Test-retest reliability and construct validity of the Pro-Environmental Behavior Task

Florian Lange & Siegfried Dewitte

Behavioral Engineering Research Group, KU Leuven, Leuven, Belgium

Correspondence:

Florian Lange

BEE – Behavioral Engineering Research Group

KU Leuven

Naamsestraat 69

3000 Leuven

Belgium

florian.lange@kuleuven.be

This work has been accepted for publication in the *Journal of Environmental Psychology* on 4 January 2021. The final version of the article can be found at <https://doi.org/10.1016/j.jenvp.2021.101550>.

Acknowledgements

This research was supported by the FWO and European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 665501 and by a FWO postdoctoral fellowship (No 12U1221N, awarded to F.L.).

Declarations of interest: none.

Abstract

The study of pro-environmental behavior has been dominated by self-report scales and field observations. Although both approaches have their merits, they also suffer from critical limitations that obstruct progress towards a better understanding of pro-environmental behavior. The Pro-Environmental Behavior Task (PEBT) was developed to address these limitations by facilitating the study of actual pro-environmental behavior under controlled laboratory conditions. In the present research, we examined whether the PEBT can also provide a valid and reliable indicator of people's propensity to show pro-environmental behavior. In support of the task's test-retest reliability, we found PEBT choice behavior to be strongly correlated across a period of approximately one month ($r = .81$). In addition, PEBT performance was significantly correlated to both objective observations and subjective reports of other pro-environmental behaviors. Our results further suggest participants' PEBT choices to be most closely related to individual differences in transportation behavior. These findings indicate that the PEBT might prove useful not only as a procedure for the experimental study of pro-environmental behavior, but also for the performance-based assessment of a person's propensity to engage in pro-environmental behavior.

Keywords: pro-environmental behavior; conservation (ecological behavior); measurement; validity; reliability; laboratory assessment

1. Introduction

Many research questions in environmental psychology require the recording of quantitative information about pro-environmental behaviors. For example, researchers might want to examine whether the frequency or duration of a pro-environmental behavior varies as a function of experimental manipulations. Information about such variables is typically obtained by either observing the behavior (or its products) in the field or by asking people to retrospectively report on their behavior (Lange & Dewitte, 2019a). Given the cost and limited experimental control of field studies and given the many reasons to question the accuracy of retrospective self-reports, Lange, Steinke, and Dewitte (2018) recently proposed a third way to study pro-environmental behavior. They developed the Pro-Environmental Behavior Task (PEBT), a laboratory procedure involving consequential choices between an objectively pro-environmental and an objectively environmentally harmful response option. The PEBT provides the user with valid and reliable recordings of a pro-environmental behavior. Recordings are valid because PEBT choices have actual consequences for the environment and thus meet the critical requirement for being a pro-environmental behavior according to impact-oriented definitions (Stern, 2000; Steg & Vlek, 2009). They are also reliable because choices are computer-recorded, which arguably renders recordings more precise and less biased than human observations.

However, there is a second way to interpret quantitative recordings of pro-environmental behaviors (Yoder, Lloyd, & Symons, 2018). Researchers recording properties of pro-environmental behaviors might not be interested in the behavior *per se*, but rather in a psychological characteristic of the person who shows this behavior. Recorded pro-environmental behavior would be considered as an indicator of environmental attitude (Kaiser, Merten, & Wenzel, 2018) or of the general propensity to engage in pro-environmental behavior (Otto, Evans, Moon, & Kaiser, 2019). Critically, using behavior

recordings for the measurement of person characteristics raises additional psychometric questions (Yoder et al., 2018). Empirical support for the construct validity, internal consistency, and retest-reliability of PEBT choice behavior is needed before it can be considered for the measurement of a person's propensity to show pro-environmental behavior. With the present set of studies, we aim to probe the psychometric properties of the PEBT in order to examine its potential role in assessing individual differences in that propensity.

Initial validation studies (Lange et al., 2018) have already shown the PEBT to have excellent split-half reliability ($r_{SB} = .98$). While this finding indicates that people make internally consistent choices on the task, little is still known about the longitudinal stability of PEBT choice behavior. In a pilot study ($N = 40$), we obtained a first estimate of the PEBT's test-retest reliability of $r = .38$ across a one-year period (see Supplementary Materials). Given the small sample size, this estimate is unlikely to be very reliable. In addition, even if PEBT choices lack one-year stability, they might be reasonably stable across shorter periods of time. The estimation might have also been affected by the fact that the one-year period covered the first prominent actions of two global climate movements (i.e., Fridays for Future and Extinction Rebellion). Therefore, we aimed to estimate the test-retest reliability of the PEBT with increased precision across a one-month period in Study 1.

In addition, Study 1 aimed to provide new insights into the task's construct validity by relating PEBT choice behavior to other pro-environmental behaviors. Previous studies have found participants' performance on the task to be moderately correlated to aggregates of their self-reported frequency to engage in other pro-environmental behaviors ($r \approx .30$, Lange et al., 2018; Lange & Dewitte, 2019b). This finding raises two important questions. First, given the limitations of subjective frequency reports (Gifford, 2014; Kormos & Gifford, 2014; Lange & Dewitte, 2019a), it remains unclear whether individual differences in PEBT performance are also associated with objectively observed pro-environmental behaviors. In Study 1, we

address this question by relating PEBT choice behavior to participants' performance on other laboratory procedures eliciting pro-environmental behavior. To our knowledge, this is the first analysis of the relationship between different pro-environmental behaviors in the laboratory. On the one hand, we examined shared inter-individual variance between PEBT choice behavior and participants' responses to a pro-environmental donation opportunity. Assessing the amount of money people give to environmental organizations has been a popular method to observe actual pro-environmental behavior in the laboratory (Vesely & Klöckner, 2018), the field (Alpizar, Carlsson, & Johansson-Stenman, 2008), and online studies (Carrico, Raimi, Barnes Truelove, & Eby, 2018). For comparison, we also related PEBT performance to behavior in an interpersonal donation game to analyze whether a large proportion of pro-environmental PEBT choices is associated with non-selfish behavior in general (rather than to the specific type of non-selfish behavior that is pro-environmental behavior, Klein, Hilbig, & Heck, 2017). On the other hand, we related PEBT choice behavior to performance on a newly developed laboratory procedure giving participants the opportunity to complete a dull number-identification task to generate funds for environmental protection. Finding PEBT performance to be correlated to objective recordings of other pro-environmental behaviors would further support the task's construct validity for the measurement of a somewhat generalized propensity to show pro-environmental behavior.

Second, while PEBT choice behavior may be indicative of such a general propensity, it is also possible that individual differences in PEBT performance relate more specifically to individual differences in particular pro-environmental behaviors. Previous examinations of the task's construct validity (Lange et al., 2018; Lange & Dewitte, 2019b) have exclusively relied on composite scores, aggregating participants' self-reported frequency to engage in multiple different pro-environmental behaviors (e.g., newspaper recycling, volunteering for an environmental group). While these studies were powerful enough to examine the relationship

between choice behavior on the PEBT and this composite score, they did not allow for an adequately powered analysis of the task's correlation with specific pro-environmental behaviors. In Study 2, we address this issue by conducting a meta-analysis on the relationship between PEBT performance and frequency reports of specific pro-environmental behaviors. Combining information from four studies using very similar materials, this meta-analysis provided largely unbiased and reasonably precise correlation coefficients. It allowed examining which pro-environmental behaviors relate to PEBT choice behavior and whether the strength of this relationship differs between behaviors. Through these analyses, we aimed to gain more specific insights into the construct validity of the PEBT.

2. Test-retest reliability and relationship to observed behaviors (Study 1)

Study 1 aimed to examine the one-month test-retest reliability of PEBT choice behavior as well as its relationship to other types of objectively recorded pro-environmental behaviors. To this end, we asked participants who had completed the PEBT in the context of another study (Lange, Brick, & Dewitte, 2020) to return to the laboratory for a retest session. In this context, we also provided participants with additional opportunities to show pro-environmental behavior in the laboratory, which allowed us to test whether people who behave pro-environmentally on the PEBT also do so on other occasions.

2.1 Methods

We confirm that we report all conditions, data exclusions, and how we determined our samples sizes. In this manuscript, we report only those measures that are relevant for the psychometric evaluation of the PEBT, but the dataset including all measures that we assessed is available at <https://osf.io/apfdv/>.

2.1.1 First session

First-session PEBT data for our test-retest reliability study was collected in the fall of 2019, in the context of laboratory experiment on the effect of observability on pro-

environmental behavior (Lange et al., 2020). The experiment was approved by the local ethics committee (G-2018 11 1394). For this study, participants completed the PEBT either in private or in the presence of another participant who completed a different procedure. As we did not find evidence for between-condition differences in any of the variables analyzed here (Lange et al., 2020), we decided to pool the data from both conditions for the present analyses (total $N = 204$). This sample size was determined based on power considerations related to the observability experiment. Sensitivity analyses (G*Power 3.1.9.2, Faul, Erdfelder, Lang, & Buchner, 2007) revealed that, given this sample size and $\alpha = .05$, correlation analyses had 80% power to detect associations of $r = .19$ and 95% power to detect associations of $r = .25$. As previous studies found correlations between PEBT choice behavior and other pro-environmental behaviors of around $r = .30$ (Lange et al., 2018), the first-session sample was likely large enough for powerful analyses of relationships among the behaviors described below.

Data from two participants were excluded from all analyses. One participant was observed with a phone in the testing room (against explicit instructions) and the other participant completed only half of the task due to a software crash. Of the final first-session sample, 124 participants were female (77 male, one preferred not to say). Participants' age ranged from 17 to 47 years ($M = 23.61$ yrs, $SD = 4.83$ yrs). Eighty-eight percent of the participants were students and 41% of them named Dutch as their first language. The majority indicated having a bicycle (82%) and a driver's license (63%), while 22% reported having a car. Participants received six euros for their participation (total study duration ≈ 30 min).

The PEBT version used in this study comprised 40 trials. Each trial involved a consequential choice between two response options, labelled "car" and "bicycle". Following their choice, participants had to endure a waiting period before they were presented with the next choice. The waiting time associated with the two options varied across trials. The waiting

time associated with choosing the car option was either 5, 10, 15, or 20 seconds and it was always shorter than the waiting time associated with choosing the bicycle option. The waiting time difference between the options (i.e., the waiting time cost added by choosing the bicycle option) varied randomly at five levels (10s, 15s, 20s, 25s, 30s). These levels were chosen based on prior studies (Lange et al., 2018; Lange & Dewitte, 2019b) finding that these waiting times produce largely non-skewed distributions in the overall proportion of bicycle choices.

In addition, every time the car was chosen, an array of USB-powered lights located on the desk of the participants was illuminated for the duration of the waiting period. In half of the trials four lights were illuminated, and in the other half twelve lights were illuminated. That is, each participant made 20 choices with relatively low environmental impact (i.e., powering four lights produced approx. 3000mg/h of CO₂ emissions) and 20 choices with relatively high environmental impact (i.e., powering twelve lights produced approx. 9000mg/h of CO₂ emissions). The order of the four-light block and the twelve-light block was counterbalanced across participants. Participants were explicitly informed about these consequences, that is, about the waiting times associated with the response options, the waiting time difference, the number of lights to be turned on by choosing the car option, and the associated CO₂ emissions. The proportion of trials on which participants chose the pro-environmental bicycle option served as the PEBT outcome measure.

After the final PEBT trial, participants proceeded to completing a short questionnaire on the computer. The questionnaire involved a second opportunity to behave pro-environmentally. Participants could donate up to two euros of their participation fee to an environmental organization that offsets carbon emissions (<https://carbonaltdelate.eu/>). They were told (truthfully) that we would multiply all donations by 1.5. To select an amount, participants first had to click on a scale before a slider appeared that could be positioned at any point of the continuous scale between zero and two euros. Along similar lines,

participants were given an opportunity to behave prosocially. They could transfer up to two euros of their fee to the participant sitting next to them (using the same slider procedure as described above for pro-environmental donations). They were told (truthfully) that we would double all transfers and that their neighbor also had the opportunity to transfer up to two euros to them. This cooperation game was presented on a separate page directly after the pro-environmental donation opportunity.

2.1.2 Second session

Two weeks after the end of data collection for the first session, all 204 participants received an email inviting them to take part in a follow-up study. The sample size for this second session was determined based on practical considerations: we offered testing slots in one particular week and included all participants that could be scheduled in that week. Ninety-four participants with valid first-session data completed the second session. The two participants whose first-session data were excluded completed the second session as well, but their data were not used in any analysis. This sample size still yielded a statistical power of 84% for detecting correlations of $r = .30$ and it allowed obtaining reasonably stable estimates of test-retest reliability (provided that the correlation between assessments was higher than $r = .70$, Schönbrodt & Perugini, 2013).

A comparison between the participants who participated vs. those who did not participate in the second session can be found in Table 1. This comparison suggests that attrition was not entirely unsystematic: participants who returned for a second session made more pro-environmental PEBT choices during the first session than participants who did not return. In addition, sample variance in PEBT choice behavior was smaller in returning participants, which might have led to underestimations of test-retest reliability and construct validity (Lakes, 2013; Polit, 2014). The mean interval between the two testing sessions was 36.71 days ($SD = 6.85$ days, range = 21 – 46 days).

