Title:
Effects of Implementing Multiple Components in a School-Wide Anti-Bullying Program: A Cluster-Randomized Controlled Trial in Elementary Schools

Hypotheses:
H1: Stronger decrease in bullying and victimization in PRIMA-L+ schools compared to PRIMA-L- schools.
H2: Stronger decrease in bullying and victimization when teachers implemented more universal program components

Analysis:
First, we conducted a power analysis based on the number of victims of bullying as an outcome measure. Since the prevalence of bullying in grades 3-6 varies between 21% and 35% in western countries (Chester et al., 2015; Modecki et al., 2014; National Center for Educational Statistics, 2019), we estimated that a minimum of 33 classes per condition was needed with a minimum of 25 students per class (assuming a response of 80%; alpha = .05, two-sided, power = .80, ICC = .032) to demonstrate a decrease of 30% of victims (i.e., from 25% to 17.5%) between the two experimental groups and the control group. With this sample size, even a small effect (Cohen's *d* = 0.20) can be demonstrated for primary and secondary measures with adequate power.

We analyzed the dichotomized data with multilevel logistic regression models (GLMM in SPSS, Version 25). Three-level hierarchical models were fitted, representing students nested in classrooms, and classrooms nested within schools. We controlled for differences in baseline levels by adding the pretest scores of the variable of interest to the models. Also, ethnicity (i.e., western or non-western), gender, and age (grand-mean centered) were included in all models, as these are well-known covariates (see Salmivalli & Voeten, 2004; Vervoort et al., 2010). We explored possible interaction effects of both PRIMA conditions with ethnicity, gender, or age. Finally, we controlled for differences across conditions on school size, urbanization level, and the number of students with special needs with dummy-coded school-level variables, distinguishing between large schools (> 500 students), urban schools (large and medium cities), and high level of students with special needs (schools above the national average of 9.31% students with special needs). Students with special needs have learning difficulties and, or, emotional-behavioral problems without an indicated disability or health care need (Smeets, et al., 2007).

To investigate the effects of both PRIMA conditions on bullying behavior and students' roles in bullying situations, we tested a model including all 3,155 students (i.e., intention-to-treat analysis). This analysis provides an estimate of the program effects in general school practices with varying levels of program implementation to establish typical effects in educational practice. We also analyzed the data with the same statistical model with the inclusion of only those students who were sufficiently exposed to at least one of the universal PRIMA

components, in order to estimate the maximum effects of the program when it was adequately implemented (i.e., a received-intervention analysis). To investigate whether classes that implemented one, two, three, or four universal program components are more effective in reducing bullying, we compared these subgroups with classes that implemented zero universal program components, including the control group.

Dependent variables:

- *Self-reported victims*: We used the global item from the revised Olweus Bully/Victim Questionnaire (OBVQ, Olweus, 1996) to measure self-reported victimization: "How often have you been bullied at school in the last couple of months?". Students answered on a 5-point scale (0 = *not at all*, 1 = *once or twice*, 2 = *two or three times a month*, 3 = about once a week, 4 = *several times a week*).
- *Self-reported bullies*: Self-reported bullying was measured by asking students whether they had engaged in a series of behaviors that are often associated with bullying in the last couple of months. Students responded to items on a 5-point scale (0 = *not at all*, 1 = *once or twice*, 2 = *two or three times a month*, 3 = about once a week, 4 = *several times a week*). These eight items were based on the OBVQ (Olweus, 1996) and had an internal consistency of $\alpha = .882$ at the pretest.
- *Peer-reported victims, bullies and other roles:* Two single items, based on the Participant Roles Questionnaire (Kärnä et al., 2013; Salmivalli et al., 1996), were used to identify peer-reported victimization and bullying. Students were asked to nominate students who were being bullied in the past couple of months from a list of classmates: *"Which children are being bullied by other children?",* and to nominate students who bullied other children: *"Which classmates bully other children?".* Students could nominate an unlimited number of classmates for each item or nominate no one. To prevent a systematic nomination bias of classmates that are on top of the list, the order of names was randomized. Also based on the Participant Roles Questionnaire (Kärna et al., 2013; Salmivalli et al., 1996), three single items were used to identify students' participant roles in bullying situations concerning the past couple of months; reinforcers of bullies: *"Which classmates reinforce bullies, for example, by laughing or giggling when someone gets bullied?"*; outsider: *"Which classmates do nothing when someone gets bullied, for example, they walk away or act like they did not see the bullying?"*; and defenders of victims: *"Which classmates help children that are being bullied, for example, by comforting, supporting, or defending them?".* Similar to the procedure for peer-reported bullies and victims, students could nominate an unlimited number of classmates or no one. The list of names was randomized, and the relative criterion was applied to assign students to one or more participant roles.

Data from both self-reported victimization and bullying-related behaviors were severely skewed, with most children not being victimized or involved in bullying-related behaviors at the pre- and posttest and a smaller number of students who were bullies or victims. Therefore, we

dichotomized these variables by classifying students as self-reported victims using the standard cut-off *'at least two or three times a month'* (Solberg & Olweus, 2003). Students were classified as self-reported bullies when they had a score equal or higher than 2 (i.e., a student indicates to be bullying at least twice a month on at least one of the eight items, or to be bullying once a month on at least two of the eight items). Adding the variables as continuous variables in the models did not produce significant differences in the results compared to the dichotomous variables.

The relative criterion was used to assign students to roles by standardizing received peer nominations within classrooms. For each class, z-scores were calculated based on the number of nominations received from classmates. Subsequently, students who scored above the class average were assigned to the corresponding role (0 = *not assigned;* 1 = *assigned*). This procedure allowed students to be assigned to a single role, multiple roles, or no role at all. In the latter case, students were 'unclassified'.

Background variables:

Students filled out questions on their date of birth, gender, grade level, and ethnicity. Ethnicity was measured by asking what the student considered his or her background with the possibility to tick multiple boxes (e.g., Dutch and Moroccan). We then dichotomized students into 'western' or 'non-western' background based on the criteria of the Dutch Central Statistical Office (CBS, n.d.).

Stacking of program components:

In order to investigate the effects of stacking components, we calculated and dichotomized the implementation level for each program component. First, teachers were asked to indicate the degree to which they implemented each part (e.g., lesson 1, lesson 2, etc.) of each PRIMA component (e.g., student lessons) separately on a four-point scale: 0 = *not at all;* 1 = *less than 50%;* 2 = *more than 50%;* 3 = *completely.* We subsequently dichotomized the scores to indicate whether students (or their teachers) were sufficiently exposed to each of the program components. Regarding the universal program components, we considered an implementation of at least 50% of the components to be a successful implementation of student lessons, e-learning, and the monitor report. The face-to-face training was considered to be completed when teachers indicated that they attended the full training session. Concerning the selective components, the protocols for specific bullying situations and the protocols for students directly involved were considered to be used when teachers indicated to have consulted at least one of the protocols for both types of protocols separately.

We determined the level of implementation for the universal components of the PRIMA program by adding the dichotomized variables of student lessons, monitor report, e-learning, and face-to-face training together into an aggregated implementation score. This resulted in the following scores: 0 = *no components implemented;* 1 = *one component;* 2 = *two components;* 3 = *three components,* or 4 = *four components.* We included these components as they are universal;

the use of the selective components (i.e., protocols resulting from the monitor and the protocols for specific situations) heavily depends on specific bullying incidents that may occur at school. Therefore, the implementation of selective components is highly context-specific, and its interpretation is, therefore, less straightforward.

Tables and figures: