

Appendix A

Instrumenten

Demografische gegevens

Demografische gegevens van leerlingen werd verzameld voor de start van het programma (T1), waaronder geboortegeslacht, gender, leeftijd, etnische identiteit, en groep/leerjaar. Geboortegeslacht werd beoordeeld door te vragen of leerlingen met het lichaam van een jongen of een meisje waren geboren, wat in drie categorieën werd onderscheiden: jongen (1), meisje (2) en anders/weet ik niet (3). Gender werd beoordeeld door te vragen hoe leerlingen zich voelden, welke we hebben gecategoriseerd in drie groepen: jongen (1), meisje (2) en anders (3, "een jongen 'en een meisje", "geen jongen en ook geen meisje", "anders", "dat weet ik nog niet", of "dit wil ik niet zeggen"). Leeftijd werd gerapporteerd in jaren (continue variabele). We beoordeelden etnische identiteit door te vragen met welke identiteit de leerlingen zich het meest verbonden voelden (ik zie mezelf als: "Nederlands, Indonesisch, Turks, Marokkaans, Surinaams, Antilliaans, anders, of een combinatie") en onderscheidde twee groepen: Westers (1; bijv. Nederlands) of mix Westers/niet-Westers en niet-Westers (2; bijv. Marokkaans of Nederlands-Turks). Groep/leerjaar werd gerapporteerd als groep 5, 6, 7 of 8.

Demografische gegevens van leerkrachten werd verzameld voor de start van het programma (T1), waaronder geboortegeslacht, gender, leeftijd, etnische identiteit en groep/leerjaar (waar ze les in geven). Geboortegeslacht werd beoordeeld door te vragen wat het geboortegeslacht van de leerkracht was, en werd onderscheiden in drie categorieën: man (1), vrouw (2) en anders/wil ik liever niet zeggen (3). Gender werd beoordeeld door te vragen hoe leerkrachten zich voelden, wat we hebben gecategoriseerd in drie groepen: man (1), vrouw (2), en anders (3, "een man en een vrouw", "geen man en geen vrouw", "anders", "dit

weet ik niet”, of “wil ik liever niet zeggen”). Voor de regressieanalyses is alleen man en vrouw gebruikt. Leeftijd werd gerapporteerd in jaren (continue variabele). We beoordeelden etnische identiteit door te vragen met welke identiteit de leerlingen zich het meest verbonden voelden (ik zie mezelf als: “Nederlands, Indonesisch, Turks, Marokkaans, Surinaams, Antilliaans, anders, of een combinatie”) en onderscheidde twee groepen: Westers (1; bijv. Nederlands) of mix Westers/niet-Westers en niet-Westers (2; bijv. Marokkaans of Nederlands-Turks). Groep/leerjaar waar ze les geven werd gerapporteerd als groep 5-6, 7-8 of anders (bijvoorbeeld groepen 5-8 of 6-7).

Primaire uitkomsten

Prosociaal gedrag werd gemeten met verschillende instrumenten die verschillende domeinen meten. Individueel prosociaal gedrag werd gemeten met de subschaal prosociaal gedrag van de *Strengths and Difficulties Questionnaire* (SDQ), Nederlandse versie (Van Widenfelt et al., 2003), aangepast voor de huidige studie. Er werden vier items gebruikt (e.g., “Ik probeer aardig te zijn tegen mijn klasgenoten”), beoordeeld op een vierpuntschaal van 1 (*nooit*) tot 4 (*altijd*). De Nederlandse versie van de SDQ had voldoende psychometrische eigenschappen in steekproeven met kinderen en jonge adolescenten, een acceptabele interne consistentie, test-hertest betrouwbaarheid en goede concurrente validiteit, met een Cronbach’s alfa van > 0.60 voor de subschaal prosociaal gedrag (Muris et al., 2003; Van Widenfelt et al., 2003). Gemiddelde scores werden berekend, waarbij hogere scores meer sociaal gedrag weerspiegelden ($\alpha = 0.63\text{-}0.69$ voor T1-T3).