The second session was identical to the first one, with the exception that all participants completed the procedure in private and that we did not administer the cooperation game. Instead, we added a self-report measure of environmental attitudes and a third opportunity to show pro-environmental behavior to generate further information regarding the construct validity of the PEBT. To assess environmental attitudes, we administered the brief version of the Environmental Attitudes Inventory (EAI-24, Milfont & Duckitt, 2010). The EAI-24 conceptualizes environmental attitudes as covering twelve dimensions (e.g., enjoyment of nature, support for interventionist conservation policies, and confidence in science and technology), each of which is assessed by two items. These first-order factors have been shown to load on a single second-order factor (i.e., generalized environmental attitudes, Milfont & Duckitt, 2010). Hence, we decided to use the sum score across all 24 attitude times ($\alpha = .85$) for further construct validation of the PEBT. EAI scores are typically found to relate to pro-environmental behavior (Milfont & Duckitt, 2004) and they were also correlated to participants' performance in previous PEBT studies (Lange et al., 2018). We aimed to replicate this correlation in the present study while focusing on the relationship between PEBT choice behavior and other laboratory observations of pro-environmental behavior. For this latter purpose, we presented participants with another opportunity to engage in pro-environmental behavior at the end of the questionnaire. They completed eight trials of the Work for Environmental Protection Task (WEPT). On each trial, participants had the option to complete a page of a monotonous number-identification task, and for each page they completed, we donated a small amount of money to an environmental organization working for the protection of tropical forests. As an outcome measure for this task, we recorded how many pages participants completed to generate pro-environmental donations ($\alpha = .86$). A detailed description of the task can be found in the Supplementary Materials.

Table 1*First-session data depending on whether participants took part in the second session (Study 1)*

	participation in second session		effect size [95% CI]
	No (<i>n</i> = 108)	Yes (<i>n</i> = 94)	
Female participants	62%	61%	OR = 0.97 [0.55, 1.71]
Students	89%	87%	OR = 0.85 [0.36, 2.00]
Dutch native speakers	48%	32%	OR = 0.51 [0.28, 0.90]
Driver's license	58%	68%	OR = 1.52 [0.86, 2.72]
Car possession	26%	18%	OR = 0.63 [0.32, 1.24]
Bicycle possession	83%	81%	OR = 0.84 [0.41, 1.74]
Age	<i>M</i> = 23.02, <i>SD</i> = 4.48	<i>M</i> = 24.29, <i>SD</i> = 5.14	<i>d</i> = 0.27 [-0.01, 0.54]
Pro-environmental PEBT choices (%)	<i>M</i> = 60.83, <i>SD</i> = 28.13	<i>M</i> = 69.26, <i>SD</i> = 25.72	<i>d</i> = 0.30 [0.02, 0.58]
Euros donated (environment)	<i>M</i> = 0.49, <i>SD</i> = 0.71	<i>M</i> = 0.58, <i>SD</i> = 0.76	<i>d</i> = 0.12 [-0.15, 0.40]
Euros donated (neighbor)	<i>M</i> = 0.47, <i>SD</i> = 0.70	<i>M</i> = 0.45, <i>SD</i> = 0.59	<i>d</i> = -0.03 [-0.31, 0.25]

Note. PEBT = Pro-Environmental Behavior Task.

2.2 Results

2.2.1 Retest reliability

The proportion of pro-environmental PEBT choices was smaller in the second than in the first session, $t(93) = 3.08$, $d = 0.34$, 95% CI [-0.63, -0.05]. PEBT choice behavior was strongly correlated between sessions (Figure 1), as indicated by a Pearson coefficient of $r = .81$, 95% CI [.73, .87]. The intraclass correlation for consistency (computed using a two-way mixed effects model, for single measurements, McGraw & Wong, 1996) was of similar strength, ICC = .80, 95% CI [.72, .86]. An analysis of test-retest correlations as a function of PEBT trial characteristics can be found in the Supplementary Materials. This analysis

indicates test-retest reliability to be higher for trials involving larger cost of choosing the pro-environmental PEBT option (i.e., a larger waiting time difference between the options). It also suggests that, when considered in isolation, the trials with the lowest waiting time difference (10 seconds) may not elicit sufficient variation to provide a reliable measurement.

Figure 1

Test-retest correlation of PEBT choice behavior (Study 1)

Note. Scatterplot of the between-session correlation of performance on the Pro-Environmental Behavior Task (PEBT) across one month. Reference lines are linear regression lines.

2.2.2 Construct validity

The proportion of pro-environmental PEBT choices was moderately correlated to the amount of money participants donated to a pro-environmental organization, $r = .28$, 95% CI [.15, .40], but not to the amount of money participants transferred to the fellow participant sitting next to them, $r = .09$, 95% CI [-.05, .23]. The difference between these two Pearson correlation coefficients was significant, $z = 2.48$, $p = .014$ (calculated according to Meng, Rosenthal, & Rubin, 1992).

Among the $n = 94$ participants who returned for the second session, the proportion of pro-environmental PEBT choices was again correlated with the amount of money participants donated to the pro-environmental organization, $r = .41$, 95% CI [.23, .57]. We further examined correlations with the number of pages completed to generate donations for the protection of tropical forests on the WEPT ($M = 4.12$, $SD = 2.20$). Pro-environmental behavior on the PEBT was positively related to pro-environmental behavior on the WEPT, $r = .23$, 95% CI [.03, .41]. Similarly, the proportion of pro-environmental PEBT choices was positively related to participants' EAI-24 scores ($M = 122.22$, $SD = 16.56$), $r = .36$, 95% CI [.17, .52], replicating earlier work (Lange et al., 2018). An overview of all correlations can be found in the Table 2.

Table 2

Pearson correlation coefficients for the evaluation of construct validity (Study 1)

	1	2	3	4	5	6
1. PEBT S ₁	-					
2. Donation (environment) S ₁	.28 ^a	-				
3. Donation (neighbor) S ₁	.09 ^a	.39 ^a	-			
4. PEBT S ₂	.81 ^b	.39 ^b	.23 ^b	-		
5. Donation (environment) S ₂	.33 ^b	.85 ^b	.44 ^b	.41 ^b	-	
6. WEPT S ₂	.29 ^b	.25 ^b	.40 ^b	.23 ^b	.26 ^b	-
7. EAI-24 S ₂	.32 ^b	.31 ^b	.22 ^b	.36 ^b	.33 ^b	.06 ^b

Note. S₁ = first session, S₂ = second session, PEBT = Pro-Environmental Behavior Task, WEPT = Work for Environmental Protection Task, EAI-24 = Environmental Attitudes Inventory – short version. ^abased on $N = 202$, ^bbased on $N = 94$.

3. Relationship to different everyday pro-environmental behaviors (Study 2)

Study 1 revealed correlations between PEBT performance and recordings of other pro-environmental behaviors in the laboratory. Study 2 further examined the task's construct validity by relating participants' choice behavior in the laboratory to self-reports of their pro-environmental activities outside the laboratory. To examine these correlations in a powerful way, we meta-analyzed data from four studies that administered both the PEBT and a 12-item self-report scale for the assessment of pro-environmental behaviors.

