NOVAFRICA
Working Paper Series

Menstrual Stigma, Hygiene, and Human Capital: Experimental Evidence from Madagascar

Karen Macours
Paris School of Economics, and INRAE.

Julieta Vera Rueda
Paris School of Economics

Duncan Webb
Nova School of Business and Economics, and NOVAFRICA

ISSN 2183-0843
Working Paper No 2605
August 2026

Menstrual Stigma, Hygiene, and Human Capital: Experimental Evidence from Madagascar

Karen Macours*‡, Julieta Vera Rueda*, Duncan Webb†

August 10, 2026

Abstract

Menstrual stigma is pervasive in low- and middle-income countries, yet we know little about its consequences for girls' human capital. In a field experiment in 140 schools in Madagascar (N=2,256), we evaluate a school hygiene program focused on reducing menstrual stigma through sanitation infrastructure, menstrual products, and teacher- and peer-led sensitization. The program generates substantial learning gains (0.15 SD) without improving attendance or health. The learning gains are instead explained by psychosocial mechanisms: reduced stigma, lower stress measured by heart rate, and a better school social climate. Our results suggest that menstruation can be a meaningful barrier to girls' learning.

JEL Codes: *C93, I12, I21, J16, O12*

Keywords: *menstrual stigma, learning outcomes, school climate, psychosocial mechanisms, social norms, adolescent girls, field experiment*

*Paris School of Economics. ‡INRAE. †Nova School of Business and Economics. Emails: karen.macours@psemail.eu, julieta.verarueda@gmail.com, dmbwebb@gmail.com. We would like to thank our partners, CARE Madagascar and CARE France for their hard work and unwavering commitment to the project – especially Emilie Vigeanel, Marina Ogier, Andrianarilaza Rakotovao, Emma Hanitriniaina, Ifalianiaina Mihantarijaona, and Hasina Razakarivelonirina. We are indebted to Eliane Ralison and the entire survey team from Consultant Associates for countless inputs throughout the survey and lab-in-the-field instrument design, excellent data collection, and for their perseverance and ingenuity in tracking girls that moved. We thank Abhijit Banerjee, Kathleen Beegle, Janet Currie, Sarah Deschênes, Esther Duflo, Pascaline Dupas, Emanuela Galasso, Seema Jayachandran, Sylvie Lambert, Suanna Oh, Betsy Levy Paluck, Jon de Quidt, Manisha Shah and participants at presentations at CEPR, CSAE, EUI, IADB, Geneva, Kent, Leuven, Milano-Bicocca, Prato, Princeton, PSE, Stockholm School of Economics, and Stockholm University for very helpful comments and feedback. We thank Emanuela Galasso for sharing validated standardized test instruments and data from related work in Madagascar. We would like to thank Oscar Diaz, Mariachiara Iannuzzi, William Amedanou, Amit Khanal, and Alex Verlet for their fantastic research assistance. This work has been generously supported by CARE, the Fonds d'Innovations pour le Developpement, J-PAL's Post-Primary Initiative, CEPREMAP, and the French National Research Agency (ANR), under Grant ANR-17-EURE-0001. Duncan Webb has been supported by Fundação para a Ciência e a Tecnologia (UID/00124/2025, UID/PRR/124/2025, Nova School of Business and Economics), and LISBOA2030 (DataLab2030 - LISBOA2030-FEDER-01314200). This study was pre-registered in the AEA RCT Registry under the unique identifying number AEARCTR-0008972. The study was approved by the Institutional Review Board at the Paris School of Economics.

1 Introduction

Menstruation is deeply stigmatized in much of the world, with an estimated one in three menstruating women facing shame and harassment (WaterAid, 2013). For adolescent girls, this stigma could constrain human capital development at a critical age: shame and anxiety about leaks or smell can inhibit concentration at school, social sanctions against discussing menstruation can restrict the flow of information about menstrual hygiene, and teasing of menstruating girls can worsen the school social environment. Yet although menstrual health and hygiene programs are widely promoted interventions for improving girls' education in developing countries, we have almost no evidence on whether reducing menstrual stigma improves girls' human capital.

School-based programs typically address menstruation and hygiene by providing menstrual products, sanitation infrastructure, and information. The implicit theory of change underlying these programs and their evaluations is that providing products and infrastructure improves health and reduces absenteeism. Evaluations have accordingly focused on attendance and health, typically finding mixed or weak effects (Oster & Thornton, 2011; Hennegan & Montgomery, 2016; Benshaul-Tolonen et al., 2020, 2021; Czura et al., 2024; Agarwal et al., 2024), and they rarely measure learning: a systematic review of hygiene interventions found that only 2 of 44 evaluations did so (Chirgwin et al., 2021).[1]

The stigma surrounding menstruation suggests that these programs could affect human capital through a second cluster of channels that is often overlooked: *psychosocial* channels. When a behavior is surrounded by shame and social sanctions, those affected face not only the physical consequences of poor hygiene, but also stress, anxiety, bullying, and social exclusion, which can inhibit concentration and learning quite apart from any effects on health or attendance (El-Gilany et al., 2005; Ali & Rizvi, 2010). Hygiene programs could also improve the school environment more broadly: encouraging students to share responsibility for sanitation facilities can foster cooperation and strengthen peer relationships, and teacher training on hygiene may improve student-teacher interactions or signal that the school cares about students' wellbeing. A growing body of evidence shows that school interventions that aim to build social cohesion, foster empathy, or improve peer relationships can improve learning without directly targeting academic skills (Alan et al., 2021; Alan & Kubilay, 2025; Ashraf et al., 2020; Edmonds et al., 2023). Yet previous research on menstrual health and hygiene in schools has focused almost entirely on attendance and health, leaving the role of psychosocial channels and stigma unexamined.

We address these gaps using a field experiment in 140 schools in rural Madagascar to evaluate a school program that aims to reduce menstrual stigma and improve menstrual hygiene and a broader set of hygiene behaviors (hand-washing, latrine usage and cleanliness). The program, developed by our partner NGO, CARE, uses a package of school-level interventions involving both students and teachers, with a particular focus on menstruation. The program's hygiene focus also serves

[1] A scoping review of 83 studies of water, sanitation, and hygiene (WASH) in schools similarly finds that learning outcomes are "rarely measured or reported," despite being stated as a motivating rationale for such programs (Bick et al., 2025). The few RCTs that do measure test scores find null effects (Oster & Thornton, 2011; Nelson et al., 2025). The Global Education Evidence Advisory Panel's review of cost-effective approaches to improving learning does not classify WASH or hygiene interventions (Banerjee et al., 2023).

as an entry point for norm change: because stigma in this context includes a strong norm against discussing menstruation, having teachers and students talk about it openly and routinely undermines the norm directly.

Menstruation is highly stigmatized in this context: in our baseline survey, only 45% of girls had ever heard classmates discuss menstruation (vs. 70% for hand-washing) and 39% believed that menstruation should not be discussed. This translates to poor knowledge and practice of optimal hygiene—for example, only 11% of girls report using a menstrual product that is not just reused cloth.[2] In our control group there is a strong correlation between menstrual stigma and learning: girls with higher stigma score approximately 0.2 SD lower on learning tests, consistent with psychosocial factors representing a binding constraint on learning in this context.

These menstruation-specific barriers exist amid severe economic and educational challenges in Madagascar. Foundational learning levels are extremely poor: only 7% of 7–14 year-olds master basic numeracy (UNICEF, 2019), and only 50% of children complete primary school (UNESCO Institute for Statistics, 2023). Moreover, Madagascar is one of the poorest countries in the world, with 80% of its population living in extreme poverty—particularly in rural areas like the region of our study (Mulangu & Oviedo Silva, 2024). Given these severe economic challenges, it was not *a priori* obvious that addressing psychosocial constraints related to menstruation would be able to drive improved educational outcomes.

Our experimental design aims to answer two questions. First, we test whether a package of hygiene interventions (the "base program") designed to simultaneously address multiple hygiene-related barriers—both physical and psychosocial—can meaningfully impact girls' human capital, and through which channels.[3] This package (i) trains school teachers and establishes school hygiene committees to promote hygiene behaviors and reduce menstrual stigma, while also (ii) relaxing hard constraints by building hygiene infrastructure (hand-washing basins and latrines) and distributing free reusable menstrual pads.

Second, we test the additional effect of an innovative approach to addressing the harmful social norm underlying menstrual stigma: identifying and amplifying the voice of *positive deviants* (Marsh et al., 2004)—people embedded within a social network who are willing to engage in prosocial behaviors in defiance of a harmful social norm. We identify approximately four "*Young Girl Leaders*" (YGLs) in each school, selected based on their willingness to actively and openly speak about menstruation despite the taboo, and on their broader leadership skills. YGLs are trained on a curriculum focusing on the reduction of menstrual stigma, key hygiene behaviors, broader female empowerment and leadership skills, and are then coached to reduce stigma and promote hygiene among their peers at school. Selecting positive deviants, rather than opinion leaders chosen for their influence or reach, reflects the nature of the norm being targeted: because menstrual stigma is sustained by silence, amplifying girls already willing to speak openly targets the norm directly. Working through them

[2]This aligns with similarly poor knowledge about menstruation documented in Madagascar-wide studies, e.g., only 48% of females above 15 years old can correctly identify the fertile period during the menstrual cycle (United Nations Children's Fund (UNICEF) & World Health Organization (WHO), 2023).

[3]As such, we follow a similar logic to evaluations of other multi-faceted interventions intentionally designed to address multiple constraints at once (e.g., Banerjee et al., 2015 or Carneiro et al., 2021).

may be more effective than individualized attempts to change attitudes, both because it leverages a *social multiplier* effect through peer networks (Banerjee et al., 2013; Beaman et al., 2021; Banerjee, Chandrasekhar, et al., 2019) and because public defiance can change not only attitudes but also beliefs about others' attitudes, thereby coordinating a change in norms (Paluck et al., 2016, 2021).

Our study randomized each of the 140 schools into one of three conditions: (i) *Control* (35 schools); (ii) *Base Only* (35 schools); and (iii) *Base+YGL* (70 schools), which receive all elements of the base program along with the YGL program. After 1.5 years of intervention, we evaluate effects on human capital (education, physical health, and mental health), menstrual stigma, and hygiene knowledge and behaviors in an endline survey (N=2,256). To credibly identify impacts, we use objective measures of learning, attendance, stress, stigma, and directly observed hygiene behavior, reducing concerns about experimenter demand effects—a common problem when evaluating norm-change interventions.

We first document that the intervention has large positive impacts on the outcomes directly targeted by the program. It substantially improves hygiene knowledge (pooled effect of *Base Only* and *Base+YGL*: 0.46 SD) and hygiene behaviors (0.48 SD), with impacts corroborated through directly-observed measures.

We then show that the program has important impacts on girls' human capital outcomes. It substantially improves girls' learning at endline: the pooled effect on our combined index of standardized achievement and cognitive tests is 0.15 SD, with the largest gains on the math and language achievement tests (0.13–0.25 SD). It also increases the probability of progressing to the next grade level (+9 p.p., 17%) and raises girls' official grade point average (0.11 SD, p=0.08). The gains on the math and language achievement tests place the intervention at the 75th (*Base Only*) and 68th percentile (*Base+YGL*) of effective learning programs evaluated in low-income country schools (Muralidharan, 2017; Evans & Yuan, 2022a). This is despite the program not directly targeting improvements in academic performance.

The improvements in learning are not driven by improved school attendance or health, in contrast to the conventional channels through which hygiene programs are typically expected to operate (Oster & Thornton, 2011). There are no effects on attendance, as measured by 5 unannounced spot checks, official school enrollment registers, and girls' self-reports. And we find no effects on an index of physical health combining girls' reported overall health and urinary tract infection symptoms. This suggests that the improvements in learning were driven by increased *quality* of time at school, rather than *quantity*.

The evidence instead points to several interrelated psychosocial factors explaining the learning improvements. First, the intervention dramatically reduces menstrual stigma. Girls become substantially more willing to speak openly about menstruation, report more progressive attitudes and perceived norms, and feel less shame. We measure stigma not only through survey responses but also through a set of lab-in-the-field exercises designed to capture revealed preferences: willingness to speak about menstruation in front of peers, willingness to request menstruation-related information, social sanctions against discussing menstruation (dictator games), and enumerator observations of shame. The combined index of these objective measures improves by 0.60 SD (p<0.001). Reducing stigma could have improved learning if shame and anxiety about menstruation inhibit concentration, or if the social sanctions surrounding menstruation worsen the school environment. Some suggestive empirical patterns are consistent with this interpretation. First, learning effects are stronger among girls who were menstruating at the time of the endline survey, when stigma-related stress would be most salient. Second, the intervention had smaller and not statistically significant impacts on boys' learning, consistent with a gender-specific mechanism, though the girl–boy difference is imprecisely estimated. Third, stigma improves rapidly, with detectable effects at the 5-month midline survey; girls' grade progression and 2022 exam pass rates imply that academic performance also improved within the first school year, whereas hygiene behavior does not show similar early improvements.

Second, the program generates broader improvements in the schools' psychosocial environment. We find an increase of 0.24 SD in an index of network connection between students and a reduction in bullying reported by girls (−0.08 SD, p=0.06).[4] These improvements could enhance learning by reducing stress, by increasing collaboration and peer learning, or by increasing students' willingness to participate in class (Alan et al., 2021; Alan & Kubilay, 2025).[5]

Third, we document changes in objectively-measured stress levels. Girls' heart rates during the endline survey are 0.12 SD lower in treatment schools, an effect approximately equivalent to a quarter of the heart rate increase driven by extreme psychological stress (Kudielka et al., 2004; Seddon et al., 2020). Given evidence that stress acts as a "bandwidth constraint" inhibiting cognitive function (Mullainathan & Shafir, 2013; Haushofer & Fehr, 2014; Schilbach et al., 2016; Kaur et al., 2025), this stress reduction may have directly improved concentration and learning while at school.[6] Reduced menstrual stigma plausibly contributes to both the improved psychosocial environment and the reduced stress, and may also drive the improved peer relationships and reduced severe bullying that we document.

After establishing the effects of the base program, we show that adding Young Girl Leaders generates additional improvements in key targeted behaviors. Compared to *Base Only* schools, *Base+YGL* schools show larger effects on hygiene behavior, faster improvements in hygiene knowledge, and more improvements in the schools' social climate. Self-reported menstrual stigma shows larger reductions in *Base+YGL* schools, though objective lab-based measures of stigma improve similarly in both arms, suggesting some of the additional self-reported gains may reflect reporting bias. Overall, these results suggest that positive deviants can be an effective tactic for embedding behavior and norm change in a social network.

Despite these additional improvements, adding YGLs does not further improve learning: there is no significant difference between *Base+YGL* and *Base Only* schools on our combined learning index (p of difference: 0.17). We find suggestive evidence that YGLs' intensive focus on menstruation and hygiene partially crowds out attention to academics. In secondary schools, however, this cost

[4] These effects may be partially driven by the program encouraging students to engage in collaborative hygiene behaviors at school, such as cleaning the courtyards, toilets, and classrooms.

[5] We find no evidence of impacts being driven by changes in teacher motivation or teaching practices.

[6] The null effects on physical health, along with the correlation between heart rate and girls' reports of stress about exams and menstruation, suggest that the heart rate changes reflect mental health improvements rather than a hygiene-related reduction in infections.

is offset by the psychosocial benefits that effective positive deviants generate: YGLs lead to larger improvements in social climate and hygiene, with no average cost to learning. The net effect of YGLs in secondary schools is thus positive on balance, highlighting that the approach works well where sufficiently mature and confident leaders can be identified.

The psychosocial effects of the program also have an unexpected dimension: we find an increase in early pregnancies over the previous year in *Base Only* schools (from 2.0% to 3.6% of girls, driven by secondary schools), possibly reflecting the increased social connections between boys and girls that the program fosters. We find no such increase in *Base+YGL* schools, plausibly because YGLs generate a counteracting increase in support networks, information and empowerment among girls. This illustrates both the breadth of the psychosocial channels through which the program operates and the importance of measuring unintended consequences.

We provide, to the best of our knowledge, the first experimental evidence on whether reducing menstrual stigma improves girls' education. Few studies focus specifically on reducing menstrual stigma (two notable exceptions are Castro & Czura, 2025; Castro & Mang, 2024), and none has evaluated impacts on schooling. Previous work has primarily evaluated the effect of menstrual products, hygiene infrastructure, or information provision, and has examined attendance and health as primary channels, typically finding mixed or weak effects (Oster & Thornton, 2011; Hennegan & Montgomery, 2016; Phillips-Howard et al., 2016; Benshaul-Tolonen et al., 2020, 2021; Chirgwin et al., 2021; Czura et al., 2024; Agarwal et al., 2024). Our results align with these null findings, but reveal large impacts on learning, together with evidence consistent with previously overlooked psychosocial mechanisms: reduced stigma and stress and an improved school social environment. This highlights the need to measure learning and diverse psychosocial pathways in impact evaluations, even for programs not explicitly focused on academic outcomes.

Our findings have broader implications for the ongoing policy debate about the global learning crisis (Akyeampong et al., 2023) and the related academic literature on the effect of school-based interventions on learning outcomes and gender differences in such impacts (Evans & Yuan, 2022b). We bring together two disparate strands of research: one showing that psychosocial interventions in schools can have important positive impacts on welfare and academic performance (e.g., Stallard et al., 2014; Alan et al., 2019, 2021, 2024; Ashraf et al., 2020; Edmonds et al., 2023; Dinarte-Diaz & Egana-delSol, 2024), and another evaluating schools as entry points for health interventions (Miguel & Kremer, 2004; Baird et al., 2016; Dupas & Miguel, 2017). The evidence pointing to psychosocial drivers of learning effects suggests that a broader set of health-focused programs could leverage such drivers. We also expand the evidence on gender-specific barriers to girls' education in low- and middle-income countries (LMICs) by examining menstrual stigma, a previously underexplored barrier.[7] More broadly, we add to a growing body of evidence that interventions not directly targeting academic skills—such as bicycle provision (Muralidharan & Prakash, 2017), deworming (Miguel & Kremer, 2004), or school feeding programs—can substantially improve educational outcomes, often through channels that standard evaluations fail to measure. We also speak to the literature on

[7]Related work targeting other constraints for adolescent girls in Sub-Saharan Africa includes Buehren et al. (2024); Bandiera et al. (2020); Bergstrom & Özler (2023); Hamory et al. (2023); Shah et al. (2023), motivated by evidence of very high returns to keeping girls in secondary schools (Duflo et al., 2024).

improving education at scale in LMICs by focusing on a very low-income setting and a program with over 15,000 beneficiaries (Evans et al., 2024).

Finally, our experimental design allows us to evaluate a novel approach to norm change: working with *positive deviants* to address harmful social norms. We show that this tactic can be effective, while highlighting the challenges of identifying effective positive deviants in all settings—particularly in primary schools. This builds on the literature examining the role of positive deviants or "trendsetters" (Paluck et al., 2016; Bicchieri, 2017; Bicchieri & Funcke, 2018), and more broadly on work addressing harmful social norms (Jayachandran, 2021; Dhar et al., 2022; Gulesci et al., 2025; Bursztyn et al., 2020; Banerjee, La Ferrara, & Orozco-Olvera, 2019; Matavelli, 2023; Webb, 2025).

2 Study design

2.1 Interventions

The study evaluated the effect of a bundle of interventions designed by our partner NGO, CARE Madagascar. The program aimed to reduce menstrual stigma and promote hygienic behaviors, and to thereby improve human capital outcomes for girls in the later grade levels of primary school and all grade levels of secondary schools.[8] Approximately 15,450 students received some intervention as part of the program.

Each of the 140 schools in our sample was randomly allocated into one of the following treatments:

1. *Control* (35 schools), which received no school-level interventions.
2. *Base Only* (35 schools), which received a package of interventions that included teacher-led sensitization, the construction of sanitation infrastructure, vouchers for free menstrual pads for female students, and the creation of a school-level sanitation committee involving parents, students, and teachers.
3. *Base program + Young Girl Leaders* (70 schools), which received all the elements of the base program, plus the Young Girl Leaders intervention.

We did not include a YGL-only arm: girls openly challenging menstrual stigma risked backlash, and the base program's teacher training, infrastructure, and institutional support provided a protective environment for such norm-defiant behavior.

Randomization was stratified by (i) whether a school was primary or secondary, (ii) whether there was existing functional water access at school at baseline, and (iii) the number of girls enrolled above the age of 12.

A summary timeline of program activities and main data collection exercises is shown in Figure 1.

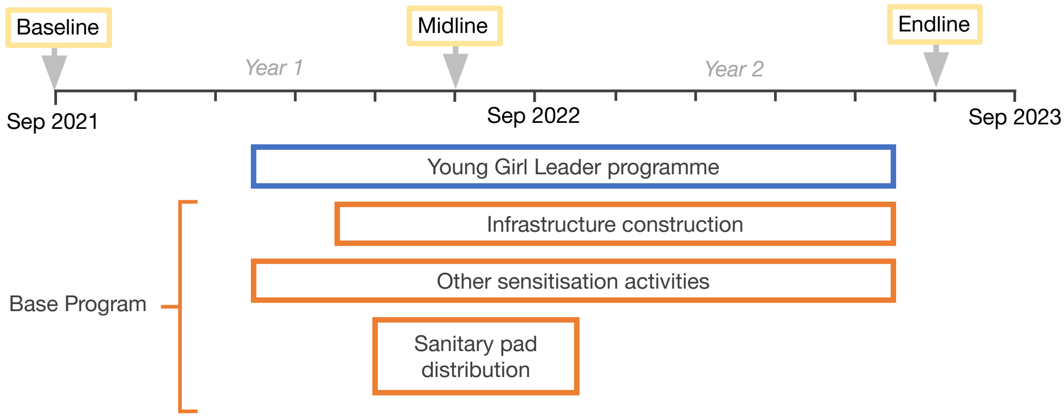
2.1.1 Base program

The base program aimed to: (i) reduce menstrual stigma and promote hygienic behaviors by sensitizing teachers (and to a lesser extent parents) and asking that they in turn sensitize students; and (ii) relax hard constraints that would otherwise prevent girls from engaging in hygienic behaviors.

Teacher training, curriculum, & hygiene committee. To sensitize the school community to the hygienic

[8] The region's secondary schools are mostly lower secondary schools, although our sample also includes 6 upper secondary schools (*lycées*).

Figure 1: *Study timeline*



Notes: Timeline of program activities and data-collection waves; not to scale.

behaviors being promoted, local representatives of the Ministry of Education and personnel from our NGO partners trained two teachers in every school, one of whom was typically the school principal. There were 2 days of training at the start of the program, and shorter refresher trainings that took place every semester. The trained teachers were asked to sensitize all students and other teachers by organizing short (5-minute) weekly sessions to talk about menstruation and hygiene. These typically took place during school assemblies, in the school yard, or at the start of classes.

The key messages and hygiene behaviors being targeted in the curriculum were as follows: (i) menstrual hygiene, including how to properly wash during menstruation, how to use and wash reusable sanitary pads, and basic information about menstruation and the body; (ii) reducing menstrual stigma, including encouraging open discussion of menstruation and explaining that menstruation is a normal biological phenomenon and should not be a source of shame; (iii) using latrines; (iv) hand-washing with soap; (v) using and keeping clean water; and (vi) food hygiene.

Two design features are worth noting. First, the program used hygiene, a topic teachers can raise with minimal embarrassment, as an entry point for discussing menstruation: menstrual health was embedded in a broader hygiene curriculum rather than treated as a standalone taboo subject. Second, because the stigma centres on the norm that menstruation should not be discussed, regular open discussion of menstruation by teachers and students was itself a stigma-reduction mechanism, beyond the content of the messages.

In addition, an 8-member "hygiene committee" composed of teachers, parents, and students was formed in every treatment school. The committee was responsible for reinforcing hygienic behaviors and reducing stigma and for maintaining the hygiene infrastructure. Adult members received an initial training session about the roles and responsibilities of the committee.[9]

[9] Together, the teacher-led curriculum and the hygiene committee meant that sensitization on hygiene behaviors was not *only* carried out by the Young Girl Leaders (YGLs). This also meant that the activities of YGLs were more likely to be perceived to be endorsed by adults and figures of authority in the school, reducing the risk of backlash against YGLs. There were also some light-touch interventions that took place at the district level, thereby affecting girls and parents of both treatment and control schools equally. These included (i) a local newspaper encouraging positive hygiene behaviors, (ii) training of local seamstresses to make menstrual pads, (iii) community-level celebrations of hygiene-related world days.

Infrastructure. To allow students to engage in the hygiene behaviors recommended by the program, the NGO financed and built hygiene and sanitation infrastructure in all treatment schools. Specifically, our partner built between 2 and 10 (mean: 3.9) new latrines on the school site, and in some cases rehabilitated existing latrines so that they were functional. These latrines were separated by gender, in line with the idea that gender-segregated toilets can reduce the risk of sexual violence for girls (Gao et al., 2025), and incorporated in-built hand-washing facilities. This infrastructure was finalized in all treatment schools by endline, but was only at very early stages by midline, so by midline the treatment consisted mainly of the sensitization component rather than the new physical infrastructure.

Vouchers for sanitary pads. To relax hard constraints related to *menstrual* hygiene, the NGO distributed vouchers which could be exchanged for a set of 6 free reusable sanitary pads from local tailors who were trained by the program. The vouchers were distributed to all secondary school girls, and to all girls in the final two grades of primary school.

2.1.2 Young Girl Leaders intervention

The Young Girl Leaders intervention involved selecting girls in each school and designating them as "Young Girl Leaders" (YGLs) who were responsible for peer-to-peer promotion of hygiene and menstrual hygiene behaviors. An average of four YGLs were selected in each treated school.[10]

YGLs were selected based on a survey of teachers and of the students whom those teachers nominated. We selected girls in the same way across all YGL schools, using a predetermined weighting of survey responses that incorporated (i) leadership qualities, (ii) measures of "positive deviance" (e.g., willingness to take prosocial actions in defiance of others, such as speaking openly about menstruation), and (iii) having started menarche themselves. We restricted to girls who were above the passing GPA, to avoid hurting the performance of girls who were already struggling academically.[11]

YGLs were asked to spread positive messages and encourage hygienic behavior among their peers at school, and to openly discuss menstruation in a way that could reduce menstrual stigma.[12] They received an initial intensive training with weekly sessions of 1.5 hours for approximately 3 months. These training sessions were delivered using a cascade model, where technicians from local NGOs were trained by our main partner NGO, and then in turn trained YGLs. The sessions initially focused on enhancing the leadership and communication skills of the YGLs, and then moved onto the same curriculum focused on menstrual stigma and hygiene used in the teachers' curriculum. After the initial 3-month training period, YGLs received coaching visits by a local NGO technician every two

[10] The number selected depended on the number of pupils in the school. There was one YGL for every 60 students in the school, apart from schools with fewer than 120 students, which always had two YGLs. Endline data confirms that treatment assignment was fully respected: the targeted number of YGLs were identified and trained in all 70 treatment schools, and there were no YGLs in control or *Base Only* schools.

[11] Endline data confirms that these selection criteria successfully identified socially influential girls: within *Base+YGL* schools, YGLs score 0.55 SD higher than non-YGL girls on a composite index of peer network connectedness ($p<0.001$), with particularly large differences in the number of peers they discuss menstruation with (Appendix Table G1).

[12] To test whether peer-to-peer sensitization was more effective if it had the authority of formal sessions, or if it arose organically during conversations, we randomized whether YGLs were asked to sensitize their peers "formally" (using classroom sessions, 35 schools), or "informally" (during organic discussions, 35 schools) (see Appendix J). Endline data revealed, however, that there were only modest differences in actual implementation (Figure J1). Consistent with limited differences in implementation, we cannot reject that the treatment effects across all our main outcomes are equal for the two modalities (Table J2; p-value of joint test = 0.71), and therefore present the pooled results throughout the paper.

weeks which aimed at giving additional guidance and encouragement.[13] Further details on YGL selection and training are in Appendix I.

2.2 Data collection

2.2.1 Baseline survey and sample

A total of 140 primary and secondary schools in rural areas of Madagascar's Amoron'i Mania region were selected to be part of the study, based on: (i) excluding schools inaccessible by vehicle; (ii) prioritizing schools with the most girls above age 12, for whom the intervention would be most relevant; and (iii) excluding schools too close together to minimize spillover effects. Only one selected school was girls-only; the mean proportion of enrolled girls was 52% (SD: 10 p.p.).

We sampled an average of 17 girls per school to be surveyed at baseline (September–October 2021), totaling N=2,393, based on a listing from official school enrollment registers. Of these students, 90% were between 4th and 9th grade, but the sample included students between 3rd grade and 12th grade.[14]

Sample characteristics (Table A2). Girls' average age in the sample at baseline was 14 years, and about half of the girls had not reached menarche at baseline. The sample is characterized by high levels of poverty, and relatively large accumulated school delays (2.2 years on average). For 77% of girls, the primary caretaker's main activity is agriculture; only 11% have access to electricity, and only 2% of households own a vehicle. Girls take on average 30 minutes to walk to school, and live in large households (mean size: 6.5), with 81% living in the same household as their mother. School dropout at early ages is also frequent: 17% of girls in control schools drop out between year 1 and year 2 of the study.

Balance. For all relevant treatment comparisons, the treatment groups are well-balanced across key characteristics (Tables A2, A3, A4), and joint tests of imbalance across all key baseline variables are insignificant (p=0.75 for *Base Only* vs. control; p=0.86 for *Base+YGL* vs. control).

2.2.2 Endline survey: girls' survey

The main source of outcomes is an endline survey with the baseline girls, initiated after 1.5 years of intervention (May–July 2023, at the end of the second school year) and completed through intensive tracking in October 2023. We surveyed 2,256 out of 2,393 girls, corresponding to an attrition rate of 5.7%. Attrition is balanced across treatment arms (Table A5, column 1). Most girls were interviewed at their homestead during the initial fieldwork period, but many had changed schools, dropped out, or migrated to other regions. We therefore carried out intensive tracking with in-person interviews at migration destinations (e.g., in neighboring regions and in the national capital, Antananarivo; see Appendix Section B.3). A random subset of schools was selected for intensive tracking, and we

[13] At the start of the second school year, a comprehensive refresher training covering all topics was provided to the YGLs. YGLs who graduated from a school were replaced with new YGLs, who were selected using the same criteria as the first school year and also received the same training.

[14] The baseline took place at the end of the summer vacation in 2021, using a sample based on school registers from the previous school year. We therefore avoided sampling girls who had just finished the final grade of a cycle (9th or 12th grade), since those who passed would no longer be enrolled in the school in the new school year. In lower secondary schools, we sampled equally from girls who had just finished 6th, 7th, and 8th grade, and in high schools we sampled equally from girls who had just finished 10th and 11th grade. In primary schools, to collect data on girls for whom the focus on menstrual hygiene and stigma was likely to be most relevant, we sampled girls who were aged 12 and above and who were in the upper grades of primary school (including 5th grade girls registered to repeat), filling out the remaining sample from other grades and younger ages in smaller schools where necessary.

show robustness of all results when using sampling weights to account for this random selection (Table B4). A further 3.7% of girls, who could not be located in person, were interviewed by phone using a shorter survey that did not include standardized learning tests; standardized learning tests were therefore completed for 2,167 of the 2,256 surveyed girls (Table 2). Since *Base Only* girls were slightly more likely to be surveyed by phone (p=0.03), we calculate Lee bounds to show further robustness.

We also added to our endline survey all girls selected as YGLs at the start of the program. This sample is used only for descriptive analysis to characterize the YGLs and their selection.

2.2.3 Endline survey: other surveys

In addition to the girls' survey, during the endline data collection we also carried out 4 other surveys:

1. *Mothers*. We conducted a home-based survey of the mother or the same female caregiver who had been surveyed in the baseline. We found 2,287, corresponding to an attrition rate of 4%, balanced across treatments (Table A5, column 4).
2. *Teachers and directors*. We surveyed all 140 school directors and the "*professeurs titulaires*" in charge of each target grade (randomly selecting 1 teacher per grade in case of multiple sections). We pool these for analysis, yielding 743 interviews in total. 52.1% of those interviewed were women. In this survey we also elicited students' official GPAs received by all the children in at least one class for every grade level.
3. *Male classmates*. Using up-to-date school registers at endline, we randomly selected 10 boys in each school to do a short survey focused on attitudes towards menstruation and a dictator game related to menstruation.
4. *Information supply experiment*. Using school registers at endline, we randomly sampled a new set of 12 girls and 4 boys in each school to take part in an interactive lab-in-the-field exercise designed to measure girls' willingness to supply information about menstruation to others. Each girl was taught a menstruation-related topic and a general hygiene topic, then asked to choose which to share with the whole group—either anonymously (via video) or by explaining it themselves in front of the group, with the latter capturing social image concerns associated with the menstruation taboo (see Appendix D).

Sample selection. Despite no effects on whether girls are enrolled in *any* school, there is some evidence that girls are less likely to remain in the *same* school as at baseline in treatment schools due to faster grade progression (Table B5). This does not threaten our main estimates, since we track all baseline girls regardless of current enrollment. However, it does imply that results from samples of students still enrolled at endline—official GPAs, male classmates, and lab-in-the-field participants—may be subject to differential selection. In general though, we find that those who remain in the same school are negatively selected (they are poorer, are more delayed in their schooling, have worse knowledge and behaviors at baseline), suggesting that any improvements on the selected samples will tend to be a lower bound (Table B6). We discuss this issue further in Appendix B.4.[15]

2.2.4 Midline survey

We carried out a midline after approximately 5 months of intervention (June–August 2022). Attrition was 11% and was slightly higher in treatment schools (Table A5, column 3, which reports the probability of a girl being interviewed at both midline and endline). Since the midline is not used for key outcomes, only to provide a broad description of dynamics over time, this does not threaten our main conclusions.

2.2.5 Spot-check absenteeism and school register data

To measure school absenteeism without having to depend on potentially unreliable official school registers, we carried out 6 waves of unannounced spot checks in all schools. Enumerators visited the schools without prior notification, and took attendance calls for all classes above 3rd grade. They also digitized the official school enrollment and attendance records covering up to the 3 months preceding each visit. The first spot check took place before the intervention started (February 2022) to act as a baseline. Subsequent waves took place in June, October, and December 2022, and January and June 2023.

2.2.6 Exploratory 2nd endline

In the first endline, we found an unexpected increase in pregnancies for girls in *Base Only* schools that was driven by only a small number of reports. To understand the robustness of this result and the mechanisms behind it, we collected additional school-level data on pregnancies in June 2024 (see Appendix F).

2.3 Outcomes

Here, we describe the main outcomes and measurement approaches used in the paper (see Appendix Section K.1 for more details). Our primary outcomes capture learning and the channels through which the program may affect it: (i) standardized learning tests; (ii) school attendance; (iii) physical health; (iv) heart rate as an objective measure of stress and anxiety; (v) menstrual stigma (both survey-based and revealed-preference measures); and (vi) school social climate. Menstrual stigma serves a dual role: it is both a direct outcome of interest, given the intervention's explicit focus on stigma reduction, and a potential channel for understanding the learning effects. Our secondary outcomes capture behaviors directly targeted by the program: (vii) hygiene knowledge; (viii) hygiene behavior; and (ix) observed cleanliness. We also examine early pregnancy as an unintended consequence. We complement these with alternative measures that highlight robustness and mechanisms.