Prosociaal gedrag als een injunctieve norm (i.e., de houding van kinderen ten opzichte van prosociaal gedrag die gemiddeld in de klas wordt aangehouden; Sentse et al., 2015) werd gemeten met drie items speciaal ontwikkeld voor deze studie (i.e., Waardoor kun je populair worden in de klas?: “Door elkaar te helpen”, “Door aardig te zijn tegen elkaar”, of “Door

samen te werken”), beoordeeld op een driepuntschaal van 1 (*niet waar*) tot 3 (*zeker waar*).

Gemiddelde scores werden berekend, waarbij hogere scores een hogere injunctieve norm van de groep ten opzichte van prosociaal gedrag weerspiegelden ($\alpha = 0.71-0.83$ voor T1-T3).

Samenwerking en conflict in de klas werden gemeten met de *Classroom Peer Context Questionnaire* (CPCQ), subschalen samenwerking en conflict, respectievelijk (Boor-Klip et al., 2014, 2016, 2017), beoordeeld op een vijfpuntschaal van 1 (*helemaal niet waar*) tot 5 (*helemaal waar*). De subschaal samenwerking bestaat uit vier items die beoordelen in hoeverre kinderen positief gedrag ervaren in de klas (e.g., “In deze klas helpen kinderen elkaar”). De subschaal conflict bestaat uit vier items die beoordelen in hoeverre kinderen negatief gedrag ervaren in de klas. Voor de huidige studie werden twee items gebruikt (e.g., “In deze klas zijn kinderen onaardig tegen elkaar”). De CPCQ had adequate psychometrische eigenschappen in een steekproef van kinderen – zoals betrouwbaarheid, construct- en concurrente validiteit, en stabiliteit – met een Cronbach’s alfa van > 0.79 voor beide subschalen (Boor-Klip et al., 2014, 2016, 2017). Gemiddelde scores werden berekend, waarbij hogere scores meer samenwerking in de klas ($\alpha = 0.74-0.81$ voor T1-T3) en meer conflicten in de klas ($\alpha = 0.71-0.76$ voor T1-T3) weerspiegelden.

Intrinsieke (prosociale) motivatie werd gemeten met vier items (e.g., “Ik vind het belangrijk om iemand te helpen”), gebaseerd op onderzoek naar sociale doelen (Rudolph et al., 2011), beoordeeld op een driepuntschaal van 1 (*niet waar*) tot 3 (*zeker waar*). Gemiddelde scores werden berekend, waarbij hogere scores meer intrinsieke motivatie voor prosociaal gedrag weerspiegelden ($\alpha = 0.65-0.71$ voor T1-T3).

Autonomie werd gemeten met items gebaseerd op onderzoek naar de Vreedzame School (Battistich et al., 1997; Lickona & Davidson, 2003; Pauw, 2013) en de *Perceived Choice and Awareness Scale* (Sheldon et al., 1996), die respectievelijk sociale

verantwoordelijkheid en autonomie maten. Er werden vijf items gebruikt (e.g., “In deze klas helpen kinderen bij het maken van groepsafspraken” of “In deze klas kun je kiezen hoe je iets wilt doen”), beoordeeld op een vierpuntschaal van 1 (*nooit*) tot 4 (*altijd*). Gemiddelde scores werden berekend, waarbij hogere scores meer autonomie weerspiegelden ($\alpha = 0.67\text{-}0.74$ voor T1-T3).

Competentie werd gemeten met vijf items (e.g., “Ik denk positief over mezelf”) gebaseerd op de *Rosenberg Self-Esteem Scale* (RSES), Nederlandse versie (Franck et al., 2008; Rosenberg, 1965), beoordeeld op een driepuntschaal van 1 (*niet waar*) tot 3 (*zeker waar*). De Nederlandse RSES toonde een hoge interne consistentie aan bij volwassenen (Franck et al., 2008) en adequate psychometrische eigenschappen bij adolescenten, zoals constructvaliditeit (Bagley, 1997; Cong & Cheong, 2022). Gemiddelde scores werden berekend, waarbij hogere scores meer competentie weerspiegelden ($\alpha = 0.77\text{-}0.81$ voor T1-T3).

