## Supporting Information

### Experiment 1, ancillary analyses

**Age and gender.** Neither age nor gender was a significant predictor of helping behavior, $\beta(N = 51) = 0.14$, $p = .32$, and $t(49) = .51$, $p = .62$, respectively. There were also no significant interactions between these demographic variables and condition, $F(2, 45) = 0.41$, $p = .67$, and $F(2, 45) = 2.07$, $p = .14$, respectively.

**Manipulation questions.** On an exploratory basis, we examined whether there were differences in children's responses to the manipulation questions as a function of wording (e.g., "How much do you want to be a helper [help when someone needs help]?"). Past research among adults has not found condition differences in response to analogous manipulation questions (Bryan et al., 2011). Combining across the two items, the condition effect showed a non-significant trend (noun: $M = 5.38$ out of 6, $SD = 1.27$; verb: $M = 4.59$, $SD = 1.53$), $F(1,32) = 2.71$, $p = .11$.

Because of this trend, we also explored whether responses to the manipulation questions were related to our dependent variable; there was no such relationship, $r(34) = .07$, $p = .71$. We also tested whether responses to these questions contributed to the condition effect on helping, and found that it did not. Controlling for responses to these questions, the difference between the noun and verb conditions (the two conditions in which the manipulation questions were asked) remained significant, $F(1,31) = 5.24$, $p = .029$.

Apart from the question of whether responses to the manipulation questions contributed to the effect on helping, it is interesting to consider why there was not a significant effect of condition on these responses. We believe the most likely reason for

this relates to Donald Campbell's (1963) "hurdles" explanation for attitude-behavior inconsistency. Campbell points out that reporting a positive attitude toward an approved but effortful behavior (e.g., helping) is much easier than actually performing that behavior. So, it is unsurprising that children in our verb condition reported attitudes about helping (a low hurdle) that were nearly as positive as those reported by our noun condition children. Only a more difficult criterion, like actually putting down interesting toys to go help an experimenter (i.e., a high hurdle), would be expected to clearly distinguish the truly committed from those who merely feel positively disposed.

**Experiment 1, additional coding of session videos**

      **Helping behavior.** The helping data were based on real-time coding by the experimenters, who followed pre-determined coding rules. To assess the reliability of this coding, in which the experimenter was aware of children's condition assignment, we enlisted an independent coder (blind to condition) to review the videos of each session and re-code helping behavior. This code was nearly identical to the experimenter's coding (kappa = .99); they only differed with respect to one task (the crayons) for one participant. In this case, the experimenter coded the behavior as not helping and the video coder coded it as helping. To resolve this discrepancy, we reviewed the video ourselves and found that the child picked up one crayon from the ground (before any prompts) and immediately used it to draw with. When the experimenter delivered the verbal prompt ("Better pick those up") the child responded "Me? But I didn't drop them." and continued drawing without helping. Because the participant did not appear to intend to help, we retained the original coding of the child's behavior as not helping. Regardless, the results are not affected by this difference.

**Task timing.** Our experimental script called for specific time delays between helping prompts delivered by the experimenters. The experimenters estimated these delays during the experimental sessions. To rule out the possibility that task timing was affected by experimenter bias, an independent video coder (blind to condition) measured the exact timing of the helping window for each task, including the timing of each prompt.

The most important question in terms of timing is whether children in the noun condition were given a longer window to help with any of the tasks than children in the verb condition. This was not the case; there were no significant differences between the noun and verb conditions in the total time window for helping for any of the four tasks.

We also examined the timing within tasks of the individual prompts. Of the 11 statistical comparisons, only one showed a significant difference between the noun and verb conditions: the first prompt in the bin task was slightly later in the noun condition ($M$ = 8.19 seconds, $SD$ = 2.74) than in the verb condition ($M$ = 6.59, $SD$ = 0.87), $t(31)$ = 2.29, $p$ = .029. To ensure that this difference did not affect helping, we also looked at children's average latency to help. Children in the noun condition did not take any longer than children in the verb condition to help with the bin task (or with any of the other three tasks), suggesting that the timing of this prompt did not affect the results of the experiment.

**Experiment 2, ancillary analyses**

**Age and gender.** Gender was not a significant predictor of helping, $F(1, 81)$ = 0.99, $p$ = .32. Age was a marginally significant predictor, $F(1, 78)$ = 3.55, $p$ =.06, with older children helping less than younger ones. There was no significant interaction

between either demographic variable and condition, $F(1, 77) = 0.002$, $p = .96$, and $F(1, 80) = 0.79$, $p = .14$, respectively.

**Manipulation questions.** We again explored whether there were differences in children's responses to the manipulation questions as a function of wording. Again, there was no significant condition difference (noun: $M = 5.61$ out of 6, $SD = 0.73$; verb: $M = 5.70$, $SD = 0.70$), $F(1,82)=0.003$, $p = .96$.

**Additional analyses based on analytic decisions.** As noted in the Participants section, we excluded data from four participants and tested our central hypothesis in the Results section controlling for school to account for variation in testing environments. Because precise significance values are affected by those decisions, here we present results from various combinations of analytical decisions. First, we describe each decision in detail:

*Experimenter error.* For two participants (one assigned to the noun condition and one assigned to the verb condition), Experimenter 1 accidentally switched between the noun wording and the verb wording. These data were excluded because the participants did not actually receive the treatment they were assigned.

*Procedure disruption.* One participant (in the noun condition) needed a bathroom break after the helping language manipulation and before the dependent variable stage. The break took 5 minutes, which we considered to be a significant disruption of the standardized procedure.

*Distraction during DVs.* For one participant (in the noun condition), Experimenter 2 made a note that there was noise during the blocks task that she "tried to talk over." There were no similar issues in any other experimental session. An

examination of the video recording revealed that the noise was caused by the child's

playing and was loud enough that one of the experimenter's prompts was inaudible. This

was considered to be a significant disruption of the standardized procedure.

  ***Controlling for variation in testing environments.*** Unlike Experiment 1, which

was conducted at a single school with dedicated research rooms and minimal distraction,

Experiment 2 was conducted in nine different settings. Therefore the analysis reported in

the main text controlled for the effect of testing environment (i.e., school) on our

dependent variable. As the Supporting Table 1 below makes clear, the effect of school on

helping behavior is significant or marginally significant for almost every possible

combination of analytical decisions, suggesting that the variation in testing environments

did indeed introduce error into our measurement.

*Supporting Table 1.* *Primary analysis reported in the main text.

| Exclusions | Controlling for school? | N | Condition p-value | School p-value |
|---|---|---|---|---|
| No exclusions | No | 96 | .13 | N/A |
| Experimenter error only | No | 94 | .09 | N/A |
| Experimenter error, procedure disruption | No | 93 | .06 | N/A |
| Experimenter error, procedure disruption, and distraction during DVs | No | 92 | .028 | N/A |
| No exclusions | Yes | 96 | .10 | .10 |
| Experimenter error only | Yes | 94 | .07 | .12 |
| Experimenter error, procedure disruption | Yes | 93 | .047 | .08 |
| Experimenter error, procedure disruption, and distraction during DVs* | Yes | 92 | .015 | .026 |

# References

Campbell, D. (1963). Social attitudes and other acquired behavioral dispositions. In S.

Koch (Ed.), *Psychology: A study of a science* (Vol. 6, pp. 94-172). New York,

NY: McGraw-Hill.