjustments highlight the negative effect on d'(RT) if either the average reaction time for Hits decreased or the average reaction time for False Alarms increased. Therefore, the final implicit moral self-image scores based on reaction times were multiplied by (-1) to interpret them analogously to the original implicit moral self-image scores based on response rates. A higher moral self-image score indicates a higher moral self-image due to faster associations of ME + GOOD.

$$Implicit\ moral\ self-image\ score(RT) = (d'(ME + GOOD) - d'(ME + BAD)) * (-1)$$

Hypotheses for Reaction Times

Hypothesis 3: The participants' implicit moral self-image scores, based on reaction times, will **correlate positively (r > 0)** with the explicit moral self-image scores (questionnaire-based) in both groups.

Hypothesis 4: The **correlation** between the participants' implicit moral self-image scores, based on reaction times, and explicit moral self-image scores will be **higher in group B than in group A.**

V.IV.II Results

The identical algorithm for the analysis of response rates was applied to the same (standard) sample as described in 4.2 Participants per Group. Thus, the identical strict exclusion criteria as in the previous analysis hold (see 4.1 Exclusion Criteria), and the sample was not corrected for perfect responses (H = 100% and FA = 0%). This is because calculating a Z-score of no False Alarms (average response time = 0ms) is mathematically unfeasible, and approximating reaction times would distort the results, making them harder to compare.

Firstly, d'(RT) for each participant and block was calculated with this formula: $d'(reaction\ times\ (RT)) = Z\left(\frac{Average\ RT\ (Hits)}{Response\ deadline}\right) - Z\left(\frac{Average\ RT\ (False\ Alarms)}{Response\ deadline}\right)$). It is important to emphasize that a negative d'(RT) demonstrates a higher sensitivity, as faster reaction times for Hits (or slower for False Alarms) indicate a stronger association. Secondly, the participant's d'(RT)(ME + BAD) was subtracted from d'(RT)(ME + GOOD) to determine each participant's implicit moral self-image score. Thirdly, the scores were corrected by (-1) and analyzed.

Table 8 shows all characteristics of the implicit moral self-image scores (based on reaction times) and the Pearson correlation for groups A and B with the strict exclusion criteria. As explained before and defined in the formula, the outcomes were already multiplied by (-1).

	Group A	Group B
Max	0.4413	1.0472
Min	-0.4057	-0.4623
Mean	0.0530	0.0211
Median	0.0517	-0.0224
Pearson correlation coefficient	0.2883	0.1974

Table 9: Characteristics of implicit moral self-image scores, based on reaction times and Pearson correlation coefficients for group A and group B with strict exclusion criteria

Group A

The implicit moral self-image scores based on reaction times for group A encompassed a range from -0.4057 to 0.4413, with a median of 0.0517 and a mean of 0.0530. The implicit moral self-image score was negative for 11 participants (36.67%).

The Pearson correlation coefficient was r = 0.2883, indicating a weak positive linear relationship between the explicit and implicit moral self-image scores. Basically, this positive correlation demonstrates that people who were faster in the ME + GOOD pairing indicated a higher moral self-image in the explicit scale. Therefore, the results in group A, from the newly introduced implicit moral self-image score based on reaction times, support hypothesis 3 that explicit and implicit moral self-image scores correlate positively, strengthening the quality of the GNAT. Figure 3 shows group A's distribution of values and correlation (dotted line).