Verbondenheid werd gemeten met items van de *Classroom Peer Context Questionnaire* (CPCQ), subschalen comfort and cohesie (Boor-Klip et al., 2014, 2017), beoordeeld op een vijfpuntschaal van 1 (*helemaal niet waar*) tot 5 (*helemaal waar*). De subschaal comfort bestaat uit vier items die beoordelen in hoeverre kinderen zich op hun gemak voelen in de klas (e.g., “In deze klas kan ik mezelf zijn”). De subschaal cohesie bestaat uit vier items die beoordelen in hoeverre kinderen in de klas tijd met elkaar doorbrengen. Voor de huidige studie werden twee items van deze subschaal gebruikt (e.g., “In deze klas hoort iedereen bij de groep”). De CPCQ had adequate psychometrische eigenschappen – zoals betrouwbaarheid, construct- en concurrente validiteit, en stabiliteit – in een steekproef van kinderen, met een Cronbach’s alfa van > 0.68 voor beide subschalen (Boor-Klip et al., 2014). Gemiddelde scores werden berekend, waarbij hogere scores een grotere mate van verbondenheid weerspiegelden ($\alpha = 0.87\text{-}0.89$ voor T1-T3).

Secondaire uitkomsten

Schoolwelzijn werd gemeten met zeven items speciaal voor deze studie ontwikkeld, gebaseerd op onderzoek naar KiVa (Kärnä et al., 2011; Kaufman et al., 2021). Deze items maten gevoelens van plezier en verbondenheid met de klas en school (e.g., “Ik ga graag naar school”), beoordeeld op een vierpuntschaal van 1 (*nooit*) tot 4 (*altijd*). Deze items zijn gebruikt in eerder onderzoek naar de effectiviteit van KiVa in Nederland, waar een hoge betrouwbaarheid werd gerapporteerd ($\alpha = 0.85$; Kaufman et al., 2021). Gemiddelde scores werden berekend, waarbij hogere scores een positiever schoolwelzijn weerspiegelden ($\alpha = 0.90-0.91$ voor T1-T3).

Veiligheid op school werd gemeten met een item speciaal ontwikkeld voor deze studie, dat het gevoel van veiligheid op school meet (e.g., “Ik voel me veilig op en rond school”), beoordeeld op een vijfpuntschaal van 1 (*heel erg onveilig*) tot 5 (*heel veilig*). Een hogere score weerspiegelt een hoger gevoel van veiligheid op school.

Antisociaal gedrag bestond uit *pesten* en *slachtofferschap van pesten*, gemeten met items van de *Olweus Bully-Victim Questionnaire*, Nederlandse versie (OBVQ; Olweus, 1996). Er werden twee vragen gesteld met betrekking tot pesten (“Hoe vaak heb je de afgelopen maanden andere kinderen gepest?”) en slachtofferschap van pesten (“Hoe vaak ben je de afgelopen maanden gepest?”). De antwoorden werden beoordeeld op een vijfpuntschaal: Ik heb niet gepest/ik werd niet gepest (1), een of twee keer (2), twee of drie keer per maand (3), ongeveer een keer per week (4), en meerdere keren per week (5). Zeven aanvullende vragen hadden betrekking op slachtofferschap van pesten en vroegen naar gedragingen die verband houden met pesten (e.g., “Ik werd bedreigd of gedwongen dingen te doen die ik niet wilde doen”), beoordeeld op een vijfpuntschaal van 1 (*helemaal niet*) tot 5 (*meerdere keren per week*). De OBVQ toonde adequate psychometrische eigenschappen in

steekproeven met kinderen – zoals betrouwbaarheid, construct- en concurrente validiteit – met een Cronbach's alfa boven de 0.80 (Gonçalves et al., 2016; Khawar, 2015; Kyriakides et al., 2006). Omdat we algemene afname van pestgedrag wilden onderzoeken, gebruikten we de maximumscore die op de specifieke items werd gegeven (Huitsing et al., 2020). Hogere scores weerspiegelen meer pestgedrag (individueel item) en slachtofferschap van pesten (individueel item en maximum score van 7 items: $\alpha = 0.89-0.88$ voor T1-T3).