3.1 Methods

We re- and meta-analyzed data from four laboratory studies that involved both the PEBT and the pro-environmental behavior self-report scale described by Schultz and colleagues (2005). The scale asked participants to report, on a 5-point scale (*never, rarely, sometimes, often, very often*), how frequently they engaged in twelve pro-environmental behaviors during the past year. In Study A (run in May 2017) and Study B (June/July 2017), participants completed the scale before completing the PEBT. A detailed description of these studies can be found in Lange and colleagues (2018). In Study C (November/December 2017) and Study D (April/May 2018), participants completed the PEBT before the self-report scale. A detailed description of these studies can be found in Lange and Dewitte (2019b). All studies involved the administration of unrelated measures (assessing, among others, personality traits) between assessments of PEBT performance and self-reported pro-environmental behavior. All studies have been approved by the local ethics committee and preregistered (with the exception of Study C, which was an exploratory study). All registrations, data, and materials are available on the Open Science Framework (<https://osf.io/apfdv/>). We confirm that we did not conduct any further studies administering the PEBT together with the pro-environmental behavior self-report scale.

In all four studies, participants (recruited from the faculty's subject pool) completed the PEBT according to the principle described in Study 1, with a few procedural details varying across studies. In Study C and D, participants completed two blocks of 24 trials each: choosing the car option turned on four lights in the first block and twelve lights in the second block. In Study A and B, participants completed an additional third block in which choosing the car option turned on eight lights and block order was counterbalanced across participants. Moreover, while Study A, C, and D used the labels "car" and "bicycle", Study B used pseudowords ("sest" and "dift") to label the pro-environmental and environmentally harmful PEBT response option. Finally, only Study D involved the administration of six PEBT practice trials. An overview of the four studies is given in Table 3.

Table 3

Overview of the four PEBT studies included in the meta-analysis (Study 2)

	Study A	Study B	Study C	Study D
number of PEBT trials	72	72	48	48
PEBT waiting time differences (seconds)	0, 5, 10, 15, 20, 30	0, 5, 10, 15, 20, 30	5, 10, 15, 20, 30, 60	5, 10, 15, 20, 30, 60
sample size	$N = 122$	$N = 122$	$N = 143$	$N = 263$
gender	51% female, 48% male, 1% don't say	61% female, 39% male	58% female, 41% male, 1% don't say	63% female, 36% male, 1% don't say
students	90%	79%	85%	92%
age (years)	$M = 24.51$, $SD = 6.73$	$M = 24.43$, $SD = 5.45$	$M = 24.38$, $SD = 6.28$	$M = 22.98$, $SD = 3.75$
pro-environmental PEBT choices (%)	$M = 66.05$, $SD = 23.95$	$M = 71.36$, $SD = 22.06$	$M = 53.72$, $SD = 26.51$	$M = 59.44$, $SD = 25.00$
average score pro-environmental behavior self-report scale	$M = 2.90$, $SD = 0.72$, $\alpha = .87$	$M = 2.94$, $SD = 0.57$, $\alpha = .76$	$M = 2.91$, $SD = 0.81$, $\alpha = .86$	$M = 2.87$, $SD = 0.66$, $\alpha = .85$

Note. PEBT = Pro-Environmental Behavior Task.

For each study, we computed Spearman correlations between the proportion of pro-environmental PEBT choices and the frequency of engaging in everyday pro-environmental behaviors, separately for the twelve behaviors assessed with the self-report scale. By this means, we obtained effect sizes for the association between PEBT performance and other pro-environmental behaviors. For comparison, we also computed, for all of the twelve self-report items, the average Spearman correlation with the other eleven behaviors assessed with the self-report scale. This provided us with an estimate of how closely the frequencies of different pro-environmental behaviors are related to each other when they are assessed using the same method (i.e., retrospective self-reports).

For each self-reported pro-environmental behavior, we then conducted two meta-analyses (one for the association with PEBT performance and one for the association with other self-reported behaviors), aggregating effect-size information across the four studies. Given the procedural differences between the studies, we decided to use a random-effects model for meta-analysis (i.e., the Hedges-Vevea model implemented by Field & Gillett, 2010). As displayed in Table 3, waiting time differences (i.e., the cost of showing pro-environmental behavior on the PEBT) varied between studies as did the distributions of PEBT choice behavior. The variance of PEBT choices was, for example, smaller in Study B than in Study C, which might lead to attenuated correlations in Study B and, thus, to substantial effect-size heterogeneity across studies. Heterogeneity of effect sizes was examined using Cochran's Q and transformed into I^2 (Higgins, Thompson, Deeks, & Altman, 2003). The I^2 index served as an estimate of between-study variability in true effect sizes, with I^2 values of about 25%, 50% and 75% indicating low, moderate, and high heterogeneity, respectively (Higgins et al., 2003).

3.2 Results

Meta-analytical results are displayed in Table 4. PEBT choice behavior was significantly related to all twelve pro-environmental behaviors assessed with the self-report scale. All correlations were positive and small-to-medium in size ($r_s = .10 - .29$). The smallest correlation was observed between PEBT choices and self-reported pro-environmental donations, whereas PEBT choice behavior was most closely related to participants' self-reported frequency to "conserve gasoline by walking or bicycling". The confidence interval of the latter effect size excluded most of the other effect sizes, indicating that individual differences in PEBT performance were more strongly related to transportation choices than to most of the other pro-environmental behaviors we assessed.

All self-reported pro-environmental behaviors were also related to other self-reported pro-environmental behaviors, with average correlations ranging from $r_s = .18$ to $r_s = .36$. For most behaviors, correlations with other self-reports were stronger than correlations with PEBT choice behavior (although some of the confidence intervals overlapped). One notable exception was participants' self-reported frequency to conserve gasoline by walking or bicycling, which was more closely related to PEBT performance, $r_s = .29$, 95% CI [.21, .39], than to other self-reported pro-environmental behaviors, $r_s = .18$, 95% CI [.10, .26]. Effect-size heterogeneity across studies was negligible for most correlations and graphical inspection of Figure 2 does not reveal any signs of small-study bias (i.e., the larger studies did not systematically produce smaller effect sizes than the smaller studies).