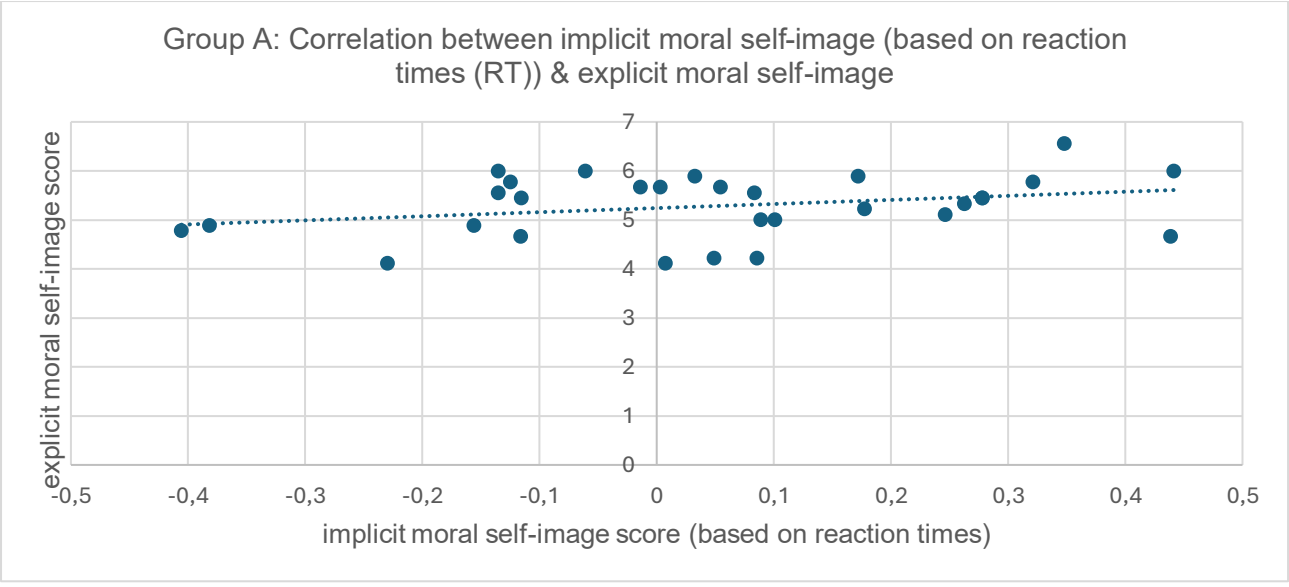

Figure 3: Correlation (dotted line) between the implicit moral self-image scores, based on reaction times, and the explicit moral self-image scores for group A

Group B

For group B, the mean was 0.0211, the median was -0.0224, and the implicit moral self-image scores ranged from -0.4623 to 1.0472. The implicit moral self-image score was negative for 21 participants (55.26%), and the Pearson correlation coefficient was r = 0.1974, which is comparable to group A and also suggests a weak positive linear relationship, confirming hypothesis 3. Figure 4 shows group B's distribution of values and correlation (dotted line).

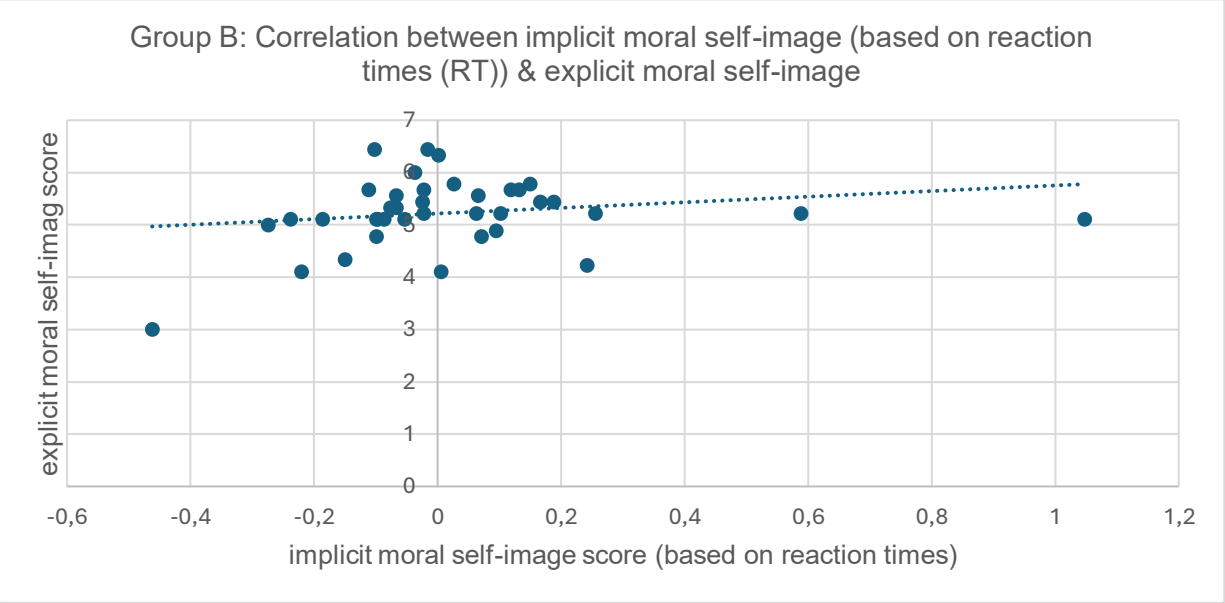

Figure 4: Correlation (dotted line) between the implicit moral self-image scores, based on reaction times, and the explicit moral self-image scores for group B