Perceptie van leerlingen over de leerkrachthouding ten opzichte van pesten werd gemeten met items gebaseerd op onderzoek naar KiVa (Veenstra et al., 2014). Drie items hadden betrekking op het bespreken van pesten ("Heeft je leerkracht het gehad over pesten tijdens de les?", met antwoordopties: nee, één keer, 2-4 keer, 5-8 keer, meer dan 8 keer), houding ten opzichte van pesten ("Wat vind je leerkracht van pesten?", met antwoordopties: het is iets goeds, het maakt niet uit, ik weet het niet, het is iets slechts, het is iets heel erg slechts), en effectiviteit om pesten te verminderen ("Wat kan je leerkracht doen om pesten te stoppen?" met antwoordopties: niets, erg weinig, een beetje, veel, heel veel). Een gemiddelde score werd berekend, waarbij hogere scores een positievere houding van leerkrachten tegenover pesten weerspiegelden ($\alpha = 0.35-0.42$ voor T1-T3). Omdat de totale schaal een zeer lage betrouwbaarheid had, hebben we de analyses per item uitgevoerd.

Leerkrachthouding ten opzichte van de stimulering van het vervullen van basispsychologische behoeften werd door leerkrachten beoordeeld, met specifieke vragen specifiek voor deze studie ontwikkeld die betrekking hadden op de belangrijkste componenten van SterkWerk. Aan leerkrachten werd gevraagd wat zij belangrijk vinden in hun werk als leerkracht om te stimuleren, of zij dit daadwerkelijk stimuleren, of zij hier goed in zijn en of zij er kennis over hebben. Er werden vier subschalen gecreëerd, voor prosociaal gedrag (7 items, bijv. "Ik vind het belangrijk om als leerkracht te stimuleren dat kinderen elkaar helpen"; $\alpha = .80-.85$ voor T1-T3), autonomie (5 items; bijv. "Ik ben er goed in om

kinderen verantwoordelijkheid te geven”; $\alpha = .77-.78$ voor T1-T3), competentie (4 items, bijv. “Ik zorg ervoor dat kinderen voelen dat ze ergens goed in zijn”; $\alpha = .82-.83$ voor T1-T3), en verbondenheid (7 items, bijv. “Ik weet hoe ik een veilige sfeer kan bieden in mijn klas, zodat iedereen erbij hoort”; $\alpha = .84-.86$ voor T1-T3). Middels een factoranalyse (PAF, *principal axis factoring*) hebben we onderzocht in hoeverre de vragenlijst inderdaad vier aparte subschalen weerspiegelde. De KMO (Kaiser-Meyer-Olkin) was $> .80$. en de *Bartlett's test of sphericity* was significant ($p < .001$), en factor 1 verklaarde tussen de 50% en 67% van de variantie voor alle subschalen. PCA (*principal component analysis*) liet hetzelfde patroon zien. Omdat de losse subschalen $> 50\%$ van de variantie verklaarden, de subschalen een hoge interne consistentie hadden, en alle *factor loadings* voldoende waren ($> .3$) bleek er goede ondersteuning voor de aparte subschalen voor dit onderzoek. Leerkrachten konden een cijfer geven van 1-10. Hogere scores weerspiegelen een positieve houding voor het stimuleren van pro sociaal gedrag en basis psychologische behoeften (autonomie, competentie, en verbondenheid).