PSYCHOMETRIC EVALUATION OF THE PEBT

Table 4

Meta-analysis of relationships among pro-environmental behaviors (Study 2)

self-report item ("Please indicate how often you have done each of the following in the past year.")	<i>N</i>	<i>M</i>	<i>SD</i>	<i>correlation with PEBT choice behavior</i>	<i>95 % CI</i>	<i>Q</i>	<i>I²</i>	<i>correlation with self-reported behaviors</i>	<i>95 % CI</i>	<i>Q</i>	<i>I²</i>
looked for ways to reuse things	642	3.72	1.01	.21	.13, .29	3.20	6%	.26	.18, .33	2.02	0%
recycled newspapers	544	3.53	1.42	.20	.12, .28	1.27	0%	.27	.18, .35	1.26	0%
recycled cans or bottles	640	4.04	1.18	.13	.06, .21	0.92	0%	.24	.17, .32	1.81	0%
encouraged friends or family to recycle	642	3.28	1.26	.16	.09, .24	0.11	0%	.36	.28, .43	3.44	13%
purchased products in reusable containers	632	3.04	1.16	.14	.06, .22	0.69	0%	.29	.22, .37	3.06	2%
picked up litter that was not your own	624	2.87	1.12	.17	.09, .24	2.41	0%	.23	.15, .30	0.26	0%
composted food scraps	594	3.07	1.39	.12	.03, .20	3.14	4%	.21	.12, .28	1.24	0%
conserved gasoline by walking or bicycling	608	3.98	1.11	.29	.21, .36	1.76	0%	.18	.10, .26	1.57	0%
wrote a letter supporting an environmental issue	605	1.32	0.78	.15	.06, .25	0.44	0%	.25	.17, .33	3.28	9%
voted for a candidate who supported environmental issues	444	2.52	1.51	.23	.09, .35	6.23	52%	.28	.17, .39	4.58	34%
donated money to an environmental group	621	1.59	0.99	.10	.02, .18	0.40	0%	.22	.14, .30	2.87	0%
volunteered time to help an environmental group	616	1.69	1.05	.16	.08, .24	2.14	0%	.24	.16, .32	0.54	0%

359 *Note.* Correlations with self-reported behaviors are average correlations with other self-report items. PEBT = Pro-Environmental Behavior Task.

360 **Figure 2**

361 *Study-level correlations between PEBT performance and self-reported behaviors (Study 2)*

362

363 *Note.* Horizontal bars are 95% confidence intervals. PEBT = Pro-Environmental Behavior

364 Task. Please see Table 4 for complete items texts.

4. Discussion

The present results shed further light on the psychometric properties of the PEBT as a measure of individual differences in the propensity to show pro-environmental behavior. PEBT choice behavior was found to be fairly stable across a period of approximately one month, with a between-session correlation of $r = .81$ and an ICC of .80. In addition, PEBT performance was correlated to both objective observations and subjective reports of other pro-environmental behaviors. Our laboratory study found PEBT choice behavior to be more closely related to pro-environmental vs. prosocial donation behavior and our meta-analysis suggested a particularly close link to self-reported pro-environmental behavior in the domain of transportation choices. In the following, we will discuss the implications of these findings for the reliability, validity, and utility of the PEBT in pro-environmental behavior research.

If choice behavior on the PEBT is to be treated as an indicator of a person characteristic, it needs to exhibit some degree of temporal stability (see Yoder et al., 2018). Using a 40-trial PEBT version in Study 1, we obtained an estimate for the task's one-month test-retest reliability of about .80. Similar estimates have been reported for measurement instruments based on self-reported pro-environmental behavior ($r = .76$, Kaiser & Wilson, 2000, $r = .85$, Markle, 2013). Yet it might still be a conservative estimate as we observed a decrease in the proportion of pro-environmental PEBT choices from the first to the second session. This decrease is reminiscent of the decrease in pro-environmental choices that we typically see within a PEBT session (Lange & Dewitte, 2019b): a large majority of participants start by choosing the bicycle option on the first trials and only switch to a more balanced choice behavior after having experienced the PEBT consequences. In Study 1, the additional variance produced by this trend likely affected the first session more strongly than the second and thus reduced the proportion of shared variance between the sessions. This notion receives tentative support from the data of $n = 16$ participants who indicated that they

had participated in a PEBT study before the first session of Study 1. These participants were already acquainted with the PEBT consequences, their average proportion of pro-environmental PEBT choices was constant (64%) across sessions, and test-retest reliability for this subsample was $ICC = .91$, 95% CI [.77, .97]. This finding indicates that the test-retest reliability of the PEBT could be further increased by adding practice trials to the task. Similarly, our supplementary analyses suggest that adding trials with a larger waiting time difference between PEBT options could further improve the task's reliability. If choosing the pro-environmental PEBT option involves only small waiting time cost (e.g., 10 seconds), a large majority of participants exclusively choose this option, which results in ceiling effects, restricted variance, and reduced reliability (see Table S2). Moreover, it should be noted that participants completed the first session of our test-retest study under different experimental conditions (observable vs. private), whereas all of them completed the second session in private. The context variation between sessions for half of our sample might have attenuated the correlation between sessions. Yet test-retest reliability was not smaller in participants who completed the first session in the observable condition ($ICC = .86$, compared to the private condition, $ICC = .73$), which renders this possibility less likely.

In order for the PEBT to play a role in the assessment of individual differences in the propensity to show pro-environmental behavior, behavior on the task also needs to be correlated to other indicators of that propensity (e.g., other pro-environmental behaviors). Study 1 found PEBT choice behavior to be moderately related to other laboratory observations of pro-environmental behavior (i.e., pro-environmental donations, behavior on the WEPT). Hence, individual differences in the behavioral trade-offs between a) saving time and conserving energy (PEBT), b) keeping money and offsetting carbon emissions (donation), and c) avoiding a boring number-identification task and contributing to rainforest preservation (WEPT) seem to overlap to a certain extent. In addition, Study 2 revealed small-to-medium

correlations between PEBT choice behavior and all of the twelve pro-environmental behaviors that we assessed via self-report. PEBT performance was significantly associated with the reported frequency to engage in behaviors as diverse as voting, volunteering, and recycling. The strength of these correlations is not sufficient to attest to the construct validity of the PEBT as a standalone measure of the propensity to show pro-environmental behavior. PEBT choice behavior is one pro-environmental behavior (just like recycling or donating for an environmental organization) and assessing individual differences in one pro-environmental behavior cannot reliably distinguish individuals regarding their tendency to show other pro-environmental behaviors (see also Weigel & Newman, 1976). Given the differences between pro-environmental behaviors (e.g., in the cost they involve), it is not surprising that people who behave pro-environmentally in one domain do not do so in another domain (Kaiser, 1998; Kaiser, Byrka, & Hartig, 2010). This is also illustrated by the imperfect correlations between self-reports of pro-environmental behavior that we observed in Study 2. Correlations among these behaviors were only slightly larger than the correlations between self-reported behaviors and PEBT performance, with the difference plausibly being due to common-method bias (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

Yet there appears to be some shared variance among self-reported pro-environmental behaviors as well as between (self-reported) pro-environmental behaviors and PEBT performance. Based on such consistent patterns of positive correlations among behaviors, information can be aggregated into composite measures reflecting a person's general propensity to engage in pro-environmental activities (Weigel & Newman, 1976). Neither PEBT choice behavior nor the question "How often have you recycled newspaper during the past year?" can be regarded as *valid measures* of that propensity, but our results suggest that they can both serve as *valid indicators*. Self-reports of pro-environmental behaviors have frequently been used in this capacity (e.g., Kaiser, 1998; Kaiser & Wilson, 2000; Markle,

2013), but propensity measurement based on objectively observed behavioral indicators has been rare, possibly because of the difficulty to observe such indicators. This is problematic because many individual-difference studies rely on self-reports not only for assessing the propensity to show pro-environmental behavior, but also for assessing its potential correlates (e.g., attitudes or personality traits). In the absence of methodological multiplicity, observed correlations with the propensity to show pro-environmental behavior are likely to be systematically inflated by common-method variance (Lange & Dewitte, 2019a). Given the empirical support of the PEBT's validity as an indicator of this propensity, the task can now be used to mitigate this problem through enabling multimethod studies. An important next step in that direction would be the development and validation of additional procedures for the objective recording of pro-environmental behavior at the individual level (e.g., Klein et al., 2017; Tarditi, Hahnel, Jeanmonod, Sander, & Brosch, 2018).