Comparison between Analysis of Response Rates and Reaction Times

Pearson Correlation coefficient	**Group A**	**Group B**
Pearson correlation coefficient based on response rates	-0.0078	0.4870
Pearson correlation coefficient based on reaction times	0.2883	0.1974

Table 10: Pearson correlation coefficients, based on response rates and reaction times, for group A and group B

Table 9 depicts the different Pearson correlation coefficients. It visualizes the differences between the groups and the analysis methods. The first observation is that the correlation coefficients, based on reaction times, are comparable within the two groups and similar to previous literature (Perugini, 2005). The second observation is that in the reaction-based

analysis, group A had a stronger correlation (between explicit and implicit measures) than group B, which contradicts hypothesis 4.

The plots, Figure 3 and Figure 4 visualize the distribution of values (based on reaction times) for groups A and B, respectively, and show a different pattern. As group A is more evenly distributed (within a range of -0.4 and 0.5), group B clusters around the implicit moral self-image score of 0, and more outliers in group B (e.g., 1.0472) can be observed. This highlights the deviations and differences between the groups.

V.IV.III Discussion

Examining the results of this alternative analysis method, introduced as an implicit moral self-image score based on reaction times, revealed contrasting findings. Primarily, it becomes evident that the correlation coefficients (corrected through multiplication by (-1)) of both groups were positive (with r = 0.2883 for group A and r = 0.1974 for group B), supporting hypothesis 3. This implies that participants who performed better (by reacting faster for Hits) in the ME + GOOD block of the GNAT indicated a rather positive moral self-image in the explicit questionnaire and vice versa.

These results could emphasize the effectiveness of the GNAT as an implicit measure when focusing on reaction times, especially because it was within the typical range of implicit measures (between 0.2 and 0.3) (Perugini, 2005). In contrast, this weak positive correlation of both groups implies a weak convergent validity (Carlson & Herdman, 2012). Interpreting these results accordingly could be construed as evidence for the double dissociation strategy (Asendorpf et al., 2002), especially as reaction times could be considered more unconscious and implicit.

In hypothesis 4, the correlation between implicit and explicit measures was expected to be stronger for group B compared to group A. However, the results contradict this hypothesis, as group A correlated stronger between both measures than group B. The same explanations as in the main analysis (based on response rates) could not be applied to this analysis. Specifically, the response-rates-based analysis argued that learning effects caused a lower correlation in this group. The reaction-time-based finding contradicts the results of the previous analysis. Here, group A, with a fourfold repetition of the stimuli words and potentially stronger learning effects (see 5.1 Key Findings), correlated more significantly than group B, where the stimuli words were not repeated. This difference could be explained by the outliers in group B, distorting the coefficients and decreasing the correlation. Group A had a more evenly distributed pattern (as seen in Figure 3), yielding a higher correlation than Group B and the previous analysis. This indicates differences between the distribution of values and raises concerns about the quality of the results.

The results obtained through this analysis should be considered carefully for the following reasons: First, the formula could be incorrect, as it was developed and primarily used in this thesis. The research body should validate and peer-review this formula before bringing it into further use. Second, the correlation coefficient might not be an appropriate measure to examine the effectiveness of the GNAT in applying the moral self-image. It can easily be biased, and its significance can be interpreted differently; hence, it is rather ambiguous (Perugini, 2005).

Limitations of this Results & Suggestions for Future Research

The same limitations, as elaborated in 5.2 Limitations of Results & Suggestions for Future Research, apply to the experiment. As this was the first approach, extending the formula derived from signal detection theory on reaction times, additional limitations of this analysis should be considered. Further tests of the reliability and validity of this formula are needed and highly recommended. Alternatively, it is strongly suggested that this reaction time analysis should be repeated with the often-used algorithm Greenwald et al. recommended (2003) and a focus on the coefficients by Cohen (2013), primarily based on means and standard deviations of reaction times.

Additionally, the calculation problem of no False Alarms could not be corrected by applying the approach of Kadlec (1999), and there were no approximations for the reaction times that should be substituted for zero False Alarms. The approximation of reaction times could be subjective and influenced by the researcher, which leads to a high arbitrary and later high variability bias in the data set. Future research could consider the single-task IAT (see Friese et al. (2007)) as an alternative experimental design, as this experiment forces a decision of either one key (assigned with the target category and an evaluative attribute) or the second key (assigned only with the opposing evaluative attribute). It thus avoids perfect responses and can be analyzed using the proposed algorithm of Greenwald et al. (2003).