Burgerschapsvaardigheden van leerlingen werden beoordeeld door leerkrachten, met specifieke vragen specifiek ontwikkeld voor deze studie, gebaseerd op de Vragenlijst Burgerschap Meten ((Daas et al., 2019). Vier domeinen werden beoordeeld door leerkrachten door te vragen hoe leerlingen in hun klas: 1) democratisch handelen (bijvoorbeeld “beslissingen accepteren die door de hele klas zijn genomen (meerderheidsbeslissingen)”; 3 items, $\alpha = .70-.74$ voor T1-T3), 2) sociaal verantwoordelijk handelen (bijvoorbeeld “komen voor elkaar op”; 2 items, $\alpha = .54-.63$ voor T1-T3), 3) omgaan met conflicten (bijvoorbeeld: “houden rekening met de andere persoon in een conflict bij het zoeken naar een oplossing”; 2 items, $\alpha = .71-.73$ voor T1-T3), en 4) omgaan met verschillen (bijvoorbeeld “zijn nieuwsgierig naar waar andere leerlingen goed in zijn (kwaliteiten)”; 3 items, $\alpha = .38-.54$ voor T1-T3). De antwoorden werden beoordeeld op een schaal van 1 ((vrijwel) nooit) tot 4

((vrijwel) altijd). Vanwege de lage betrouwbaarheid voor sommige subschalen werd een totale schaal gecreëerd waarbij een hogere score positievere burgerschapsvaardigheden in de klas weerspiegelt (10 items; $\alpha = .81-.85$ voor T1-T3). Een factoranalyse (zowel PAF als PCA) liet ook zien dat de vragenlijst uit één factor bestond (met $> 30\%$ verklaarde variantie).

Referentielijst

- Bagley, C. (1997). Norms and construct validity of the rosenberg self-esteem scale in canadian high school populations: Implications for counselling floyd bolitho lome bertrand. *Canadian Journal of Counselling*, 31, 82–92.
- Battistich, V., Solomon, D., Watson, M., & Schaps, E. (1997). Caring school communities. *Educational Psychologist*, 32(3), 137–151. https://doi.org/10.1207/s15326985ep3203_1
- Boor-Klip, H. J., Segers, E., Hendrickx, M. M. H. G., & Cillessen, A. H. N. (2014). Perceptions of classroom peer context: Associations with social status, academic achievement, and self-esteem | Beleving van de peer context in de klas: Samenhang met sociaal functioneren, academisch functioneren en zelfbeeld. *Pedagogische Studien*, 91(5).
- Boor-Klip, H. J., Segers, E., Hendrickx, M. M. H. G., & Cillessen, A. H. N. (2016). Development and Psychometric Properties of the Classroom Peer Context Questionnaire. *Social Development*, 25(2), 370–389. <https://doi.org/10.1111/sode.12137>
- Boor-Klip, H. J., Segers, E., Hendrickx, M. M. H. G., & Cillissen, A. H. N. (2017). The moderating role of classroom descriptive norms in the association of student behavior with social preference and popularity. *Journal of Early Adolescence*, 37(3), 387–413.
- Cong, C. W., & Cheong, J. Y. (2022). Validation of Rosenberg self-esteem scale for Malaysian adolescents. *Current Psychology*, February, 8–11. <https://doi.org/10.1007/s12144-022-02960-z>
- Daas, R., Dijkstra, A. B., & Kort, I. De. (2019). *Verkenning instrumenten opbrengstgericht burgerschapsonderwijs*.
- Franck, E., De Raedt, R., Barbez, C., & Rosseel, Y. (2008). Psychometric properties of the Dutch Rosenberg Self-Esteem Scale. *Psychologica Belgica*, 48(1), 25–35. <https://doi.org/10.5334/pb-48-1-25>
- Gonçalves, F. G., Heldt, E., Peixoto, B. N., Rodrigues, G. A., Filipetto, M., & Guimarães, L. S. P. (2016). Construct validity and reliability of Olweus Bully/Victim Questionnaire - Brazilian version. *Psicologia: Reflexao e Critica*, 29(1). <https://doi.org/10.1186/s41155-016-0019-7>
- Huitsing, G., Lodder, G. M. A., Browne, W. J., Oldenburg, B., Van der Ploeg, R., & Veenstra, R. (2020). A Large-Scale Replication of the Effectiveness of the KiVa