Of further relevance for behavior-based propensity measures, we found PEBT choice behavior to be more closely related to pro-environmental than to prosocial donation behavior in Study 1. Given that the propensity to show pro-environmental behavior is often confounded with the propensity to show prosocial behavior (Klein et al., 2017), such a differential correlation may be considered as valuable (although preliminary) evidence for the discriminant validity of the PEBT. Rather than relating to all kinds of non-selfish behaviors, PEBT choice behavior seems to relate to the relatively specific propensity to show pro-environmental behavior. This hypothesis, however, requires confirmation from future studies involving a more comprehensive assessment of the propensity to show prosocial behavior.

Another interesting pattern of differential associations was observed in the meta-analysis of Study 2. PEBT choice behavior was most closely related to participants' self-reported transportation choices, which, in turn, were more closely related to PEBT performance than to other self-reported pro-environmental behaviors. This finding indicates

that the PEBT may be particularly useful for studying individual differences in environmentally relevant transportation studies. Arguably, PEBT task contingencies are more similar to everyday transportation choices than to some other pro-environmental behaviors. Showing pro-environmental behavior on the task and using environmentally friendly modes of transportation in everyday life involve costs in terms of waiting time and environmental benefits in terms of saved energy and CO₂ emissions. In contrast, the costs of pro-environmental donation behavior, for example, are mostly financial. Individual differences in time resources may thus be relevant for PEBT performance and transportation choices, but to a lesser degree for donation behavior. Future systematic examinations are required to clarify the role of such contingency differences in accounting for differential correlations among pro-environmental behaviors. More work is also needed to extend the results obtained in the present studies to other populations. The vast majority of participants included in all studies were university students living in Belgium and the psychometric properties of the PEBT may vary across populations. Use of the PEBT in other populations should thus ideally be preceded by additional validation studies (e.g., Lange & Iwasaki, 2020).

In sum, our studies support the use of the PEBT as a reliable and valid indicator of the propensity to show pro-environmental behavior. Previous studies have already demonstrated the choice of the pro-environmental PEBT option to be an instance of pro-environmental behavior (Lange et al., 2018). However, this demonstration alone does not imply that the task can be used for the meaningful assessment of individual differences. To play a role in behavior-based propensity measurement, the observed behavior also has to be somewhat stable over time and to converge with other indicators of the propensity in question. The present results suggests these criteria to be largely met by PEBT choice behavior. Therefore, the PEBT might not only serve as a procedure for the valid experimental study of pro-

489 environmental behavior *per se*, but also provide a valid behavioral indicator for a person's
490 propensity to engage in pro-environmental behavior.

ACCEPTED MANUSCRIPT

References

- Alpizar, F., Carlsson, F., & Johansson-Stenman, O. (2008). Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics*, 92, 1047-1060.
<https://doi.org/10.1016/j.jpubeco.2007.11.004>
- Carrico, A. R., Raimi, K. T., Truelove, H. B., & Eby, B. (2018). Putting your money where your mouth is: an experimental test of pro-environmental spillover from reducing meat consumption to monetary donations. *Environment and Behavior*, 50, 723-748.
<https://doi.org/10.1177/0013916517713067>
- Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63, 665-694. <https://doi.org/10.1348/000711010X502733>
- Gifford, R. (2014). Environmental psychology matters. *Annual Review of Psychology*, 65, 541-579. <https://doi.org/10.1146/annurev-psych-010213-115048>
- Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327, 557-560.
<https://doi.org/10.1136/bmj.327.7414.557>
- Kaiser, F. G. (1998). A general measure of ecological behavior. *Journal of Applied Social Psychology*, 28, 395-422. <https://doi.org/10.1111/j.1559-1816.1998.tb01712.x>
- Kaiser, F. G., & Wilson, M. (2000). Assessing People's General Ecological Behavior: A Cross-Cultural Measure. *Journal of Applied Social Psychology*, 30, 952-978.
<https://doi.org/10.1111/j.1559-1816.2000.tb02505.x>
- Kaiser, F. G., Merten, M., & Wetzels, E. (2018). How do we know we are measuring environmental attitude? Specific objectivity as the formal validation criterion for measures of latent attributes. *Journal of Environmental Psychology*, 55, 139-146.
<https://doi.org/10.1016/j.jenvp.2018.01.003>

- 516 Kaiser, F. G., Byrka, K., & Hartig, T. (2010). Reviving Campbell's paradigm for attitude
517 research. *Personality and Social Psychology Review*, 14, 351-367.
518 <https://doi.org/10.1177/1088868310366452>
- 519 Klein, S. A., Hilbig, B. E., & Heck, D. W. (2017). Which is the greater good? A social
520 dilemma paradigm disentangling environmentalism and cooperation. *Journal of*
521 *Environmental Psychology*, 53, 40-49. <https://doi.org/10.1016/j.jenvp.2017.06.001>
- 522 Kormos, C., & Gifford, R. (2014). The validity of self-report measures of proenvironmental
523 behavior: A meta-analytic review. *Journal of Environmental Psychology*, 40, 359-371.
524 <https://doi.org/10.1016/j.jenvp.2014.09.003>
- 525 Lakes, K. D. (2013). Restricted sample variance reduces generalizability. *Psychological*
526 *Assessment*, 25, 643-650. <https://doi.org/10.1037/a0030912>
- 527 Lange, F., Brick, C., & Dewitte, S. (2020). Green when seen? No support for an effect of
528 observability on environmental conservation in the laboratory: a registered report.
529 *Royal Society Open Science*, 7, 190189. <https://doi.org/10.1098/rsos.190189>
- 530 Lange, F., & Dewitte, S. (2019). Cognitive Flexibility and Pro-environmental Behaviour: A
531 Multimethod Approach. *European Journal of Personality*, 33, 488-505.
532 <https://doi.org/10.1002/per.2204>
- 533 Lange, F., & Dewitte, S. (2019). Measuring pro-environmental behavior: Review and
534 recommendations. *Journal of Environmental Psychology*, 63, 92-100.
535 <https://doi.org/10.1016/j.jenvp.2019.04.009>
- 536 Lange, F., & Iwasaki, S. (2020). Validating the Pro-Environmental Behavior Task in a
537 Japanese Sample. *Sustainability*, 12, 9534. <https://doi.org/10.3390/su12229534>
- 538 Lange, F., Steinke, A., & Dewitte, S. (2018). The Pro-Environmental Behavior Task: A
539 laboratory measure of actual pro-environmental behavior. *Journal of Environmental*
540 *Psychology*, 56, 46-54. <https://doi.org/10.1016/j.jenvp.2018.02.007>

