# Supplementary Material

**Details on exclusion criteria and decisions**

Subjects who did not directly witness the recalled conflict (e.g., conflicts that took place via phone or internet) or who witnessed the conflict at a distance that discouraged or prevented them from intervening (e.g., cases of road rage or overhearing an argument in someone else’s house) were excluded. Subjects reporting conflicts involving guns were also excluded because the steep physical costs present in such a situation to deter intervention are likely vastly greater than situations not involving guns, and we wanted to ensure that the effects of WTR and the likelihood of intervening in everyday situations were not washed out by these events, and we didn’t have enough cases to be able to generalize to situations involving weapons (but see Phillips & Cooney, 2005). We also excluded subjects who reported conflicts involving multiple transgressors because they introduce complex group dynamics that are beyond the scope of the present paper (i.e., focusing on situations analogous to the third-party punishment game). In addition, we excluded subjects who reported parent/child conflicts (e.g., scolding) due to social norms regarding both parenting styles and interfering with another’s parenting, and we excluded conflicts arising during participation in a sport because of the team dynamics involved. Finally, we excluded subjects who reported conflicts in which transgressors and victims had asymmetric institutional power (e.g., boss and employee, professor and student) because of the strong inherent disincentives for subjects to intervene in such situations.

We did not code responses for those subjects that were excluded a priori on the basis of most of the exclusion criteria outlined above. Specifically, the following types of cases were not coded:

- physical barriers that would have made intervention impossible (32)

- conflicts involving a gun (3)

- reported "conflicts" that from the descriptions were actually jokes (4) or not conflicts (12)

- conflicts involving more than one attacker and/or victim (36)

- conflicts that were not witnessed firsthand (23)

- conflicts that were not single events (e.g., "my friend used to get bullied a lot"; 25)

- conflicts between parents/children (7)

- conflict in which a disputant was a police officer (1)

- conflicts in which the subject was either the attacker or the victim (54)

- conflicts that took place during a sports match (19)

- cases where subjects did not respond to the question describing the conflict and/or their own actions, or where it was impossible for the coders to discern from the description what had actually occurred (132)

However, we did code responses for cases involving bosses and teachers/professors before we decided to exclude them (after coding but prior to initial data analyses) due to their inherent power asymmetries

- boss (35)

- teacher/professor (6; but one of these cases [who "did nothing"] did not have a valid wtr\_transgressor, so it will be excluded from model containing that term).

The analyses reported in Tables S9 and S10 reinclude these 41 cases. There were 35 cases of doing nothing, 5 interventions, and one punishment. There were no substantive differences between these models and the ones reported in the main text.

**Demographic predictors of exclusion**

Additionally, we checked whether any of the demographic variables we had available (age, sex, and dataset) predicted meeting exclusion criteria using 3 logistic regression models predicting being excluded.

Sex (dummy coded as 1 = male) did not predict meeting exclusion criteria, b = .04, OR = 1.04, p = .750. Age did predict meeting exclusion criteria, such that older subjects were slightly more likely to be excluded, b = .03, OR = 1.03, p < .001. Finally, dataset also predicted meeting exclusion criteria, with both US students (b = 1.65, OR = 3.91, p < .001) and US MTurkers (b = 1.36, OR = 5.21, p < .001) being significantly more likely to meet exclusion criteria than Japanese students. Additionally, US MTurkers were somewhat more likely to meet exclusion criteria than were US Students, b = .29, OR = 1.33, p = .023.