- Antibullying Program: a Randomized Controlled Trial in the Netherlands. *Prevention Science*, 21(5), 627–638. <https://doi.org/10.1007/s11121-020-01116-4>
- Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Kaljonen, A., & Salmivalli, C. (2011). A Large-Scale Evaluation of the KiVa Antibullying Program: Grades 4-6. *Child Development*, 82(1), 311–330. <https://doi.org/10.1111/j.1467-8624.2010.01557.x>
- Kaufman, T. M. L., Huitsing, G., Bloemberg, R., & Veenstra, R. (2021). The Systematic Application of Network Diagnostics to Monitor and Tackle Bullying and Victimization in Schools. *International Journal of Bullying Prevention*, 3(1), 75–87. <https://doi.org/10.1007/s42380-020-00064-5>
- Khawar, R. (2015). Validation of the Urdu Version of Olweus Bully/Victimization Questionnaire (OBVQ). *Journal of Pakistan Psychiatric Society*, 12(1), 24–28.
- Kyriakides, L., Kaloyirou, C., & Lindsay, G. (2006). An analysis of the Revised Olweus Bully/Victim Questionnaire using the Rasch measurement model. *British Journal of Educational Psychology*, 76(4), 781–801. <https://doi.org/10.1348/000709905X53499>
- Lickona, D., & Davidson, M. (2003). School as a caring community profile-II. *Unpublished Manuscript*.
- Muris, P., Meesters, C., & Van den Berg, F. (2003). The Strengths and Difficulties Questionnaire (SDQ) further evidence for its reliability and validity in a community sample of Dutch children and adolescents. *European Child and Adolescent Psychiatry*, 12(1), 1–8. <https://doi.org/10.1007/s00787-003-0298-2>
- Olweus, D. (1996). *The Revised Olweus Bully/Victim Questionnaire*. Mimeo. Bergen, Norway: Research Center for Health Promotion (HEMIL Center), University of Bergen. <https://doi.org/10.1037/t09634-000>
- Pauw, L. M. J. (2013). *Onderwijs en burgerschap: wat vermag de basisschool? Onderzoek naar De Vreedzame School*.
- Rosenberg, M. (1965). Society and the adolescent self-image. *Princeton University Press*, Princeton, NJ.
- Rudolph, K. D., Abaied, J. L., Flynn, M., Sugimura, N., & Agoston, A. M. (2011). Developing relationships, being cool, and not looking like a loser: Social goal orientation predicts children's responses to peer aggression. *Child Development*, 82(5), 1518–1530. <https://doi.org/10.1111/j.1467-8624.2011.01631.x>
- Sentse, M., Veenstra, R., Kiuru, N., & Salmivalli, C. (2015). A Longitudinal Multilevel Study of Individual Characteristics and Classroom Norms in Explaining Bullying Behaviors. *Journal of Abnormal Child Psychology*, 43(5), 943–955. <https://doi.org/10.1007/s10802-014-9949-7>
- Sheldon, K. M., Ryan, R., & Reis, H. T. (1996). What makes for a good day? Competence and autonomy in the day and in the person. *Personality and Social Psychology Bulletin*, 22(12), 1270–1279.

Van Widenfelt, B. M., Goedhart, A. W., Treffers, P. D. A., & Goodman, R. (2003). Dutch version of the Strengths and Difficulties Questionnaire (SDQ). *European Child and Adolescent Psychiatry*, 12(6), 281–289. <https://doi.org/10.1007/s00787-003-0341-3>

Veenstra, R., Lindenberg, S., Huitsing, G., Sainio, M., & Salmivalli, C. (2014). The role of teachers in bullying: The relation between antibullying attitudes, efficacy, and efforts to reduce bullying. *Journal of Educational Psychology*, 106(4), 1135–1143. <https://doi.org/10.1037/a0036110>