**Online appendix**

We also performed the main analyses with only paternal reports of child disruptive behavior instead of combining maternal and paternal reports. This allowed us to ensure that the results were consistent even if we completely ruled-out the possibility of shared rater variance between ratings of maternal negativity and of child disruptive behavior.

The cross-lagged model (see Figure A1) showed excellent fit, χ2 (23) = 33.02, *p* < .05, CFI = 1.00, TLI = 1.00, RMSEA = .02. With respect to hypothesis 1, by looking at the within family component of the model shown in Figure A1, it is possible to see that children’s disruptive behavior (T1) predicts higher maternal negativity at T2 and maternal negativity at T2 in turn predicts more child disruptive behavior at T3, showing that hypothesis 1 was confirmed. Our second hypothesis, where children’s disruptive behavior contributed to itself via maternal negativity, was also confirmed. The completely standardized effect was significant, β = .05, SE = .02, *p* < .05, indicating that children’s disruptive behavior at T1 predicted higher disruptive behavior at T3 via maternal negativity at T2. Thus, results were identical to those using a combination of maternal and paternal reports of child disruptive behavior.

Third, by looking at Figure 1 it is possible to see that one between-family cross-lag from the family average of children’s disruptive behavior at T2 to the family average for maternal negativity at T3 was significant. None of the other cross-lags were significant. In order to test for a context effect the difference in the between and within family cross-lag parameter estimates was examined. The between-family coefficient was not significantly larger that its within-family counterpart, Wald’s test (1) = 1.47, *p* > .05. Consistent with results based on maternal and paternal reports of child disruptive behavior, there was no support for hypothesis 3 which pertained to the presence of a between-family context effect. Contrary to results using combined maternal and paternal reports of child disruptive behavior, however, the cross-lag paths from sibling disruptive behavior to mother negativity was not significant at the between family level.

With respect to hypothesis 4, we tested the hypothesis that stability coefficients would be stronger between families than within families. Stability in disruptive behavior was not significantly larger at the between-family level than at the within-family level between T1 and T2, Wald’s test (1) = 1.40, *p* > .05, whereas this was significant between T2 to T3, Wald’s test (1) = 2.52, *p* < .05. The stability of maternal negativity was significantly larger at the between-family level compared to the within-family level at T1 and T2, Wald’s test (1) = 43.59, *p* < .05, and between T2 and T3, Wald’s test (1) = 3.08, *p* < .05. Thus, results were largely consistent with hypothesis 4 as well as with results based on maternal and paternal reports, with the exception of the stability in disruptive behavior between T1 and T2 which did not reach significance here.



*Figure A1.* Multilevel cross-lagged model with standardized path estimates and their 95% confidence intervals with paternal reports of child disruptive behavior. Neg. = maternal negativity. Disrupt. = disruptive behavior. Parameters with a full line are significant at *p* < .05. Parameters with a dashed line are not significant at *p* < .05. Parameter estimates for the effects of child age, child sex and maternal education are not shown to reduce visual clutter.