Table S1. Examples of responses that were coded as “punishment,” “intervention,” or “nothing.”

|  |  |  |
| --- | --- | --- |
| Code | Conflict | Third Party’s Response |
| Punish | 1. Stranger robbed another stranger | Chased after transgressor |
|  | 2. Stranger attempted sexual assault on friend | Yelled at transgressor |
|  | 3. Stranger insulted another stranger | Insulted transgressor |
|  | 4. Friend cut hair off of stranger | Hit and scolded transgressor |
|  | 5. Stranger pushed friend at a club | Fought the transgressor |
|  |  |  |
| Intervene | 1. Acquaintance drunkenly attacked friend | Broke up confrontation |
|  | 2. Acquaintance started verbal argument with friend | Removed friend from situation |
|  | 3. Fistfight between acquaintance and friend | Called the police |
|  | 4. Argument between acquaintance and friend over a romantic partner | Calmed the situation down by facilitating discussion |
|  | 5. Stranger insulting another stranger | Verbally stood up for victim |
|  |  |  |
| Nothing | 1. Two strangers in a gang-related fight | Took no action |
|  | 2. Stranger insulted friend | Consoled friend afterward |
|  | 3. Witnessed stranger tackle another stranger in cafeteria fight | Took no action |
|  | 4. Witnessed stranger attack a homeless person | Took no action |
|  | 5. Witnessed stranger mug stranger in parking lot | Asked if victim was okay afterward |

Note. The wording of these reported conflicts and responses has been edited from subject’s exact responses to preserve privacy—subjects were told on the consent form for the experiment that their answers would not be directly quoted in any public dissemination. The edited examples preserve the nature of the situation and response. The majority of responses coded as “nothing” were self-reported by the subject as taking no action. Some cases were coded as “nothing” despite subjects saying they took action, such as the two examples in the table that did not meet our definitions of intervention or punishment for the current paper. “Stranger,” “acquaintance,” and “friend” labels in the table refer to the subject’s relationship with the person involved.

Table S2. Full Ordinal Logistic Regression Model for Intervention and Punishment

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter | b | 95% CI | OR | p |  |
| Intercept 1 | -1.96 | [-2.43, -1.49] | 0.14 | < .001 |  |
| Intercept 2 | -5.07 | [-5.66, -4.48] | 0.01 | < .001 |  |
| WTRvictim | 1.05 | [0.41, 1.68] | 2.85 | .001 |  |
| WTRtransgressor | -0.78 | [-3.12, 1.56] | 0.46 | .514 |  |
| US MTurkers | 0.39 | [-0.28, 1.06] | 1.48 | .256 |  |
| US Students | 1.12 | [0.50, 1.74] | 3.07 | < .001 |  |
| WTRvictim\*Transgressor WTR | 0.44 | [-2.19, 3.06] | 1.55 | .745 |  |
| WTRvictim\*US MTurkers | 0.65 | [-0.23, 1.53] | 1.92 | .148 |  |
| WTRvictim\*US Students | 0.00 | [-0.82, 0.82] | 0.99 | .999 |  |
| WTRtransgressor\*US MTurkers | 1.37 | [-1.33, 4.06] | 3.92 | .320 |  |
| WTRtransgressor\*US Students | 0.62 | [-2.11, 3.34] | 1.85 | .659 |  |
| WTRvictim\*WTRtransgressor\*US MTurkers | -1.25 | [-4.26, 1.77] | 0.29 | .418 |  |
| WTRvictim\*WTRtransgressor\*US Students | -0.75 | [-3.84, 2.34] | 0.47 | .633 |  |

Note. WTR = welfare trade-off ratio. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. Reference group = JPN Students.

Table S3. Ordinal Logistic Regression Model for Intervention and Punishment (interactions removed)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter | b | 95% CI | OR | p |  |
| Intercept 1 | -2.09 | [-2.41, -1.76] | 0.12 | < .001 |  |
| Intercept 2 | -5.19 | [-5.68, -4.70] | 0.01 | < .001 |  |
| WTRvictim | 1.20 | [.90, 1.50] | 3.33 | < .001 |  |
| WTRtransgressor | -0.26 | [-.67, .15] | 0.77 | .211 |  |
| US MTurkers | 0.89 | [.54, 1.24] | 2.44 | < .001 |  |
| US Students | 1.12 | [.79, 1.45] | 3.07 | < .001 |  |

Note. WTR = welfare trade-off ratio. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. Reference group = JPN Students.