Our data collection prioritized objective and revealed-preference measures over self-reports. This addresses concerns about social desirability bias and experimenter demand effects, which are particularly relevant since respondents likely knew the program focused on hygiene and menstruation. We obtained objective measures for key welfare outcomes (learning, attendance, stress, and stigma) and prioritize these results; we rely on self-reported measures to understand mechanisms and sec-

[15] As for other possible sources of selection, Table A5 shows that there was (i) no differential attrition for teachers (column 7), though there is a lower probability of successfully eliciting school GPAs among *Base Only* schools (column 6); (ii) no difference in the number of sampled boys successfully interviewed (column 8); (iii) no differential sample selection for girls in the lab-in-the-field (column 9); and (iv) no differential availability of official absenteeism data from school registers (column 10).

ondary outcomes. For these measures, we test for differential reporting bias by examining whether treatment effects are driven by girls with high social desirability scores, using an adapted Crowne & Marlowe (1960) index from baseline (see Table A6; Appendix Section K.3).

Index construction. Unless otherwise specified, we use two aggregation approaches. First, when combining questions from the same survey module that measure a single underlying construct (e.g., depression symptoms, attitudes towards menstruation), we use factor analysis to determine weights based on how strongly each item loads onto a common factor. Second, when combining conceptually distinct sub-indices from different survey modules into a single aggregate index (e.g., combining hygiene knowledge with menstruation knowledge into an overall knowledge index), we use unweighted averages of Z-scores following Kling et al. (2007). Knowledge and behavior indices are exceptions: they are constructed as sums of correct answers (then standardized), since all items have equal importance from a programmatic perspective. Further details on index construction for each outcome are provided in Appendix Section K.1.

Unless otherwise specified, outcomes are standardized to have a mean of 0 and standard deviation of 1 in the control group to ease interpretation. Below we describe all outcomes in more detail, with further information in Appendix K.1.

2.3.1 Primary outcomes

Learning. Our key outcome is a set of standardized tests, administered by the survey team at endline, that evaluates academic skills learned at school and cognitive ability. It includes modules on (i) math fluency, based on a timed written math test involving simple arithmetic calculations; (ii) Malagasy reading fluency; (iii) a listening exercise; (iv) a reading comprehension test; and (v) a digit-span test that measures working memory (forward and backward). All modules were validated using pilot data, collected by the same survey team before the endline on a population of the same age and region as our study population. Most modules had also been earlier tested and validated for prior studies in Madagascar, with the remaining having been validated elsewhere in Sub-Saharan Africa. For the main analysis, we combine all 5 modules into a single index using factor analysis. We also consider the results of each test separately, to explore the differences between the academic achievement tests (math and reading fluency) and the other tests that capture attention, memory, and cognition.

We complement these tests with two administrative measures of learning and a measure of grade progression. First, we collect official grade point averages (GPAs) from schools' student-specific "bulletins" and class registers, which enumerators collected during the teacher or director survey. Students receive an overall GPA out of 20 for each trimester.[16] Second, we use school-level results on national exams taken at the end of primary and lower secondary school. Because grading is centralized, these address concerns about teacher-driven effects on classroom grades. Finally, we measure grade progression by comparing girls' grade level at baseline and endline (one academic year apart), allowing us to identify whether they advanced a grade, repeated a grade, or dropped out. As well as being a key outcome of interest in itself, this measure also has the advantage of

[16]This data is available for all students in target grades (not just our sample of 17 girls). However, since it is not available at endline for students who have dropped out, it is vulnerable to composition effects driven by selective dropout. We discuss this issue further in Appendix B.4.

being available for all girls in the endline survey, including those interviewed by phone. Together, these measures provide multiple objective indicators of academic achievement that allow for cross-validation.

School attendance. Our key measure of school attendance uses spot check data that measures the number of students physically present at school when the spot check is carried out. For this outcome we include fixed effects for the 5 follow-up waves of spot checks to account for time trends between spot check waves. We complement this with (i) official school registers to measure the number of students present and enrolled each day (controlling for day-level fixed effects); and (ii) girls' reports of whether they were still enrolled in school and absent from school in the last 3 months.

Physical health. Our key measure uses a combined index of two survey modules. First, we measure generalized health by asking girls the physical functioning sub-scale of the Pediatric Quality of Life Inventory (PedsQL) (see Benshaul-Tolonen et al., 2021). Girls are asked whether they had difficulty carrying out any daily activities, such as running or completing tasks at home. Second, since improvements in menstrual hygiene are associated with reductions in urinary tract infections (UTIs) (Phillips-Howard et al., 2016; Czura et al., 2019), we also asked girls to report symptoms of UTIs, based on Alidjanov et al. (2014).

Stress and anxiety. Our key objective measure of stress and anxiety is girls' heart rates during the endline survey. This was incorporated to measure a combination of (i) general anxiety levels experienced by girls, and (ii) the anxiety levels experienced specifically when discussing topics related to hygiene and menstruation. We used wristband monitors that included optical heart rate sensors that recorded the girls' heart rate in beats per minute on a second-by-second basis. Heart rate reflects the arousal of the autonomic nervous system, a primary biological response to psychological stressors. Biomedical research suggests that average heart rate is a good predictor of self-reported anxiety, e.g., in one of the few large-scale studies examining interpersonal correlations between heart rate data and anxiety, average heart rate was one of the best available predictors of self-reported anxiety (Zhang et al., 2025). The monitors we used have been demonstrated to have similar accuracy to industry-standard chest strap monitors (Reece et al., 2021), but were substantially less invasive and more comfortable for girls to wear.[17] We discuss the interpretation and quality of the heart rate data further in Appendix C.1. We complement this with self-reported psychosocial well-being by asking girls (i) an adapted version of the CESD depression index, and (ii) questions about self-esteem, measuring whether girls are confident in themselves and their abilities.

Menstrual stigma. To understand whether the constraints imposed by menstrual stigma are relaxed by the program, we elicit a rich set of stigma-related measures, which we aggregate by distinguishing between reported (survey-based) and observed (revealed-preference) measures. For all measures of stigma, a reduction in stigma is coded positively.

Our aggregate survey-based measure of stigma combines the 5 sub-indexes described here. First, we measure whether the program lifts the taboo on discussing menstruation. We ask (i) how often

[17] In some cases, heart rate monitors were not fixed properly to girls' wrists, or the battery of monitors ran out on the field. Because of this, heart rate data is available for only 1,904 girls. However, the proportion of data available is equal across treatment arms (Table A5, column 5).

girls heard classmates, teachers, and parents discuss menstruation in the last 3 months and how often they asked others about menstruation; and (ii) how willing girls are to discuss menstruation in front of others. We then measure (iii) how progressive girls' attitudes are towards menstruation using Likert-scale questions measuring agreement with statements like "Girls should be allowed to come to school when they menstruate"; and (iv) perceived norms by asking girls to estimate how much people in their social network (other girls, boys, teachers, and mothers) agree with a subset of these attitude questions. Finally, to measure shame associated with menstruation, we (v) describe hypothetical vignettes about embarrassing scenarios related to menstruation to girls and ask about shame-related emotions in response.

The aggregate observed stigma measure combines revealed-preference measures from three incentivized lab-in-the-field experiments, and enumerator observations of girls' discomfort during endline questions about menstruation. The experiments were designed to evaluate how stigmatizing attitudes translated to behaviors, and whether the effects of the program are unique to stigmatized topics like menstruation or whether they also hold for other hygiene-related topics. Specifically, they measure willingness to speak about menstruation in front of others (*information supply*), interest in learning about menstruation (*information demand*), and prosocial behavior towards those who discuss menstruation (*dictator game*). Each experiment includes parallel questions about non-stigmatized hygiene topics, allowing us to test whether effects are specific to menstruation. See Appendix D for design details. To complement these experiments, enumerators also recorded visible indicators of shame when girls discussed menstruation during the endline survey (e.g., looking at the floor, speaking quietly, appearing uncomfortable). Combining these measures into a composite index, we show that the observed stigma index has good predictive validity: it is strongly correlated with both learning test scores and self-reported stigma (both $p<0.01$; Appendix Table D1).

School social climate. To evaluate the social dynamics at school, we ask girls and teachers to report on two distinct measures: *teasing* (light-hearted laughing or mocking) and *severe bullying* (including intimidation, harassment, or violence). We also measure teasing specifically related to menstruation. Girls report on their own experiences and on a randomly selected subset of 2-3 girls in the same grade level; teachers report on a randomly selected subset of girls. We also create a *network integration index* by asking girls about their social connections with their peers—how many friends they have, how many children they play with, how many they speak with at break, and how often they do homework together. For a subset of these questions, we ask separately about other girls and other boys. Our key measure of social climate combines teasing, severe bullying, and network connection into one index.

2.3.2 Secondary outcomes

The secondary outcomes capture knowledge and behaviors directly targeted by the program: hygiene knowledge, hygiene behavior, and observed cleanliness. These serve as first-stage outcomes, documenting whether the program successfully shifted the knowledge and behaviors it was designed to change.

Knowledge. We ask girls questions about their knowledge of good hygienic practices. We create one index based on knowledge of overall hygiene practices (e.g., methods for cleaning water and how many times a day they should wash their hands), and another on their knowledge of menstruation and menstrual hygiene (e.g., naming hygienic menstrual products, and naming average length of a period). The sum of correct answers to the questions, normalized to have a mean of 0 and standard deviation of 1 in the control group, is used as the outcome. Our key measure combines the two indexes into an overall knowledge index using an unweighted average of Z-scores.

Behavior. We ask girls questions related to whether they engage in the types of hygienic behaviors recommended by the training curriculum. One index is again related to overall hygiene behaviors (e.g., did they use soap when washing their hands at home last week?; did they use a latrine or defecate openly the last time?), while another is related specifically to menstruation (e.g., did they use a hygienic menstrual product?; did they wash at least 3 times per day the last time they menstruated?). Our key measure combines the two indexes into an overall behavior index.

Observed cleanliness. To address concerns about experimenter demand effects leading to over-reporting of hygienic behaviors in treatment schools, enumerators directly observed cleanliness in three contexts: (i) girls themselves during the endline survey (e.g., was there visible dirt on the girl's clothes? were her hands clean?); (ii) girls' households (e.g., was there soap available at a visible hand-washing station with water? were the dishes clean?); and (iii) schools during the teacher survey (e.g., was soap available at the basins? was there a trash receptacle?). These directly observed measures provide an objective indicator of behavior, since general cleanliness was targeted by the program—even though some menstrual hygiene practices cannot be directly observed. For our main combined index, we use an equal-weighted average of Z-scores across 22 individual items.[18]

2.4 Empirical specification

In our main specification, we estimate the following regression:

$$Y_{ij} = \beta_0 + \beta_1 \text{BaseOnly}_j + \beta_2 \text{BaseYGL}_j + \mathbf{X}'_{ij}\mathbf{\Gamma} + \varepsilon_{ij} \tag{1}$$

Y_{ij} is the outcome variable measured at endline for student i in school j. For outcomes measured at the school or school-by-wave level (e.g., the attendance spot checks and register counts analyzed below), we estimate the analogous specification with the outcome and controls defined at the school or school-by-wave level rather than at the student level. BaseOnly_j is a binary variable, equalling 1 when school j is in the *Base Only* treatment and 0 otherwise. BaseYGL_j equals 1 when school j is in the *Base+YGL* treatment and is 0 otherwise. The main average treatment effects of interest are therefore given by β_1 and β_2. In some analyses, we also report a *pooled* treatment effect: the coefficient on a single indicator equal to 1 in either treatment arm, which replaces the two arm indicators in Equation 1. The tables report the two arm-specific coefficients together with the p-value of this pooled effect (the *p: Treated=0* row), while pooled magnitudes are given in the text. $\mathbf{X}_{ij}$ is a vector of controls, which always includes stratum fixed effects. In cases where it exists, it also includes Y^0_{ij}, the baseline analog of the outcome. We use post-double selection LASSO (Belloni et al., 2014) to select an additional set of controls, taking the union of the variables that predict the outcome and those that predict the treatment (see Appendix K.2). When further controls are

[18] We use equal weighting because the school subindex is based on only 3 questions that have very low control means and standard deviations, leading to very large Z-score coefficients. A combination of subindices would thus place undue weight on these 3 questions. For completeness, we report the effects on a combination of subindices in Table 1's notes.

included (e.g., wave fixed effects for spot check outcomes), we specify directly in the table notes. Standard errors are clustered at the school level unless otherwise specified.

We also test for treatment heterogeneity by (i) primary versus secondary school, (ii) school size, (iii) whether girls had started menstruating at baseline, and (iv) the positive deviance of the selected YGL. These heterogeneity analyses focus on the primary and secondary outcomes (see Appendix Figure A1).[19]

2.5 Preregistration

Our study was pre-registered on the AEA RCT registry (unique ID: AEARCTR-0008972). The intervention and sampling were completed as described in the registry. The pre-registered primary outcomes were school absenteeism and psychosocial wellbeing; pre-registered secondary outcomes included pad take-up, knowledge, exposure to discussions on hygiene/menstruation, social stigma and norms, and caregiver outcomes (mothers' knowledge and willingness to pay). Results on all pre-registered outcomes are reported in the paper regardless of statistical significance, with the exception of pad take-up and mothers' willingness to pay, which we dropped for the reasons noted below. After the pre-registration, we refined the set of key outcome variables collected at endline, motivated by qualitative insights and operational constraints.[20]

3 Results

The program was implemented with high fidelity and generated large shifts in hygiene knowledge and behavior. It also substantially improved girls' learning, despite not targeting academic skills directly. We investigate the channels underlying these learning effects, distinguishing between conventional channels (attendance, physical health, menstrual management) and psychosocial channels (stress, stigma, school social climate), and find that the evidence consistently favors the latter.

3.1 First stage: program implementation, hygiene knowledge, and behavior

3.1.1 Program implementation

The program was implemented with high fidelity to the assigned treatment conditions. First, enumerators observed improvements in hygiene infrastructure at endline in treatment schools. On average, treatment schools had 4.5 functional latrines (vs. 2.3 in control schools), 85% had at least one separate latrine for girls (vs. 23% in control), and 69% had a functional water basin (vs. 37% in control). There were no statistically significant differences between *Base Only* and *Base+YGL* schools in these outcomes. Second, menstrual pads were successfully distributed in treatment schools: 100% of directors reported a distribution (vs. 0% in control), and 67% of girls reported at endline that they

[19] We exclude the attendance outcome when examining heterogeneity by whether girls had started menstruating at baseline, since attendance is measured at the school-level and cannot be disaggregated to the individual level.

[20] We made four changes to the outcome set after pre-registration. First, we added standardized learning tests as a primary outcome after qualitative fieldwork during the midline revealed that the program appeared to be affecting classroom engagement and learning—an outcome not typically measured in menstrual hygiene evaluations and therefore not anticipated at the time of registration. Second, we adopted heart rate as an objective measure of the pre-registered "psychosocial wellbeing" outcome. Third, we elevated menstrual stigma from a secondary to a primary outcome given the intervention's explicit focus on stigma reduction. Fourth, we dropped pad take-up and willingness to pay since take-up was close to 100% in treatment schools.

had received pads (vs. 2% in control).[21] 49% of girls reported using the pads themselves; others kept them in storage, and only 1% said that a sister or mother used them. Third, directors reported that teachers received training about menstruation and hygiene in 99% of treatment schools (vs. 31% in control), and that teachers sensitized children on these topics every week in 93% of treatment schools (vs. 46% in control). Finally, the *Young Girl Leader* program was carried out successfully: when asking girls at endline whether certain girls in their school carried out sensitization activities about menstruation and health, in 97% of schools (vs. 9% in control schools) at least 20% of the girls said yes.

3.1.2 Hygiene knowledge and behavior

The base program led to substantial improvements in hygiene knowledge and behavior. The YGLs led to significant additional improvements in hygiene behavior, but not in knowledge (Table 1).

Knowledge. The program led to large improvements of 0.3–0.4 SD in girls' knowledge of both general hygiene (pooled effect: 0.41 SD, p<0.001, column 2), and of menstruation and menstrual hygiene (pooled effect 0.33 SD, p<0.001, column 3).[22] For both indexes, the point estimates are slightly larger for *Base+YGL* than *Base Only* schools, hinting that *YGLs* may have been effective at increasing hygiene knowledge, but we cannot reject equality of the same effects across arms (p of difference: 0.11 and 0.51).

Reported hygiene behavior. Girls in *Base Only* schools report engaging in more hygienic behaviors related to menstruation (p<0.001, column 6), but not more general hygiene behaviors (p=0.13, column 5).[23] By contrast, girls in *Base+YGL* schools report large improvements in both general (0.29 SD, p<0.001, column 5) and menstrual hygiene behaviors (0.56 SD, p<0.001, column 6). The coefficients are significantly larger in *Base+YGL* schools than *Base Only* schools, suggesting that YGLs play an important role in correcting the knowledge-behavior gap by embedding and promoting correct hygienic behaviors in their social networks.

Observed hygiene. We examine whether the reported improvements in hygienic behaviors translate to measures of hygiene observed directly by enumerators. While a number of behaviors targeted by the program cannot be easily directly observed (e.g., the act of hand-washing), we can observe cleanliness outcomes targeted by the program, which are not subject to reporting biases. We find that *home environments* of girls from *Base+YGL* schools are significantly cleaner (p<0.001, Panel B column 3), and treatment schools show very large improvements in observable measures of cleanliness at school, such as the availability of soap and collection of trash (p<0.001, Panel B column 4), with further improvements generated by the YGLs (p of difference: 0.02). We find no improvements in girls' personal cleanliness during surveys (pooled p=0.99, Panel B column 2), though note these measures exclude menstrual hygiene. Overall, the observed measures all show significantly stronger

[21] Lower primary school grades were not targeted to receive sanitary pad distributions. Administrative data show take-up of the vouchers among girls in grades targeted for distribution to be 91%.

[22] Boys' knowledge about menstruation, along with their exposure to communication about hygiene and menstruation, also increased substantially in treatment schools (Table A16). We find mixed results on whether knowledge and behavior changes are transmitted vertically from girls to their mothers (Table A17): mothers' menstruation knowledge improves by 0.22 SD in *Base+YGL* schools (column 2, p<0.01), but their self-reported behaviors and general hygiene knowledge are not affected (columns 1, 3, and 4).

[23] While self-reported hygiene behaviors are potentially subject to social desirability bias, we find no evidence that the effects are stronger for girls who are more likely to report socially desirable answers (Table A6).

Table 1: *Hygiene knowledge and behavior and observed hygiene*

Panel A. Self-reported measures

	(1) Combined knowledge index (Z)	(2) Hygiene knowl. (Z)	(3) Menstr. knowl. (Z)	(4) Combined behavior index (Z)	(5) Hygiene behavior (Z)	(6) Menstr. behavior (Z)
Base Only	0.414***	0.345***	0.310***	0.319***	0.123	0.360***
	(0.068)	(0.077)	(0.057)	(0.078)	(0.089)	(0.069)
Base + YGL	0.483***	0.442***	0.339***	0.550***	0.294***	0.560***
	(0.056)	(0.062)	(0.049)	(0.066)	(0.076)	(0.054)
Observations	2,167	2,167	2,167	2,256	2,167	2,256
Data source	Girls	Girls	Girls	Girls	Girls	Girls
p: Treated=0	<0.001	<0.001	<0.001	<0.001	0.001	<0.001
p: Base Only=Base+YGL	0.221	0.117	0.507	<0.001	0.022	0.002

Panel B. Enumerator-observed cleanliness

	(1) Combined observ. clean. (Z)	(2) Cleanliness: Girls (Z)	(3) Cleanliness: Home (Z)	(4) Cleanliness: School (Z)
Base Only	0.077	-0.071	0.051	1.841***
	(0.175)	(0.063)	(0.070)	(0.270)
Base + YGL	0.470***	0.044	0.173***	2.459***
	(0.138)	(0.055)	(0.052)	(0.237)
Observations	140	2,167	2,287	140
Data source	Schools	Girls	Mothers	Schools
p: Treated=0	0.013	0.990	0.010	<0.001
p: Base Only=Base+YGL	0.011	0.026	0.049	0.021

Notes: Panel A: Combined knowledge index (column 1) is a Kling-style index (unweighted average of Z-scores) combining columns 2 and 3. *Hygiene knowledge* (column 2) and *Menstruation knowledge* (column 3) are each constructed as sums of correct answers (then standardized). *Combined behavior index* (column 4) is a Kling-style index combining columns 5 and 6. *Hygiene behavior* (column 5) and *Menstruation behavior* (column 6) are each constructed as sums of recommended behaviors (then standardized). *Panel B: Combined observed cleanliness* (column 1) is an equal-weighted average of Z-scores across 22 individual cleanliness items from girls, home, and school surveys (excludes "garbage separated" due to zero variance in control). We use equal weighting rather than a Kling-style average of the three sub-indices because the school-level sub-index has very low control group variance, leading to inflated coefficients; for comparison, a Kling-style index yields coefficients of 0.93 SD for Base Only (p<0.001) and 1.53 SD for Base+YGL (p<0.001). *Cleanliness: Girls* (column 2), *Cleanliness: Home* (column 3), and *Cleanliness: School* (column 4) are each constructed using factor analysis on enumerator observations within each survey.

Panel A uses data from the Girls survey; standard errors clustered at school level in parentheses; controls include stratum fixed effects and baseline controls selected by double LASSO. Panel B column (1) and (4) use school-level data with robust standard errors; columns (2) and (3) use individual-level data with clustered standard errors. All outcomes in control group standard deviations. The bottom of each panel shows p-values for treatment effect (*Treated=0*) and equality of treatment arms (*Base Only = Base+YGL*). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

improvements in cleanliness for *Base+YGL* than for *Base Only* schools.

3.2 Learning effects

Having established that the program successfully shifted its targeted behaviors, we now turn to its effects on learning, which was not directly targeted by the intervention.

Table 2: *Education: learning*

	(1) Combined index: Achievement + cognitive tests (Z)	(2) Achievement tests only (Z)	(3) Cognitive tests only (Z)	(4) Official GPA (Z): girls	(5) Official GPA (Z): boys	(6) Progressed one class (=1)
Base Only	0.208***	0.248***	0.142*	0.160**	0.075	0.112***
	(0.072)	(0.064)	(0.075)	(0.079)	(0.071)	(0.040)
Base + YGL	0.117**	0.131**	0.086	0.086	0.085	0.076**
	(0.053)	(0.051)	(0.058)	(0.064)	(0.069)	(0.036)
Observations	2,167	2,167	2,167	7,586	6,993	2,256
Data source	Girls	Girls	Girls	GPAs	GPAs	Girls
Control mean	0	0	0	0	0	.515
Control SD	1	1	1	1	1	.5
p: Treated=0	0.004	0.001	0.052	0.080	0.195	0.011
p: Base Only=Base+YGL	0.169	0.019	0.432	0.276	0.875	0.206
p: Base Only (Girls)=Base Only (Boys)				.184		
p: Base+YGL (Girls)=Base+YGL (Boys)				.866		
p: Treated (Girls)=Treated (Boys)				.702		

Notes: In column (1), all learning tests elicited at endline (math fluency, reading fluency, listening, comprehension, digit span forward and backward) are combined into one index using factor analysis, which weights each test by how strongly it loads onto a common learning factor. In columns (2) and (3), sub-indices are constructed using factor analysis: column (2) combines reading and math fluency tests (academic achievement); column (3) combines listening, comprehension, digit span, and reverse digit span tests (cognitive ability). The official school GPAs (columns 4 and 5) were collected from grades notebooks provided by teachers and include all students in the class. *Progressed one class* (column 6) = 1 when the girl moved up one grade level in between the baseline and endline survey, and is coded as 0 if the girl is in the same grade level at endline or is not enrolled at school at endline. Standard errors are clustered at the school level and are in parentheses; controls include stratum fixed effects. Columns (1), (2), (3) and (6) also include the baseline controls selected by double LASSO. The bottom of the table denotes the p-values on the test of being treated *Treated=0* and the equality of the coefficients *Base Only = Base+YGL*. The cross-gender p-values test whether treatment effects differ between girls and boys (e.g., *Base Only (Girls)=Base Only (Boys)* tests if the Base Only effect on girls' GPA equals the effect on boys' GPA). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Learning. The base program substantially improved girls' academic and cognitive skills at endline (Table 2). The pooled effect of the *Base Only* and *Base+YGL* arms on the combined index (capturing tests of academic skills, memory, and attention) was 0.15 SD (p=0.004, column 1).[24] The effects are strongest for academic achievement tests (language and math fluency) (0.13–0.25 SD, pooled p=0.001, column 2), while effects are smaller and less consistently significant for tests capturing memory, attention and other cognitive skills (0.09–0.14 SD, pooled p=0.05, column 3; see Table A8 for results on each individual test).

These learning gains are economically significant. The math and language effects place the *Base Only* program at the 75th percentile and *Base+YGL* at the 68th percentile of learning programs evaluated in low- and middle-income countries (Muralidharan, 2017; Evans & Yuan, 2022a), and at the 90th and 80th percentiles respectively among programs reaching at least 5,000 students. These gains come despite the intervention not directly targeting learning. For the *Base Only* program,

[24] The effects are not driven by (i) outliers, since the learning test results shifted fairly uniformly across the distribution (Figure A7), or (ii) baseline imbalance, since although we did not elicit learning tests at baseline, the number of years of school delay (a proxy for academic performance) was balanced at baseline (Table A2), and the effects are essentially unchanged when controlling for schools' pre-treatment (2021) official exam pass rates (Table A9).

the academic improvements correspond to an estimated gain of 0.31 Learning-Adjusted Years of Schooling (LAYS); the corresponding gain for *Base+YGL* is 0.16 LAYS.[25]

The improvement in learning translated into higher official GPAs for girls in treatment schools (pooled effect: 0.11 SD, p=0.08; Table 2, column 4). Treatment also substantially increased the probability that girls progressed to the next grade rather than repeating or dropping out, from 51% in control schools to 60% in the pooled treatment schools (9 p.p., p=0.01, column 6). This faster progression, together with higher rates of school transfers, likely biases girls' measured GPA downward, making our estimate a lower bound on the true effect on GPA for girls.[26]

The effect on boys' official GPA is positive but not significant (pooled p=0.20, column 5), and we cannot reject that girls' and boys' effects are equal (p=0.70). However, administrative data on exit exam pass rates suggest a meaningful gender gap after the first year: girls' pass rates rose in 2022, while boys' pass rates did not (p of difference: 0.05; Table H1, see Appendix H). The observed GPA gender gap likely understates the true difference because girls progress through grades faster and transfer schools more often than boys.

The improvements in learning across the different measures appear to be driven by the base program rather than the additional *YGL* component. For all measures, the point estimate is larger for *Base Only* schools than *Base+YGL* schools, although the difference between the two arms is only significant for the subset of academic achievement tests (p=0.02). This pattern is consistent with the base program's reductions in stigma and stress being sufficient to generate the learning effects, with the *YGL* program adding no further learning benefits on average. Below, we investigate the channels underlying these learning effects (Sections 3.3 and 3.4).

Robustness. The learning results, along with the other results on primary outcomes, are robust to (i) using randomization inference to calculate p-values (Table B1), (ii) using sampling weights or Lee bounds to adjust for attrition (Table B4), and (iii) adjusting for multiple hypothesis testing (Table B8). Additional analysis shows that the results are not driven by (iv) composition effects of who remains in school (Appendix B.4), (v) geographic spillovers (Appendix B.2), or (vi) the program crowding in alternative school-level investments such as additional teachers or spending (Table B7). See Appendix B for details.

3.3 Conventional channels

The learning improvements were unexpected: the program targeted hygiene, not academic skills. Why would sanitary pads, latrines, and sensitization sessions improve math and reading scores? We use measured outcomes to distinguish two broad classes of channels: conventional channels (increased attendance, improved physical health, better menstrual management) and psychosocial channels (reduced stress, lower stigma, improved school social climate). We find little evidence for conventional channels and consistent evidence for psychosocial ones.

[25] This calculation assumes effects last 1 year, and compares to a "high-performing benchmark" of 0.8 SD improvement in academic performance per year of schooling, following Angrist et al. (2025). The base program cost \$92 per girl, 55% of which was for sanitation infrastructure; its non-infrastructure components cost \$41 per girl. Even at full cost, these gains compare favorably to programs specifically targeting academic outcomes.

[26] Faster progression biases measured GPA down because grading is tougher in higher grades. Transfers bias it down further because girls who progress out of the final year of a cycle often move to a new school and drop out of the GPA sample, leaving behind a negatively selected set of stayers (Table B5, Table B6). See Appendix B.4.

3.3.1 *Attendance*

Table 3: *Education: attendance*

	(1) # girls present (spot-checks)	(2) # girls present (registers)	(3) # girls enrolled (registers)	(4) P(absent|enrolled)	(5) Enrolled (=1)	(6) Absent in last 3 months (=1)
Base Only	-0.457	-0.849	2.143	0.016	-0.019	-0.009
	(3.403)	(2.854)	(2.554)	(0.015)	(0.024)	(0.030)
Base + YGL	-1.361	-3.080	1.005	0.021*	-0.021	-0.038
	(2.988)	(2.715)	(2.291)	(0.011)	(0.022)	(0.029)
Observations	698	34,455	698	698	2,256	1,824
Data source	Spot checks	Registers	Registers	Spot checks	Girls	Girls
Control mean	68.2	64.8	75.4	.107	.826	.442
Control SD	51.9	95.1	55.0	.103	.38	.497
p: Treated=0	0.720	0.332	0.530	0.071	0.330	0.285
p: Base Only=Base+YGL	0.696	0.421	0.552	0.722	0.926	0.256

Notes: Column 1: *# girls present* denotes the number of girls physically present at the school; data is at the *wave* × *school* level with wave fixed effects and controls for the outcome at baseline. Column 2: *# girls present* denotes how many girls were present according to the registers on that day; data is at the *day* × *school* level with date (day) fixed effects and controls for the outcome at baseline. Column 3: *# girls enrolled* is the number of girls enrolled according to the official registers in that wave of the spot-check. Column 4: *P(absent|enrolled)* is the proportion of enrolled girls absent during unannounced spot checks (i.e., 1 − present/enrolled), at the *wave* × *school* level with wave fixed effects and controls for the baseline proportion absent. For the last two outcomes, girls were asked about their school attendance at endline: *Enrolled (=1)* (column 5) refers to whether the girl was enrolled in the 2022-2023 academic year, and *Absent in last 3 months (=1)* (column 6) is an indicator of whether the girl had been absent in the last 3 months, conditioned on being enrolled. Standard errors are clustered at the school level and are in parentheses; controls include stratum fixed effects. LASSO-selected controls are included for columns 5 and 6. The bottom of the table denotes the p-values on the test of being treated *Treated=0* and the equality of the coefficients *Base Only = Base+YGL*. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

School attendance. There were no improvements in school attendance driven by the interventions (Table 3). The program had no detectable effect on (i) the number of girls physically present at school during the unannounced spot checks (pooled p=0.72, column 1); (ii) the number of girls physically present in the months preceding the spot checks, according to digitized official school registers (pooled p=0.33, column 2); (iii) the number of girls enrolled in the endline academic year 2022-23, according to those school registers (pooled p=0.53, column 3); or (iv) the proportion of enrolled girls absent during spot checks (column 4). When asking girls at endline, we find no effect on whether girls reported being enrolled in school (pooled p=0.33, column 5), or whether they were absent in the last 3 months, conditional on being enrolled (pooled p=0.29, column 6). There was also no effect on boys' attendance (Table A10).

The null effects on attendance are relatively precise: our standard errors imply a minimum detectable effect of 0.07 SD (9.1 p.p., using 80% power) on the pooled effect in column 2.

A plausible explanation for the lack of effects on school attendance is that absenteeism and dropout in this low-income context are driven by other hard economic factors that act as binding constraints. The overall rate of dropout is high (19% over the two-year period between baseline and endline). Directors reported that the main reasons for female dropout were economic factors, including (i) girls having to find a job (38% of schools), (ii) not being able to pay school fees (44% of schools), or (iii) not having enough food to eat to go to school (13% of schools).

3.3.2 *Physical health*

Physical health. There are no consistent measurable impacts on self-reported health outcomes (Table 4, pooled p on combined index=0.59, column 1). Our measure of general health based on the Pediatric Quality of Life index shows no significant effect (pooled p=0.90, column 2). There are also no clear decreases in the symptoms of urinary tract infections (UTIs) (pooled p=0.42, column 3).[27] While we cannot rule out that the overall null effects are partly driven by measurement noise on self-reported health outcomes (given that we do find impacts on objective measures of observed hygiene), the results nevertheless suggest relatively small effects on health that are likely insufficient to explain the impacts on learning. Our standard errors imply a small minimum detectable effect of 0.12 SD on the combined index, so the null is precise.

Table 4: *Physical health*

	(1) Combined index: Physical health (Z)	(2) General health (PedsQL) (Z)	(3) Urinary tract health (Z)
Base Only	-0.078	-0.037	-0.089*
	(0.057)	(0.060)	(0.054)
Base + YGL	0.004	0.009	-0.004
	(0.045)	(0.048)	(0.043)
Observations	2,256	2,256	2,256
Data source	Girls	Girls	Girls
p: Treated=0	0.588	0.902	0.416
p: Base Only=Base+YGL	0.130	0.392	0.089

Notes: Physical health index (column 1) is a Kling-style combined index (unweighted average of Z-scores) of *General health* and *UTI health*. *General health (PedsQL)* (column 2) is constructed using factor analysis on Pediatric Quality of Life Inventory (PedsQL) questions about experiencing symptoms of physical pain and general health during the last month; a higher value indicates *better* health. *UTI health* (column 3) is a simple average (then standardized) of questions about urinary symptoms during the last month (a higher value indicates fewer symptoms, i.e., better health). Standard errors are clustered at the school level and are in parentheses; controls include stratum fixed effects and the baseline controls selected by double LASSO. ** $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.*

3.3.3 Menstrual management and hygiene

We examine two additional channels through which the program could have improved learning: better menstrual management increasing time in class, and improved hygiene infrastructure and behavior. While the program affected both, evidence suggests neither is the primary driver of the learning effects.

Menstrual management and time in class. Better menstrual management might facilitate learning by increasing time spent in class. Girls with access to private toilets or sanitary pads may need to leave class less frequently, affecting the “intensive margin” of attendance. In line with this channel, we find treatment reduced menstruation-related disruptions among the 1,053 girls who had reached menarche at baseline (Table A28): girls were less likely to report being absent due to menstruation (3.1 p.p., p=0.11, column 2), leaving school during the day (3.8 p.p., p=0.05, column 4), or being unable to fully participate in school activities (3.8 p.p., p=0.10, column 5).[28]

[27] While the point estimate for *Base Only* schools on urinary tract health indicates a slight worsening, and is significant at the 10% level, the effect is not robust to alternative variable constructions, including the threshold for indicating acute cystitis that is recommended by the survey module creators (Table A11, see Alidjanov et al., 2014).

[28] We use the baseline measure of menarche because reduced stigma in *Base+YGL* schools made girls more likely to reveal menarche at endline, making it endogenous (3.5 p.p., p=0.045).

While this mechanism may have driven part of the learning effects, evidence suggests it is unlikely to be the major driver. First, overall disruption levels driven directly by menstruation are low: only 11% of menstruating girls in control schools report ever having left class due to menstruation. Second, learning effects are not larger for girls who had reached menarche at baseline (Table A13) or between baseline and endline (Table A29). Third, learning effects are no weaker for girls who already used a hygienic menstrual product at baseline (p=0.61 for the interaction), even though these girls had the least to gain from improved product access.