- 541 Markle, G.L. (2013). Pro-environmental behavior: Does it matter how it's measured?
 542 Development and validation of the Pro-Environmental Behavior Scale (PEBS).
 543 *Human Ecology*, 41, 905-914. <https://doi.org/10.1007/s10745-013-9614-8>
- 544 McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation
 545 coefficients. *Psychological Methods*, 1, 30-46. [https://doi.org/10.1037/1082-](https://doi.org/10.1037/1082-989X.1.1.30)
 546 989X.1.1.30
- 547 Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation
 548 coefficients. *Psychological Bulletin*, 111, 172-175.
- 549 Milfont, T. L., & Duckitt, J. (2004). The structure of environmental attitudes: A first-and
 550 second-order confirmatory factor analysis. *Journal of Environmental Psychology*, 24,
 551 289-303. <https://doi.org/10.1016/j.jenvp.2004.09.001>
- 552 Milfont, T. L., & Duckitt, J. (2010). The environmental attitudes inventory: A valid and
 553 reliable measure to assess the structure of environmental attitudes. *Journal of*
 554 *Environmental Psychology*, 30, 80-94. <https://doi.org/10.1016/j.jenvp.2009.09.001>
- 555 Otto, S., Evans, G. W., Moon, M. J., & Kaiser, F. G. (2019). The development of children's
 556 environmental attitude and behavior. *Global Environmental Change*, 58, 101947.
 557 <https://doi.org/10.1016/j.gloenvcha.2019.101947>
- 558 Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method
 559 biases in behavioral research: A critical review of the literature and recommended
 560 remedies. *Journal of Applied Psychology*, 88, 879-903.
 561 <http://dx.doi.org/10.1037/0021-9010.88.5.879>
- 562 Polit, D. F. (2014). Getting serious about test-retest reliability: a critique of retest research
 563 and some recommendations. *Quality of Life Research*, 23, 1713-1720.
 564 <https://doi.org/10.1007/s11136-014-0632-9>

- Steg, L., & Vlek, C. (2009). Encouraging pro-environmental behaviour: An integrative review and research agenda. *Journal of Environmental Psychology*, 29, 309-317.
<https://doi.org/10.1016/j.jenvp.2008.10.004>
- Stern, P. C. (2000). New environmental theories: toward a coherent theory of environmentally significant behavior. *Journal of Social Issues*, 56, 407-424.
<https://doi.org/10.1111/0022-4537.00175>
- Tarditi, C., Hahnel, U.J., Jeanmonod, N., Sander, D., & Brosch, T. (2018). Affective Dilemmas: The Impact of Trait Affect and State Emotion on Sustainable Consumption Decisions in a Social Dilemma Task. *Environment and Behavior*.
<https://doi.org/10.1177/0013916518787590>
- Vesely, S., & Klöckner, C. A. (2018). How anonymity and norms influence costly support for environmental causes. *Journal of Environmental Psychology*, 58, 27-30.
<https://doi.org/10.1016/j.jenvp.2018.07.013>
- Weigel, R. H., & Newman, L. S. (1976). Increasing attitude-behavior correspondence by broadening the scope of the behavioral measure. *Journal of Personality and Social Psychology*, 33(6), 793-802. <https://doi.org/10.1037/0022-3514.33.6.793>
- Yoder, P., Lloyd, B. P., & Symons, F. (2018). *Observational measurement of behavior*. Baltimore, MD: Paul H. Brookes Publishing.

Supplementary Materials

I. Pilot study

A pilot study was conducted to evaluate the one-year stability of choice behavior on the Pro-Environmental Behavior Task (PEBT). Participants who had taken part in a PEBT individual differences study were asked to return to the lab twelve months later for a second session to obtain a first estimate of the test-retest reliability of the procedure. For comparison, we also analyze the one-year test-retest reliability of a pro-environmental self-report scale (Schultz et al., 2005).

Methods

First session. A total of 263 participants completed the PEBT as part of a study on the relationship between personality traits and pro-environmental behavior in the spring of 2018 (Lange & Dewitte, 2019b). Before the PEBT, all participants completed a cognitive functioning task and after the last PEBT trial, participants completed multiple self-report scales. The entire procedure took about 60 minutes to complete and participants received twelve euros for their participation.

The PEBT version used in this study comprised 48 trials, preceded by six practice trials. Each trial involved a consequential choice between two response options, labelled “car” and “bicycle”. Following their choice, participants had to endure a waiting period before they were presented with the next choice. The waiting time associated with choosing the bicycle option was always longer than the waiting time associated with choosing the car option. The waiting time difference between the options varied randomly at six levels (5s, 10s, 15s, 20s, 30s, 60s). In addition, every time the car was chosen, an array of USB-powered lights located on the desk of the participants was illuminated for the duration of the waiting period. In first half of the trials four lights were illuminated, and in the second half twelve lights were illuminated. Participants were explicitly informed about these consequences, that is, about the

waiting times associated with the response options, the waiting time difference, the number of lights to be turned on by choosing the car option, and the associated CO₂ emissions.

The self-report questionnaire involved a pro-environmental behavior scale (Schultz et al., 2005). On this scale, participants indicated how frequently they engaged in 12 pro-environmental behaviors during the past year (1 *never* – 5 *very often*). They could choose the option ‘not applicable’ when they thought that they have had no opportunity to perform the behavior. We computed an average score across the twelve items of the pro-environmental scale ($\alpha = .85$).

Second session. Approximately one year after the end of data collection for the first session, we attempted to contact all participants via email to invite them to take part in a follow-up study. Many participants had finished their studies during the past year and thus could not be reached via their old email address anymore. Similarly, many of the participants who could be reached indicated that they had moved away in the meantime. We offered testing slots over a period of two weeks to give every participant who wanted to participate the possibility to do so and obtained second-session data from 40 participants. A comparison of the participants who participated vs. did not participate in the second session can be found in Table S1. The mean interval between the two testing sessions was 355.08 days ($SD = 17.52$ days, range = 329 – 386 days). The second session was identical to the first one.