Table S4. Ordinal Logistic Regression for Intervention and Punishment with Conflict Type Added

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | b | 95% CI | OR | p |
| Intercept 1 | -2.31 | [-2.69, -1.93] | 0.10 | < .001 |
| Intercept 2 | -5.38 | [-5.90, -4.86] | 0.00 | < .001 |
| WTRvictim | 1.17 | [0.87, 1.47] | 3.22 | < .001 |
| US MTurkers | 1.02 | [0.61, 1.45] | 2.80 | < .001 |
| US Students | 1.27 | [0.86, 1.68] | 3.55 | < .001 |
| Physical | 0.60 | [0.04, 1.15] | 1.81 | .037 |
| US MTurkers\*Physical | -0.34 | [-1.12, 0.44[ | 0.71 | .390 |
| US Students\*Physical | -0.36 | [-1.05, 0.33] | 0.70 | .310 |

Note. WTR = welfare trade-off ratio. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. Physical is a dummy code for conflict type (1 = physical; 0 = verbal). Reference Group = JPN Students

Table S5. Ordinal Logistic Regression for Intervention and Punishment with Social Norm Violation Added (JPN Students and US MTurkers only)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | (1) | | | |  | (2) | | | |
| Parameter | b | 95% CI | OR | p |  | b | 95% CI | OR | p |
| Intercept 1 | -2.40 | [-2.86, -1.94] | 0.09 | < .001 |  | -2.36 | [-2.77, -1.97] | 0.09 | < .001 |
| Intercept 2 | -6.01 | [-6.82, -5.20] | 0.00 | < .001 |  | -5.98 | [-6.76, -5.21] | 0.00 | < .001 |
| WTRvictim | 1.23 | [0.84, 1.62] | 3.43 | < .001 |  | 1.24 | [0.85, 1.62] | 3.44 | < .001 |
| US MTurkers | 0.90 | [0.29, 1.50] | 2.45 | .003 |  | 0.84 | [0.48, 1.19] | 2.31 | < .001 |
| Social Norm | 0.45 | [-0.10, 1.00] | 1.56 | .113 |  | 0.40 | [0.02, 0.78] | 1.49 | .037 |
| US MTurkers\*Social Norm | -0.09 | [-0.83, 0.65] | 0.91 | .810 |  | - | - | - | - |

Note. WTR = welfare trade-off ratio. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. Social Norm is a dummy code for a social norm violation (1 = social norm violated; 0 = social norm not violated). Reference Group = JPN Students. Model 2 dropped the nonsignificant interaction.

Table S6. Full OLS Regression Model for Anger

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | b | 95% CI | p |  |
| Intercept | 2.54 | [2.31, 2.77] | <.001 |  |
| WTRvictim | 1.31 | [0.94, 1.69] | <.001 |  |
| WTRtransgressor | -2.59 | [-3.60, -1.58] | <.001 |  |
| US MTurkers | 0.05 | [-0.32, 0.42] | .791 |  |
| US Students | -0.58 | [-0.94, -0.22] | .002 |  |
| WTRvictim\* WTRtransgressor | 0.87 | [-0.40, 2.13] | .178 |  |
| WTRvictim\*US MTurkers | -0.14 | [-0.69, 0.41] | .615 |  |
| WTRvictim\*US Students | 0.41 | [-0.11, 0.94] | .123 |  |
| WTRtransgressor\*US MTurkers | 1.02 | [-0.31, 2.36] | .132 |  |
| WTRtransgressor\*US Students | 1.54 | [0.15, 2.93] | .030 |  |
| WTRvictim \* WTRtransgressor\*US MTurkers | -0.85 | [-2.47, 0.77] | .304 |  |
| WTRvictim \* WTRtransgressor\*US Students | -0.48 | [-2.18, 1.22] | .578 |  |

Note. WTR = welfare trade-off ratio. Reference group = JPN Students.