Infrastructure and hygiene. The timing of impacts suggests hygiene infrastructure is unlikely to be the main driver of learning. Girls were more likely to progress to the next grade during the study period (Table 2, column 6), implying learning had started improving by the end of the first school year—before infrastructure construction was completed. A comparison across treatment arms reinforces this conclusion: observed hygiene (cleanliness) improved significantly more in *Base+YGL* schools than in *Base Only* schools, yet learning did not. However, this does not rule out that infrastructure and sanitary pads could have driven learning through *psychosocial* channels (e.g., knowing infrastructure was coming could increase motivation, and access to pads can reduce stress about leaks).

3.4 Psychosocial channels

The evidence points to psychosocial channels as the most likely drivers of the learning improvements. We document reductions in objectively-measured stress, large declines in menstrual stigma, and improvements in the school social environment.

3.4.1 Heart rate and stress

Heart rate. The interventions lead to a significant decrease in an objective biological measure of anxiety or stress: girls' heart rate during the endline survey. Heart rate is on average 2.3 bpm lower for girls in treatment schools (Figure 2; pooled p=0.01; effect size: 0.12 SD when using the between-girl standard deviation), and is consistently so throughout the entire survey duration (panel B).[29] The effects are very similar across *Base Only* and *Base+YGL* schools (p of difference: 0.99). To contextualize this 2.3 bpm reduction, it is equivalent to the heart rate difference associated with a large (0.82 SD) increase in baseline household wealth index (Table C5), and corresponds to about one quarter of the 10 bpm increase typically induced by extreme laboratory-induced psychological stressors (Kudielka et al., 2004; Seddon et al., 2020).[30]

We see two main interpretations of this reduction in heart rate (that are not mutually exclusive). First, it could capture a *general* improvement in stress or anxiety levels among the girls. Given the evidence that reductions in stress and anxiety can relax cognitive bandwidth constraints and thereby

[29]In Appendix C.2, we explore the robustness of the heart rate treatment effects, showing that (i) the results are robust to dropping outliers (Table C3); (ii) the results do not appear to be driven by girls being more familiar with enumerators or program technicians in treatment schools (Table C4) or by enumerator assignment (Table A12); and (iii) the strong correlations of heart rate with baseline wealth, endline depression, and endline bullying suggest that the results are not driven by measurement error (Table C5).

[30]Heart rate is one of the best cross-person predictors of anxiety (Zhang et al., 2025). *A priori*, however, the effects on heart rate could also reflect improvements in physical health (e.g., physical fitness). The null effects on the general health index point away from a physical-health explanation. Further, in the exploratory second endline, a 1 SD increase in heart rate is associated with 0.09 SD higher reported stress about exams and 0.03 SD higher reported stress about menstruation (p=0.001 and 0.06, respectively).

improve decision-making in a variety of domains (Haushofer & Fehr, 2014; Schilbach et al., 2016; Kaur et al., 2025), this therefore points to an important potential mechanism that could underlie the effects on learning. A reduction in stress or anxiety also represents an important improvement in welfare in and of itself.

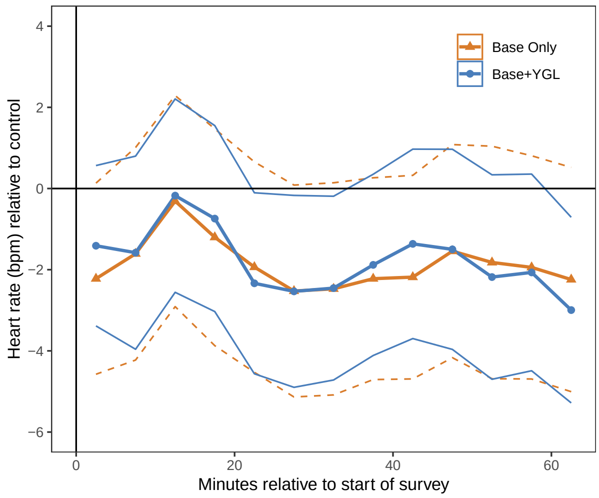
Second, the reduction in heart rate could be driven by a reduction in stress or anxiety *specifically* when discussing menstruation, thus capturing a reduction in menstrual stigma. We find mixed evidence regarding this channel, since heart rate does not seem to be differentially affected when actually speaking about menstruation during the endline (Table C7, columns 1 and 2), although there is some evidence that the treatment effect grows larger as the survey progresses (Table C7, column 4).[31]

Figure 2: *Stress: heart rate during endline survey*

(a) *Regression*

	Heart rate	
	(bpm)	(Z)
Base only	−2.346**	−0.116**
	(1.133)	(0.056)
Base+YGL	−2.338**	−0.115**
	(0.998)	(0.049)
N Obs	9,105,420	9,105,420
N girls	1904	1904
Control mean	87.6	0.0
Control SD (between)	20.3	1.0
p : Treat = 0	0.014	0.014
p : Base=Base+YGL	0.993	0.993

(b) *Dynamics*



Notes: Panel (a) shows the effect of the intervention on heart rate for girls in the endline survey. The first column reports effects in beats per minute (bpm); the second column reports effects in standard deviations (Z), calculated by dividing heart rate by the between-girl control group standard deviation. All girls who were included in the baseline sample and for whom heart rate data is available are included. *Time window*: observations are included until up to 120 minutes after the start of the survey. The unit of observation is the *girl × second (relative to the start of the survey)*. Results are robust to using a girl-level regression (Table C3, column 1). The regression controls for 30-second window fixed effects, stratum fixed effects, and variables selected by double LASSO. Standard errors are clustered at the school level and are in parentheses. Control SD (between) indicates the standard deviation of the girl-level average heart rate. Panel (b) shows dynamics over the course of the first 60 minutes of the survey (when almost all girls are still involved in the survey, see Appendix Section C.1). It plots the treatment effect relative to control for every 5 minute window (e.g., 0-5 minutes, 5-10 minutes) controlling for LASSO and stratum FEs, clustering by school, with 95% confidence intervals. ** $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.*

Stress as a channel for learning. The reduction in heart rate provides objective evidence that the program reduced anxiety and stress, which could plausibly improve learning by relaxing cognitive bandwidth constraints. In this context, girls face multiple sources of stress: high exam failure rates, economic and social pressures at home, and the challenges of managing menstruation without adequate information or resources. While only a subset of girls directly faced practical menstrual management difficulties on a given day, any related incidents could influence classroom dynamics and psychologically affect all girls, adding a gender-specific component to already high baseline stress

[31] There is also some suggestive evidence that effects on heart rates are larger for girls who had started menarche by baseline, although we do not have enough power to detect a significant interaction effects (Table A13), consistent with both the *generalized* and *menstruation-discussion-specific* reduction in stress interpretations. But effects are not larger among girls reporting menstruating on the day of the endline survey, suggesting the stress of discussing menstruation during menstruation itself is not a key driver (Table C6).

levels. By addressing this stress factor, the intervention plausibly lifted some of these bandwidth constraints, allowing girls to concentrate more effectively while at school.

Self-reported mental health. While we find impacts on objectively-measured stress, we do not find effects on self-reported depression based on the CESD index (p=0.54) or on self-esteem (p=0.78), suggesting the intervention primarily reduces anxiety rather than depression (Table A15).

3.4.2 Menstrual stigma

Here, we document the program's effects on menstrual stigma. We first discuss the survey measures (Table 5), which allow us to examine how the program changed different dimensions of stigma. We then turn to revealed-preference lab-in-the-field measures (see Appendix D) and enumerator observations of shame, which provide more objective evidence of changes in stigma.

Survey measures. The base program significantly improved a wide range of self-reported stigma measures, with YGLs generating additional gains (Table 5). In *Base Only* schools, girls reported a 0.42 SD improvement in the combined stigma index (p<0.001, column 1), reflecting 0.20–0.47 SD improvements (all p<0.01) in: (i) hearing others discuss menstruation (column 2); (ii) willingness to speak about menstruation (column 3); (iii) progressive personal attitudes (column 4); (iv) perceived progressiveness of others' attitudes, i.e., norms (column 5); and (v) shame and guilt in response to menstruation-related vignettes (column 6).[32] *Base+YGL* schools showed significantly larger improvements across all measures (0.45–0.73 SD), with a combined index effect of 0.80 SD (p<0.001). These results indicate that the program substantially reduced menstrual stigma as perceived by girls, with YGLs amplifying these effects.

Observed measures: lab-in-the-field and enumerator observations. We use lab-in-the-field measures to test whether (i) stated reductions in stigma also translate to changes in revealed-preference behaviors; and (ii) some of the program's effects are *particularly* strong for stigmatized topics, as compared to unstigmatized hygiene behaviors (e.g., hand-washing). Overall, the combined index of revealed-preference endline measures (Table 6, column 1, pooled effect: 0.60 SD, p<0.001) provides clear observed evidence of a reduction in menstrual stigma, and suggests it is of similar magnitude with and without the YGL.

[32]Girls typically *underestimate* how progressive others' attitudes are (Figure A18), a form of "pluralistic ignorance" (Katz et al., 1931; Bursztyn et al., 2020) likely driven by reluctance to discuss menstruation openly. Across all treatment groups, these misperceptions about students' and teachers' attitudes are large. The program works through two channels: it shifts actual attitudes and updates girls' beliefs about others' attitudes.

Table 5: *Menstrual stigma: survey measures*

	(1) Combined index: Menstrual stigma (Z)	(2) Received info. (Z)	(3) Willingness to speak (Z)	(4) Attitudes (Z)	(5) Norms (Z)	(6) Shame response to vignettes (rev.) (Z)
Base Only	0.421***	0.202***	0.252***	0.469***	0.280***	0.308***
	(0.068)	(0.078)	(0.076)	(0.067)	(0.060)	(0.059)
Base + YGL	0.805***	0.539***	0.730***	0.725***	0.473***	0.445***
	(0.058)	(0.065)	(0.065)	(0.060)	(0.049)	(0.055)
Observations	2,256	2,256	2,250	2,256	2,167	2,165
Data source	Girls	Girls	Girls	Girls	Girls	Girls
p: Treated=0	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
p: Base Only=Base+YGL	<0.001	<0.001	<0.001	<0.001	<0.001	0.008

Notes: Combined index (column 1) is a Kling-style index (unweighted average of Z-scores) combining the five sub-indices in columns 2–6. Each sub-index is constructed using factor analysis: *Received info* (column 2) combines questions about how many times girls heard peers, family, or teachers speak about menstruation in the last 3 months. *Willingness to speak* (column 3) combines questions about how likely girls are to discuss topics related to menstruation. *Attitudes* (column 4) combines Likert-scale questions about girls' attitudes towards menstruation. *Norms* (column 5) combines questions about the menstruation-related attitudes girls think their teachers and mothers have. *Shame response to vignettes (rev.)* (column 6) combines emotional responses to hypothetical menstruation-related scenarios; a higher value indicates less shame. All outcomes are in control group standard deviations. Standard errors are clustered at the school level (in parentheses); controls include stratum fixed effects and baseline controls selected by double LASSO. The bottom of the table shows p-values for pooled treatment effect (*Treated=0*) and equality of coefficients (*Base Only = Base+YGL*). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: *Menstrual stigma: observed measures*

	Aggregated	Information supply		Information demand		Dictator game		Observed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Combined menstrual stigma (lab, rev.) (Z)	Explanation Chose menstr. (=1)	Anonymous Chose menstr. (=1)	Midline: P(chose menstr.)	Endline: P(chose menstr.)	Girls: Prop. cookies menstr.-handw.	Boys: Prop. cookies menstr.-handw.	Observed shame in (survey, rev.) (Z)
Base Only	0.639***	-0.001	0.046**	0.037**	0.012	0.023**	0.027*	0.296***
	(0.177)	(0.026)	(0.021)	(0.019)	(0.016)	(0.010)	(0.015)	(0.061)
Base + YGL	0.585***	0.046**	0.034	0.034**	0.023	-0.010	0.023*	0.229***
	(0.160)	(0.021)	(0.021)	(0.015)	(0.016)	(0.010)	(0.013)	(0.053)
Observations	140	1,634	1,634	1,972	2,161	2,161	1,325	2,256
Control mean	0	.208	.291	.479	.571	.002	-.01	0
Control SD	1	.406	.455	.302	.274	.222	.206	1
p: Treated=0	<0.001	0.125	0.047	0.013	0.214	0.954	0.040	<0.001
p: Base Only=Base+YGL	0.710	0.059	0.522	0.857	0.420	<0.001	0.772	0.184

Notes: Column (1) presents a combined index of six observed endline stigma measures (columns 2-3, 5-8). The index is constructed by: (i) standardizing each measure to a Z-score using the control group mean and SD; (ii) collapsing each Z-scored measure to the school level; (iii) averaging the six school-level Z-scores into a Kling-style index; and (iv) re-standardizing the combined index so that the control group has mean zero and SD of one. Columns (2)-(3) present results from the *information supply experiment* where girls were informed they would need to explain or show (anonymously) a video on a topic in front of their schoolmates. They were asked to choose between a menstruation-related topic and a broader hygiene-related topic. *Explanation, chose menstr.* (column 2) indicates whether the girl chose to provide an in-person explanation about a menstruation-related topic. *Anonymous, chose menstr.* (column 3) indicates that the girl chose to anonymously show a video about a menstruation-related topic. Columns (4)-(5) display results from the information demand experiment where girls had to choose between envelopes containing multiple topics. *Information demand: P(chose menstr.)* is the proportion of times out of 4 that girls chose to learn about a menstruation-related topic instead of a non-stigmatizing topic. Column (4) is from midline and column (5) is from endline. Columns (6)-(7) show results from a dictator game. Participants (girls in column 6, boys in column 7) decided how many cookies to share with a girl after watching a video of that girl providing either a menstruation or hand-washing-related explanation. The outcome *Dictator game: Prop. cookies shared, menstr.–handw.* is the within-person difference in the proportion of cookies shared: (proportion shared with girl who provided information on menstruation) - (proportion shared with girl who provided information on hand-washing). Positive values indicate more sharing with girls talking about menstruation. Column (8) presents *Observed shame in survey (rev.)*, an index constructed from enumerator observations about the girls' behaviors during the menstruation survey questions. A higher value indicates less shame in their behavior.

In columns (2)-(8), standard errors are clustered at the school level and are in parentheses; controls include stratum fixed effects. Controls selected by double LASSO are included for columns (4)-(5) and (8), as we have baseline measures for those samples. Column (1) is estimated at the school level (N=140) with robust standard errors and stratum fixed effects. Controls in columns (2)-(3) include fixed effects for choice order. The bottom of the table denotes the p-values on the test of being treated *Treated=0* and the equality of the coefficients *Base Only = Base+YGL*. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

In the *information supply experiment*, girls chose between explaining a menstruation-related topic or an unstigmatized topic in front of their schoolmates. Girls in *Base+YGL* schools were 4.6 p.p. (22%, 0.11 SD) more likely to choose menstruation topics (Table 6, p=0.03, column 2), while there was no effect in *Base Only* schools (Table 6, p=0.98, column 2; p of difference = 0.06). YGLs are thus effective at making girls more willing to supply information about menstruation openly, indicating a reduction in taboo. The additional effect of YGLs is specific to a choice involving speaking in front of others (rather than anonymously delivering information, column 3), in line with a reduction in stigma associated with lower social image concerns.

In the *information demand experiment*, girls are more likely to ask to open an informative envelope about menstruation than about another topic at midline (+3.7 and +3.4 p.p. in *Base Only* and *Base+YGL* schools, pooled p=0.013, column 4). This is no longer significant at endline, however, possibly because treatment girls may have remembered the answers from the previous year and thus reduced their demand for information (column 5).

We further find evidence of an increased prosocial response to girls speaking about menstruation in the *dictator games* (columns 6 and 7). Girls in *Base Only* schools, and boys in all treatment schools, share relatively more cookies with a girl who has spoken about menstruation compared to one who has spoken about unstigmatized hygiene. This suggests weaker sanctions for speaking openly, in line with reductions of menstrual stigma. Finally, enumerators report fewer visual indicators of shame when girls discuss menstruation in the endline survey (pooled effect: 0.25 SD, p<0.001), e.g., they were less likely to look at the floor when discussing the subject.

Attitudes towards menstruation also improve among others in girls' networks: male classmates' attitudes become more progressive (pooled effect: 0.35 SD, p<0.001), and mothers' attitudes improve by 0.34 SD (p<0.001; see Table A16 and Table A17).

Summary. Both treatment arms substantially reduced menstrual stigma among girls and others in their networks. This is confirmed by survey measures, objective lab-in-the-field exercises, and enumerator observations. While self-reported measures show larger effects in *Base+YGL* schools (0.80 vs. 0.42 SD), the objective measures show similar improvements in both arms. This discrepancy suggests either that survey measures capture additional nuanced effects that objective measures miss or that some of the additional self-reported gains from YGLs reflect reporting bias. We find no evidence that overall treatment effects are driven by social desirability bias (Table A6), but this does not rule out differential reporting across arms.

Stigma as a channel for learning. Five empirical patterns support the interpretation that reductions in menstrual stigma are a key driver of the learning improvements documented in Section 3.2.[33] First, menstrual stigma predicts learning in the control group: girls with higher stigma score ap-

[33] A correlational mediation exercise is also broadly consistent with this interpretation: controlling for self-reported menstrual stigma strongly attenuates the learning treatment effects (Table A26). We place limited weight on this exercise, since it does not yield causal estimates and does not cleanly separate channels: controlling for the objectively-measured stigma index attenuates the effect less, while controlling for hygiene knowledge attenuates it at least as strongly. Controlling for physical health does not attenuate the treatment effect at all, and controlling for school social climate leaves the effect largely unchanged, likely because social climate improvements are concentrated in secondary schools while learning gains occur in both. The intervention's theory of change suggests that many of these factors move simultaneously, which further cautions against overinterpreting these correlations. Cross-sectional correlations between learning and the main outcome indices further illustrate the association between menstrual stigma and learning (Table A27).

proximately 0.2 SD lower on learning tests (Table A27), consistent with stigma acting as a binding constraint. Second, learning effects are stronger for girls who report menstruating on the day of the survey (Table A14), suggesting that girls for whom menstruation is most salient benefit most from destigmatization. Third, the magnitude of the intervention's effects on stigma (0.42–0.80 SD) makes stigma a plausible first-order channel for knock-on impacts on learning. Fourth, the intervention had more modest impacts on boys' learning (Table 2, column 5), consistent with a gender-specific mechanism, though the girl–boy difference is imprecisely estimated. Fifth, the speed of girls' progression to the next grade (detectable after only 5 months of intervention) parallels the rapid effects on menstrual stigma (Appendix E). Destigmatization could improve learning through multiple pathways: by reducing the anxiety and shame associated with hiding menstruation at school, consistent with theories of "minority stress" (Frost & Meyer, 2023); by increasing girls' confidence to participate in class discussions, since the repeated practice of speaking openly about a taboo topic may generalize to other subjects; and by contributing to reductions in bullying and improvements in peer relationships.

3.4.3 School social climate

Table 7: *School social climate*

Outcome group:		Teasing and bullying (reversed)				Network connection		
	(1) Combined index: school social climate (Z)	(2) Teasing (Z)	(3) Teasing about menstr. (Z)	(4) Bullying (Z)	(5) Teacher-report Combined (Z)	(6) Overall (Z)	(7) w/ Girls only (Z)	(8) w/ Boys only (Z)
Base Only	0.053	-0.108	-0.136**	0.083*	-0.067	0.174**	0.071	0.205**
	(0.071)	(0.066)	(0.069)	(0.046)	(0.056)	(0.086)	(0.076)	(0.084)
Base + YGL	0.175***	-0.020	-0.270***	0.069*	-0.028	0.267***	0.183***	0.266***
	(0.057)	(0.058)	(0.061)	(0.040)	(0.052)	(0.074)	(0.062)	(0.076)
Observations	2,167	2,167	2,167	2,167	3,637	2,167	2,167	2,167
Data source	Girls	Girls	Girls	Girls	Teachers	Girls	Girls	Girls
p: Treated=0	0.013	0.386	<0.001	0.062	0.434	0.001	0.012	<0.001
p: Base Only=Base+YGL	0.046	0.109	0.040	0.735	0.391	0.186	0.074	0.375

Notes: All teasing and bullying measures are reversed so that higher values indicate less teasing/bullying (better outcomes). Column 1 (*Combined index: school social climate*) is a Kling-style index (unweighted average of Z-scores) combining columns 2 (teasing), 4 (bullying), and 6 (overall network). Columns 2–4 are each constructed using factor analysis on girls' self-reports about teasing and bullying: column 2 includes only teasing; column 3 includes only menstruation-related teasing; column 4 includes only severe bullying. Column 5 uses factor analysis on teacher reports about teasing and bullying in their class (each teacher reports on 3 randomly selected girls; observations are at the *teacher* × *girl* level). Table A19 shows additional results including peer reports. Columns 6–8 are constructed using factor analysis on network questions about girls' awareness, friendship, interaction, cooperation, and communication with school peers: column 6 refers to all peers; column 7 to female peers only; column 8 to male peers only. Standard errors are clustered at the school level (in parentheses); controls include stratum fixed effects. All columns except column 5 have baseline controls selected by double LASSO. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We find evidence that the program generated improvements in the overall social climate at schools (Table 7). Teachers, parents, and students reported changes of these kinds during qualitative work that took place 1 year after the start of the program, and these findings are corroborated by two outcomes at endline.

First, girls report reductions in bullying. While there is no measurable effect on light teasing (Table 7, pooled p: 0.39, column 2), severe bullying falls by 0.08 SD (pooled p=0.06, column 4).[34] This reduction appears to reflect improved student relationships rather than purely reduced menstrual stigma: although overall teasing remains constant (column 2), its composition shifts towards menstruation-related content (0.23 SD increase, p<0.001, column 3). The program's focus on menstruation likely increased both actual teasing about menstruation and girls' willingness to report it. The increase in menstruation-related teasing comes entirely from boys (Table A21, column 5), though from a low base—reported teasing about menstruation remains below 12% in all schools.

Second, the program led to a significant increase in a measure of network integration among students (Table 7, column 6, pooled p<0.001), based on asking girls how many social connections they have of different types at endline (e.g., how many friends, people they play with, people they share with, people they do projects with). Social connections with both girls and boys increase significantly.[35]

The effects on social climate are significantly larger in *Base+YGL* schools than in *Base Only* schools (p=0.046), with the latter showing no significant impact on the combined index (p=0.45). However, we do see the reduction in severe bullying and improvement in social networks in *Base Only* schools, suggesting that these could partially underlie some of the improvements in learning and reductions in stress in these schools, e.g., by reducing the stress associated with bullying and permitting better concentration in schools. For *Base+YGL* schools, we later show that improvements in social climate are particularly driven by secondary schools, where effective "positive deviant" YGLs could be found more easily (Section 3.6).

One reason the program may have improved school climate was its explicit encouragement of student collaboration on hygiene tasks (Section 2.1.1), e.g., cleaning classrooms, toilets, and the courtyard. In line with qualitative fieldwork reports of increased "solidarity" among students, boys in treatment schools increased their cleaning collaboration substantially (pooled effect: 0.41 SD, p<0.01, Table A16, column 5). Girls, by contrast, increased cleaning only in *Base+YGL* schools (β=0.23 SD, p=0.001, Table A20, column 8). Thus, girls may have benefited from improved peer collaboration in all treatment schools, but in YGL schools they took on additional cleaning duties that may have crowded out a focus on academic performance and explain weaker learning results.

Teachers. Changes in teacher academic effort do not appear to drive the learning effects: teacher absenteeism rates are unaffected (Table A30, column 4, p=0.44), and we find no impacts on teaching practices in the second endline (Table F1). However, the program generates large improvements in teachers' progressive attitudes towards menstruation (0.68 SD, p<0.001) and smaller improvements in menstruation knowledge (0.16 SD, p=0.08). Teachers who became more attentive to girls' menstruation concerns, more responsive to bullying, and generally supportive may have contributed

[34] Results are consistent across girls' self-reports and peer reports about severe bullying. Teachers report directionally similar patterns, but the effects are not statistically significant (Table A19). Some bullying and network questions (which reference "classmates" or ask about experiences "in the last month") may be less applicable to the 372 girls who had dropped out by endline. We still prioritize results on the full sample of girls to avoid sample selection concerns, but results are similar when restricting to girls still enrolled at endline (Table A20).

[35] Parts of the measured effects are driven by large increases in the number of peers girls speak with specifically about menstruation (Figure A22). But even after removing this question from the index, there are still 0.13-0.17 SD increases in the network integration index with boys (pooled p-value = 0.03), suggesting a more general improvement in social dynamics between girls and boys. Measures of doing homework together with boys and girls also significantly improve.

to the improved school environment.

School climate as a channel for learning. The improvements in social climate documented above could plausibly contribute to the learning effects. Reduced bullying and stronger peer networks can improve learning by increasing students' willingness to participate in class, encouraging peer-to-peer collaboration, and improving motivation (Alan et al., 2021; Alan & Kubilay, 2025). The effects on social climate are consistent with this channel, though they are concentrated in *Base+YGL* schools while learning gains occur in both treatment arms, suggesting that social climate improvements are not the sole driver. In *Base Only* schools, the more modest reductions in severe bullying and improvements in social networks may nevertheless have contributed to better concentration and reduced stress, complementing the stigma and stress channels discussed above.

Backlash. Since the program encourages the YGLs to speak out against the harmful taboo surrounding menstruation, an obvious concern is that it could expose them and other girls to negative reactions from others. The positive effects on the school's psychosocial environment and on self-reported stigma suggest that backlash was uncommon. Descriptive results also indicate that the school community supported YGLs: 92% of teachers and directors say that YGLs had a positive impact on the school atmosphere, and 80% of girls said that they *admire* the YGLs in an open-ended question. And while the YGLs in our sample experience 0.16 SD more light teasing in the cross section than non-YGLs (p=0.03), they do not experience more severe bullying (p=0.52), suggesting a lack of strong negative reactions.

3.5 Unintended consequences: early pregnancies in the *Base Only* program

The psychosocial changes documented above also produced an unintended consequence. The base program increased the probability of mothers reporting their daughter had become pregnant (Table A23, column 2), from 2.0% in the control group to 3.5% in *Base Only* (p=0.02), driven entirely by secondary schools. While overall levels of pregnancies are low, this result appears robust: (i) the p-value using randomization inference is similar; (ii) the effect is not driven by a few outlier schools with many pregnancies (Figure A24); and (iii) girls themselves report higher pregnancy rates among schoolmates in *Base Only* schools in the exploratory second endline (Table A23, columns 3 and 4).[36]

What could explain this result? Two mechanisms appear most plausible: (i) increased social connections with boys may have created more opportunities for sexual encounters (Table A25, panel B, column 7); and (ii) normalizing discussion of menstruation may have also made sexual activity feel less taboo.[37] In a context where girls lack information about and access to contraception, either could have increased pregnancies for some girls. Regardless of the mechanism, these results provide an important cautionary note for programs that reduce stigma around menstrual and reproductive health.

[36] Consistent with these findings, mothers report more behavioral problems for their daughters in *Base Only* schools (effect: -0.15 SD, p=0.05, column 5, Table A17), but not for *Base+YGL schools* (p-value of difference: 0.02).

[37] Qualitative interviews carried out by the research team after the endline supported this explanation: mothers reported that their main message to daughters at menarche was to start avoiding boys to prevent pregnancy, and some stopped having these conversations because they thought schools were now addressing reproductive health topics. Consistent with this interpretation, mothers who report discussing menstruation with their daughters report fewer pregnancies (coefficient: -1.8 p.p., p=0.03). This negative correlation also suggests the pregnancy increase is real, not a reporting artifact: if reduced stigma only increased willingness to *report* pregnancies, we would expect mothers more openly discussing menstruation to report *more* pregnancies, not fewer.

Importantly, we do not observe a corresponding increase in pregnancy in *Base+YGL* schools. Three factors may explain the protective effect of YGLs: (i) stronger girl-to-girl support networks (Table A25, panel B, column 6); (ii) training modules for YGLs that taught leadership skills, self-confidence, and how to "say no" to boys—an approach that appears effective given stronger improvements in boys' progressive attitudes in YGL schools (Table A16, column 4); and (iii) better access to information about contraceptives, since YGLs reported fielding questions about sexual activity from classmates and could have passed on knowledge gained from their regular meetings with program technicians.[38] Overall, this protective effect is an important benefit of the YGL program that counterbalances the weaker effects on learning.

3.6 Effects of the YGL component on learning

We further examine why YGLs do not generate additional learning improvements, despite stronger reductions in survey-based measures of menstrual stigma and social climate. YGLs may even slightly harm learning, with point estimates lower for *Base+YGL* schools than for *Base Only* schools across all learning measures (Table 2), and significantly lower for achievement tests (column 2; p=0.02). These negative impacts are solely found in primary schools (Table 8): YGLs have a significantly worse impact on learning in primary than in secondary schools (p=0.03), whereas the base program's effects are similar across both levels (p of primary vs. secondary: 0.71). We find evidence for three factors that explain these patterns: (i) YGLs crowding out a focus on academics, (ii) worse bullying and teasing dynamics in YGL primary schools, and (iii) difficulties identifying effective YGLs in primary schools.

Learning crowd-out by YGLs. Two patterns suggest that YGLs' focus on menstruation and hygiene may have displaced a focus on academics. First, girls in *Base+YGL* schools took on more cleaning duties (Table A20, column 8). While this likely reflects an improved collaborative environment, time spent on cleaning tasks may have crowded out academic activities. Second, YGLs' own academic performance declined: their GPA fell by 1.2 points (out of 20) between baseline and endline, and their class rank worsened by 6.4 places (Table A31; this effect is similar in primary and secondary schools). This second pattern is only suggestive, since (i) YGLs were not randomly selected, so we cannot experimentally compare their performance to a credible counterfactual in control schools; (ii) the decline may partly reflect regression to the mean, since YGLs were selected for above-average scores (and at endline they still score 0.34 points higher than non-YGLs, p<0.01); and (iii) the analysis is vulnerable to selection bias because grades are only observed for girls still in school. Notwithstanding these caveats, the results are consistent with YGLs shifting their own focus away from academics towards hygiene and menstrual issues. This shift may also have influenced their peers, generating a broader crowding out of academic focus in *Base+YGL* schools. Moreover, since some YGLs are in our sample, their reduced performance would directly lower the average learning outcomes we measure in treatment schools.

[38] In line with this interpretation, secondary school girls in *Base+YGL* schools were more likely to report knowing someone who uses modern contraceptives in the exploratory second endline (42% vs. 31% in control).

Table 8: *Heterogeneity by primary vs. secondary schools*

	(1) Learning tests	(2) Attendance (spot checks)	(3) Physical health	(4) Mental health (heart rate rev.)	(5) Menstrual stigma (rev.)	(6) Menstrual stigma (lab, rev.)	(7) Hygiene knowledge	(8) Hygiene behavior	(9) Observed cleanliness†	(10) School social climate
Panel A. Primary Schools										
Base Only	0.207*	0.007	-0.095	0.176*	0.452***	0.741***	0.470***	0.184	0.056	0.049
	(0.106)	(0.102)	(0.074)	(0.103)	(0.114)	(0.252)	(0.118)	(0.121)	(0.280)	(0.102)
Base + YGL	0.054	-0.077	-0.105*	0.151*	0.825***	0.699***	0.562***	0.329***	0.641***	-0.008
	(0.074)	(0.102)	(0.063)	(0.089)	(0.078)	(0.252)	(0.091)	(0.104)	(0.280)	(0.078)
Observations	875	330	895	778	895	66	875	895	66	875
Control mean	-.59	-0.47	.093	-.18	-.68	-0.66	-.59	-.58	-0.66	-.04
p: Treated=0	0.140	0.550	0.072	0.058	<0.001	0.003	<0.001	0.006	0.033	0.881
p: Base Only=Base+YGL	0.140	0.333	0.895	0.782	0.002	0.838	0.348	0.131	0.024	0.538
Panel B. Secondary Schools										
Base Only	0.226**	-0.002	-0.060	0.113*	0.320***	0.544**	0.330***	0.421***	0.097	0.106
	(0.098)	(0.166)	(0.080)	(0.067)	(0.078)	(0.250)	(0.083)	(0.091)	(0.216)	(0.097)
Base + YGL	0.204***	0.143	0.065	0.134**	0.744***	0.481**	0.391***	0.757***	0.319	0.304***
	(0.071)	(0.166)	(0.063)	(0.059)	(0.075)	(0.250)	(0.074)	(0.078)	(0.216)	(0.078)
Observations	1,292	368	1,361	1,126	1,361	74	1,292	1,361	74	1,292
Control mean	.429	0.45	-.06	.135	.476	0.62	.433	.407	0.62	.029
p: Treated=0	0.003	0.509	0.690	0.023	<0.001	0.013	<0.001	<0.001	0.175	0.001
p: Base Only=Base+YGL	0.802	0.262	0.095	0.710	<0.001	0.766	0.347	<0.001	0.217	0.015
Panel C. Interaction term *p*-values										
p: Base Only × Secondary = 0	0.714	0.965	0.905	0.619	0.541	0.582	0.202	0.225	0.907	0.692
p: Base+YGL × Secondary = 0	0.028**	0.195	0.067*	0.876	0.525	0.501	0.055*	0.012**	0.244	0.006***
p: (Base Only × Sec.)=(Base+YGL × Sec.)	0.205	0.139	0.096*	0.669	0.916	0.939	0.803	0.273	0.240	0.044**
p: Treated × Secondary = 0	0.093*	0.381	0.204	0.758	0.500	0.482	0.062*	0.026**	0.476	0.035**

Notes: All outcomes are indices measured in control group standard deviations (Z-scores), and are the same as in Figure A1. Panel A shows effects for primary schools only (grades 4-7) and Panel B shows effects for secondary schools only (grades 8-12). Standard errors are clustered at the school level and are in parentheses; controls include stratum fixed effects and baseline controls selected by double LASSO. Panel C reports p-values from pooled regressions that include school level interaction terms. The first row tests whether the Base Only treatment effect differs between primary and secondary schools. The second row tests whether the Base+YGL treatment effect differs between primary and secondary schools. The third row tests whether the differential effect of Base+YGL versus Base Only varies by school level. The fourth row tests whether the pooled treatment effect differs by school level. P-values in Panel C come from regressions with stratum fixed effects and clustered standard errors at the school level. Column 2 (Attendance) uses spot check data at school×wave level (excluding baseline wave), with wave fixed effects, stratum fixed effects, and standard errors clustered at school level. ** $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.*

Bullying and teasing. YGLs' effects on social climate differ sharply by school level. In secondary schools, YGLs strongly improve social climate (0.30 SD, $p<0.001$, Table 8, panel B, column 10). In primary schools, by contrast, there is no improvement in the combined index (0.01 SD, $p=0.92$, panel A, column 10), and some indicators worsen: teasing about menstruation increases significantly, with no offsetting reduction in intimidation or improvement in between-girl networks (Table A25).

This pattern may help explain the lack of learning effects in *Base+YGL* primary schools. In secondary schools, YGLs' improvements to social climate may offset any crowding out of academic focus. In primary schools, where YGLs fail to improve—and may even worsen—the social environment, no such offset occurs. Younger YGLs may lack the maturity or social standing to shift norms. In line with this, qualitative reports from our implementing partner did confirm greater difficulty in training and coaching YGLs in primary schools.