Table S1*First-session data depending on whether participants took part in the second session (pilot study)*

	participation in second session		effect size [95% CI]
	No (<i>n</i> = 223)	Yes (<i>n</i> = 40)	
Female participants	64%	60%	OR = 0.84 [0.42, 1.67]
Students	89%	87%	OR = 0.70 [0.22, 2.20]
Dutch native speakers	67%	58%	OR = 0.67 [0.34, 1.33]
Driver's license	72%	63%	OR = 0.66 [0.33, 1.33]
Car possession	27%	20%	OR = 0.66 [0.29, 1.52]
Bicycle possession	93%	90%	OR = 1.79 [0.43, 5.19]
Age	<i>M</i> = 22.85, <i>SD</i> = 3.46	<i>M</i> = 23.70, <i>SD</i> = 5.10	<i>d</i> = 0.23 [-0.11, 0.56]
Bicycle choices (%)	<i>M</i> = 59.55, <i>SD</i> = 24.78	<i>M</i> = 58.85, <i>SD</i> = 26.54	<i>d</i> = -0.03 [-0.37, 0.31]
Self-reported PEB	<i>M</i> = 2.85, <i>SD</i> = 0.64	<i>M</i> = 3.02, <i>SD</i> = 0.74	<i>d</i> = 0.26 [-0.08, 0.60]

Results

The proportion of pro-environmental PEBT choices did not change substantially between the first (*M* = 58.85, *SD* = 26.54) and second session (*M* = 57.17, *SD* = 28.04), $M_{\text{diff}} = -1.68$, 95% CI [-8.04, 11.41]. In contrast, we observed a slight increase in the average score of the pro-environmental behavior self-report scale (session one: *M* = 3.02, *SD* = 0.74, session two: *M* = 3.23, *SD* = 0.64, $M_{\text{diff}} = 0.21$, 95% CI [0.02, 0.39]). PEBT choice behavior was moderately correlated between sessions, $r = .38$ [.08, .62]. The between-session correlation of participants' average scores on the pro-environmental behavior self-report scale seemed stronger, $r = .65$, 95% CI [.42, .80] (see Figure S1). Test-retest reliability of individual self-reported pro-environmental behaviors ranged between $r = .01$ ("wrote a letter supporting an environmental issue") and $r = .72$ ("conserved gasoline by walking or bicycling").

Figure S1. Scatterplots of between-session correlations of performance on the Pro-Environmental Behavior Task (PEBT, left) and of average scores on the pro-environmental behavior (PEB) self-report scale (right) across one year. Reference lines are linear regression lines.

II. Work for Environmental Protection Task (WEPT)

At the very end of the questionnaire administered in the second session, participants were presented with a number-identification task. They saw twenty two-digit numbers and were instructed to identify all those numbers that consisted of an even first digit and an odd second digit. This task was based on the concentration test developed by Jacobs (2014). After participants had identified all target numbers (with automatic corrective feedback, if required), they were thanked and informed that they were done with the standard part of the study. However, they were then given the option to “continue with the number-identification task for a little longer”. The instructions continued:

“For each page of numbers that you correctly complete, we will donate a small amount of money (0.20 Euro) to OroVerde, an environmental non-profit organization working on the preservation of tropical forests. OroVerde invests, for example, in reforestation and environmental education in Latin America.

There are eight pages of numbers in total. For each page, we will tell you how many numbers you will have to check (so that you can estimate the effort). You can then decide, for each page separately, if you want to do this additional effort or not.

Doing this task is completely voluntary. If you want, you can decline checking the numbers (by clicking "no") every time and go directly to the end of the study.”

They were then presented with the first choice trial, asking them whether they wanted to check ten numbers for a donation of 0.20 euros. If participants decided to take on this task, they were presented with a page containing that number of numbers. They could take as much time as they wanted to identify the correct target numbers before proceeding to the next choice trial (without corrective feedback). The trials demanded progressively increasing amounts of effort (10, 20, 50, 100, 150, 200, 250, 300 numbers). Participants were encouraged to work thoroughly as we would “only count pages that are at least 90% correct”.

671 We recorded the time that participants spent on each page of numbers. When their time was
672 more than two standard deviations below the sample mean for that page, we did not count this
673 page as completed. The number of completed pages served as outcome measure for this task,
674 reflecting how participants navigate the trade-off between minimizing the effort and time
675 spent in our laboratory and generating environmental benefits through pro-environmental
676 donations.

ACCEPTED MANUSCRIPT

III. Cost-dependent test-retest reliability

Table S2 displays an analysis of the one-month test-retest reliability of PEBT choice behavior as a function of trial characteristics. With increasing waiting time differences (WTDs), the cost of choosing the pro-environmental bicycle option increase (i.e., participants have to endure larger amounts of additional waiting time). It can be seen that the 95% confidence interval around the reliability estimate for the 10-second-WTD trials does not include the estimates for the trials with the longest WTDs.

Table S2

Test-retest correlation as a function of PEBT trial characteristics

	trials	session one	session two	$r_{1,2}$ [95% CI]	ICC [95% CI]
WTD = 10s	8	$M = 93.22$, $SD = 16.46$	$M = 89.49$, $SD = 23.21$	$r = .59$ [.44, .71]	.56 [.41, .69]
WTD = 15s	8	$M = 82.45$, $SD = 26.14$	$M = 78.32$, $SD = 31.06$	$r = .70$ [.58, .79]	.69 [.57, .78]
WTD = 20s	8	$M = 67.69$, $SD = 33.07$	$M = 63.03$, $SD = 36.57$	$r = .70$ [.58, .79]	.70 [.58, .79]
WTD = 25s	8	$M = 56.91$, $SD = 35.48$	$M = 51.20$, $SD = 37.88$	$r = .78$ [.69, .85]	.78 [.68, .84]
WTD = 30s	8	$M = 46.01$, $SD = 36.35$	$M = 37.10$ $SD = 38.60$	$r = .74$ [.63, .82]	.74 [.63, .82]
Four lights	20	$M = 63.14$, $SD = 29.40$	$M = 58.67$, $SD = 30.61$	$r = .75$ [.65, .83]	.75 [.65, .83]
Twelve lights	20	$M = 75.37$, $SD = 26.13$	$M = 68.99$, $SD = 28.70$	$r = .70$ [.58, .79]	.70 [.58, .79]
Total	40	$M = 69.25$, $SD = 25.72$	$M = 63.83$, $SD = 28.47$	$r = .81$ [.73, .87]	.80 [.72, .86]

Note. PEBT = Pro-Environmental Behavior Task, WTD = waiting-time difference between the pro-environmental (bicycle) and the environmentally harmful (car) option, CI = confidence interval, ICC = intraclass correlation for consistency (computed using a two-way mixed effects model, for single measurements)

References (Supplementary Materials)

- Jacobs, B. (2014). Analyse von Testgütekriterien und Übungseffekten zweier Online-Konzentrationstests [Analysis of test quality and practice effects of two online-concentration tests]. <http://hdl.handle.net/20.500.11780/3371>
- Schultz, P. W., Gouveia, V. V., Cameron, L. D., Tankha, G., Schmuck, P., & Franěk, M. (2005). Values and their relationship to environmental concern and conservation behavior. *Journal of Cross-Cultural Psychology*, 36(4), 457-475. <https://doi.org/10.1177/0022022105275962>