Table S7. OLS Model for Anger with Conflict Type Added

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | b | 95% CI | p |  |
| Intercept | 2.50 | [2.27, 2.73] | <.001 |  |
| WTRvictim | 1.43 | [1.08, 1.77] | <.001 |  |
| WTRtransgressor | -2.01 | [-2.54, -1.47] | <.001 |  |
| US MTurkers | 0.05 | [-0.32, 0.42] | .529 |  |
| US Students | -0.71 | [-1.08, -0.35] | <.001 |  |
| WTRvictim\*US MTurkers | -0.26 | [-0.76, 0.23] | .302 |  |
| WTRvictim\*US Students | 0.39 | [-0.10, 0.87] | .117 |  |
| WTRtransgressor\*US MTurkers | 0.45 | [-0.25, 1.15] | .210 |  |
| WTRtransgressor\*US Students | 1.25 | [0.54, 1.97] | .001 |  |
| Physical | -0.06 | [-0.40, 0.27] | .720 |  |
| Physical\*US MTurkers | -0.05 | [-0.57, -0.47] | .857 |  |
| Physical\*US Students | 0.41 | [-0.04, -0.86] | .073 |  |

Note. WTR = welfare trade-off ratio. Physical is a dummy code for conflict type (1 = physical; 0 = verbal). Reference Group = JPN Students

Table S8. OLS Model for Anger with Social Norm Violation Added (JPN Students and US MTurkers only)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | b | 95% CI | p |  |
| Intercept | 2.05 | [1.81, 2.29] | <.001 |  |
| WTRvictim | 1.07 | [0.73, 1.41] | <.001 |  |
| WTRtransgressor | -1.53 | [-2.06, -1.00] | <.001 |  |
| US MTurkers | -0.31 | [-0.73, 0.10] | .135 |  |
| WTRvictim\*US MTurkers | -0.11 | [-0.60, 0.37] | .648 |  |
| WTRtransgressor\*US MTurkers | 0.20 | [-0.49, 0.88] | .570 |  |
| Social Norm | 1.06 | [0.75, 1.36] | <.001 |  |
| Social Norm\*US MTurkers | 0.25 | [-0.20, 0.70] | .277 |  |

Note. WTR = welfare trade-off ratio. Social Norm is a dummy code for a social norm violation (1 = social norm violated; 0 = social norm not violated). Reference Group = JPN Students

Table S9. Ordinal Logistic Regression for Intervention and Punishment with Excluded Subjects Reincluded

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | b | 95% CI | OR | p |
| Intercept 1 | -2.17 | [-2.48, -1.85] | 0.11 | < .001 |
| Intercept 2 | -5.24 | [-5.71, -4.76] | 0.01 | < .001 |
| WTRvictim | 1.16 | [0.87, 1.45] | 3.20 | < .001 |
| US MTurkers | 0.93 | [0.59, 1.27] | 2.53 | < .001 |
| US Students | 1.20 | [0.88, 1.52] | 3.32 | < .001 |

Note. Model includes 41 additional cases that were excluded from the main text because they were from conflicts containing either a boss or a teacher/professor. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. WTRvictim (welfare trade-off ratio toward the victim) is a continuous predictor ranging from 0-1.13; sample variables are dummy codes. Reference group = JPN Students