Which YGLs are effective? The third factor plausibly explaining YGLs' negative effects in primary schools is the difficulty of identifying effective YGL candidates. The intervention aimed to select *positively deviant* girls—those willing to speak out against harmful norms—on the theory that such girls would be especially effective at changing peers' behavior. This theory finds support in the data: schools with more positively deviant YGLs (based on the initial selection process) exhibit significantly better learning and social climate outcomes (Table G2; see details in Appendix G). But finding such girls proves difficult in primary schools, where few students score high on positive deviance measures (Figure G3). Without mature, confident candidates, primary schools end up selecting less effective YGLs who do not appear to generate the psychosocial improvements needed to offset academic crowding out. Secondary schools, by contrast, can more often identify true positive deviants who improve social climate sufficiently to offset any costs to learning.[39]

4 Discussion

Menstrual stigma affects adolescent girls worldwide, yet evidence on whether reducing it improves girls' human capital outcomes is limited. Policymakers have developed various menstrual hygiene interventions in schools in low- and middle-income countries. Some address practical constraints such as information and access to sanitary products and infrastructure. Fewer explicitly target the stigma itself—despite stigma limiting discussion of menstruation and harming girls' wellbeing. This paper evaluates an intervention that combines destigmatization efforts with hygiene sensitization, sanitary infrastructure, and pads. The intervention also identifies "positive deviants," girls already willing to challenge menstrual stigma, and trains them as peer ambassadors to promote norm change.

The base program substantially improved learning after 1.5 years, despite targeting hygiene rather than academics, and showing no effects on physical health or school attendance. Learning gains move together with reductions in menstrual stigma and lower stress (measured by heart rate). This suggests that research and policy on menstrual hygiene that focuses only on health and attendance may miss important welfare outcomes like learning and stress, while neglecting the psychosocial

[39]In line with this, YGLs have better relative effects on learning in large schools (Table A32), plausibly because larger candidate pools increase the probability of finding effective positive deviants.

channels underlying them.

The Young Girl Leaders component generated further improvements in hygiene behavior and self-reported menstrual stigma, demonstrating the promise of using positive deviants as entry points for social change. Future research could evaluate this approach in other settings to address discrimination, restrictive gender norms, or other damaging behaviors. The intervention also posed challenges: YGLs appear to dampen learning effects in primary schools, most likely because they crowd out attention to traditional academics and because effective positive deviants who can avoid backlash were harder to find there. In secondary schools, the welfare effects of YGLs appear positive on balance: greater norm and behavior change, a stronger social climate, no learning costs, and importantly no evidence of the pregnancy increase observed in *Base Only* schools.

More broadly, this paper shows that a program addressing gender-specific psychosocial barriers can improve girls' schooling, even in contexts with severe economic and educational challenges like Madagascar. Following these results, Madagascar's Ministry of Education plans to scale up the *Base+YGL* intervention to 1,750 secondary schools, reaching half a million students by 2028.

References

Aboud, F. E., & Singla, D. R. (2012, August). Challenges to changing health behaviours in developing countries: A critical overview. *Social Science & Medicine (1982)*, *75*(4), 589–594. doi: 10.1016/j.socscimed.2012.04.009

Agarwal, S., Chia, L. E., & Ghosh, P. (2024, August). Do Sanitary Pads Improve Girls' Educational Outcomes? *Review of Economics and Statistics*, 1–54. doi: 10.1162/rest_a_01479

Akyeampong, K., Andrabi, T., Banerjee, A., Banerji, R., Dynarski, S., Glennerster, R., ... Yoshikawa, H. (2023). *Cost-Effective Approaches to Improve Global Learning - What does recent evidence tell us are "Smart Buys" for improving learning in low- and middle-income countries?* London, Washington D.C., New York: FCDO, the World Bank, UNICEF, and USAID.

Alan, S., Baysan, C., Gumren, M., & Kubilay, E. (2021, November). Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking. *The Quarterly Journal of Economics*, *136*(4), 2147–2194. doi: 10.1093/qje/qjab009

Alan, S., Boneva, T., & Ertac, S. (2019, April). Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit*. *The Quarterly Journal of Economics*. doi: 10.1093/qje/qjz006

Alan, S., Carlana, M., & Leone, M. (2024, September). *Inclusive Teaching: Spotting Social Isolation in the Classroom* (Working Paper No. 32954). National Bureau of Economic Research. doi: 10.3386/w32954

Alan, S., & Kubilay, E. (2025, February). Empowering Adolescents to Transform Schools: Lessons from a Behavioral Targeting. *American Economic Review*, *115*(2), 365–407. doi: 10.1257/aer.20240374

Ali, T. S., & Rizvi, S. N. (2010, August). Menstrual knowledge and practices of female adolescents in urban Karachi, Pakistan. *Journal of Adolescence*, *33*(4), 531–541. doi: 10.1016/j.adolescence.2009.05.013

Alidjanov, J. F., Abdufattaev, U. A., Makhsudov, S. A., Pilatz, A., Akilov, F. A., Naber, K. G., & Wagenlehner, F. M. (2014). New Self-Reporting Questionnaire to Assess Urinary Tract Infections and Differential Diagnosis: Acute Cystitis Symptom Score. *Urologia Internationalis*, *92*(2), 230–236. doi: 10.1159/000356177

Anderson, M. L. (2008, December). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495. doi: 10.1198/016214508000000841

Angrist, N., Evans, D. K., Filmer, D., Glennerster, R., Rogers, H., & Sabarwal, S. (2025, January). How to improve education outcomes most efficiently? A review of the evidence using a unified metric. *Journal of Development Economics*, *172*, 103382. doi: 10.1016/j.jdeveco.2024.103382

Ashraf, N., Bau, N., Low, C., & McGinn, K. (2020). Negotiating a Better Future: How Interpersonal Skills Facilitate Intergenerational Investment*. *The Quarterly Journal of Economics*. doi: 10.1093/qje/qjz039

Baird, S., Hicks, J. H., Kremer, M., & Miguel, E. (2016, November). Worms at Work: Long-run Impacts of a Child Health Investment*. *The Quarterly Journal of Economics*, *131*(4), 1637–1680. doi: 10.1093/qje/qjw022

Bandiera, O., Buehren, N., Burgess, R., Goldstein, M., Gulesci, S., Rasul, I., & Sulaiman, M. (2020, January). Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa. *American Economic Journal: Applied Economics*, *12*(1), 210–259. doi: 10.1257/app.20170416

Banerjee, A., Andrab, T., Banerji, R., Dynarski, S., Glennerster, R., Grantham-Mcgregor, S., ... Schmelkes, S. (2023, June). 2023 Cost-effective Approaches to Improve Global Learning - What does Recent Evidence Tell Us are "Smart Buys" for Improving Learning in Low- and Middle-income Countries?

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013, July). The Diffusion of Microfinance. *Science*, *341*(6144), 1236498–1236498. doi: 10.1126/science.1236498

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2019, November). Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials. *The Review of Economic Studies*, *86*(6), 2453–2490. doi: 10.1093/restud/rdz008

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Pariente, W., ... Udry, C. (2015, May). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, *348*(6236), 1260799–1260799. doi: 10.1126/science.1260799

Banerjee, A., La Ferrara, E., & Orozco-Olvera, V. H. (2019). The Entertaining Way to Behavioral Change: Fighting HIV with MTV. *NBER Working Paper*(26096).

Beaman, L., BenYishay, A., Magruder, J., & Mobarak, A. M. (2021, June). Can Network Theory-Based Targeting Increase Technology Adoption? *American Economic Review*, *111*(6), 1918–1943. doi: 10.1257/aer.20200295

Belloni, A., Chernozhukov, V., & Hansen, C. (2014, April). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, *81*(2), 608–650. doi: 10.1093/restud/rdt044

Benshaul-Tolonen, A., Zulaika, G., Nyothach, E., Odour, C., Mason, L., Obor, D., & Laserson, K. F. (2021). *Sanitary products, absenteeism and psychosocial well-being: Evidence from a three-arm cluster randomized controlled feasibility study in Western Kenya* (Working Paper No. WP No.93). Columbia University, SIPA, CDEP-CGEG.

Benshaul-Tolonen, A., Zulaika, G., Sommer, M., & Phillips-Howard, P. A. (2020). Measuring menstruation-related absenteeism among adolescents in low-income countries. In *The Palgrave Handbook of Critical Menstruation Studies* (pp. 705–723). Palgrave Macmillan.

Bergstrom, K., & Özler, B. (2023, June). Improving the Well-Being of Adolescent Girls in Developing Countries. *The World Bank Research Observer*, *38*(2), 179–212. doi: 10.1093/wbro/lkac007

Bicchieri, C. (2017). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press. doi: 10.1093/acprof:oso/9780190622046.001.0001

Bicchieri, C., & Funcke, A. (2018, March). Norm Change: Trendsetters and Social Structure. *Social Research: An International Quarterly*, *85*(1), 1–21. doi: 10.1353/sor.2018.0002

Bick, S., Davies, K., Mwenge, M., MacLeod, C., Braun, L., Chipungu, J., ... Dreibelbis, R. (2025). WASH and learn: A scoping review of health, education and gender equity outcomes of school-based water, sanitation and hygiene in low-income and middle-income countries. *BMJ Global Health*, *10*(5), e018059. doi: 10.1136/bmjgh-2024-018059

Borusyak, K., & Hull, P. (2023). Nonrandom Exposure to Exogenous Shocks. *Econometrica*, *91*(6), 2155–2185. doi: 10.3982/ECTA19367

Bruns, B., Schuh-Moore, A., & Adelman, E. (2017, July). *Conducting classroom observations: Stallings "classroom snapshot" observation system for an electronic tablet* (User Guide). Washington, DC:

World Bank Group, Education Global Practice.

Buehren, N., Goldstein, M., Gulesci, S., Salcher, I., Sulaiman, M., & Yam, V. (2024, August). What We Can Learn When Impacts Fail to Replicate: Lessons From an Empowerment Programme for Young Women in Tanzania. *The Journal of Development Studies*, *60*(8), 1200–1210. doi: 10.1080/00220388.2024.2337370

Bursztyn, L., González, A. L., & Yanagizawa-Drott, D. (2020, October). Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia. *American Economic Review*, *110*(10), 2997–3029. doi: 10.1257/aer.20180975

Carneiro, P., Kraftman, L., Mason, G., Moore, L., Rasul, I., & Scott, M. (2021, August). The Impacts of a Multifaceted Prenatal Intervention on Human Capital Accumulation in Early Life. *American Economic Review*, *111*(8), 2506–2549. doi: 10.1257/aer.20191726

Castro, S., & Czura, K. (2025, April). Cultural taboos and misinformation about menstrual health management in rural Bangladesh. *World Development*, *188*, 106871. doi: 10.1016/j.worlddev.2024.106871

Castro, S., & Mang, C. (2024, June). Breaking the silence – Group discussions and the adoption of menstrual health technologies. *Journal of Development Economics*, *169*, 103264. doi: 10.1016/j.jdeveco.2024.103264

Chirgwin, H., Cairncross, S., Zehra, D., & Sharma Waddington, H. (2021, December). Interventions promoting uptake of water, sanitation and hygiene (WASH) technologies in low- and middle-income countries: An evidence and gap map of effectiveness studies. *Campbell Systematic Reviews*, *17*(4), e1194. doi: 10.1002/cl2.1194

Crowne, D. P., & Marlowe, D. (1960). Marlowe-Crowne social desirability scale. *Journal of Consulting Psychology*.

Czura, K., Menzel, A., & Miotto, M. (2019). Menstrual Health, Worker Productivity and Well-being among Female Bangladeshi Garment Workers. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3519895

Czura, K., Menzel, A., & Miotto, M. (2024, January). Improved menstrual health and the workplace: An RCT with female Bangladeshi garment workers. *Journal of Development Economics*, *166*, 103174. doi: 10.1016/j.jdeveco.2023.103174

Dhar, D., Jain, T., & Jayachandran, S. (2022, March). Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India. *American Economic Review*, *112*(3), 899–927. doi: 10.1257/aer.20201112

Dinarte-Diaz, L., & Egana-delSol, P. (2024, June). Preventing Violence in the Most Violent Contexts: Behavioral and Neurophysiological Evidence from El Salvador. *Journal of the European Economic Association*, *22*(3), 1367–1406. doi: 10.1093/jeea/jvad068

Duflo, E., Dupas, P., & Kremer, M. (2024, June). *The impact of Secondary School Subsidies on Career Trajectories in a Dual Labor Market: Experimental evidence from Ghana* (Working Paper). MIT, Princeton University, and University of Chicago.

Dupas, P., & Miguel, E. (2017). Impacts and determinants of health levels in low-income countries. In *Handbook of economic field experiments* (Vol. 2, pp. 3–93). Elsevier.

Edmonds, E., Feigenberg, B., & Leight, J. (2023, July). Advancing the Agency of Adolescent Girls. *The Review of Economics and Statistics*, *105*(4), 852–866. doi: 10.1162/rest_a_01074

El-Gilany, A.-H., Badawi, K., & El-Fedawy, S. (2005, January). Menstrual Hygiene among Adolescent Schoolgirls in Mansoura, Egypt. *Reproductive Health Matters*, *13*(26), 147–152. doi: 10.1016/S0968-8080(05)26191-8

Evans, D. K., Mendez Acosta, A., & Yuan, F. (2024, February). Girls' Education at Scale. *The World Bank Research Observer*, *39*(1), 47–74. doi: 10.1093/wbro/lkad002

Evans, D. K., & Yuan, F. (2022a, September). How Big Are Effect Sizes in International Education Studies? *Educational Evaluation and Policy Analysis*, *44*(3), 532–540. doi: 10.3102/01623737221079646

Evans, D. K., & Yuan, F. (2022b, February). What We Learn about Girls' Education from Interventions That Do Not Focus on Girls. *The World Bank Economic Review*, *36*(1), 244–267. doi: 10.1093/wber/lhab007

Frost, D. M., & Meyer, I. H. (2023, June). Minority stress theory: Application, critique, and continued relevance. *Current Opinion in Psychology*, *51*, 101579. doi: 10.1016/j.copsyc.2023.101579

Gao, P., Kothari, A., & Lei, Y.-H. (2025, February). Safe spaces for children: School sanitation and sexual violence. *European Economic Review*, *172*, 104952. doi: 10.1016/j.euroecorev.2025.104952

Gulesci, S., Jindani, S., La Ferrara, E., Smerdon, D., Sulaiman, M., & Young, P. (2025, July). A Stepping Stone Approach to Norm Transitions. *American Economic Review*, *115*(7), 2237–2266. doi: 10.1257/aer.20230458

Hamory, J., Baird, S., Das, S., Jones, N., Woldehanna, T., & Yadete, W. (2023). Do layered adolescent-centric interventions improve girls' capabilities? Evidence from a mixed-methods cluster randomised controlled trial in Ethiopia.

Haushofer, J., & Fehr, E. (2014, May). On the psychology of poverty. *Science*, *344*(6186), 862–867. doi: 10.1126/science.1232491

Hennegan, J., & Montgomery, P. (2016, February). Do Menstrual Hygiene Management Interventions Improve Education and Psychosocial Outcomes for Women and Girls in Low and Middle Income Countries? A Systematic Review. *PLOS ONE*, *11*(2), e0146985. doi: 10.1371/journal.pone.0146985

Jayachandran, S. (2021, September). Social Norms as a Barrier to Women's Employment in Developing Countries. *IMF Economic Review*, *69*(3), 576–595. doi: 10.1057/s41308-021-00140-w

Katz, D., Allport, F. H., & Jenness, M. B. (1931). *Students' attitudes; a report of the Syracuse University reaction study*. Oxford, England: Craftsman Press.

Kaur, S., Mullainathan, S., Oh, S., & Schilbach, F. (2025, February). Do Financial Concerns Make Workers Less Productive? *The Quarterly Journal of Economics*, *140*(1), 635–689. doi: 10.1093/qje/qjae038

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007, January). Experimental Analysis of Neighborhood Effects. *Econometrica*, *75*(1), 83–119. doi: 10.1111/j.1468-0262.2007.00733.x

Kudielka, B. M., Buske-Kirschbaum, A., Hellhammer, D. H., & Kirschbaum, C. (2004, June). Differential heart rate reactivity and recovery after psychosocial stress (TSST) in healthy children, younger adults, and elderly adults: The impact of age and gender. *International Journal of Behavioral Medicine*, *11*(2), 116–121. doi: 10.1207/s15327558ijbm1102_8

Laajaj, R., & Macours, K. (2019, October). Measuring Skills in Developing Countries. *Journal of Human Resources*, 1018-9805R1. doi: 10.3368/jhr.56.4.1018-9805R1

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *REVIEW OF ECONOMIC STUDIES*, 32.

Marsh, D. R., Schroeder, D. G., Dearden, K. A., Sternin, J., & Sternin, M. (2004, November). The power of positive deviance. *BMJ*, *329*(7475), 1177–1179. doi: 10.1136/bmj.329.7475.1177

Matavelli, I. (2023). We Don't Talk About Boys: An Experiment on Masculinity Norms in Brazil.

Miguel, E., & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, *72*(1), 159–217. doi: 10.1111/j.1468-0262.2004.00481.x

Mulangu, F. M., & Oviedo Silva, A. M. (2024). *Navigating two decades of high poverty and charting a course for change in madagascar : Poverty and equity assessment (english)* (Tech. Rep.). Washington, D.C.: World Bank Group.

Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why Having Too Little Means So Much*. Henry Holt and Company.

Muralidharan, K. (2017). Field Experiments in Education in Developing Countries. In A. Banerjee & E. Duflo (Eds.), *Handbook of Field Experiments* (pp. 323–385). doi: 10.1016/bs.hefe.2016.09.004

Muralidharan, K., & Prakash, N. (2017). Cycling to school: Increasing secondary school enrollment for girls in India. *American Economic Journal: Applied Economics*, *9*(3), 321–350. doi: 10.1257/app.20160004

Nelson, K. A., Lagony, S., Kansiime, C., Torondel, B., Tanton, C., Ndekezi, D., . . . Weiss, H. A. (2025). Effects and costs of a multi-component menstrual health intervention (MENISCUS) on mental health problems, educational performance, and menstrual health in Ugandan secondary schools: An open-label, school-based, cluster-randomised controlled trial. *The Lancet Global Health*, *13*(5), e888–e899. doi: 10.1016/S2214-109X(25)00007-5

Orr, L., Ferris, J. D., Jacob, R., Beecroft, E., Sanbonmatsu, L., Katz, L. F., ... Kling, J. R. (2003). *Moving to opportunity: Interim impacts evaluation* (Report). Washington D.C: U.S. Dept. of HUD.

Oster, E., & Thornton, R. (2011, January). Menstruation, Sanitary Products, and School Attendance: Evidence from a Randomized Evaluation. *American Economic Journal: Applied Economics*, *3*(1), 91–100. doi: 10.1257/app.3.1.91

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021, January). Prejudice Reduction: Progress and Challenges. *Annual Review of Psychology*, *72*(1), 533–560. doi: 10.1146/annurev-psych-071620-030619

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016, January). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, *113*(3), 566–571. doi: 10.1073/pnas.1514483113

Paradkar, N., & Chowdhury, S. R. (2015). Primary study for detection of arterial blood pressure waveform components. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2015*, 1959–1962. doi: 10.1109/EMBC.2015.7318768

Phillips-Howard, P. A., Nyothach, E., Ter Kuile, F. O., Omoto, J., Wang, D., Zeh, C., ... Laserson, K. F. (2016, November). Menstrual cups and sanitary pads to reduce school attrition, and sexually transmitted and reproductive tract infections: A cluster randomised controlled feasibility study in rural Western Kenya. *BMJ open*, *6*(11), e013229. doi: 10.1136/bmjopen-2016-013229

Radloff, L. S. (1977, June). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, *1*(3), 385–401. doi: 10.1177/014662167700100306

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment*, *25*(4), 1137–1145. doi: 10.1037/a0033323

Reece, J. D., Bunn, J. A., Choi, M., & Navalta, J. W. (2021, June). Assessing Heart Rate Using Consumer Technology Association Standards. *Technologies*, *9*(3), 46. doi: 10.3390/technologies9030046

Schilbach, F., Schofield, H., & Mullainathan, S. (2016, May). The Psychological Lives of the Poor. *American Economic Review*, *106*(5), 435–440. doi: 10.1257/aer.p20161101

Seddon, J. A., Rodriguez, V. J., Provencher, Y., Raftery-Helmer, J., Hersh, J., Labelle, P. R., & Thomassin, K. (2020, June). Meta-analysis of the effectiveness of the Trier Social Stress Test in eliciting physiological stress responses in children and adolescents. *Psychoneuroendocrinology*, *116*, 104582. doi: 10.1016/j.psyneuen.2020.104582

Shah, M., Seager, J., Montalvao, J., & Goldstein, M. (2023, August). *Sex, Power, and Adolescence: Intimate Partner Violence and Sexual Behaviors* (Working Paper No. 31624). National Bureau of Economic Research. doi: 10.3386/w31624

Stallard, P., Skryabina, E., Taylor, G., Phillips, R., Daniels, H., Anderson, R., & Simpson, N. (2014). Classroom-based cognitive behaviour therapy (FRIENDS): A cluster randomised controlled trial to Prevent Anxiety in Children through Education in Schools (PACES). *The Lancet Psychiatry*, *1*(3), 185–192. doi: 10.1016/S2215-0366(14)70244-5

UNESCO Institute for Statistics. (2023). *SDG 4 country profile: Madagascar* (Country Profile). Montreal, Canada: UNESCO Institute for Statistics.

UNICEF. (2019). *Enquête par grappes à indicateurs multiples (MICS) 2018: Rapport final des résultats* (Technical Report). UNICEF.

United Nations Children's Fund (UNICEF), & World Health Organization (WHO). (2023). *Progress on household drinking water, sanitation and hygiene 2000–2022: Special focus on gender* (Tech. Rep.). New York: United Nations Children's Fund (UNICEF) and World Health Organization (WHO).

WaterAid. (2013). *We can't wait: A report on sanitation and hygiene for women and girls.* WaterAid International. Retrieved from: http://www.zaragoza.es/contenidos/medioambiente/onu/1325-

eng_We_cant_wait_sanitation_and_hygiene_for_women_and%20girls.pdf.

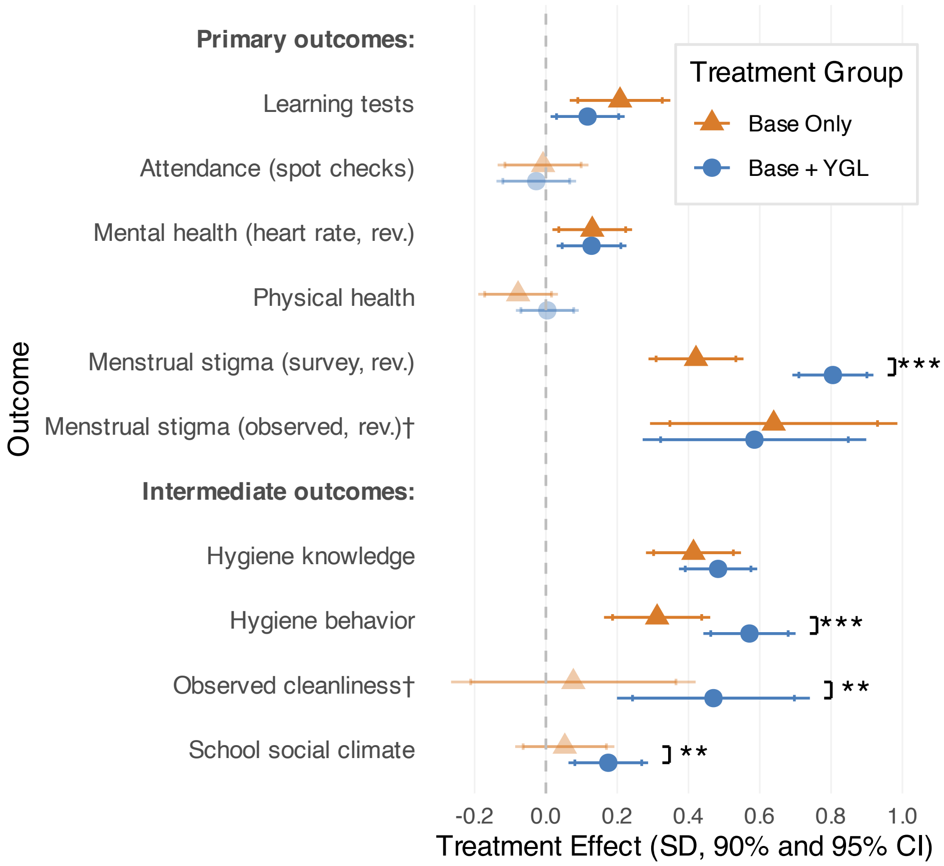
Webb, D. (2025). Silence to Solidarity: How Communication About a Minority Affects Discrimination. *mimeo, Paris School of Economics*.

Zhang, Y., Stewart, C., Ranjan, Y., Conde, P., Sankesara, H., Rashid, Z., … Folarin, A. A. (2025, April). Large-scale digital phenotyping: Identifying depression and anxiety indicators in a general UK population with over 10,000 participants. *Journal of Affective Disorders*, *375*, 412–422. doi: 10.1016/j.jad.2025.01.124

Zwane, A. P., Zinman, J., Van Dusen, E., Pariente, W., Null, C., Miguel, E., … Banerjee, A. (2011, February). Being surveyed can change later behavior and related parameter estimates. *Proceedings of the National Academy of Sciences*, *108*(5), 1821–1826. doi: 10.1073/pnas.1000776108

A Supplementary tables and figures

Figure A1: *Effect of treatments on key outcomes for girls*



Notes: The Figure shows treatment effects of *Base Only* and *Base+YGL* on all key outcomes, using Equation 1. All outcomes are standardized to Z-scores by subtracting control group mean and dividing by control group standard deviation. *Learning tests* correspond to Table 2, column 1; *Attendance (spot checks)* to Table 3, column 1; *Mental health (heart rate, rev.)* to the standardized measure of Figure 2, panel A; *Physical health* to Table 4, column 1; *Menstrual stigma (survey, rev.)* to Table 5, column 1; *Menstrual stigma (observed, rev.)* to Table 6, column 1; *Hygiene knowledge* to Table 1, column 1; *Hygiene behavior* to Table 1, column 4; *Observed cleanliness* to Table 1, Panel B column 1; *School social climate* to Table 7, column 1). All apart from *Attendance (spot checks)*, *Menstrual stigma (observed, rev.)*, and *Observed cleanliness* use data from girls' surveys, control for LASSO and stratum fixed effects, and cluster at school level. *Attendance (spot checks)* uses data from 5 spot check waves, controlling for baseline spot check attendance and stratum fixed effects, clustering at school level. (+) *Menstrual stigma (observed, rev.)* combines data from girls, boys, and information supply experiment, aggregated to school level (N=140), hence wider confidence intervals. *Observed cleanliness* aggregates individual cleanliness observations to the school level (N=140). Confidence intervals are 95% (bars) and 90% (tails). Stars indicate p-values for H_0: *BaseOnly=Base+YGL* for that outcome; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
†: *Menstrual stigma (observed, rev.)* combines measures from different surveys and lab-in-the-field exercises; *Observed cleanliness* aggregates individual cleanliness observations to the school level. Both outcomes are only defined at the school level (N=140).

Table A2: *Baseline balance: girl demographics*

Variable	(1) Total Mean/(SE)	(2) Control Mean/(SE)	(3) Base + YGL Mean/(SE)	(4) Base Only Mean/(SE)	(3)-(2) Pairwise t-test P-value	(4)-(2) Pairwise t-test P-value
Girl's age (years)	14.000	13.741	14.104	14.033	0.292	0.483
	(0.139)	(0.290)	(0.187)	(0.299)		
Girl already had her period (=1)	0.471	0.447	0.483	0.468	0.606	0.807
	(0.028)	(0.058)	(0.038)	(0.061)		
Girl has a brother (=1)	0.548	0.574	0.556	0.508	0.512	0.033**
	(0.011)	(0.022)	(0.016)	(0.021)		
Distance to school (km)	2.320	2.260	2.379	2.259	0.518	0.996
	(0.072)	(0.149)	(0.107)	(0.126)		
Time to school (min)	30.167	29.722	30.755	29.404	0.637	0.896
	(0.832)	(1.871)	(1.153)	(1.564)		
Was absent last 3 months (=1)	0.552	0.546	0.573	0.518	0.497	0.517
	(0.016)	(0.031)	(0.025)	(0.029)		
Days absent last 3 months	1.412	1.391	1.322	1.613	0.671	0.289
	(0.074)	(0.125)	(0.104)	(0.167)		
Father passed away (=1)	0.085	0.077	0.083	0.096	0.687	0.256
	(0.007)	(0.009)	(0.010)	(0.013)		
Father in household (=1)	0.696	0.711	0.696	0.682	0.592	0.401
	(0.012)	(0.023)	(0.017)	(0.027)		
Mother passed away (=1)	0.049	0.040	0.050	0.053	0.331	0.282
	(0.004)	(0.008)	(0.006)	(0.008)		
Mother in household (=1)	0.816	0.835	0.805	0.820	0.208	0.595
	(0.009)	(0.020)	(0.013)	(0.019)		
Guardian knows how to read and write (=1)	0.828	0.829	0.836	0.810	0.799	0.474
	(0.010)	(0.020)	(0.015)	(0.017)		
Guardian has no education (=1)	0.095	0.085	0.096	0.102	0.465	0.310
	(0.007)	(0.011)	(0.012)	(0.014)		
Guardian attended secondary or higher education (=1)	0.303	0.308	0.293	0.318	0.648	0.794
	(0.013)	(0.029)	(0.018)	(0.028)		
Guardian occupation: Agriculture (=1)	0.773	0.768	0.781	0.762	0.696	0.899
	(0.014)	(0.030)	(0.018)	(0.028)		
Observations	2390	568	1216	606		
Schools	140	35	70	35		

Notes: Columns 1-4 show means with standard errors in parentheses. Columns 5-6 show p-values from pairwise t-tests comparing Base+YGL and Base Only to Control, respectively. Sample includes all girls at baseline. Joint F-test of all variables in balance tables (Table A2– Table A4): Base+YGL vs. Control, F=1.31, p=0.86; Base Only vs. Control, F=1.45, p=0.75. Joint-test p-values are calculated using randomization inference with school-level permutations, following stratification. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: *Baseline balance: outcome indexes at baseline*

Variable	(1) Total Mean/(SE)	(2) Control Mean/(SE)	(3) Base + YGL Mean/(SE)	(4) Base Only Mean/(SE)	(3)-(2) Pairwise t-test P-value	(4)-(2) Pairwise t-test P-value
Wealth index (Z)	0.000	–0.136	0.083	–0.039	0.208	0.632
	(0.071)	(0.143)	(0.100)	(0.145)		
Years of school delay	2.224	2.160	2.233	2.264	0.628	0.598
	(0.067)	(0.121)	(0.092)	(0.156)		
WASH knowledge index (Z)	0.093	0.000	0.107	0.151	0.317	0.206
	(0.041)	(0.091)	(0.056)	(0.077)		
WASH behaviour index (Z)	0.008	–0.000	–0.011	0.054	0.906	0.606
	(0.034)	(0.080)	(0.046)	(0.067)		
Menstruation knowledge index (Z)	0.047	–0.000	0.099	–0.014	0.504	0.937
	(0.061)	(0.121)	(0.085)	(0.130)		
Menstruation behaviour index (Z)	0.070	–0.000	0.094	0.086	0.475	0.591
	(0.055)	(0.107)	(0.077)	(0.119)		
Mental health index (Z)	0.020	–0.000	0.061	–0.042	0.412	0.604
	(0.028)	(0.063)	(0.040)	(0.050)		
Menstrual health problems index (Z)	0.018	0.000	0.042	–0.013	0.699	0.919
	(0.043)	(0.092)	(0.059)	(0.089)		
Menstruation inhibiting school activity index (Z)	0.032	0.000	0.042	0.041	0.316	0.330
	(0.010)	(0.042)	(0.001)	(0.001)		
Received info. on hygiene (Z)	0.037	–0.000	0.055	0.036	0.455	0.687
	(0.027)	(0.067)	(0.032)	(0.058)		
Received info. on menstruation (Z)	0.021	–0.000	0.066	–0.049	0.520	0.694
	(0.043)	(0.084)	(0.060)	(0.091)		
Willingness to speak about hygiene (Z)	–0.006	–0.000	–0.011	–0.001	0.899	0.989
	(0.035)	(0.074)	(0.051)	(0.063)		
Willingness to speak about menstruation (Z)	0.040	–0.000	0.058	0.043	0.415	0.648
	(0.032)	(0.056)	(0.044)	(0.076)		
Exposure to bullying (self-reported) (Z)	0.032	0.000	0.021	0.082	0.757	0.357
	(0.031)	(0.055)	(0.042)	(0.070)		
Attitudes towards menstruation (Z)	–0.072	0.000	–0.134	–0.014	0.213	0.907
	(0.042)	(0.091)	(0.057)	(0.085)		
Shame response to vignettes (Z)	0.020	–0.000	0.015	0.048	0.880	0.692
	(0.043)	(0.080)	(0.060)	(0.093)		
Perceived norms related to menstruation (Z)	0.076	–0.000	0.089	0.120	0.277	0.222
	(0.036)	(0.064)	(0.052)	(0.075)		
Social desirability index (Z)	0.006	0.000	0.037	–0.051	0.538	0.456
	(0.024)	(0.051)	(0.033)	(0.045)		
Observed hygiene: home (Z)	–0.024	0.000	–0.065	0.035	0.407	0.721
	(0.033)	(0.066)	(0.043)	(0.073)		
Observations	2390	568	1216	606		
Schools	140	35	70	35		

Notes: Sample includes all girls at baseline. Columns (1)-(4) show means and standard deviations of indexes calculated using the girls' baseline survey. Columns (5)-(6) show the p-value of a pairwise test comparing *Base Only* and *Base+YGL* with *control*, respectively. Standard errors cluster at the school level. ** p < 0.1, ** p < 0.05, *** p < 0.01.*

Table A4: *Baseline balance: school-level characteristics*

Variable	(1) Total Mean/(SE)	(2) Control Mean/(SE)	(3) Base + YGL Mean/(SE)	(4) Base Only Mean/(SE)	(3)-(2) Pairwise t-test P-value	(4)-(2) Pairwise t-test P-value
Share of girls enrolled	0.515	0.511	0.518	0.513	0.554	0.895
	(0.005)	(0.011)	(0.007)	(0.010)		
Secondary School	0.529	0.514	0.543	0.514	0.785	1.000
	(0.042)	(0.086)	(0.060)	(0.086)		
Any projects ongoing in school	0.293	0.371	0.243	0.314	0.172	0.621
	(0.039)	(0.083)	(0.052)	(0.080)		
Total number of students	212.821	208.486	218.629	205.543	0.671	0.907
	(9.315)	(19.102)	(13.883)	(16.358)		
Number of girls	108.971	105.886	112.529	104.943	0.612	0.945
	(5.084)	(10.566)	(7.587)	(8.715)		
Number of girls aged 8 or above	90.336	92.143	91.871	85.457	0.983	0.629
	(4.995)	(10.843)	(7.323)	(8.476)		
Number of girls aged 12 or above	67.886	70.371	67.414	66.343	0.816	0.784
	(5.047)	(11.084)	(7.065)	(9.535)		
School is accessible	0.100	0.086	0.114	0.086	0.656	1.000
	(0.025)	(0.048)	(0.038)	(0.048)		
Years since school was opened	29.229	30.000	28.429	30.057	0.726	0.992
	(1.826)	(3.810)	(2.520)	(3.761)		
Functional water access point	0.532	0.529	0.529	0.543	0.994	0.912
	(0.042)	(0.087)	(0.060)	(0.085)		
School has a latrine	0.921	0.886	0.929	0.943	0.464	0.400
	(0.023)	(0.055)	(0.031)	(0.040)		
Latrine is functional	0.886	0.857	0.914	0.857	0.372	1.000
	(0.027)	(0.060)	(0.034)	(0.060)		
School has a shower facility	0.036	0.029	0.043	0.029	0.722	1.000
	(0.016)	(0.029)	(0.024)	(0.029)		
Shower is functional	0.029	0.000	0.043	0.029	0.218	0.321
	(0.014)	(0.000)	(0.024)	(0.029)		
School has a handwashing station	0.550	0.486	0.600	0.514	0.270	0.814
	(0.042)	(0.086)	(0.059)	(0.086)		
Handwashing station is functional	0.457	0.400	0.486	0.457	0.411	0.635
	(0.042)	(0.084)	(0.060)	(0.085)		
School has a water point	0.250	0.286	0.200	0.314	0.329	0.798
	(0.037)	(0.077)	(0.048)	(0.080)		
Water point is functional	0.150	0.171	0.100	0.229	0.299	0.557
	(0.030)	(0.065)	(0.036)	(0.072)		
Distance to commune headquarters (km)	7.196	7.771	6.833	7.346	0.472	0.809
	(0.567)	(1.123)	(0.728)	(1.353)		
Private School	0.050	0.086	0.029	0.057	0.199	0.648
	(0.018)	(0.048)	(0.020)	(0.040)		

Notes: Columns 1-4 show means with standard errors in parentheses. Columns 5-6 show p-values from pairwise t-tests comparing *Base+YGL* and *Base Only* to control, respectively. Sample includes all 140 schools at baseline. Randomization was stratified by school characteristics. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: *Probability of attrition and sample availability*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	In Endline (=1)	Endline in person (=1)	In Midline and Endline (=1)	Mother in Endline (=1)	Heart rate available (=1)	Class with marks (=1)	Class with teacher int. (=1)	Boy from main sample list (=1)	Girl from main sample list (=1)	P(register available)
Base Only	-0.028	-0.045**	-0.059**	-0.018	-0.048	-0.093**	-0.019	0.034	-0.021	-0.007
	(0.019)	(0.020)	(0.029)	(0.012)	(0.030)	(0.046)	(0.048)	(0.022)	(0.025)	(0.013)
Base + YGL	0.001	-0.020	-0.044*	-0.013	-0.026	-0.028	-0.004	-0.016	-0.012	0.006
	(0.016)	(0.017)	(0.024)	(0.011)	(0.026)	(0.034)	(0.040)	(0.023)	(0.019)	(0.011)
Observations	2,390	2,390	2,390	2,390	2,390	500	500	1,325	1,627	3,626
Data source	Girls	Girls	Girls	Girls	Heart Rate	Teachers	Teachers	Boys	Lab in field	Registers
Control mean	.951	.928	.856	.963	.802	.919	.887	.837	.853	.901
p: Treated=0	0.570	0.079	0.038	0.183	0.178	0.149	0.818	0.963	0.404	0.894
p: Base Only=Base+YGL	0.090	0.158	0.518	0.526	0.387	0.101	0.726	0.012	0.711	0.256

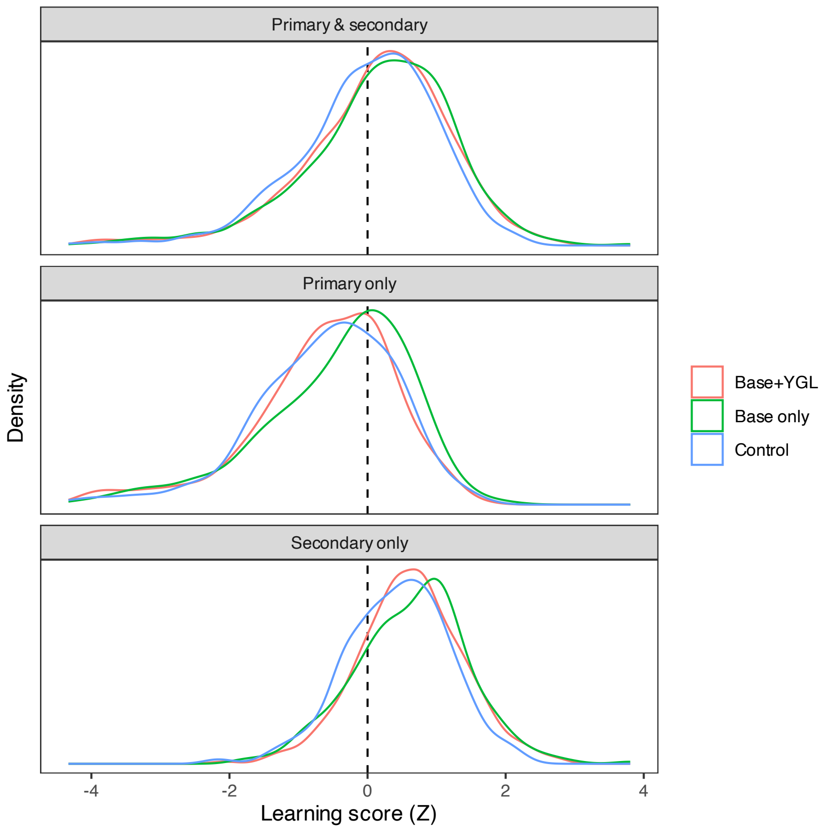
Notes: The table displays the probability of each observation from the sampling frame being in the data. *Outcomes:* Column (1): a girl from the baseline sample was interviewed at endline. Column (2): girl from the baseline sample was interviewed in person (rather than through phone). Column (3): girl was interviewed both at midline and endline. Column (4): the girl's mother or caretaker from the baseline sample was also interviewed at endline. Column (5): heart rate data was available for the girl's interview. Columns (6)-(7): we aimed to digitize the official school GPAs and interview teachers from a sample of 500 classes that included all target grade levels (from 3rd grade to 12th grade), randomly selecting one class per grade level if there were multiple. Column (6) shows probability of collecting school GPAs for each class; column (7) shows probability of interviewing the teacher. Column (8): the proportion of boys interviewed from the main sampling frame of 10 randomly selected boys based on the school registers (rather than a randomly selected replacement list). Column (9): for the lab-in-the-field, the proportion of girls interviewed based on a sampling frame of 12 randomly selected girls based on school registers (instead of a randomly selected replacement list). Column (10): at the *class* × *wave* level, the proportion of measured days for which official absenteeism data was available. Standard errors are clustered at the school level and are in parentheses; controls include stratum fixed effects.

Table A6: *Are effects driven by participants with high social desirability bias?*

Outcome	Base Only	Base+YGL	High SDB	Base × High SDB	(Base+YGL) × High SDB	N
Learning tests	0.207***	0.168***	0.055	0.003	–0.092	2167
	(0.077) [0.007]	(0.063) [0.008]	(0.093) [0.552]	(0.090) [0.978]	(0.088) [0.295]	
Physical health	–0.070	0.004	0.071	–0.017	0.000	2256
	(0.086) [0.414]	(0.069) [0.957]	(0.120) [0.553]	(0.138) [0.904]	(0.107) [0.997]	
Mental health (heart rate rev.)	0.091	0.150*	0.031	0.081	–0.044	1904
	(0.099) [0.358]	(0.080) [0.059]	(0.120) [0.797]	(0.147) [0.583]	(0.116) [0.705]	
Menstrual stigma (rev.)	0.485***	0.870***	0.001	–0.130	–0.121	2256
	(0.081) [0.000]	(0.066) [0.000]	(0.088) [0.994]	(0.093) [0.161]	(0.084) [0.151]	
Menstrual stigma (lab rev.)	0.573***	0.591***	0.026	0.073	–0.095	2256
	(0.165) [0.001]	(0.145) [0.000]	(0.067) [0.695]	(0.083) [0.377]	(0.074) [0.197]	
Hygiene knowledge	0.436***	0.516***	–0.068	–0.043	–0.057	2167
	(0.077) [0.000]	(0.067) [0.000]	(0.086) [0.430]	(0.093) [0.648]	(0.081) [0.480]	
Hygiene behavior	0.288***	0.623***	–0.001	0.048	–0.102	2256
	(0.093) [0.002]	(0.080) [0.000]	(0.105) [0.990]	(0.119) [0.689]	(0.104) [0.327]	
School social climate	0.155*	0.204***	0.052	–0.124	–0.033	2167
	(0.093) [0.096]	(0.069) [0.003]	(0.102) [0.612]	(0.108) [0.252]	(0.096) [0.733]	

Notes: Outcomes are the same indices as in Figure A1. Each row is a separate regression. *High SDB* = 1 if the girl's baseline Crowne-Marlowe social desirability score is above median. Columns 4-5 show interactions testing whether treatment effects differ for high-SDB girls. Standard errors clustered at school level in parentheses; p-values in brackets. Controls include stratum fixed effects and LASSO-selected baseline controls. ** p < 0.1, ** p < 0.05, *** p < 0.01.*

Figure A7: *Distribution of learning test scores at endline*



Notes: Density plot of learning test scores (academic achievement and cognitive tests combined) in each treatment arm. First panel includes both primary and secondary, second panel is only primary schools, and third panel is only secondary schools.

Table A8: *Effects on each individual learning test*

VARIABLES	(1) Math fluency	(2) Reading fluency	(3) Listening	(4) Comprehension	(5) Digit span	(6) Reverse digit span
Base Only	0.254***	0.159***	0.110	0.080	0.114*	0.048
	(0.082)	(0.049)	(0.080)	(0.076)	(0.062)	(0.068)
Base + YGL	0.150**	0.073*	0.044	0.041	0.033	0.107**
	(0.062)	(0.041)	(0.062)	(0.058)	(0.051)	(0.050)
Observations	2,167	2,167	2,167	2,167	2,167	2,167
Data source	Girls	Girls	Girls	Girls	Girls	Girls
p: Treated = 0	.0036	.0106	.2596	.3388	.2386	.0575
p: Base Only = Base + YGL	.1093	.0372	.3621	.5816	.1534	.3625

Notes: The outcomes are the individual components of the learning tests implemented during the girls' endline survey. All scores are in control group standard deviations. Sample includes all girls at endline interviewed in person. Standard errors are clustered at the school level and are in parentheses; controls include stratum fixed effects and the baseline controls selected by double LASSO. ** p < 0.1, ** p < 0.05, *** p < 0.01.*

Table A9: *Robustness: controlling for schools' pre-treatment (2021) exam pass rates*

	(1) Combined index: Achievement + cognitive tests (Z)	(2) Combined index: Achievement + cognitive tests (Z)
Base Only	0.219***	0.218***
	(0.074)	(0.074)
Base + YGL	0.132***	0.125**
	(0.051)	(0.051)
Observations	1,968	1,968
Data source	Girls	Girls
Control mean	-.02	-.02
Control SD	.987	.987
p: Treated=0	0.001	0.002
p: Base Only=Base+YGL	0.204	0.168

Notes: The outcome in both columns is the combined learning index (achievement and cognitive tests, Z-scored). Column (1) reports the main specification; column (2) additionally controls for the school's official exit-exam pass rate in 2021, the year before the intervention began. Both columns are estimated on the subsample of girls in the 125 schools (of 140) with available 2021 exam data, so that the added control is the only difference between columns. Standard errors clustered at school level in parentheses. Controls include stratum fixed effects and LASSO-selected baseline controls. ** p < 0.1, ** p < 0.05, *** p < 0.01.*

Table A10: *Boys' school attendance*

	(1) # boys present (spot-checks)	(2) # boys present (registers)
Base Only	-1.522	1.906
	(3.288)	(3.103)
Base + YGL	-2.115	-0.057
	(2.583)	(2.615)
Observations	698	34,455
Control mean	62.9	57.3
Control SD	45.2	78.7
p: Treated=0	0.458	0.814
p: Base Only=Base+YGL	0.814	0.532

Notes: Column 1: outcome is the number of boys physically present during the spot checks. Observations are at the school × wave level. Specification controls for spot check wave fixed effects. Column 2: outcome is the number of boys present on a given day according to official school registers. Observations are at the school × day level. Specification controls for day fixed effects. Both columns control for stratum fixed effects and cluster standard errors at the school level.

Table A11: *Urinary tract health: robustness*

	(1) Any UTI symptoms (=1)	(2) # UTI symptoms (0–7)	(3) Cystitis symptom threshold (Alidjanov et al.)
Base Only	0.027	0.028	-0.008
	(0.023)	(0.040)	(0.011)
Base + YGL	-0.019	-0.037	-0.004
	(0.019)	(0.032)	(0.007)
Observations	2,256	2,256	2,256
Data source	Girls	Girls	Girls
Control mean	.207	.35	.041
Control SD	.406	.776	.198
p: Treated = 0	0.863	0.624	0.423
p: Base Only = Base + YGL	0.018	0.069	0.731

Notes: Column 1: binary indicator for any UTI symptoms with moderate/severe intensity (severity ≥ 2). Column 2: count of symptoms at moderate/severe levels (0-7 scale, higher = worse health). Column 3: acute cystitis threshold (sum of severities for symptoms u1-u6 ≥ 6, following the threshold recommended in Alidjanov et al., 2014). Standard errors clustered at school level in parentheses. Controls include stratum fixed effects and LASSO-selected baseline controls. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: *Main results controlling for enumerator fixed effects*

	Base Only	Base+YGL	N
Learning tests	0.181***	0.117***	2167
	(0.055) [0.001]	(0.042) [0.006]	
Physical health	–0.011	0.043	2256
	(0.051) [0.821]	(0.042) [0.307]	
Mental health (heart rate rev.)	0.117**	0.116**	1904
	(0.056) [0.035]	(0.050) [0.020]	
Menstrual stigma (rev.)	0.454***	0.840***	2256
	(0.068) [0.000]	(0.054) [0.000]	
Menstrual stigma (lab rev.)	0.644***	0.563***	2256
	(0.139) [0.000]	(0.127) [0.000]	
Hygiene knowledge	0.427***	0.509***	2167
	(0.056) [0.000]	(0.047) [0.000]	
Hygiene behavior	0.343***	0.611***	2256
	(0.074) [0.000]	(0.063) [0.000]	
School social climate	0.102	0.192***	2167
	(0.065) [0.115]	(0.045) [0.000]	

Notes: Outcomes are the same indices as in Figure A1. All specifications include stratum fixed effects, LASSO-selected baseline controls, and fixed effects for 32 enumerators. Standard errors clustered at school level in parentheses; p-values in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: *Heterogeneity by whether girls had reached menarche at baseline*

	Base Only	Base+YGL	Menarche at baseline	Base × Menarche at base.	Base+YGL × Menarche at base.	N
Learning tests	0.200**	0.056	0.007	–0.128	0.024	2165
	(0.094) [0.033]	(0.067) [0.398]	(0.140) [0.962]	(0.126) [0.309]	(0.088) [0.785]	
Physical health	–0.085	–0.007	0.351**	0.033	0.040	2254
	(0.075) [0.259]	(0.061) [0.909]	(0.175) [0.046]	(0.114) [0.772]	(0.088) [0.650]	
Mental health (heart rate rev.)	0.066	0.050	–0.231	0.103	0.135	1902
	(0.079) [0.407]	(0.068) [0.462]	(0.168) [0.168]	(0.113) [0.363]	(0.103) [0.190]	
Menstrual stigma (rev.)	0.460***	0.884***	0.214	–0.106	–0.180*	2254
	(0.089) [0.000]	(0.065) [0.000]	(0.157) [0.173]	(0.120) [0.379]	(0.095) [0.058]	
Hygiene knowledge	0.461***	0.594***	0.261*	–0.138	–0.263***	2165
	(0.092) [0.000]	(0.067) [0.000]	(0.139) [0.061]	(0.119) [0.246]	(0.100) [0.009]	
Hygiene behavior	0.256***	0.491***	0.229	0.158	0.137	2254
	(0.093) [0.006]	(0.081) [0.000]	(0.148) [0.122]	(0.127) [0.214]	(0.091) [0.130]	
School social climate	–0.077	0.066	0.011	0.258**	0.217**	2165
	(0.088) [0.381]	(0.067) [0.328]	(0.161) [0.945]	(0.120) [0.031]	(0.089) [0.015]	

Notes: Outcomes are the same indices as in Figure A1. Each row is a separate regression. *Menarche at baseline* = 1 if the girl had reached menarche at baseline. Columns 4-5 show interactions testing whether treatment effects differ for girls who had reached menarche. Standard errors clustered at school level in parentheses; p-values in brackets. Controls include stratum fixed effects and LASSO-selected baseline controls. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: *Learning Effects by Menstruation Status During Survey*

	(1) Learning tests (Z)	(2) Learning tests (Z)	(3) Learning tests (Z)	(4) Learning tests (Z)
Base Only			0.177**	0.181**
			(0.082)	(0.081)
Base + YGL			0.107*	0.121**
			(0.060)	(0.059)
Menst. during survey	-0.129	-0.110	-0.129	-0.111
	(0.138)	(0.130)	(0.138)	(0.130)
Base Only × Menst. during survey			0.309*	0.312*
			(0.183)	(0.173)
Base + YGL × Menst. during survey			0.140	0.139
			(0.157)	(0.151)
Treated	0.130**	0.140**		
	(0.058)	(0.057)		
Treated × Menst. during survey	0.192	0.192		
	(0.152)	(0.146)		
Observations	2,166	2,166	2,166	2,166
Age FE	No	Yes	No	Yes
Data source	Girls	Girls	Girls	Girls
Control mean	0	0	0	0

Notes: Post-double-selection LASSO regressions of the learning test index (Z-scored) on treatment indicators interacted with an indicator for menstruation during the survey day. Columns (1)–(2) pool treatment arms; columns (3)–(4) split by Base Only and Base+YGL. Even-numbered columns add age fixed effects. Girls who had never menstruated are coded as not menstruating during the survey. All regressions include randomization stratum fixed effects, LASSO-selected baseline controls, and cluster standard errors at the school level. * p<0.10, ** p<0.05, *** p<0.01.

Table A15: *Self-reported mental health: depression and self-esteem*

	(1) Midline: Depression (rev.) (Z)	(2) Endline: Depression (rev.) (Z)	(3) Endline: Self-esteem (Z)
Base Only	-0.039	0.069	0.049
	(0.045)	(0.064)	(0.069)
Base + YGL	-0.003	0.067	0.067
	(0.041)	(0.062)	(0.053)
Observations	1,963	1,964	1,964
Data source	Girls	Girls	Girls
p: Treated=0	0.700	0.241	0.227
p: Base Only=Base+YGL	0.390	0.967	0.764

Notes: Only girls interviewed at both midline and endline are included. *Depression (rev.)* (columns 1–2) is constructed using factor analysis on CESD depression items, reversed so that a higher value indicates better mental health (i.e., less depression). *Self-esteem* (column 3) is constructed using factor analysis on questions about how the girl perceived her own abilities. Standard errors are clustered at the school level and are in parentheses; controls include stratum fixed effects and the baseline controls selected by double LASSO. ** $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.*

Table A16: *Boys' outcomes: knowledge, attitudes, and collaboration*

	(1) Menstruation knowledge (Z)	(2) Health info. diffusion (Z)	(3) Menstr. info. diffusion (Z)	(4) Attitudes index (Z)	(5) Cleaning collaboration (Z)
Base Only	0.246***	0.259*	0.509***	0.218**	0.443***
	(0.086)	(0.141)	(0.145)	(0.100)	(0.162)
Base + YGL	0.323***	0.700***	1.071***	0.415***	0.398***
	(0.067)	(0.116)	(0.125)	(0.079)	(0.146)
Observations	1,323	1,325	1,325	1,322	1,325
Data source	Boys	Boys	Boys	Boys	Boys
p: Treated=0	<0.001	<0.001	<0.001	<0.001	0.003
p: Base Only=Base+YGL	0.346	<0.001	<0.001	0.024	0.745

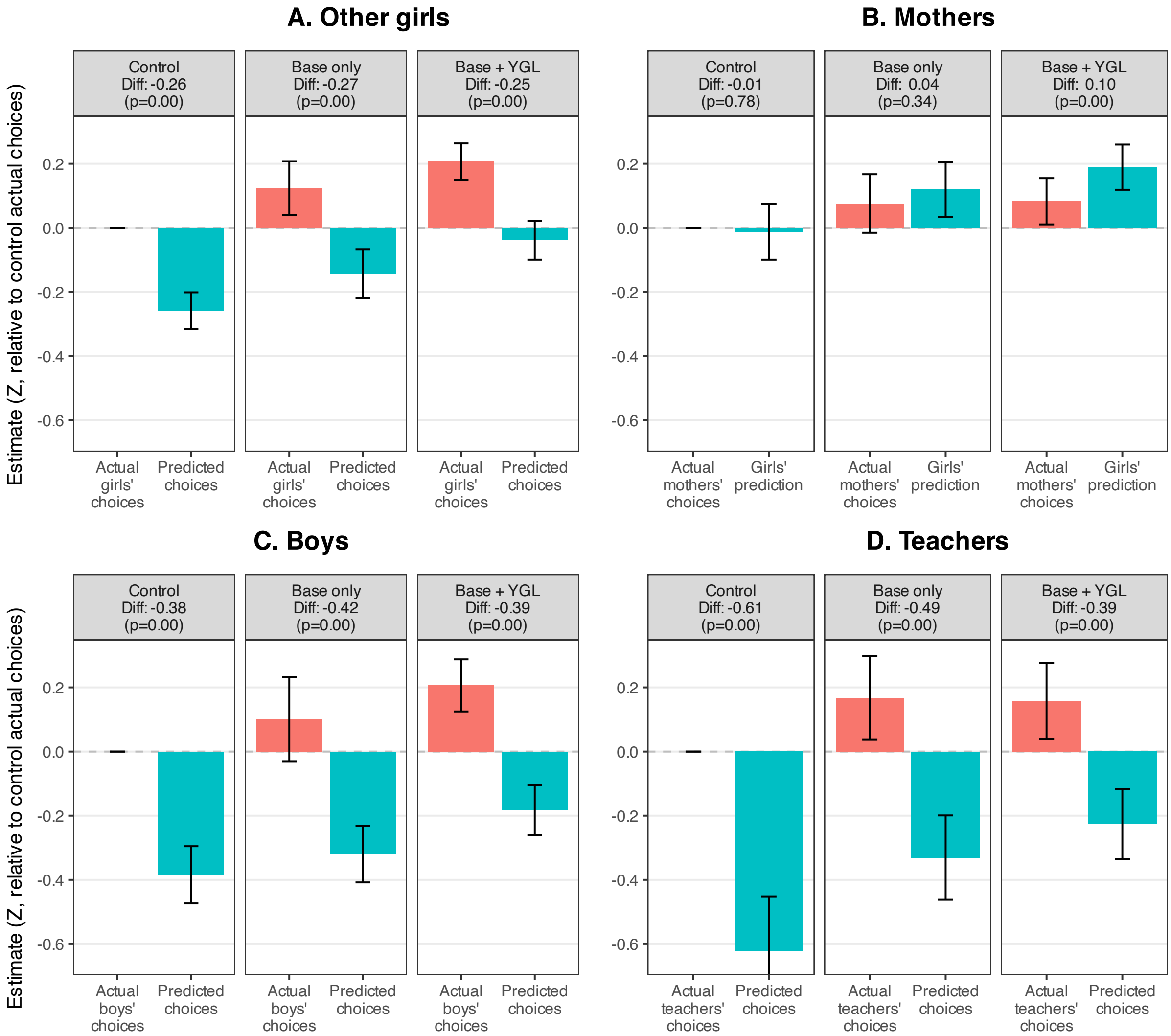
Notes: All outcomes are indices in Z-scores: menstruation knowledge, health information diffusion, menstruation information diffusion, progressive menstruation attitudes, and cleaning collaboration at school. Standard errors clustered at school level in parentheses. Controls include stratum fixed effects. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A17: *Mothers' knowledge, behavior, and perceptions of girls' behavior*

VARIABLES	(1) Hygiene knowl. (Z)	(2) Menstr. knowl. (Z)	(3) Hygiene behavior (Z)	(4) Menstr. behavior (Z)	(5) Attitudes (Z)	(6) Girls' behavior (Z)
Base Only	0.010	0.093	-0.104	0.028	0.333***	-0.143*
	(0.085)	(0.073)	(0.076)	(0.061)	(0.065)	(0.077)
Base + YGL	0.029	0.220***	-0.046	0.007	0.350***	0.017
	(0.073)	(0.059)	(0.064)	(0.057)	(0.055)	(0.061)
Observations	2,287	2,286	2,287	2,287	2,287	2,260
Data source	Mothers	Mothers	Mothers	Mothers	Mothers	Mothers
p: Treated = 0	.7483	.0018	.2798	.8039	<0.001	.5778
p: Base Only = Base + YGL	.7853	.0463	.3866	.6174	.7849	.0176

Notes: Hygiene knowledge index (column 1) combines questions about WASH knowledge related to the program. *Menstruation knowledge index* (column 2) is constructed from questions that test mothers' knowledge about menstruation. *Hygiene behavior index* (column 3) compiles questions about hygiene behavior during the last week. *Menstruation behavior index* (column 4) includes questions about mothers' behavior during the last period, such as the use of serviettes. *Attitudes* (column 5) includes questions about attitudes towards menstruation, where a higher value indicates more progressive attitudes. *Girls' behavior* (column 6) includes mothers' reports about the psychosocial behavior of their daughters using a Strengths and Difficulties Questionnaire, measuring problematic emotions and behaviors of the girls, as reported by mothers. Standard errors are clustered at the school level and are in parentheses; controls include stratum fixed effects and the baseline controls selected by double LASSO. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure A18: *Misperceptions about others' menstruation attitudes*



Notes: This figure shows treatment effects on second-order beliefs—girls' predictions about others' attitudes towards menstruation. Left bar in each panel is actual attitude responses made by that group, right bar is girls' predictions of that group's attitude responses. All are normalized by control group actual attitudes. Second-order beliefs were asked for a subset of all attitude questions (see Appendix K.1) and actual choices shown in this Figure are for the same set of questions, explaining the difference in point estimates compared to Table A16, column 4 and Table A17, column 5. Panel A: predictions about other girls. Panel B: predictions about mothers. Panel C: predictions about boys. Panel D: predictions about teachers. Higher values indicate more (predicted) progressive attitudes. All specifications include stratum fixed effects and LASSO-selected baseline controls, with standard errors clustered at school level. Bars show 95% confidence intervals.

Table A19: *Effects on teasing and bullying (reversed), disaggregated*

	(1) Self-report Combined (rev.) (Z)	(2) Self-report Teasing only (rev.) (Z)	(3) Self-report Bullying only (rev.) (Z)	(4) Peer-report Combined (rev.) (Z)	(5) Peer-report Teasing only (rev.) (Z)	(6) Peer-report Bullying only (rev.) (Z)	(7) Teacher-report Combined (rev.) (Z)	(8) Teacher-report Teasing only (rev.) (Z)	(9) Teacher-report Bullying only (rev.) (Z)
Base Only	-0.029	-0.108	0.083*	0.074	0.029	0.081*	0.067	0.023	0.074
	(0.059)	(0.066)	(0.046)	(0.052)	(0.061)	(0.046)	(0.056)	(0.067)	(0.049)
Base + YGL	0.026	-0.020	0.069*	0.061	0.006	0.070	0.028	-0.032	0.052
	(0.053)	(0.058)	(0.040)	(0.050)	(0.059)	(0.044)	(0.052)	(0.061)	(0.047)
Observations	2,167	2,167	2,167	3,042	3,042	3,042	3,637	3,637	3,637
Data source	Girls	Girls	Girls	Girls	Girls	Girls	Teachers	Teachers	Teachers
p: Treated=0	0.893	0.386	0.062	0.146	0.786	0.065	0.434	0.776	0.192
p: Base Only=Base+YGL	0.293	0.109	0.735	0.779	0.694	0.787	0.391	0.327	0.579

Notes: All measures are reversed so that higher values indicate less teasing/bullying (better outcomes). Columns (1)-(3) are girls' reports about teasing and bullying towards themselves. Columns (4)-(6) are girls' reports about teasing and bullying towards other girls, using a randomly selected subset of 2 or 3 girls in their grade level (and removing cases where the respondent did not know the other girl). Columns (7)-(9) are teachers' reports about randomly selected girls in their class. Each teacher reports on a randomly selected sample of 3 girls from their class; the observations are at the *teacher* × *girl* level. Columns (1), (4) and (7) combine reports of both teasing and severe bullying. Columns (2), (5), and (8) include only teasing. Columns (3), (6) and (9) include only severe bullying. Combined indexes are constructed using loadings from exploratory factor analysis. The same relative loadings are used for teasing and bullying. Standard errors are clustered at the school level and are in parentheses. Sample includes all girls at endline interviewed in person. Controls include stratum fixed effects and the baseline controls selected by double LASSO. All outcomes are in control group standard deviations.

Table A20: *School social climate: robustness to excluding girls who dropped out*

Outcome group:		Bullying (reversed)			Network connection			
	(1) Combined index: school psychosocial climate (Z)	(2) Teasing (Z)	(3) Intimidation/ harassment (Z)	(4) About menstr. (Z)	(5) Overall (Z)	(6) w/ Girls only (Z)	(7) w/ Boys only (Z)	(8) Cleaning collaboration (Z)
Base Only	0.086	-0.134*	0.096*	-0.158**	0.150*	0.070	0.179**	-0.017
	(0.067)	(0.071)	(0.050)	(0.067)	(0.087)	(0.081)	(0.086)	(0.067)
Base + YGL	0.172***	-0.036	0.063	-0.227***	0.245***	0.173***	0.246***	0.226***
	(0.058)	(0.063)	(0.042)	(0.062)	(0.079)	(0.066)	(0.083)	(0.066)
Observations	1,795	1,795	1,795	1,795	1,795	1,795	1,795	1,795
Data source	Girls	Girls	Girls	Girls	Girls	Girls	Girls	Girls
p: Treated=0	0.008	0.262	0.082	<0.001	0.003	0.024	0.003	0.022
p: Base Only=Base+YGL	0.154	0.102	0.456	0.264	0.199	0.134	0.358	<0.001

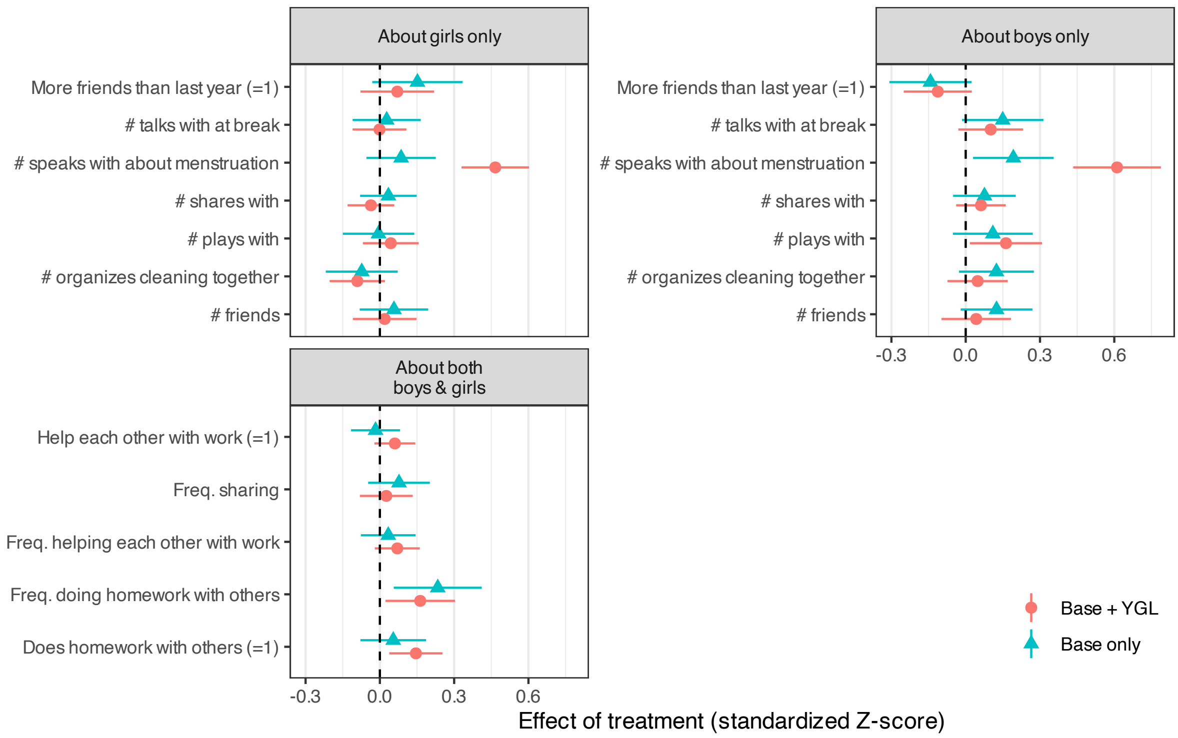
Notes: This table replicates Table 7 but restricts to girls who were still enrolled in school at endline (N=1,795). The main table includes 372 additional girls who had dropped out of school, for whom some bullying and network questions (which reference "classmates") may be less applicable. Results are similar in magnitude and significance, suggesting the main findings are robust to this sample restriction. All bullying measures are reversed so positive values indicate less bullying (better outcomes). The cleaning collaboration index (column 8) is shown separately and is not included in the combined social climate index in Table 7, as these questions were not asked to girls who had dropped out of school. Standard errors clustered at school level in parentheses. Controls include stratum fixed effects and LASSO-selected baseline controls. ** $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.*

Table A21: *Teasing by boys or by other girls*

VARIABLES	(1) Teased by boy(s) (Self-reported) (=1)	(2) Teased by girls(s) (Self-reported) (=1)	(3) Teased by boy(s) (Peer-reported) (=1)	(4) Teased by girls(s) (Peer-reported) (=1)	(5) Teased about menstr. by boy(s) (=1)	(6) Teased about menstr. by girls(s) (=1)
Base Only	0.024	0.022	-0.009	-0.013	0.034**	-0.003
	(0.026)	(0.022)	(0.026)	(0.020)	(0.016)	(0.012)
Base + YGL	0.006	0.009	0.013	0.007	0.068***	0.007
	(0.022)	(0.020)	(0.024)	(0.018)	(0.014)	(0.011)
Observations	2,167	2,167	3,042	3,042	2,256	2,256
Data source	Girls	Girls	Girls	Girls	Girls	Girls
Control mean	.241	.222	.252	.157	.05	.048
Control SD	.428	.416	.435	.364	.218	.214
p: Treated = 0	.5892	.4896	.8341	.9663	<0.001	.6125
p: Base Only = Base + YGL	.419	.5294	.382	.3633	.0241	.3718

Notes: Columns (1), (3), (5): outcome = 1 when girl was teased by a boy. Columns (2), (4), (6): outcome = 1 when girl was teased by a girl. Columns (1) and (2) are girls' reports about teasing towards themselves. Columns (3) and (4) are girls reports about randomly selected girls in their class. Columns (5) and (6) are self-reports about teasing specifically about menstruation. Sample includes all girls at endline interviewed in person. Controls include stratum fixed effects and the baseline controls selected by double LASSO. Standard errors are clustered at the school level and are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure A22: *Individual network questions*



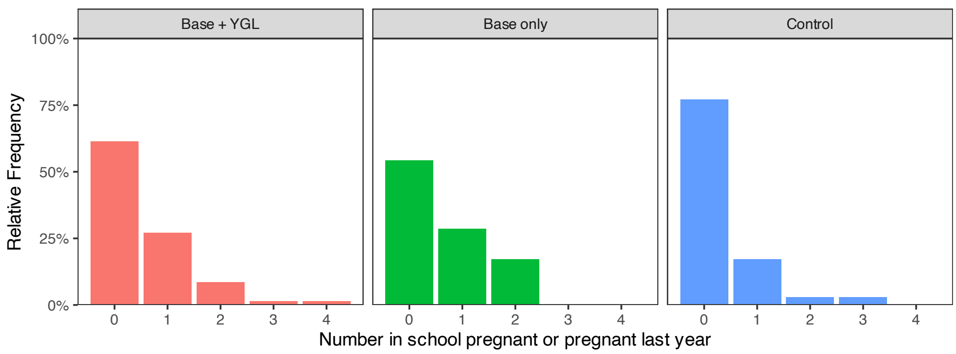
Notes: This figure shows the effects on each outcome that is used in the network index in Table 7. Only girls included in the baseline sample were used. All outcomes are normalized using the control mean and standard deviation from the endline. All regressions control for stratum fixed effects and confidence intervals are calculated based on standard errors clustered at the school level.