Table S10. OLS Model for Anger with Excluded Subjects Reincluded.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Reference Group = JPN Students | | | | |
| Parameter | b | 95% CI | p |  |
| Intercept | 2.51 | [2.30, 2.72] | <.001 |  |
| WTRvictim | 1.45 | [1.12, 1.78] | <.001 |  |
| US MTurkers | 0.05 | [-0.28, 0.39] | .749 |  |
| US Students | -0.59 | [-0.92, -0.25] | .001 |  |
| JPN Students | - | - | - |  |
| WTRtransgressor | -2.12 | [-2.63, -1.60] | <.001 |  |
| WTRvictim \*US MTurkers | -0.22 | [-0.70, 0.26] | .368 |  |
| Transgressor WTR\*US MTurkers | 0.57 | [-0.11, 1.25] | .103 |  |
| WTRvictim \*US Students | 0.33 | [-0.14, 0.80] | .170 |  |
| WTRtransgressor \*US Students | 1.35 | [0.64, 2.05] | <.001 |  |
| WTRvictim \*JPN Students | - | - | - |  |
| WTRtransgressor \*JPN Students | - | - | - |  |

Note. Model includes 40 additional cases that were excluded from the main text because they were from conflicts containing either a boss or a teacher/professor. WTR = welfare trade-off ratio; WTRvictim and WTRtransgressor are continuous predictors ranging from 0-1.13; sample variables are dummy codes. The two models are identical but recoded with different reference groups.

A screenshot of a cell phone

Description automatically generated

Figure S1. WTR (welfare trade-off ratio) toward the victim as a function of relationship category, broken apart by sample. Boxes represent the inner quartile range (IQR), whiskers extend to the furthest values within 1.5\*IQR, and the horizontal lines correspond to the median. Note that US students were not given the option of selecting the “enemy or rival” category.

A picture containing photo, wall, showing

Description automatically generated

Figure S2. WTR (welfare trade-off ratio) toward the transgressor as a function of relationship category, broken apart by sample. Boxes represent the inner quartile range (IQR), whiskers extend to the furthest values within 1.5\*IQR, dots correspond to values beyond 1.5\*IQR, and the horizontal lines correspond to the median. Note that US students were not given the option of selecting the “enemy or rival” category.