Table A23: *Effects on marriage and pregnancy*

Data source:	Mothers, Endline 1		Girls, Endline 2	
	(1)	(2)	(3)	(4)
Dependent variable:	Has been married (=1)	Pregnant now or last year (=1)	At least one girl pregnant in school, most recent school year (=1)	Number of girls pregnant in school, most recent school year
Base Only	0.009	0.016**	0.100***	0.136**
	(0.008)	(0.007)	(0.035)	(0.064)
Base + YGL	0.002	0.009	0.006	0.000
	(0.007)	(0.007)	(0.029)	(0.052)
Observations	2,287	2,287	2,702	2,702
Data source	Mothers	Mothers	Girls	Girls
Control mean	.024	.02	.12	.176
Control SD	.152	.141	.325	.598
p: Treated=0	0.502	0.066	0.195	0.381

Notes: Columns 1-2: mothers' reports about daughters at endline 1. Column 1: girl has been married (binary). Column 2: girl pregnant now or in last year (binary). Columns 3-4: girls' reports at endline 2 about pregnancies in their school. Column 3: at least one girl pregnant in most recent school year (binary). Column 4: number of girls pregnant in most recent school year. Standard errors clustered at school level in parentheses. Controls include LASSO-selected baseline controls for endline 1 and stratum fixed effects for all columns.

Figure A24: *Histogram: number of pregnant girls in each school*



Notes: Observation is at the school level. Shows the histogram of the number of pregnant girls in a school at endline, divided by the total number of schools in each treatment group in order to maintain comparability between *Base+YGL* (70 schools), *Base Only* (35 schools) and *Control* schools (35 schools). Pregnancy = 1 if mothers responded that their daughter was currently pregnant or had been pregnant in the last year.

Table A25: *Heterogeneous effects on school social climate by primary vs. secondary*

	(1) Teasing (Self-report)	(2) Teasing about menstr. (Self-report)	(3) Bullying (Self-report)	(4) Teasing and bullying (Teacher report)	(5) Network connection index	(6) Network connection w. girls only	(7) Network connection w. boys only
Panel A. Primary Schools							
Base Only	-0.015	-0.076	0.085	-0.174*	0.019	-0.061	0.076
	(0.102)	(0.089)	(0.072)	(0.089)	(0.105)	(0.098)	(0.100)
Base + YGL	-0.157*	-0.321***	-0.018	-0.024	0.160**	0.097	0.167**
	(0.092)	(0.073)	(0.069)	(0.092)	(0.072)	(0.070)	(0.074)
Observations	875	875	875	1,365	875	875	875
Control mean	.185	.055	-.04	.049	-.21	-.03	-.29
p: Treated=0	0.218	<0.001	0.787	0.471	0.096	0.521	0.040
p: Base Only=Base+YGL	0.087	0.012	0.118	0.023	0.181	0.077	0.388
Panel B. Secondary Schools							
Base Only	-0.164**	-0.179*	0.082	-0.004	0.287**	0.167	0.299**
	(0.077)	(0.098)	(0.057)	(0.069)	(0.124)	(0.107)	(0.120)
Base + YGL	0.064	-0.236***	0.130***	-0.029	0.344***	0.245***	0.333***
	(0.068)	(0.090)	(0.046)	(0.062)	(0.112)	(0.092)	(0.114)
Observations	1,292	1,292	1,292	2,272	1,292	1,292	1,292
Control mean	-.13	-.04	.032	-.03	.155	.024	.216
p: Treated=0	0.889	0.009	0.011	0.711	0.002	0.013	0.003
p: Base Only=Base+YGL	<0.001	0.501	0.299	0.669	0.554	0.350	0.706
Panel C. Interaction term *p*-values							
p: Base Only x Secondary = 0	0.362	0.435	0.969	0.133	0.104	0.116	0.164
p: Base+YGL x Secondary = 0	0.065*	0.462	0.075*	0.960	0.154	0.200	0.198
p: (Base Only x Sec.) = (Base+YGL x Sec.)	0.002***	0.144	0.061*	0.045**	0.601	0.517	0.754
p: Treated x Secondary = 0	0.356	0.819	0.170	0.688	0.112	0.135	0.163

Notes: All teasing and bullying measures are reversed: positive values indicate less teasing/bullying (better outcomes). Panel A: primary schools. Panel B: secondary schools. Columns 1-3: teasing and bullying measures from girls' self-reports. Column 4: teachers' reports. Columns 5-7: network connection indices. Panel C: p-values from pooled regressions testing school level interactions. All outcomes in Z-scores. Standard errors clustered at school level in parentheses. Controls include stratum fixed effects and LASSO-selected baseline controls (columns 1-3, 5-7); column 4 uses OLS without LASSO. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A26: *Learning mediation analysis: controlling for main outcome indexes*

	(1) Learning tests	(2) Learning tests	(3) Learning tests	(4) Learning tests	(5) Learning tests	(6) Learning tests	(7) Learning tests	(8) Learning tests	(9) Learning tests
Base Only	0.199**	0.202**	0.143*	0.091	0.154*	0.068	0.158**	0.171**	0.195**
	(0.081)	(0.081)	(0.080)	(0.073)	(0.084)	(0.077)	(0.079)	(0.073)	(0.081)
Base + YGL	0.117**	0.114*	0.104*	-0.091	0.077	-0.040	0.047	0.009	0.108*
	(0.058)	(0.058)	(0.057)	(0.058)	(0.063)	(0.059)	(0.059)	(0.065)	(0.057)
Physical health		0.086***							
		(0.021)							
Mental health (heart rate, rev.)			0.037*						
			(0.022)						
Menstrual stigma, self-report (rev.)				0.254***					
				(0.025)					
Menstrual stigma, observed (rev.)					0.072*				
					(0.042)				
Hygiene knowledge						0.316***			
						(0.029)			
Hygiene behavior							0.105***		
							(0.021)		
Observed cleanliness								0.206***	
								(0.047)	
School social climate									0.047**
									(0.019)
Observations	2,167	2,167	1,904	2,167	2,167	2,167	2,167	2,167	2,167

Notes: Outcome is learning test scores (Z). Each column includes the Base Only and Base+YGL treatment indicators plus one of the main outcome indexes (Z-scores, as in Figure A1) as a potential mediator: (1) no mediator, (2) physical health, (3) heart rate (reversed), (4) menstrual stigma, self-report (reversed), (5) menstrual stigma, observed (reversed), (6) hygiene knowledge, (7) hygiene behavior, (8) school psychosocial climate. Standard errors clustered at school level in parentheses. Controls include stratum fixed effects. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A27: *Correlations Between Learning Test Scores and Main Outcome Indices*

	Control Group		Full Sample	
Outcome	Learning	N	Learning	N
Physical health	0.171**	527	0.103***	2167
	(0.082) [0.045]		(0.027) [0.000]	
Mental health (heart rate, rev.)	-0.018	468	0.041*	1904
	(0.075) [0.815]		(0.024) [0.097]	
Menstrual stigma (rev.)	0.135***	527	0.228***	2167
	(0.037) [0.001]		(0.022) [0.000]	
Hygiene knowledge	0.141***	527	0.250***	2167
	(0.052) [0.010]		(0.022) [0.000]	
Hygiene behavior	0.079	527	0.130***	2167
	(0.047) [0.102]		(0.024) [0.000]	
School social climate	0.042	527	0.063**	2167
	(0.052) [0.418]		(0.026) [0.016]	

Notes: Each row reports the coefficient from a separate OLS regression of the outcome index (Z-scored) on the learning test index (Z-scored). Control group regressions include stratum fixed effects only; full sample regressions additionally control for treatment indicators. All regressions cluster standard errors at the school level. Standard errors in parentheses; p-values in brackets. * p<0.10, ** p<0.05, *** p<0.01.

Table A28: *Restrictions to girls' activities due to menstruation (among girls who reached menarche at baseline)*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Any (=1)	Absent (=1)	Absent because ill (=1)	Left school during day (=1)	Couldn't participate fully (=1)	Missed social activities (=1)	Other activities (=1)
Base Only	-0.005	-0.021	-0.000	-0.047*	0.002	0.001	-0.008
	(0.046)	(0.027)	(0.019)	(0.025)	(0.035)	(0.025)	(0.026)
Base + YGL	-0.074**	-0.036*	-0.017	-0.033*	-0.059**	-0.029	0.005
	(0.034)	(0.019)	(0.017)	(0.020)	(0.023)	(0.022)	(0.025)
Observations	1,055	1,055	1,052	1,055	1,055	1,055	1,055
Data source	Girls	Girls	Girls	Girls	Girls	Girls	Girls
Control mean	.375	.117	.05	.108	.225	.096	.113
Control SD	.485	.322	.219	.311	.418	.295	.317
p: Treated = 0	.1174	.1069	.5066	.0541	.1049	.3843	.9719
p: Base Only = Base + YGL	.1234	.5282	.2358	.4908	.0647	.1067	.5638

Notes: Shows girls' reports of restrictions to their activities because of menstruation at endline. Sample only includes 1,053 girls who reported having started menarche at *baseline* and interviewed in person at endline. We use baseline menarche rather than endline because at endline there is a small (5%) but significant difference in the share of girls report having started menarche, possibly related to reduced menstrual stigma leading to less underreporting in treatment schools. Standard errors are clustered at the school level and are in parentheses. Controls include stratum fixed effects and the baseline controls selected by double LASSO. Column (1): outcome is whether any of columns (2)-(7) are true. Column (2): has the girl ever been absent from school because of menstruation? Column (3): has the girl ever been absent because they were ill because of menstruation? Column (4): has the girl ever left school during the school day because of menstruation? Column (5): does the girl report not being able to participate fully in school activities because of menstruation? Column (6): has the girl missed any social activities because of menstruation? Column (7): has the girl been restricted from any other activities due to menstruation? The recall period used is whether these restrictions have *ever* occurred to them, so the questions were asked to all girls including those who had dropped out of school in the last year.

Table A29: *Heterogeneity by whether girls started menarche between baseline and endline*

	Base Only	Base+YGL	Newly menstr.	Base × Newly menstr.	Base+YGL × Newly menstr.	N
Learning tests	0.135*	0.060	0.016	0.072	0.070	2167
	(0.071) [0.057]	(0.055) [0.279]	(0.094) [0.862]	(0.147) [0.626]	(0.100) [0.481]	
Physical health	–0.089	–0.031	–0.203*	0.077	0.128	2256
	(0.066) [0.174]	(0.051) [0.543]	(0.109) [0.063]	(0.160) [0.632]	(0.123) [0.297]	
Mental health (heart rate rev.)	0.102*	0.108**	0.058	0.040	0.021	1904
	(0.060) [0.089]	(0.055) [0.048]	(0.106) [0.586]	(0.129) [0.755]	(0.115) [0.856]	
Menstrual stigma (rev.)	0.371***	0.735***	0.096	0.002	0.081	2256
	(0.073) [0.000]	(0.059) [0.000]	(0.090) [0.284]	(0.115) [0.987]	(0.102) [0.430]	
Hygiene knowledge	0.374***	0.412***	0.289***	–0.046	0.056	2167
	(0.071) [0.000]	(0.060) [0.000]	(0.101) [0.004]	(0.114) [0.688]	(0.100) [0.574]	
Hygiene behavior	0.248***	0.423***	1.046***	0.203*	0.311***	2256
	(0.078) [0.001]	(0.062) [0.000]	(0.083) [0.000]	(0.123) [0.099]	(0.097) [0.001]	
School social climate	0.051	0.124*	–0.006	–0.091	0.159	2167
	(0.076) [0.500]	(0.063) [0.052]	(0.127) [0.963]	(0.168) [0.589]	(0.132) [0.227]	

Notes: Outcomes are the same indices as in Figure A1. Each row is a separate regression. *Newly menstr.* = 1 if the girl started menarche between baseline and endline (not menstruating at baseline but menstruating by endline). Columns 4-5 show interactions testing whether treatment effects differ for girls that entered menarche between baseline and endline. Standard errors clustered at school level in parentheses; p-values in brackets. Controls include stratum fixed effects and LASSO-selected baseline controls. ** p < 0.1, ** p < 0.05, *** p < 0.01.*

Table A30: *Effects on teachers*

	(1) Progressive Menstruation Attitudes (Z)	(2) Mens. Knowledge Index (Z)	(3) Progressive Gender Attitudes (Z)	(4) Absent Last Month (=1)
Base Only	0.668***	0.135	0.093	-0.089
	(0.085)	(0.113)	(0.123)	(0.055)
Base + YGL	0.681***	0.166*	0.141	-0.007
	(0.086)	(0.093)	(0.111)	(0.048)
Observations	743	743	726	743
Data source	Teachers	Teachers	Teachers	Teachers
Control mean	0	0	0	.497
Control SD	1	1	1	.501
p: Treated=0	<0.001	0.079	0.226	0.441
p: Base Only=Base+YGL	0.815	0.760	0.649	0.111

Notes: Teacher survey outcomes. Column 1: progressive menstruation attitudes (Z-score combining attitudes about whether girls should attend school during menstruation). Column 2: menstruation knowledge (Z-score of true/false questions). Column 3: progressive gender attitudes (Z-score, disagreement with statements like "Boys are better leaders than girls"). Column 4: teacher reported being absent at least once last month (binary). Sample includes both male and female teachers. Standard errors clustered at school level in parentheses. Controls: stratum fixed effects and primary school dummy.

Table A31: *Changes in academic performance of selected YGLs*

	Change in grade (/20)	Change in class rank
Intercept	−1.173***	6.429***
	(0.205)	(0.820)
Num.Obs.	191	191
Grade/rank at baseline	11.7	4.8

Notes: Sample includes all girls who were selected as YGLs (in January 2022) *and* interviewed at endline (in summer 2023) and who were in school at endline. 255 YGLs were selected, 46 were not found at endline, and 18 of those found were no longer in school and thus did not have school GPAs to elicit. During YGL nomination and at endline, girls were asked about their school GPAs (out of a maximum of 20) and their rank in class. Column (1) shows the change in school GPAs from nomination to endline. Column (2) shows the change in class rank from nomination to endline (with a higher rank indicating relatively worse academic performance). Regressions control for stratum fixed effects, and standard errors are clustered at the school level. ** p < 0.1, ** p < 0.05, *** p < 0.01.*

Table A32: *Heterogeneity by school size*

Outcome	Base Only	Base+YGL	Large School	Base Only × Large	Base+YGL × Large	N
Learning tests	0.382***	0.060	0.516***	–0.309**	0.138	2167
	(0.093) [0.000]	(0.063) [0.345]	(0.179) [0.004]	(0.132) [0.019]	(0.099) [0.165]	
Attendance (spot checks)	–0.088	–0.101	–0.066	0.169	0.274	698
	(0.128) [0.494]	(0.111) [0.363]	(0.193) [0.735]	(0.202) [0.403]	(0.174) [0.118]	
Physical health	–0.067	–0.029	–0.108	–0.010	0.053	2256
	(0.085) [0.429]	(0.069) [0.674]	(0.193) [0.577]	(0.111) [0.926]	(0.092) [0.563]	
Mental health (heart rate rev.)	0.172	0.111	0.329	–0.073	0.057	1904
	(0.105) [0.102]	(0.089) [0.213]	(0.229) [0.152]	(0.124) [0.559]	(0.105) [0.587]	
Menstrual stigma (rev.)	0.599***	0.863***	–0.115	–0.273*	–0.110	2256
	(0.120) [0.000]	(0.076) [0.000]	(0.167) [0.491]	(0.146) [0.061]	(0.111) [0.321]	
Menstrual stigma (lab rev.)	0.981***	1.014***	0.844	–0.670**	–0.823***	140
	(0.243) [0.000]	(0.229) [0.000]	(0.679) [0.216]	(0.336) [0.048]	(0.298) [0.007]	
Hygiene knowledge	0.520***	0.562***	–0.030	–0.166	–0.142	2167
	(0.128) [0.000]	(0.088) [0.000]	(0.192) [0.877]	(0.150) [0.270]	(0.114) [0.215]	
Hygiene behavior	0.345***	0.512***	–0.034	–0.052	0.105	2256
	(0.106) [0.001]	(0.085) [0.000]	(0.245) [0.891]	(0.149) [0.726]	(0.125) [0.400]	
Observed cleanliness	0.199	0.349	–0.114	–0.217	0.240	140
	(0.277) [0.475]	(0.226) [0.124]	(0.465) [0.807]	(0.360) [0.548]	(0.283) [0.399]	
School social climate	0.023	0.000	–0.388**	0.069	0.289***	2167
	(0.110) [0.836]	(0.077) [1.000]	(0.174) [0.026]	(0.142) [0.629]	(0.109) [0.008]	

Notes: Outcomes are the same indices as in Figure A1. *Large school* = 1 if school has ≥ median number of girls aged 12+ at baseline. Columns 4-5: interactions testing whether treatment effects differ by school size. Standard errors clustered at school level in parentheses; p-values in brackets. Controls include stratum fixed effects and LASSO-selected baseline controls. The Attendance (spot checks) row uses school×wave level data (excluding baseline wave), with wave fixed effects and OLS. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B Robustness of main results

This appendix shows that our main conclusions are not altered when using randomization inference p-values (Section B.1), or threatened by spillovers across girls (Section B.2), attrition (Section B.3), composition effects of who remains in school (Section B.4), multiple hypothesis testing (Section B.6), crowd-in of other investments (Section B.5).

B.1 Randomization inference *p*-values

Table B1: *Robustness: randomization inference p-values*

	Base+YGL	Base Only	Pooled	Base Only vs. Base+YGL
Learning tests	.043	.023	.015	.201
Attendance (spot checks)	.686	.92	.717	.707
Mental health (heart rate, rev.)	.03	.067	.011	.983
Physical health	.939	.229	.658	.161
Menstrual stigma (rev.)	<0.001	<0.001	<0.001	<0.001
Menstrual stigma (observed, rev.)†	.001	<0.001	.001	.690
Hygiene knowledge	<0.001	<0.001	<0.001	.278
Hygiene behavior	<0.001	.001	<0.001	<0.001
Observed cleanliness†	.001	.66	.02	.007
School social climate	.005	.564	.043	.063

Notes: Randomization inference p-values from 1000 replications. Outcomes are the same as in Figure A1. Columns show p-values for: (1) Base+YGL effect, (2) Base Only effect, (3) Pooled treatment effect, (4) Test of equality between Base Only and Base+YGL. Controls include stratum fixed effects and LASSO-selected baseline controls (except attendance which uses OLS with wave fixed effects).

B.2 Spillovers

This section shows evidence that spillovers between schools do not drive our main results.

Our sample of 140 schools was selected to avoid cases where schools were very close to each other. Nevertheless, since these schools were selected from only 3 districts in the region of Amoron'i Mania, schools and households are in some cases close enough that there is a risk of spillover effects from treatment schools to control schools.

Methodology. To evaluate whether such spillovers threaten our main results, we follow the methodology of Borusyak & Hull (2023) to test for spillover effects based on geographic proximity between treatment and control schools. We use school GPS coordinates from administrative data and the Endline 2 survey to calculate distances between all school pairs.

We construct spillover exposure measures based on baseline enrollment in nearby treated schools. Specifically, for each school j, we calculate:

$$\text{ActualNearbyTreated}_{jk} = \sum_{s \in \mathcal{T}_j, d_{js} \leq k} \text{Enrollment}_{s,\text{baseline}}/100$$

$$\text{ExpectedNearbyTreated}_{jk} = \frac{1}{R} \sum_{r=1}^{R} \sum_{s \in \mathcal{T}_r, d_{js} \leq k} \text{Enrollment}_{s,\text{baseline}}/100$$

where $\text{Enrollment}_{s,\text{baseline}}$ denotes the number of students enrolled in school s at baseline, $\mathcal{T}_j$ denotes the set of actually treated schools that excludes j, d_{js} is the distance in kilometers between schools j and s, $k \in \{3, 5\}$ km is the distance threshold, and $\mathcal{T}_r$ represents the set of schools assigned to treatment in randomization r, again excluding school j itself so that the actual and expected exposure measures are directly comparable. The expected enrollment is calculated as the average over $R = 10,000$ rerandomizations of treatment assignment within strata, following the original randomization protocol. Enrollment measures are scaled by dividing by 100 to ease coefficient interpretability.

This approach addresses the concern that schools may be spatially clustered in ways correlated with both treatment assignment and outcomes. By controlling for the expected spillover exposure under random assignment, we isolate the causal effect of actual spillover exposure net of any geographic clustering (e.g., this controls for differences in population density that would confound spillover estimates).

There is substantial variation in potential spillover exposure: across the whole sample 28% of girls have zero treated students from another school within 3km, the median number of treated students from another school within 3km is 254, and the 90th percentile is 784. The equivalent figures for the control group girls are: 26% with 0, median of 257, and 90th percentile of 727. To interpret these numbers, note a) that distances are measured as the crow flies, which in the mountainous terrain translates to much larger actual distances; and b) given that drop-out rates in primary school are high, many children that treatment or control group children interact with in their villages are out of school and hence not treated.

We then estimate the following specification for girl i in school j:

$$Y_{ij} = \beta_0 + \beta_1 \text{BaseOnly}_j + \beta_2 \text{BaseYGL}_j + \gamma \text{ActualNearbyTreated}_{jk} + \delta \text{ExpectedNearbyTreated}_{jk} + \mathbf{X}'_{ij}\mathbf{\Gamma} + \varepsilon_{ij} \tag{2}$$

where γ captures the spillover effect per 100 students enrolled in nearby treated schools, controlling for the expected geographic exposure captured by δ.

Results. Table B2 presents the results for our main outcomes using both 3km and 5km distance thresholds. We find no consistent evidence of spillovers that would undermine our main results. The spillover coefficients are generally small in magnitude and not statistically significant, providing reassurance that our main treatment effects are not confounded by spillovers between nearby schools. The only significant estimate is a negative effect of treated students within 5km (0.024 SD per 100 students) on physical health (p=0.013). Since this estimate is not robust to the change in distance threshold, and would lead to a positive bias on the treatment effect estimates on health that would not explain the null overall results, we do not see this as a threat to our main conclusions.

Stigma trends in control vs. treatment schools. A related concern is whether stigma measures in control schools change over time in response to nearby program implementation. Figure B3 plots a composite menstrual stigma index for control and treatment (pooled) groups at baseline, midline, and endline. The composite averages eight Z-scored items measured identically at all three waves. The control group shows a gradual improvement over time, while treatment schools show a larger and steeper decline in stigma. The modest improvement in control schools suggests some secular trend, limited positive spillovers, or a response to the light-touch district-level awareness activities that reached treatment and control schools alike (see footnote 9); if anything, this implies our treatment effect estimates are a lower bound on the true program impact.

B.3 Tracking, attrition weights and Lee bounds

Tracking strategy. Regular data collection ("phase 1") occurred between May and July 2023. This phase included repeated visits to absent girls, mothers, and teachers. An intensive tracking phase ("phase 2") was then organized in October 2023 following a multi-pronged approach to minimize attrition.

We revisited all schools where there was still a relatively large share of girls to track after the regular data collection for any of the following reasons: (i) schools where many girls had been absent during regular tracking because of seasonal work in mines, and where girls were expected to return to their homestead in October; (ii) schools where there were many refusals because the regular survey visit was too close to exams, and girls asked the team to return later; (iii) schools for which the survey team did not have enough time for the girls' survey during the regular phase; and (iv) schools where fewer than two teachers were surveyed in the regular tracking phase.

In addition to this set of schools, we randomly selected 19 of the remaining 62 schools where at least one girl was missing, and included these 19 schools for intensive tracking. This resulted in a total sample of 83 schools to be revisited.

Table B2: *Spillover effects*

	Treated in 3km (per 100 students) (1)	Treated in 5km (per 100 students) (2)	N
Learning tests	0.012	0.010	2167
	(0.018) [0.500]	(0.010) [0.324]	
Physical health	–0.005	–0.024**	2256
	(0.018) [0.787]	(0.010) [0.013]	
Mental health (heart rate rev.)	–0.009	0.006	1904
	(0.017) [0.576]	(0.009) [0.522]	
Menstrual stigma (rev.)	0.027	0.005	2256
	(0.021) [0.201]	(0.011) [0.659]	
Menstrual stigma (lab rev.)	–0.007	0.003	2256
	(0.041) [0.855]	(0.027) [0.914]	
Hygiene knowledge	–0.017	–0.001	2167
	(0.021) [0.430]	(0.010) [0.948]	
Hygiene behavior	0.027	–0.011	2256
	(0.023) [0.246]	(0.014) [0.440]	
School social climate	0.012	–0.021	2167
	(0.024) [0.623]	(0.013) [0.108]	

Notes: This table shows the estimates of spillover effects using methodology based on Borusyak & Hull (2023). Each coefficient is the estimate of γ from Equation 2. For example, column (1), row (1) indicates that each 100 treated students within 3km of a girl's school is associated with 0.012 SD higher learning test scores. 28% of girls have no treated schools within 3km; median is 254 enrolled treated students within 3km (control: 257, treatment: 251); 90th percentile is 784 (control: 727, treatment: 790). All regressions include stratum fixed effects and baseline controls selected by post-double-selection LASSO. Standard errors are clustered at the school level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

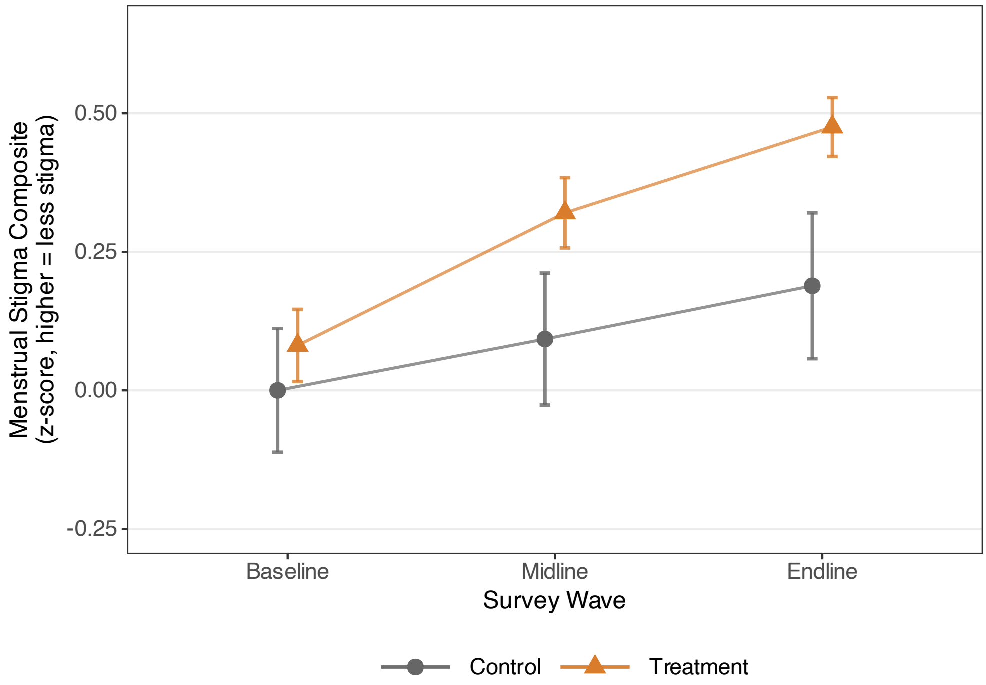


Figure B3: *Menstrual Stigma Trends: Control vs. Treatment Schools*

Notes: Mean composite stigma index at each survey wave for control and pooled treatment groups, with 95% confidence intervals. The composite averages eight Z-scored items measured identically at all three waves: two attitude items ("girls should be allowed at school during their period" and "girls should hide their period at school," reverse-coded; Likert 1–5), two appropriateness-of-discussion items ("is it appropriate to discuss periods with classmates/teachers?"; binary), and four second-order belief items (what classmates and guardians would think about school attendance during periods and hiding menstruation; Likert 1–5, negative items reverse-coded). Each item is Z-scored using the baseline control group mean and standard deviation. Higher values indicate less stigma. Sample restricted to girls surveyed at baseline; midline point conditions on midline participation. Standard errors clustered at the school level.

In the revisited schools, all remaining girl, mother, and teacher surveys were targeted for completion. Some teams targeted the original school locality, while in parallel, other teams looked for girls who had moved in their migration destination. This latter step involved in-person interviews in common migration destinations (e.g., the capital city Antananarivo, and regions nearby the study region of Amoron'i Mania), along with phone surveys (for 3.7% of the sample) in the remaining destinations.

Table B4: *Robustness: using sampling weights*

	(1) Base Only	(2) Base+YGL	(3) *p*-value: Treated=0	(4) *p*-value: Base=Base+YGL	(5) N
Learning tests	0.213***	0.132**	.0022	.2155	2167
	(0.072) [0.003]	(0.053) [0.013]			
Physical health	–0.094	–0.010	.3903	.1306	2256
	(0.058) [0.107]	(0.046) [0.833]			
Mental health (heart rate rev.)	0.107*	0.110**	.0214	.9511	1904
	(0.057) [0.061]	(0.050) [0.029]			
Menstrual stigma (rev.)	0.414***	0.806***	<0.001	<0.001	2256
	(0.067) [0.000]	(0.059) [0.000]			
Menstrual stigma (lab, rev.)	0.615***	0.549***	<0.001	.609	2256
	(0.151) [0.000]	(0.137) [0.000]			
Hygiene knowledge	0.409***	0.483***	<0.001	.1879	2167
	(0.067) [0.000]	(0.055) [0.000]			
Hygiene behavior	0.310***	0.563***	<0.001	<0.001	2256
	(0.077) [0.000]	(0.065) [0.000]			
School social climate	0.064	0.175***	.0081	.0904	2167
	(0.072) [0.377]	(0.055) [0.002]			

Notes: Robustness check using sampling weights to account for random selection into intensive tracking. Outcomes are the same indices as in Figure A1. Sampling weights are 62/19 for girls in the 19 schools randomly selected for intensive tracking (out of 62 eligible schools), and 1 for all other girls. Standard errors clustered at school level in parentheses; p-values in brackets. Controls include stratum fixed effects and LASSO-selected baseline controls.

Sampling weights. Given the random selection of a subset of schools for intensive tracking, we can calculate *effective tracking rates* (Orr et al., 2003). Intensive tracking was successful in finding a very large share of targeted girls, leading to an effective tracking rate of 99%.

We calculate sampling weights that account for the random selection of schools for intensive tracking. Sampling weights are 1 for all girls found during regular tracking, or in high-priority schools that were always revisited during intensive tracking. For the remaining 62 schools with missing girls, 19 were randomly selected for intensive tracking. Girls in these 19 schools receive a sampling weight of 62/19—the inverse of the selection probability—so that they represent all 62 schools in weighted analyses.

The results adjusting for sampling weights are shown in Table B4.

Lee bounds. Alternatively, Lee bounds (Lee, 2009) for the learning results for both treatment groups together show an upper bound of 0.18 SD (p<0.001) and a lower bound of 0.08 SD (p=0.135). Lee bounds for impacts on achievement tests show 0.35 SD (p<0.001) as upper bound and 0.16 SD as lower bound (p=0.056).

B.4 Composition effects

This section explores whether composition effects resulting from *who* remains in treated and untreated schools could explain some of the results. While composition effects do not threaten our main conclusions from girls' surveys (because girls were tracked and interviewed independently of whether they were still in school), they could confound some of the treatment effect estimates on (i) GPAs, (ii) boys; (iii) lab-in-the-field measures; and (iv) psychosocial climate in schools.

Since the intervention improves girls' academic performance, they more often progress to the next grade (Table B5, column 2). Especially for girls in the final year of primary or lower secondary school, this means they are more likely to have transitioned to a different school by endline in the

Table B5: *Composition effects: do girls stay in the same school as baseline?*

	(1) In School (=1)	(2) Progressed one class (unc.) (=1)	(3) Same school as baseline (unc.) (=1)
Base Only	-0.019	0.112***	-0.082**
	(0.024)	(0.040)	(0.032)
Base + YGL	-0.021	0.076**	-0.105***
	(0.022)	(0.036)	(0.030)
Observations	2,256	2,256	2,256
Data source	Girls	Girls	Girls
Control mean	.826	.515	.613
p: Treated=0	0.330	0.011	0.001
p: Base Only=Base+YGL	0.926	0.206	0.396

Notes: Column (1) shows whether girls are still enrolled in any school at endline; column (2) shows whether girls progressed to the next grade between baseline and endline (coded as zero for girls no longer enrolled); column (3) shows whether girls were enrolled in the same school at endline as at baseline (coded as zero for girls who transferred or dropped out). Standard errors clustered at school level in parentheses. Controls include stratum fixed effects and baseline controls selected by double LASSO. ** p < 0.1, ** p < 0.05, *** p < 0.01.*

treatment group, thus being less likely to be in the same school as at baseline (column 3). Moreover, in this context, drop-out after finishing a cycle (primary or lower secondary) is relatively high, and can help explain the negative (though insignificant) effect on overall enrollment (column 1).

Main outcomes. In general, this will not compromise the identification of our endline results because we measure final outcomes on all girls whom we find, regardless of whether they are in the same school as at baseline or have dropped out of school, and find no imbalances across treatments (Tables A2–A4). It does imply, however, that treatment girls will on average have been exposed to the intervention for less than 1.5 years, so the estimates should be considered as intention-to-treat estimates.

The selection moreover presents an important caveat to the results on grade point average (GPA) (Table 2, columns 4–5), since treatment girls on average received GPA scores from a higher grade at endline. This is likely to bias treatment effect estimates downwards if it is in general more difficult to receive higher GPAs in higher grades. In addition, among girls still enrolled at endline, those who remain at their baseline school are largely negatively selected relative to those who transferred to a different school (e.g., they are poorer, more delayed in their schooling, and have worse hygiene and menstruation knowledge at baseline; see Table B6). This similarly implies that our measure of GPA that only focuses on girls who did not switch schools will tend to underestimate the treatment effects.

In a similar way, this negative selection of who remains in school is likely to lead to underestimates of treatment effects on the lab-in-the-field outcomes and on boys. It also complicates the interpretation of the indicators of the schools' psychosocial environment.

Exploratory 2nd endline. We drew the sample for the exploratory second endline from enrollment registers in 2024. This sample is thus subject to composition effects. We discuss this further in Appendix F.

B.5 Crowd-in of other investments

We ask school directors to report on other school-level investments that could have been encouraged or displaced by the intervention. We find no evidence that the program crowded in or out other types of school-level investments, e.g., it did not change the student-teacher ratio, the types of contracts held by teachers, or school-level spending (Table B7, p=0.92 on pooled effect on overall index of

Table B6: *Baseline correlates of remaining at same school by endline*

	Outcome: Same school in baseline & endline (=1)			
	Control Group Only (1)	N	All Sample (2)	N
Girl's age (years)	–0.583*** (0.197) [0.006]	446	–0.700*** (0.097) [0.000]	1824
Girl already had her period (=1)	–0.088** (0.043) [0.047]	446	–0.125*** (0.021) [0.000]	1822
Wealth index (Z)	–0.246 (0.180) [0.181]	446	–0.338*** (0.101) [0.001]	1824
Years of school delay	0.378* (0.196) [0.063]	446	0.124 (0.086) [0.151]	1817
WASH knowledge index (Z)	–0.211** (0.090) [0.025]	446	–0.209*** (0.043) [0.000]	1822
WASH behaviour index (Z)	–0.256* (0.128) [0.053]	446	–0.093* (0.050) [0.062]	1822
Menstruation knowledge index (Z)	–0.198*** (0.072) [0.009]	446	–0.335*** (0.036) [0.000]	1822
Menstruation behaviour index (Z)	–0.265*** (0.082) [0.003]	446	–0.304*** (0.044) [0.000]	1824
Mental health index (Z)	0.242** (0.112) [0.039]	443	0.169*** (0.050) [0.001]	1819
CESD index (Z)	–0.002 (0.135) [0.988]	444	0.074 (0.051) [0.150]	1822
Menstrual health problems index (Z)	–0.143 (0.095) [0.143]	446	–0.181*** (0.044) [0.000]	1824
Menstruation inhibiting school activity index (Z)	–0.036 (0.037) [0.341]	446	–0.021 (0.018) [0.245]	1824
Received info. on hygiene (Z)	–0.181 (0.110) [0.110]	446	–0.157*** (0.045) [0.001]	1822
Received info. on menstruation (Z)	–0.035 (0.087) [0.692]	446	–0.121*** (0.040) [0.003]	1822
Willingness to speak about hygiene (Z)	–0.109 (0.106) [0.312]	446	–0.111** (0.045) [0.014]	1821
Willingness to speak about menstruation (Z)	0.093 (0.102) [0.368]	445	0.060 (0.046) [0.194]	1808
Exposure to bullying (self-reported) (Z)	–0.220* (0.130) [0.099]	446	–0.090* (0.053) [0.090]	1824
Attitudes towards menstruation (Z)	0.184** (0.091) [0.050]	445	0.075* (0.040) [0.064]	1814
Perceived norms related to menstruation (Z)	0.059 (0.095) [0.541]	446	–0.020 (0.050) [0.691]	1822

Notes: This table examines composition effects by regressing a baseline characteristic (shown in each row) on whether the girl is in the same school at endline as in baseline, conditional on being enrolled in any school at endline (so the comparison is between girls who stayed at their baseline school and girls who transferred to a different school). Column (1) shows the coefficient for the control group only; column (2) for the full sample, controlling for treatment dummies. For example, the first row indicates that being in the same school as baseline was associated with being 0.58 years younger on average in the control group. Standard errors clustered at school level in parentheses. ** $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.*

crowd-in).

Table B7: *No evidence that intervention crowded in or out other school-level investments*

	(1)
	School crowd-in index
Base Only	-0.019
	(0.219)
Base+YGL	-0.017
	(0.194)
Observations	140
Data source	Schools
Control mean	0
Control SD	1
p: Treated = 0	0.922
p: Base Only = Base+YGL	0.993

Notes: The *school crowd-in index* is constructed using factor analysis from 12 standardized variables: (1) teachers per student, (2) proportion of teachers paid by government, (3) proportion of teachers with permanent contracts, (4) number of parent association meetings, (5) parent association collection last month per student, (6) parent association collection last year per student, (7) infrastructure spending per student, (8) materials spending per student, (9) training spending per student, (10) activities spending per student, (11) other spending per student, and (12) indicator for infrastructure funded by non-project sources. The index is standardized (Z-score). All regressions include stratum fixed effects. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.6 Multiple hypothesis testing

Multiple hypothesis testing. Given that we evaluate multiple primary outcomes, we check robustness to multiple hypothesis testing. We calculate Anderson (2008) sharpened q-values that control for the false discovery rate for our set of key outcomes, treating our primary and secondary outcomes as two separate families.

Our substantive conclusions are not affected by this adjustment (Table B8): the pooled coefficient on learning has a q-value of 0.009 (compared to a p-value of 0.004); the pooled coefficient on heart rate is still significant at the 5% level; and the pooled coefficient on the stigma index is still significant at the 1% level.

Table B8: *Robustness: adjusting for multiple hypothesis testing*

Panel A: Pooled Treatment Effects

Family	Outcome	Estimate	p-val	q-val
Main	Learning tests	0.147	0.004	0.006
	Attendance (spot checks)	−0.020	0.721	0.317
	Physical health	−0.023	0.588	0.308
	Mental health (heart rate, rev.)	0.129	0.006	0.007
	Menstrual stigma (survey, rev.)	0.675	< 0.001	< 0.001
	Menstrual stigma (observed, rev.)	0.603	< 0.001	< 0.001
Intermediate	Hygiene knowledge	0.459	< 0.001	< 0.001
	Hygiene behavior	0.489	< 0.001	< 0.001
	Observed cleanliness	0.339	0.013	0.007
	Social climate	0.136	0.013	0.007

Panel B: Individual Treatment Arms

Family	Treatment	Outcome	Estimate	p-val	q-val
Main	Base Only	Learning tests	0.208	0.004	0.006
		Attendance (spot checks)	−0.008	0.908	0.257
		Physical health	−0.078	0.170	0.073
		Mental health (heart rate, rev.)	0.130	0.022	0.017
		Menstrual stigma (survey, rev.)	0.421	< 0.001	< 0.001
		Menstrual stigma (observed, rev.)	0.639	< 0.001	< 0.001
	Base + YGL	Learning tests	0.117	0.027	0.021
		Attendance (spot checks)	−0.027	0.643	0.347
		Physical health	0.004	0.922	0.444
		Mental health (heart rate, rev.)	0.128	0.010	0.014
		Menstrual stigma (survey, rev.)	0.805	< 0.001	< 0.001
		Menstrual stigma (observed, rev.)	0.585	< 0.001	< 0.001
Intermediate	Base Only	Hygiene knowledge	0.414	< 0.001	< 0.001
		Hygiene behavior	0.312	< 0.001	< 0.001
		Observed cleanliness	0.077	0.661	0.494
		Social climate	0.053	0.454	0.435
	Base + YGL	Hygiene knowledge	0.483	< 0.001	< 0.001
		Hygiene behavior	0.571	< 0.001	< 0.001
		Observed cleanliness	0.470	0.001	0.001
		Social climate	0.175	0.002	0.002

Notes: q-values are calculated using the Anderson (2008) sharpened process that controls for the false discovery rate across the 10 outcome variables shown. Raw p-values are shown alongside FDR-adjusted q-values. Panel A presents pooled treatment effects (any treatment vs. control), while Panel B shows effects for each individual treatment arm (Base Only and Base+YGL). In Panel A, all analyses in the same outcome *Family* are treated as a family to adjust for multiple hypothesis testing. In Panel B, all analyses with the same (*Family* × *Treatment*) are treated as a family. All estimates are standardized by the control group mean (Z-scores). All models are otherwise the same as the main specification.

C Heart rate

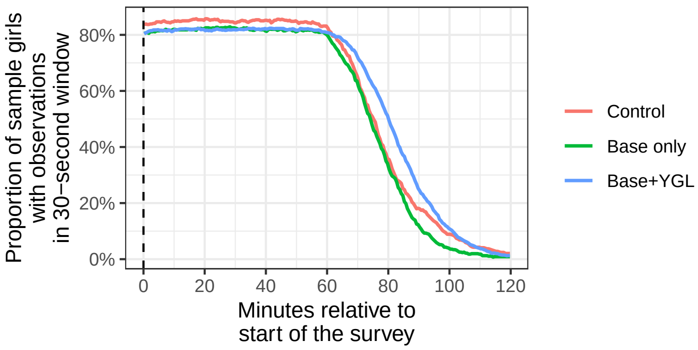
C.1 Heart rate data

We use the *Scosche Rhythm 24* heart rate wristband monitor to measure girls' heart rate during the endline survey. This has been shown to have comparable accuracy to chest-strap monitors (Reece et al., 2021), but has the advantage of being substantially less invasive.[40] Qualitative reports from enumerators indicate that girls frequently forgot that they were wearing the wristband by the end of the survey, suggesting that girls were not overly distracted or stressed by wearing the wristband itself.

Sample. In some cases, heart rate monitors were not fixed properly to girls' wrists, or the battery of monitors ran out on the field. Because of this, heart rate data is available for only 1,904 girls. However, the proportion of data available is equal across treatment arms (Table A5, column 6).

Time-window restrictions. For the main results in Figure 2, panel (a), we include all observations less than 120 minutes after the start of the survey. This includes almost all observations across all arms (see Figure C1). We do not cut off at an earlier window in the main specification, because the average length of the survey differs across treatments (e.g., because girls in *Base+YGL* schools are asked about the YGLs in their schools in a separate module), so any earlier cutoff could generate composition effects in the heart rate data. However, for tests of the *dynamics* over time (Figure 2 panel (b), and Table C7), we use a 60 minute cutoff window. This avoids highly noisy time periods with few observations driving the results, since the number of girls with data available drops rapidly after 60 minutes because many girls finish the survey (see Figure C1).

Figure C1: *Proportion of girls with heart rate observations available in each 30-second window*



Notes: Shows the proportion of sample girls in each treatment arm who have heart rate data available in each 30-second window of the survey. The number of girls with data drops after approximately 60 minutes as girls finish the survey.

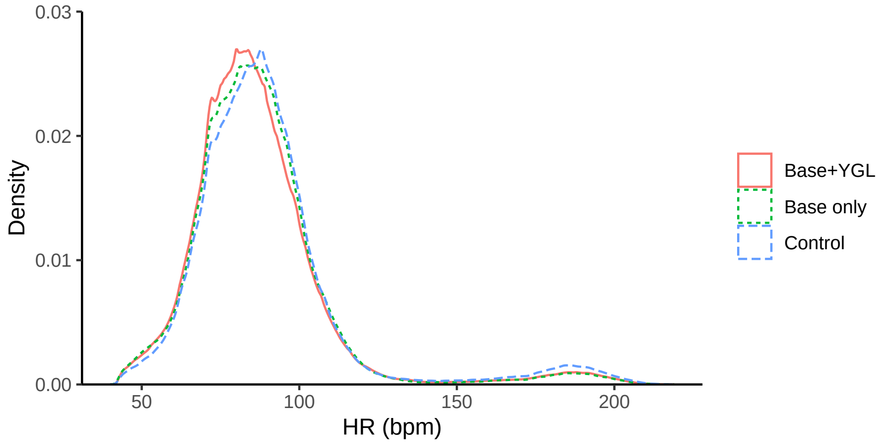
C.2 Heart rate robustness

Outliers. The distribution of heart rates shows an overall shift to the left in treatment groups (Figure C2). However, these distributions also show a grouping of large heart rate values at approximately double the magnitude of the mode. These could reflect measurement error, since they are above the typical range of heart rate that could be driven by psychological stress, possibly driven by the heart rate monitors erroneously counting *both* of the two main peaks in arterial pressure (called the systolic and diastolic peaks) that occur within one heartbeat (see e.g. Paradkar & Chowdhury (2015) for discussion of peak detection issues in arterial pressure). As robustness tests, we drop or halve any observations above 175 bpm, and show that this still yields significant pooled effects, although the coefficient magnitude drops (Table C3; p=0.09 and 0.07, columns 2 and 3). Results are also robust to using girl's mean heart rate during the interview rather than all the heart rates in each 30 second window (p=0.02, column 1) and to controlling for age fixed effects (p=0.02, column 4).

Enumerator familiarity effects. One concern is that girls in treatment schools have lower heart rates

[40] We initially piloted using these wristbands to measure heart rate variability, but pilot results indicated that this measure was highly noisy given the accuracy required to measure the *variation* in heart rate timing compared to the *mean* heart rate.

Figure C2: *Distribution of heart rate at the girl × second level*



Notes: Shows the distribution of heart rate observations in each treatment group, measured in beats per minute. Unit of observation is the *girl* × *second*. Sample includes all girls with heart rate data at endline.

Table C3: *Robustness of effects on heart rate to dealing with outliers*

	HR (bpm)			
	Individual-level	Drop outliers	Divide outliers	w/ Age FEs
Base only	−2.212*	−1.007	−1.068	−2.351**
	(1.169)	(0.800)	(0.789)	(1.129)
Base+YGL	−2.261**	−1.202*	−1.262*	−2.313**
	(1.018)	(0.695)	(0.690)	(0.993)
Num.Obs.	1904	8 877 622	9 105 420	9 105 420
N girls	1904	1904	1904	1904
Control mean	87.9	85	85.3	87.9
Control SD (between)	19.0	14.6	12.2	19.0
p : Treat = 0	0.021	0.087	0.069	0.021
p : Base=Base+YGL	0.961	0.77	0.768	0.961

Notes: Outcome is heart rate measured in beats per minute. Column (1): Observations are amalgamated to the *girl* level by taking the mean heart rate throughout the first 120 minutes of the endline survey. Column (2): observations are at the *girl* × *second* level, with observations with heart rate above 175 bpm dropped from the sample. Column (3): observations are at the *girl* × *second* level, with observations with heart rate above 175 bpm divided by 2 (based on the assumption that the heart rate monitor may be erroneously double counting both the systolic and diastolic peaks within a single heartbeat). This division creates a unimodal distribution of heart rate observations. Observations are at the girl × second level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the school level and are in parentheses. Sample includes all girls in endline interviewed with heart rate data. Controls include stratum fixed effects and the baseline controls selected by double LASSO. Columns 2 and 3 also include 30-second time-window fixed effects. Columns 4 additionally adds age (year) fixed effects.

during endline surveys because they are more familiar with strangers (program technicians and surveys) visiting schools and talking about menstruation-related topics, and are thus less stressed (in line with Zwane et al., 2011). While surveyors visit the schools an equal number of times across treatment groups, program technicians do indeed visit treatment schools more (e.g., *Base+YGL* schools received 9.2 visits on average between January and June 2023; *Base Only* schools: 4.3 visits; control schools: 0 visits). While we cannot rule out that this drives some of the effects, this story is made less plausible by the fact that the effects are similar in *Base+YGL* and *Base Only* schools despite almost double the visits in *YGL* schools, and because overall treatment effects are not driven by girls in schools with more technician visits (Table C4).

Table C4: *Relationship between technician visits and heart rate*

	(1) Heart rate (Z)	(2) Heart rate (Z)
Base Only	-0.131*	-0.148**
	(0.069)	(0.066)
Base + YGL	-0.125	-0.149***
	(0.082)	(0.057)
Treated × Technician visits (N)	-0.002	
	(0.006)	
Treated × Above med. visits (=1)		0.018
		(0.049)
Observations	1,904	1,904

Notes: This table shows that increased exposure to program technicians is unlikely to explain the effects on heart rate. *Technician visits (N)* is the number of times a program technician visited a school between January and June 2023, according to the NGO's own administrative tracking data. Uninteracted *Technician visits (N)* term is excluded because it is 0 in the control group. *Above med. visits (=1)* is 1 if number of visits is above the median within each treatment group. This is always 0 in the control group, so the uninteracted *Technician visits* term is not included. Stratum fixed effects are included but no LASSO-selected controls.

Raw correlations. The correlations between heart rate and other outcomes also provide some indication that they are measuring signal and not just picking up measurement error (Table C5). Girls from wealthier households have significantly lower heart rates (p<0.001), and at endline having a lower heart rate is correlated with higher learning scores (p=0.09), lower depression (p=0.05), and less teasing (p=0.03).

Table C5: *Correlations between heart rate and girls' characteristics*

Variable	Control Group	N	All Sample	N
Panel A: Baseline demographics				
Girl's age (years)	–0.068	468	–0.127***	1904
	(0.081) [0.403]		(0.040) [0.002]	
Girl already had her period (=1)	0.027	468	–0.008	1902
	(0.016) [0.101]		(0.009) [0.365]	
Wealth index (Z)	–0.150	468	–0.145***	1904
	(0.091) [0.108]		(0.036) [0.000]	
Years of school delay	–0.044	468	–0.080**	1897
	(0.080) [0.585]		(0.038) [0.038]	
Panel B: Main outcomes measured at endline				
Learning tests	0.013	468	–0.037*	1904
	(0.053) [0.816]		(0.022) [0.090]	
Physical health	0.078	468	0.038	1904
	(0.059) [0.198]		(0.025) [0.126]	
Menstrual stigma (rev.)	–0.017	468	–0.004	1904
	(0.036) [0.632]		(0.022) [0.849]	
Hygiene knowledge	0.006	468	–0.024	1904
	(0.041) [0.879]		(0.019) [0.213]	
Hygiene behavior	–0.009	468	–0.055**	1904
	(0.039) [0.817]		(0.026) [0.036]	
School psychosocial climate	0.079*	468	0.023	1904
	(0.040) [0.057]		(0.030) [0.448]	
Panel C: Mental health and bullying measured at endline				
Mental health (Z)	–0.096	468	–0.051**	1904
	(0.057) [0.103]		(0.026) [0.049]	
Self-esteem (Z)	–0.104*	468	–0.014	1904
	(0.061) [0.096]		(0.026) [0.593]	
Positive emotions (Z)	–0.097*	468	–0.045*	1904
	(0.052) [0.071]		(0.024) [0.061]	
Negative emotions (Z)	0.034	468	0.036	1904
	(0.078) [0.664]		(0.033) [0.274]	
Bullying: teasing (Z)	0.093***	468	0.057**	1904
	(0.032) [0.006]		(0.024) [0.017]	
Bullying: intimidation/harassment (Z)	0.057	468	0.002	1904
	(0.059) [0.338]		(0.029) [0.941]	

Notes: Each row shows coefficient from regressing heart rate on the standardized variable indicated in the row. Panel A: baseline demographics. Panel B: main outcome indices measured at endline. Panel C: mental health and bullying indicators at endline. Negative coefficients indicate the variable is associated with lower heart rate. Column 1: control group only. Column 2: full sample.

Table C6: *Is heart rate affected by whether girls report menstruating during the survey?*

	(1) Avg. Heart Rate	(2) Avg. Heart Rate
Base Only		-2.539**
		(1.241)
Base + YGL		-3.056***
		(1.099)
Menst. During Survey	-1.700	-1.683
	(3.346)	(3.350)
Base Only × Menst. During Survey		-3.858
		(3.834)
Base + YGL × Menst. During Survey		2.436
		(3.946)
(Pooled) Treatment	-2.875***	
	(1.041)	
Treatment × Menst. During Survey	0.608	
	(3.694)	
Observations	1,904	1,904

Notes: Outcome is girls' average heart rate (BPM) measured during endline survey. *Menst. During Survey* = 1 if girl reported menstruating at time of survey. Column 1: pooled treatment with interaction. Column 2: separate treatment arms with interactions. Interaction terms test whether stress reduction is larger for girls discussing menstruation while actively menstruating. No significant interactions found, suggesting treatment reduces general stress rather than stress specific to currently menstruating. Controls include stratum fixed effects.

Table C7: *Heart rate: dynamics during the survey*

	(1)	(2)	(3)	(4)
Treated (=1)	−1.705	−1.720	−1.372	−1.374
	(1.094)	(1.093)	(1.103)	(1.103)
After first question about menstr. (=1)	−1.219*	1.520*		
	(0.667)	(0.893)		
Treated * After first question about menstr.	−0.026	−0.016		
	(0.758)	(0.772)		
Time since start of survey (m)			−0.038*	
			(0.020)	
Treated * Time since start of survey			−0.012	−0.077***
			(0.022)	(0.022)
FEs: 30-second window		X		X
Num. obs.	6 678 438	6 678 438	6 678 438	6 678 438

Notes: Outcome is *heart rate (bpm)* measured at endline. All specifications use stratum fixed effects and LASSO-selected controls. Standard errors are clustered at the school level. Observations are at the girl × second level. Only observations in the first 60 minutes of the survey are included, since girls' interviews typically last between 60 and 90 minutes and so the number of observations begins to drop rapidly and become a selected sample. *Treated*=1 when a girl is in a *Base Only* or *Base+YGL* school, and 0 otherwise. Columns (1) and (2) examine the role of the first question about menstruation ("Have you ever had your period?"). Column (1) does not control for 30-second time-window fixed effects, but column (2) does. Columns (3) and (4) examine a linear time trend (measured in minutes). Column (4) also controls for time-window fixed effects. Since *Time since start of survey* is perfectly correlated with the time-window fixed effects, it is not included in the model in column (4).

D Lab-in-the-field measures of stigma

D.1 Design

Information supply experiment. This was an interactive lab-in-the-field exercise designed to measure girls' *revealed-preference* willingness to supply information about menstruation to others. In each school, this activity was carried out twice, with 6 girls and 2 boys participating in each group. All 8 students sat in a classroom. Each girl was taken out one by one, and had two new topics explained to them by enumerators: one topic related broadly to hygiene/health (e.g., what is pneumonia?), and the other related to menstruation (e.g., what are menstrual cramps?).

Girls then faced two binary choices. In one choice ("*anonymous*"), they were told that a video would be shown to the rest of the class about one of the topics, and they could choose which one — the menstruation-related topic, or the hygiene-related topic. No one would be told which video the girl picked. Since this choice was anonymous, it was designed to minimize concerns about taboo and stigma, and to measure the desire to supply the information to the rest of the class. In the other binary choice ("*explanation*"), by contrast, girls were told that they would have to *themselves* explain the topic to the rest of the class.[41] Since this involved talking openly about a topic in front of others, the choice was likely to be strongly affected by social image concerns and the taboo around menstruation, allowing us to test for changes in the level of taboo driven by the treatments. In each group (with 6 girls), one girl was randomly selected to have their "explanation" choice implemented, and another was selected to have their "anonymous video" choice implemented.

Information demand experiment. An alternative revealed-preference measure of stigma we use is whether girls choose to find out information about menstruation or another topic, intended as a measure of the demand for information. During both the midline and endline surveys, girls were shown a series (4) of pairs of envelopes, each of which showed a question on the outside, some of which related to menstruation (e.g., "Can stress affect the menstrual cycle?") and some which related to broader hygiene and health (e.g., "Are mosquitos more attracted by certain blood types?"). Girls make binary choices over whether they would prefer to open the menstruation-related envelope or the hygiene-related envelope, and get to see the answer to the question inside the envelope. The set of 4 question-pairs were the same for midline and endline. Ex post, we realize that this may have reduced the signal at endline, in case some girls remembered the answers from their midline choice and therefore may have chosen a different envelope at endline.

Dictator game. Finally, a standard prediction of theories of taboo is that those who *break* the taboo should face social sanctions. To test this, we ask endline girls and the 10 randomly selected endline boys to play two dictator games. Students are shown two 1-minute videos in a random order, both depicting a teenage girl from the region explaining a topic. One explains what menstruation is and how to maintain menstrual hygiene, and the other talks about the use of soap in hand-washing and the spread of germs. After each video, the student is shown 5 cookies, and asked how many they want to share with the person in the video. Just as with the lab-in-the-field, we can test whether the taboo around menstruation has decreased by looking at whether students are more likely to share cookies with the girl speaking about menstruation (relative to the girl speaking about hand-washing). Using this relative measure as our outcome increases power by allowing us to control for individual fixed effects.

D.2 Predictive and convergent validity

Table D1 assesses predictive and convergent validity. Columns (1)–(2) regress learning and self-reported stigma on an individual-level composite of the three measures in the main girl survey (dictator game, envelope experiment, enumerator-observed shame). Columns (3)–(4) add the information-supply experiments and boys' dictator game, aggregating all six measures to the school level. In the full sample, the individual-level index predicts both learning (0.089 SD, $p<0.001$) and self-reported stigma (0.144 SD, $p<0.001$). The school-level index also predicts self-reported stigma in both the control group (0.43 SD, $p<0.001$) and the full sample (0.15 SD, $p=0.001$).

[41] We randomly varied the order of explanation and anonymous choices, and the topics being proposed to each girl.

Table D1: *Combined Observed Stigma Index: Predictive and Convergent Validity*

Outcome	Combined lab & observed stigma (excl. info. supply, individual, Z) Coeff.	N	Combined lab & observed stigma (incl. info. supply, school-level, Z) Coeff.	N
Panel A: Control group only				
Learning (Z)	0.042	527	0.426***	35
	(0.034) [0.215]		(0.075) [0.000]	
Self-reported stigma (rev., Z)	0.085*	540	0.421***	35
	(0.044) [0.055]		(0.074) [0.000]	
Panel B: Full sample (controlling for treatment)				
Learning (Z)	0.089***	2,167	0.092*	140
	(0.023) [0.000]		(0.048) [0.058]	
Self-reported stigma (rev., Z)	0.144***	2,256	0.152***	140
	(0.024) [0.000]		(0.043) [0.001]	

Notes: Columns (1)–(2) report individual-level regressions of each outcome on the observed stigma index available in the main girl survey: a Kling-style composite (equal-weighted average of Z-scores) of three behavioral/observational measures collected from each girl—the dictator game difference, the envelope experiment, and enumerator-observed shame. Columns (3)–(4) incorporate three additional measures that were collected from separate samples and are therefore only available at the school level: the two lab-in-the-field (LiF) information-supply experiments (explanation choice and anonymous video choice, administered to a subsample) and the boys' dictator game. School-level regressions regress school-mean outcomes on the combined 6-measure lab & observed stigma index (average of Z-scored school means). Panel A restricts to the control group; Panel B uses the full sample and includes treatment arm indicators and stratum fixed effects. Individual-level regressions include stratum fixed effects and cluster standard errors at the school level; school-level regressions use robust standard errors. Standard errors in parentheses; p-values in brackets. * p<0.10, ** p<0.05, *** p<0.01.

D.3 Results

The series of revealed-preference measures show suggestive evidence that the reported reductions in menstrual stigma translate into changes in behavior (Table 6). In the *information supply experiment*, when girls were informed they would be asked to explain a topic in front of their schoolmates, and were asked whether to choose a menstruation-related topic or a broader hygiene-related topic, they were 4.6 p.p. (22%, 0.11 SD) more likely to choose the menstruation topic in the *Base+YGL* schools (p=0.03, column 2). By contrast, there is no effect on *Base Only* schools (p=0.98, column 2; p of difference = 0.06). The *YGLs* in schools are therefore effective at making girls more willing to supply information about menstruation openly, indicating a reduction in the taboo on the topic. In line with this, the additional effect of *YGLs* is specific to a choice that involves speaking *in front of other people* (when social image concerns are more salient). When girls instead choose *anonymously* which video should be shown to their schoolmates (column 3), the pooled treatment effect is significant (+3.8 p.p., 13%, 0.08 SD, p=0.05), and *Base+YGL* girls do not choose the menstruation topics more often than *Base Only* girls (p of difference = 0.52, column 3).[42]

In the *information demand experiment*, there are no detectable effects on the probability of selecting a menstruation-related topic when presented with envelopes with information about various topics contained inside (column 4, pooled p=0.21) at endline. However, at *midline*, we found a stronger pooled effect of 4.1 p.p. on the probability of choosing a menstruation-related envelope (8.6%, 0.14 SD, p=0.003), with no difference between *Base Only* and *Base+YGL* schools (p=0.88), suggesting that information demand does go up at least early on in the program.

In the *dictator game*, we examine the relative proportion of cookies shared with the girl discussing menstruation compared to hand-washing, controlling for individual fixed effects. We find that girls share relatively more with the menstruation-discusser in *Base Only* schools but not *Base+YGL* schools (Table 6, column 5). Since undoing menstrual stigma also requires that girls face fewer social sanctions from an "outgroup" when menstruating or discussing menstruation, we also examine whether the program affects the behavior of boys. We find that boys share relatively more cookies with menstruation-discussers in both types of treatment schools (Table 6, column 6).

Overall, the combined set of revealed-preference measures provide evidence of a reduction in menstrual stigma driven by the program, with the combined index showing a reduction of 0.08 SD at the school level. This observed effect appears smaller in magnitude than the large effects on self-reported measures, with results on individual measures also being less consistent. This suggests that while attitudes and girls' perceptions of stigma changed substantially, this did not always translate into shifts in observed behavior. The set of observed menstrual stigma results together indicate that the *YGL* component did not necessarily improve some of the outcomes more than the base intervention, which lines up with the finding on observed shame. Taken together, there is clear evidence of reduction in stigma for both the *Base Only* and the *Base+YGL* programs, while the even larger change in self-reported stigma in the *YGL* component does not translate in further changes in observed behavior.

[42] The double difference (comparing the difference between *Base Only* and *Base+YGL* across the *explanation* and *anonymous* outcomes) has p-value of 0.06, in line with the claim that the *YGL* intervention's impact is specific to the *explanation* outcome.

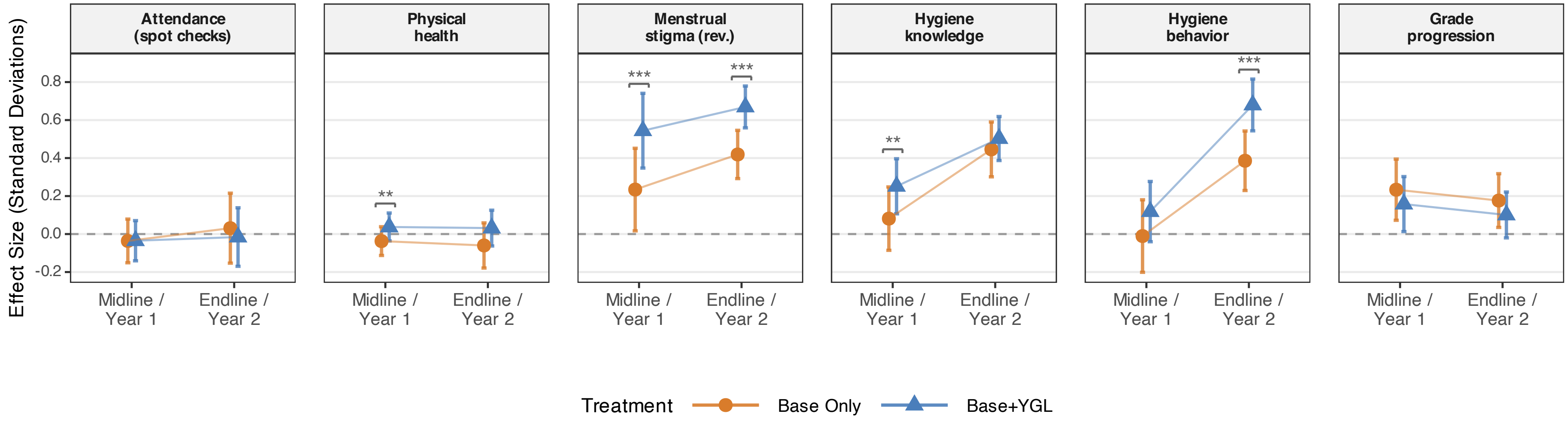
E Dynamics: comparing midline and endline

To understand the dynamic process through which the interventions affected learning and other outcomes, Figure E2 shows the effect sizes of the two treatment arms at both midline (after approximately 5 months of the program) and endline (after approximately 1.5 years of the program).[43] Four key patterns stand out. First, the null results on school attendance and physical health also hold at midline, with no significant effects at the 5% level across any treatment (Figure E2, panels 1-2). Second, the program reduced menstrual stigma relatively rapidly, showing significant effects at midline (panel 3; p=0.03 for *Base Only* and p<0.001 for *Base+YGL*). Third, knowledge about menstruation and hygiene improved more rapidly in the *Base+YGL* schools compared to the *Base Only* schools at midline (p of difference = 0.04). But by endline the differences had diminished, suggesting that *Base Only* schools caught up (Figure E2, panel 4). Finally, reported *hygiene behavior* was slow to move, with no significant effects at midline (pooled p=0.31). Only by endline do we see strong and significant effects (Figure E2, panel 5). This suggests that behavior takes longer to change than knowledge and reported stigma, in line with a literature showing that behavior change can be challenging in the field of hygiene and health (Aboud & Singla, 2012).

While we did not measure learning tests or heart rate at midline, the timing of girls' progression to the next grade between the first and second academic years (Table 2, column 6; Figure E2, panel 6) implies that academic performance had already improved after only 5 months of intervention. Consistent with this, administrative data on exit-exam pass rates in 2022 show that girls in treatment schools were significantly more likely to pass than girls in control schools (61% vs. 47%, p<0.001; Appendix H, Table H1, column 2). These rapid improvements are consistent with the learning results being driven by reductions in menstrual stigma rather than hygiene behavior, which was slower to change.

[43]To ensure comparability across waves, we (i) estimate only using individuals found at both midline and endline waves for this figure; (ii) construct indexes so that they are only composed of questions that were common across midline and endline waves, and use the same factor loadings for factor analysis-based indexes; and (iii) normalize all outcomes using the control mean and standard deviation *at endline*.

Figure E2: *Dynamics: comparing effects at midline and endline*



Notes: This figure compares treatment effects at midline and endline. For the first five panels, indices are constructed using only questions asked in both surveys for comparability: all indices use factor analysis with factor loadings calculated from endline data, then applied to both midline and endline responses. Normalization uses endline control group mean and standard deviation for both time periods, and the sample is restricted to girls present in both surveys. In the *Grade progression* panel, the "Midline" point uses an indicator for whether a girl had progressed at least one grade from baseline, as a proxy for midline progression, and the "Endline" point uses an indicator for whether a girl had progressed at least one grade from baseline *and* plans to be enrolled in school next year, as a proxy for endline progression. Each point is a separate post-double-selection LASSO regression with stratum fixed effects; error bars show 95% confidence intervals and standard errors are clustered at school level. Brackets with stars test equality of the *Base Only* and *Base+YGL* effects: $*$ p<0.10, $**$ p<0.05, $***$ p<0.01.

F Exploratory second endline

Motivation and methodology. In the first endline, we found an unexpected increase in pregnancies for girls in *Base Only* schools, but this was driven by very few girls. To understand the robustness of this result and the mechanisms behind it, we revisited schools again in June 2024. By this time, YGLs had been added to schools originally in the *Base Only* arm for the preceding academic year (2023/2024), but we used longer recall periods and the timing of conception to understand pregnancies occurring in 2022/2023. Due to budget constraints and the difficulty of relocating girls who have moved to other schools or other regions of Madagascar, we were only able to carry out surveys with girls currently enrolled in school, and took a new sample of 19 currently-enrolled girls per school from the same target grades as the baseline sample.