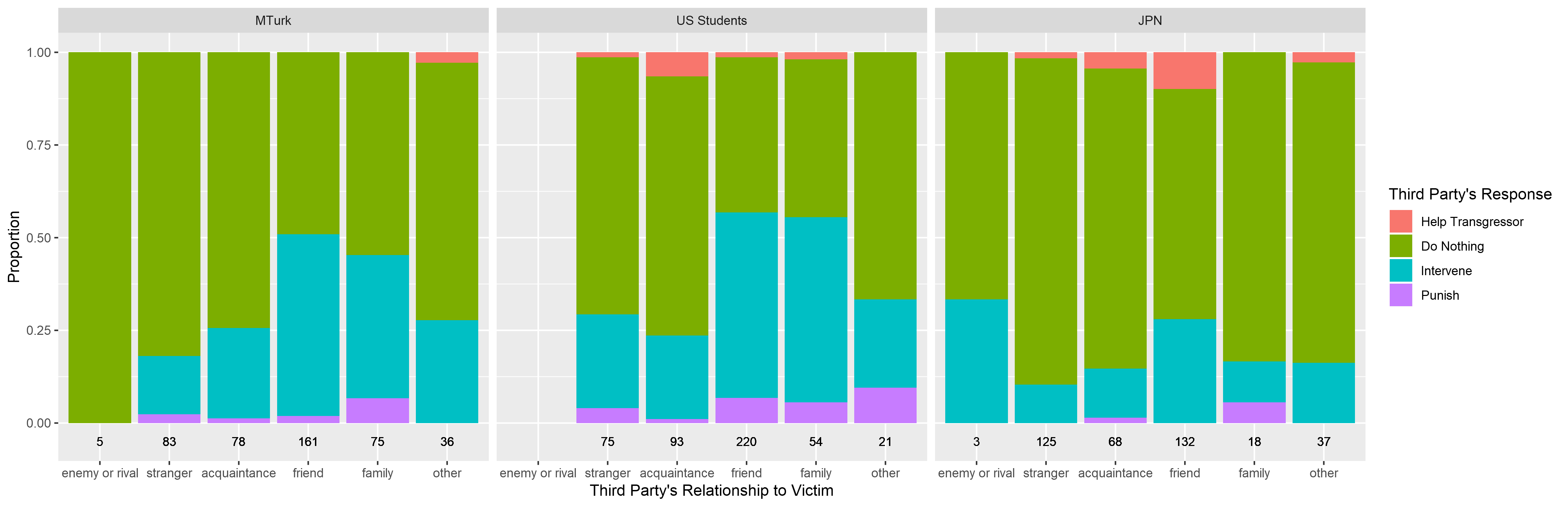


Figure S3. Third party’s response as a function of relationship category of victim, broken apart by sample. Colored bars represent the proportion of responses within a category of relationship. Numbers on the x-axis are the cell counts for each category. In all analyses reported in the paper, helping the transgressor was categorized as doing nothing. Note that US students were not given the option of selecting the “enemy or rival” category.

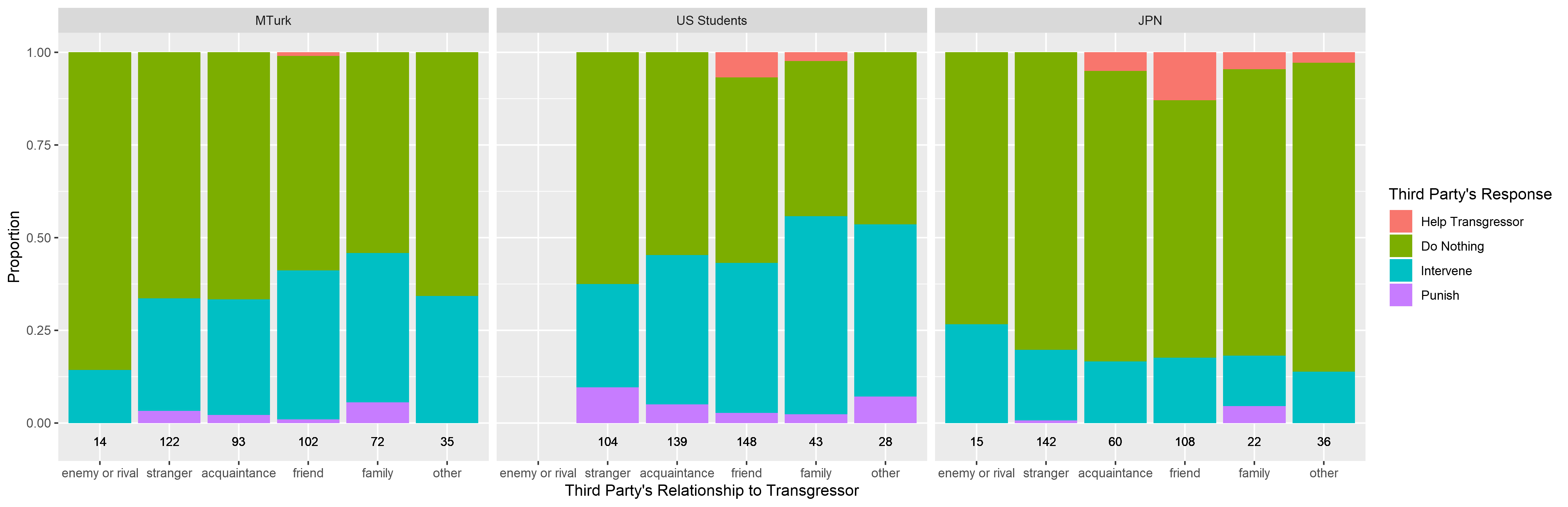


Figure S4. Third party’s response as a function of relationship category of transgressor, broken apart by sample. Colored bars represent the proportion of responses within a category of relationship. Numbers on the x-axis are the cell counts for each category. In all analyses reported in the paper, helping the transgressor was categorized as doing nothing. Note that US students were not given the option of selecting the “enemy or rival” category.