Sample selection would confound treatment effect estimates on our main outcome, given the evidence of negative selection into remaining enrolled in the same school (see further discussion in Appendix B.4), and confirmatory reports from directors at the 2nd endline that total enrollment in the 2023-24 school year, which accounts for new entries, increased substantially more in treatment schools than in control schools (Table F2). We therefore do not seek to measure effects on our main human capital outcomes using this survey.

Instead, we primarily use this 2nd endline to analyze the robustness of the effects on pregnancy, because the salience of pregnancies in school means that we can ask girls to report any pregnancies at their school or in their class over the last year, and the timing of pregnancies implies that this will refer to pregnancies conceived primarily in 2023, over the final year of program implementation. The use of such school-level peer reports reduces concerns about selectivity in who remains in school. In addition, since teachers are not subject to the same churn and selection concerns, we also elicit a wider array of measures of teacher motivation for 401 teachers in target grades, and use 395 classroom observations of these teachers to observe whether teaching practices are changed by the program.

Measurement. To measure pregnancies, we ask girls in our 2nd endline sample how many girls have been pregnant in her school over the most recent school year (implying conception mostly in final year of program implementation in 2023). To measure teacher motivation, we ask a series of modules to teachers, primarily using peer reports about other teachers to reduce social desirability bias, focusing on: (i) absenteeism (e.g., % teachers absent in the last week); (ii) extra-curricular volunteering (e.g., % participating in general assemblies and parental meetings); (iii) learning support for students (e.g., % giving extra remedial courses to students); (iv) academic preparation (e.g., time spent on preparing lessons); (v) punctuality (e.g., % teachers who are on time to class); (vi) motivation (e.g., % who encourage students, care for students); and (vii) retention (e.g., % who want to change schools or leave teaching). These are combined into a single index using factor analysis. To measure teaching practices, enumerators conducted observations based on an adapted form of the Stallings classroom observation method (Bruns et al., 2017). For 395 1-hour classroom sessions, enumerators sat at the back of class and every 5 minutes noted down which type of activity the teacher was engaging in at that moment (with an average of 10.4 observations per class). We divide activities into learning activities (e.g., lecturing in front of class, students working on worksheet, students presenting) and non-learning activities (e.g., roll-call, disciplining students).

Results. Results on pregnancy are presented and discussed in the main text (Table A23). We find no significant impacts on teacher motivation (Table F1, column 1), or on teaching dynamics (Table F1, columns 2–3).

Table F1: *Effects on teacher motivation and teaching practices at endline 2*

	(1) Teacher motivation index (Z)	(2) Classroom obs.: Learning activity (=1)	(3) Classroom obs.: Non-learning activity (=1)
Treated school (=1)	0.213	0.018	0.007
	(0.139)	(0.018)	(0.015)
Observations	401	4,100	4,100
Data source	Teachers	Class Obs.	Class Obs.
Control mean	0	.865	.105
Control SD	1	.342	.306

Notes: The Teacher Motivation Index is measured at the individual teacher level (N=401 teachers from 140 schools, approximately 3 teachers per school) and is constructed as a standardized composite index combining six sub-indices: (1) absenteeism (reverse coded), (2) voluntary learning activities (extra classes, parent engagement about student learning), (3) class preparation time, (4) intent to change schools or leave teaching (reverse coded), (5) general motivation (caring for students, encouraging students, providing extra help), and (6) punctuality and time management. The index is standardized to have mean 0 and standard deviation 1 in the control group. Classroom observation variables are binary indicators from direct observations of teacher activities during class (N=4,100 observations across multiple time points). Learning activities include teaching, homework, and active instruction; non-learning activities include discipline, administrative tasks, and teacher absence from class. The treatment variable pools Base Only and Base+YGL schools. All regressions include stratum fixed effects. Standard errors are clustered at the school level (id_ecole). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table F2: *Directors' reports of changes in enrollment over the last year (endline 2)*

	(1) Enrollment increased (=1)	(2) Enrollment decreased (=1)	(3) Enrollment increased more for boys (=1)	(4) Enrollment increased more for girls (=1)
Treated school (=1)	0.248**	-0.204**	0.092	0.097
	(0.097)	(0.089)	(0.067)	(0.076)
Observations	137	137	137	137
Data source	Schools	Schools	Schools	Schools
Control mean	.375	.438	.063	.188

Notes: Column 1: director reports enrollment increased in last year (binary). Column 2: enrollment decreased (binary). Column 3: enrollment increased more for boys than girls (binary). Column 4: enrollment increased more for girls than boys (binary). Treatment schools show higher enrollment increases and lower decreases, with similar patterns for boys and girls. Standard errors clustered at school level in parentheses. Controls include stratum fixed effects.

G Which YGLs are effective?

We first verify that the YGL selection process identified girls who are highly connected in peer networks at endline. Table G1 compares YGLs to non-YGL girls within *Base+YGL* schools on self-reported network measures; the items measure girls' degree (number of reported ties), not their position in the full network. YGLs score 0.55 SD higher on the composite network index (p<0.001), with large differences in both girl-specific and boy-specific sub-indices, indicating broad connectedness across both genders rather than popularity within a single gender. The single largest difference is in the number of peers with whom they discuss menstruation, consistent with their designated role as menstrual hygiene ambassadors.

Table G1: *YGL Social Network Connectedness at Endline*

	BL Sample	Full Sample
Network index (Z)	0.590***	0.549***
	(0.092) [0.000]	(0.070) [0.000]
Network – girls (Z)	0.420***	0.407***
	(0.074) [0.000]	(0.060) [0.000]
Network – boys (Z)	0.567***	0.515***
	(0.099) [0.000]	(0.069) [0.000]
# girl friends (Z)	0.123*	0.121**
	(0.070) [0.084]	(0.059) [0.045]
# boy friends (Z)	0.296***	0.230***
	(0.086) [0.001]	(0.061) [0.000]
# girls plays with (Z)	0.147*	0.083
	(0.076) [0.057]	(0.059) [0.159]
# girls talks with at break (Z)	0.148*	0.083
	(0.074) [0.051]	(0.058) [0.153]
# girls discusses menstruation with (Z)	0.736***	0.833***
	(0.123) [0.000]	(0.109) [0.000]
Does homework with others (Z)	0.177**	0.157***
	(0.068) [0.011]	(0.056) [0.006]
Help each other with work (Z)	0.018	0.024
	(0.069) [0.791]	(0.052) [0.645]
N	1103	1526
YGLs	297	501

Notes: Each row reports the coefficient on a YGL indicator (=1 if the girl was nominated as a Young Girl Leader) from a separate OLS regression, estimated within Base+YGL treatment schools only. All outcome variables are Z-scored using the control group mean and standard deviation. Column (1) restricts to girls surveyed at baseline; column (2) includes all girls surveyed at endline (including YGLs added to the endline sample who were not in the baseline). The composite network index is a factor-weighted average of individual network items. All regressions include randomization stratum fixed effects and cluster standard errors at the school level. Standard errors in parentheses; p-values in brackets. * p<0.10, ** p<0.05, *** p<0.01.

Methodology. We analyze heterogeneity with respect to YGLs' positive deviance. To do this, we construct an index of each YGL's *positive deviance*, using the data collected and selection criteria

Table G2: *Heterogeneous effects by YGL positive deviance*

	Base Only	Base+YGL	Base+YGL*Pos. Deviance Index	N
Learning tests	0.207***	0.111**	0.095**	2167
	(0.071) [0.004]	(0.052) [0.032]	(0.038) [0.013]	
Attendance (spot checks)	0.007	–0.013	0.024	698
	(0.045) [0.872]	(0.033) [0.702]	(0.015) [0.121]	
Physical health	–0.078	0.002	0.034	2256
	(0.056) [0.167]	(0.044) [0.961]	(0.036) [0.344]	
Mental health (heart rate rev.)	0.129**	0.126**	0.044	1904
	(0.056) [0.022]	(0.050) [0.012]	(0.033) [0.178]	
Menstrual stigma (rev.)	0.421***	0.806***	–0.018	2256
	(0.068) [0.000]	(0.058) [0.000]	(0.040) [0.659]	
Menstrual stigma (observed rev.)	0.641***	0.584***	–0.097	140
	(0.177) [0.000]	(0.159) [0.000]	(0.122) [0.426]	
Hygiene knowledge	0.414***	0.484***	–0.012	2167
	(0.068) [0.000]	(0.056) [0.000]	(0.040) [0.775]	
Hygiene behavior	0.311***	0.568***	0.055	2256
	(0.075) [0.000]	(0.065) [0.000]	(0.040) [0.165]	
Observed cleanliness	0.078	0.469***	–0.090	140
	(0.176) [0.656]	(0.138) [0.001]	(0.104) [0.389]	
School social climate	0.053	0.171***	0.065**	2167
	(0.071) [0.453]	(0.056) [0.002]	(0.032) [0.042]	

Notes: Outcomes are the same indices as in Figure A1. Positive deviance index (Z-score) is a school-level measure of selected YGLs' characteristics from initial selection data: e.g., based on willingness to adopt new behaviors, discuss sensitive topics, convince others, and comfort explaining sanitary pad use. Column 3: interaction between Base+YGL treatment and positive deviance index. Standard errors clustered at school level in parentheses; p-values in brackets. Controls include stratum fixed effects and LASSO-selected baseline controls.

used during the selection process.[44] The index measures girls' willingness to speak out or defy the existing social norm (see Appendix Section K.3 for the questions used).[45] We then average across the YGL of a school to obtain a school-level indicator of positive deviance of the selected YGLs. The variation in this index does not only capture individual YGLs' characteristics, but is also likely to capture features of the baseline school environment, since the probability of finding more positive deviant girls does not vary randomly across schools.

We then parametrically analyze heterogeneity using this school-level indicator of YGL positive deviance (Table G2).

Schools with positively deviant YGLs have significantly better learning test scores and better school social climate, compared to YGL schools with less positively deviant YGLs ($p<0.05$). This suggests that schools with more positively deviant YGLs were more effective at generating learning, health, and psychosocial improvements. Plausibly, girls who are willing and experienced in speaking out in defiance of social norms are better able to speak openly about menstruation without damaging the overall social climate of the school. Conversely, schools where YGLs score lower on positive deviance have significantly weaker human capital outcomes, and are likely to drive the overall lower point estimates on learning found in *Base+YGL* schools compared to *Base Only* schools.[46]

This pattern can explain the heterogeneity of the YGLs' effects between primary and secondary schools. The girls selected as YGLs score significantly lower on the index of positive deviance in primary schools (Figure G3), e.g., they score 0.25 lower on a 4-point scale when asked "When talk-

[44]In the design phase, we considered randomizing the YGL selection criteria - selecting positive deviants in some schools and *popular* (network-central) girls in others. We stepped away from this design after piloting, because in many schools there were not enough candidates who satisfied other key criteria (having started menarche and having good school grades) to generate variation in the type of selected YGL *within* a school. Given that there was, however, variation *between* schools in positive deviance, we instead decided at the design stage to test for heterogeneity with respect to the selected YGLs' positive deviance.

[45]Our measure of positive deviance does not simply pick up the overall "quality" of the YGL, e.g., in terms of leadership skills, confidence, popularity, or intelligence. When we construct a quality measure based on the first factor from an exploratory factor analysis of the YGL selection data, we do not find similar heterogeneity results.

[46]We find no significant interaction effect between Base+YGL and positive deviance for pregnancy (p=0.75), although this is perhaps unsurprising given the low power on the pregnancy outcome.

ing about delicate topics (like menstruation or sexual health), do you feel comfortable and open?" (p=0.003). And having positively deviant YGLs does seem to be crucial for not harming learning; even when controlling for school level (primary vs. secondary), having more positively deviant YGLs is still associated with better learning outcomes (Table G4). This suggests that in primary schools, it is often not possible to find girls who are sufficiently confident and open to act as effective positive deviants, and that this damages their effectiveness at improving school-level outcomes.

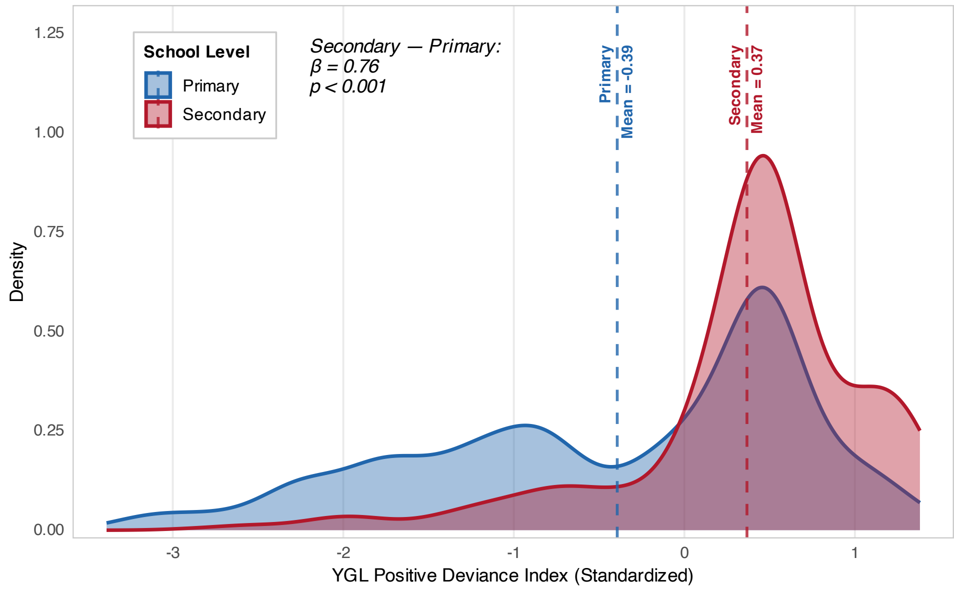


Figure G3: *Distribution of YGL positive deviance index, primary vs. secondary*

Notes: Distribution of YGL positive deviance index (standardized) by school level. The index combines YGL selection criteria: willingness to adopt new behaviors, discuss sensitive topics, convince others, and comfort explaining sanitary pad use. Dashed lines show group means. Secondary schools have significantly higher average positive deviance scores than primary schools (difference = 0.761, p < 0.001), indicating greater availability of suitable YGL candidates in secondary schools.

Table G4: *Heterogeneous effects by YGL positive deviance, controlling for secondary school*

	Base Only	Base+YGL	YGL × Pos. Dev	YGL × Secondary	N
Learning tests	0.208***	0.076	0.083*	0.061	2167
	(0.071) [0.004]	(0.085) [0.368]	(0.045) [0.069]	(0.114) [0.596]	
Attendance (spot checks)	0.006	0.017	0.034**	–0.059	698
	(0.044) [0.890]	(0.034) [0.614]	(0.015) [0.030]	(0.067) [0.386]	
Physical health	–0.077	–0.070	0.009	0.124	2256
	(0.056) [0.169]	(0.065) [0.279]	(0.038) [0.820]	(0.085) [0.145]	
Mental health (heart rate rev.)	0.129**	0.158**	0.056	–0.056	1904
	(0.056) [0.022]	(0.079) [0.045]	(0.037) [0.134]	(0.094) [0.551]	
Menstrual stigma (rev.)	0.421***	0.821***	–0.013	–0.025	2256
	(0.068) [0.000]	(0.091) [0.000]	(0.049) [0.797]	(0.124) [0.843]	
Menstrual stigma (observed rev.)	0.641***	0.595***	–0.094	–0.020	140
	(0.178) [0.000]	(0.224) [0.009]	(0.136) [0.491]	(0.278) [0.943]	
Hygiene knowledge	0.413***	0.573***	0.019	–0.154	2167
	(0.068) [0.000]	(0.077) [0.000]	(0.042) [0.646]	(0.096) [0.108]	
Hygiene behavior	0.313***	0.397***	–0.004	0.293***	2256
	(0.074) [0.000]	(0.091) [0.000]	(0.043) [0.920]	(0.113) [0.010]	
Observed cleanliness	0.078	0.632***	–0.034	–0.306	140
	(0.175) [0.659]	(0.195) [0.001]	(0.121) [0.778]	(0.270) [0.260]	
School social climate	0.055	0.026	0.015	0.251**	2167
	(0.070) [0.435]	(0.080) [0.751]	(0.037) [0.683]	(0.106) [0.017]	

Notes: Heterogeneity by YGL positive deviance, also controlling for secondary school interaction term. Specification is otherwise the same as Table G2.

H Administrative exam data

Methodology. We validate our grade progression results using administrative data on the binary pass rates in primary and lower-secondary school exit exams from the Regional Directorate of National Education, taken only in the final year of primary and secondary. Since this data excludes *lycées* (upper-secondary) schools and was matched imperfectly to our sample, only 123–125 schools were available, with a small imbalance in pass rates in the baseline year of 2021 (53% pass rate in control schools, vs. 46% in *Base+YGL* schools, p of difference: 0.06). We control for baseline pass rates to account for this imbalance.

Results. In line with our result that girls progress more quickly from one grade to the next in treatment schools, exit exam pass rates in 2022 after approximately 5 months of program implementation are significantly greater for girls (going from 47% in control schools to 61% in treatment schools, p<0.001, Table H1, column 2). We see weaker, but still significantly positive, effects on girls' pass rates in 2023, after 1.5 years of program implementation (pooled effect: 8 p.p., p=0.05, column 5). The weaker effects in 2023 are plausibly driven by a composition effect, since the best girls pass to the next grade in 2022 and do not repeat in 2023.

While boys' pass rates do not improve in 2022 (p=0.24, column 3), they experience substantial catch-up by 2023 (pooled effect: 13, p<0.001, column 6). However, because of the compositional changes and differential pass rates in 2022, the 2023 results reflect different student compositions and cannot be directly compared across genders.

Table H1: *Pass rates in primary and secondary school exit exams (administrative data)*

	Exam data: 2022			Exam data: 2023		
	(1) Passed (=1)	(2) Girl Passed (=1)	(3) Boy Passed (=1)	(4) Passed (=1)	(5) Girl Passed (=1)	(6) Boy Passed (=1)
Base Only	0.089**	0.143***	0.041	0.117***	0.090*	0.148***
	(0.042)	(0.041)	(0.052)	(0.042)	(0.050)	(0.041)
Base+YGL	0.089**	0.129***	0.045	0.102***	0.076	0.118***
	(0.038)	(0.040)	(0.047)	(0.038)	(0.047)	(0.034)
Primary school exam (=1)	0.336***	0.317***	0.407***	0.189***	0.228**	0.212***
	(0.079)	(0.086)	(0.085)	(0.071)	(0.088)	(0.078)
Observations	3,906	2,098	1,808	3,833	2,017	1,814
N schools	124	124	123	125	125	124
Control mean	.493	.472	.514	.606	.614	.597
p: Treated=0	0.006	<0.001	0.240	0.002	0.054	<0.001
p: Base Only=Base+YGL	0.992	0.728	0.936	0.642	0.711	0.392
p: Base Only (Girls)=Base Only (Boys)		.023			.270	
p: Base+YGL (Girls)=Base+YGL (Boys)		.202			.261	
p: Treated (Girls)=Treated (Boys)		.053			.210	

Notes: Columns 1-3: 2022 exams (after 5 months of program). Columns 4-6: 2023 exams (after 1.5 years). Combines primary (CEPE) and secondary (BEPC) exit exams. Outcome is whether student passed. Columns 2 and 5 restrict to girls only; columns 3 and 6 restrict to boys only. Controls include 2021 school-level pass rates, secondary school dummy, share of girls enrolled, total enrollment, and functional water access at baseline. Standard errors clustered at school level in parentheses. Controls include stratum fixed effects.

I YGL intervention details

I.1 YGL selection

YGLs were selected in *Base+YGL* schools in January 2022 in three steps.

1. *Teacher nomination survey*. Teachers selected a long-list of potential YGLs from target grades in a focus group session. They were asked to nominate girls who dare to speak up, who can generate change, and who are natural leaders.

2. *Girl survey*. From these nominated girls, the technicians of the local implementing NGOs then asked girls a series of questions about their own behaviors. These questions included questions about "positive deviance" (their willingness to openly defy an existing norm), leadership and communication skills.

3. *Ranking*. Based on the teacher and girl survey, only girls who satisfied basic eligibility criteria were eligible (above 10/20 average grade this year; has started menarche; consented to take part in the program). Then among these, we ranked candidates according to the weighting scheme in Table I2 that was developed by the partner NGO. Eligible girls with the highest ranking were selected as YGLs.

In each school, there was one finalized YGL per 60 students, with a minimum of 2 YGLs.

I.2 Young Girl Leader Curriculum

The Young Girl Leader program trains selected adolescent girls to serve as peer educators and role models within their schools. The curriculum combines personal development with technical knowledge on Water, Sanitation, and Hygiene (WASH), with particular emphasis on menstrual health management and reducing menstrual stigma. Initial training is delivered by local civil society organization technicians in weekly 1.5-hour meetings over approximately three months, working through the six curriculum sessions below (roughly one session every two weeks).

Session 1: Leadership and Self-Knowledge. This foundational session builds self-confidence and leadership identity. Participants explore what it means to be a leader, identify personal qualities, and develop self-awareness. Key topics include positive self-perception, public speaking confidence, and understanding their role as agents of change.

Session 2: Communication and Sensitization Techniques. This session equips YGLs with practical skills for peer education, covering effective communication, public speaking strategies, and techniques for delivering sensitization messages. Practical exercises help participants practice perspective-taking and message delivery.

Session 3: Menstrual Hygiene Management, Gender, and Action Planning. The central session of the curriculum addresses menstruation as a natural phenomenon while tackling associated taboos and misinformation. Girls learn proper use of reusable sanitary pads, discuss challenges during menstruation, and explore the link between menstrual stigma and school performance. Gender dimensions are integrated throughout, including discussion of how boys and fathers can support menstrual health management. Examples include emphasizing that menstruation is not taboo and can be discussed openly, that boys and men have supportive roles to play (e.g., fathers purchasing sanitary products, boys helping peers manage visible stains).

Session 4: Handwashing and Latrine Use. This session covers two core WASH messages: proper hand-washing with soap (including the five critical moments and correct technique) and hygienic latrine use. YGLs learn to demonstrate these practices using visual aids and hands-on exercises.

Session 5: Safe Water and Food Hygiene. This session completes the five key WASH messages by addressing water treatment methods, safe water storage, and food hygiene principles. Participants learn practical demonstrations they can replicate in peer sensitization sessions.

Session 6: Assertiveness and Personal Protection. The final session focuses on personal empowerment and protection from gender-based harm. Girls learn when and how to refuse requests that could put them at risk, building assertiveness skills relevant to their role as leaders.

Cross-Cutting Themes. Gender and stigma are integrated across all sessions.

Table I2: *Young Girl Leader (YGL) Selection Criteria and Weights*

Variable	Survey Question(s)	Notes	Weight
Good behavior (no misconduct)	Last year at school, was it noted that [name] had repeated or significant bad behavior?	Reversed: original asks about bad behavior	8.0%
Trusted with responsibility	To what extent do you agree with the following statement: 'She is someone we would be comfortable entrusting with a responsibility.'	—	8.0%
Communicates clearly	In your opinion, when she speaks in class, do her peers understand her easily?	—	8.0%
Age	How old are you?	—	8.0%
Open to adopting new practices*	To what extent do you agree: 'If I learn new knowledge that I believe is good for me, I am ready to adopt it even if it is very different from my habits and those of my peers.'	—	6.7%
Comfortable with sensitive topics*	If someone talks to you about sensitive topics (like menstruation or sexual health), do you feel comfortable and open?	—	6.7%
Persistence discussing sensitive topics*	Imagine you discuss and express your opinions to your peers on a sensitive topic like menstruation or sexual health. If people around you feel uncomfortable or show embarrassment, frustration, or anger, what action(s) would you take?	+2 for often convincing, +1 for sometimes, −1 for giving up	6.7%
Persuasive with peers	If she proposes an idea or expresses her opinion, in general, how many of her peers are persuaded and follow her?	—	6.0%
Conducted awareness activities	Have you conducted awareness activities in the last 3 school years? If yes, what is the maximum number of people you sensitized in a group?	Number sensitized if conducted activities, else 0	6.0%
Asks questions in class	During class or school activities, how often does she dare to ask questions or express her ideas on general topics (e.g., the subject matter or organization) in front of the class?	—	5.0%
Curiosity index	When there is something you don't understand during class, what do you do?	+2 for asking teacher, +1 for options 2–4, −1 for options 5–6	5.0%
Shares ideas with class	If you have an idea or opinion, in general, how often do you spontaneously propose it to the rest of the class?	—	5.0%
Comfort explaining menstrual products*	Would you feel comfortable explaining how to use a sanitary pad in front of: (a) all the girls in your class? (b) a smaller group of girls? (c) a larger group including girls from other classes?	Index: 3 = whole class & large group; 2 = class only; 1 = small group; 0 = neither	5.0%
Discussed sensitive topics with peers*	During the previous year, have you discussed and expressed your opinions in front of those around you about a sensitive subject such as menstruation or sexual health?	—	3.0%
Felt comfortable during discussion*	During these moments of discussion, overall, did you feel comfortable discussing the sensitive subject?	—	3.0%
Volunteers for activities	How often does she volunteer for activities or tasks in the classroom or at school?	—	2.7%
Shares new knowledge with friends*	How often do you tell your friends about new things you have learned and your new experiences?	—	2.7%
Held extracurricular responsibility	Have you held an extracurricular/school/extra-school responsibility in the last 3 school years?	—	2.7%
Attended awareness events	In the last 3 years, have you attended an awareness event or an event that allows you to learn new things (e.g., cultural and environmental awareness events)?	—	2.0%
		Total	**100%**

Notes: Each variable was standardized to Z-scores using Year 1 means and standard deviations, then weighted and summed to create a composite YGL index. Girls were ranked by this index within each school, with the top-ranked girls selected as YGLs. The first 3 questions were answered by teachers about each girl; all other questions were self-reported by the girls. * denotes variables used to construct the positive social deviance index.

Assessment and Quality Assurance. YGLs undergo three written evaluations throughout training:

1. *Behavioral change*: Pre-test before Session 1, post-test after Session 6
2. *Menstrual hygiene knowledge*: Pre- and post-test around Session 3
3. *WASH key messages*: Pre-test before Session 4, post-test after Session 5

Girls scoring at least 75% on two of three assessments are confirmed as YGLs. Those not meeting this threshold receive additional support until achieving the required competency levels.

Post-Training Implementation. After completing training, YGLs conduct two types of peer sensitization activities (formal and informal, see Appendix J). Program technicians provide ongoing support visits every two weeks to coach the YGL, monitor progress, collect documentation, and address implementation challenges.

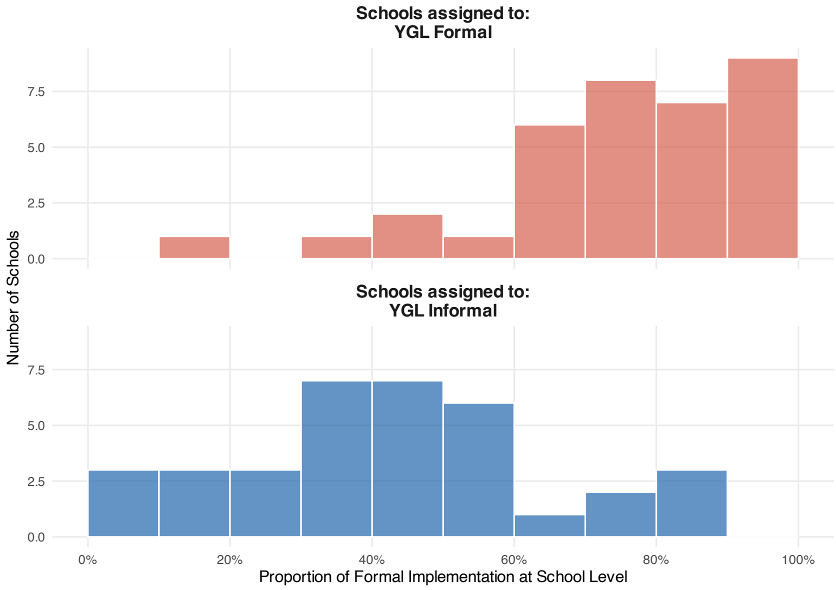
J YGL formal vs. informal

We randomized 35 schools to have YGLs carry out "formal" classroom sensitization sessions with their peers, and 35 other schools to have YGLs carry out "informal" sensitization during organic conversations with their peers.

Figure J1 shows that the differences in de facto implementation were only moderate.

Table J2 shows the differences in key endline outcomes across the two arms. A joint test of differences across the two arms on all key outcomes yields an F-statistic of 1.02, with a randomization inference p-value of 0.71, so we cannot reject the null that the arms have the same effects overall.

Figure J1: *Type of implementation of YGL sensitization program (formal vs. informal)*



Notes: This figure shows the distribution of the proportion of formal implementation at the school level for YGL Formal (structured) and YGL Informal (organic) schools. For each school, we calculate the proportion of students who reported formal implementation approaches versus informal approaches. **Formal implementation** includes: (1) specific time slots for YGL sessions, (2) sessions conducted in classroom settings, or (3) sessions organized with teachers' help. **Informal implementation** includes: (4) sessions without organizing, (5) sessions without specific time slots, (6) sessions conducted outside classroom, or (7) sessions without adults' help. Only schools with at least 3 valid responses are included. The histogram shows that YGL Formal schools tend to have higher proportions of formal implementation compared to YGL Informal schools. Data source: girls' reports in YGL schools, only including those who reported involvement in a girl-led sensitization session.

Table J2: *Effects of formal vs. informal YGL program*

Outcome	(1) Base+YGL (Informal)	(2) Base+YGL (Formal)	(3) *p*-val. of diff.	(4) N
Learning tests	0.105*	0.129**	.702	2167
	(0.059) [0.076]	(0.063) [0.040]		
Attendance (spot-checks, Z)	–0.046	–0.006	.428	698
	(0.062) [0.460]	(0.064) [0.927]		
Physical health	–0.020	0.028	.417	2256
	(0.054) [0.715]	(0.052) [0.598]		
Mental health (heart rate rev.)	0.179***	0.078	.065	1904
	(0.056) [0.001]	(0.058) [0.180]		
Menstrual stigma (rev.)	0.734***	0.874***	.035	2256
	(0.065) [0.000]	(0.068) [0.000]		
Menstrual stigma (observed, rev.)	0.664***	0.505***	.366	140
	(0.182) [0.000]	(0.183) [0.007]		
Hygiene knowledge	0.469***	0.497***	.629	2167
	(0.057) [0.000]	(0.068) [0.000]		
Hygiene behavior	0.612***	0.532***	.258	2256
	(0.073) [0.000]	(0.075) [0.000]		
Observed cleanliness (Z)	0.527***	0.413**	.449	140
	(0.145) [0.000]	(0.169) [0.016]		
School social climate	0.117*	0.232***	.067	2167
	(0.063) [0.064]	(0.066) [0.000]		

Notes: The table compares the effects of the *YGL* informal and formal modalities using girls' outcomes at endline. All outcomes are standardized to Z-scores using control group mean and standard deviation. Each row represents a regression of the outcome on *Base Only*, *Base+YGL (Informal)* and *Base+YGL (Formal)*. Column (1) shows the coefficient on *Base+YGL (Informal)*; column (2) shows the coefficient on *Base+YGL (Formal)*; column (3) shows the *p*-value for the test of *Base+YGL (Informal) = Base+YGL (Formal)*. Standard errors are clustered at the school level and are in parentheses; coefficient p-values are in brackets; controls include stratum fixed effects and the baseline controls selected by double LASSO. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

K Data

K.1 Construction of outcomes

This section documents the questions used to construct the main outcome indexes used in the paper.

Index aggregation methods
We use two primary aggregation approaches:

Factor analysis is used when combining questions from the same survey module that measure a single underlying construct (e.g., depression symptoms, attitudes towards menstruation, teasing experiences). We conduct exploratory factor analysis using a single factor with oblimin rotation. Each question is weighted by how strongly it loads onto the common factor. The resulting index is then standardized to have mean zero and standard deviation one in the control group.

Kling-style aggregation (Kling et al., 2007) is used when combining distinct sub-indices into a single aggregate index (e.g., combining hygiene knowledge with menstruation knowledge into an overall knowledge index, or combining multiple stigma sub-indices into a combined stigma index). This involves: (i) standardizing each sub-index to a Z-score using the control group mean and standard deviation; (ii) taking the unweighted average of these Z-scores; and (iii) re-standardizing the combined index so that the control group has mean zero and standard deviation one.

Knowledge and behavior indices are exceptions to factor analysis. These are constructed as sums of correct answers or recommended behaviors (then standardized), since all items have equal programmatic importance.

Below we document the specific questions used for each outcome.

General health *(factor analysis)*. Following Benshaul-Tolonen et al. (2021) in using standard questions from the physical functioning subscale of the Pediatric Quality of Life Inventory, girls are asked how much the following things have been difficult for them over the last month, with possible answers as (i) never; (ii) almost never; (iii) sometimes; (iv) often; or (v) almost always: (i) it's hard for me to walk more than a block; (ii) it's hard for me to run; (iii) it is difficult for me to play sports or exercise; (iv) it's hard for me to lift something heavy; (v) it's difficult for me to take a bath or shower by myself; (vi) it is difficult for me to do household chores; (vii) I have pain or aches; (viii) I have little energy. The index is constructed so that a higher value indicates *better* health (i.e., fewer difficulties).

UTI health *(simple average, standardized)*. Following Phillips-Howard et al. (2016) and using the standardized scale from Alidjanov et al. (2014) girls are asked to report symptoms of urinary tract infections (UTIs): "Please indicate if you have had the following symptoms for a few days in the last month." (Answer options: (i) No, (ii) Yes, light symptoms, (iii) Yes, moderate symptoms, (iv) Yes, serious symptoms.): (i) frequent urination of small volumes of urine; (ii) urgent urination; (iii) painful or burning sensation when urinating; (iv) sensation of incomplete emptying of the bladder after urination; (v) feeling of pain or discomfort in the lower abdomen; (vi) presence of blood in the urine. They are also asked "Did these symptoms occur at the same time (on the same days)?". The index is constructed so that a higher value indicates *better* health, i.e., fewer symptoms of urinary tract infections.

Mental health *(factor analysis)*. Based on an adapted CESD index (Radloff, 1977), girls are asked how many days in the last 7 they experienced each of the following: (i) bothered by things that don't usually bother you; (ii) not wanting to eat as usual; (iii) unable to shake sadness off, even with help from family and friends (rev.); (iv) happy (rev.); (v) able to concentrate on tasks (rev.); (vi) discouraged; (vii) everything felt difficult; (viii) full of hope for the future (rev.); (ix) too moody to talk to anyone; (x) afraid; (xi) restless sleep; (xii) happy (rev., second item); (xiii) spoke less than usual; (xiv) felt alone; (xv) others sought you out to chat (rev.); (xvi) wanted to do daily activities (rev.); (xvii) cried uncontrollably; (xviii) felt sad; (xix) told yourself that other girls don't like you; (xx) lacked the desire to do things. The index is constructed so that a higher value indicates *better* mental health (i.e., fewer symptoms of depression). We use reverse-coded questions to correct for acquiescence bias.[47]

Self esteem index *(factor analysis)*. Based on the Rosenberg (1965) standardized scale for self-esteem, adapted and reduced to 7 items after piloting based on internal validity analysis (taking out questions with negative or low factor weights). Girls are asked "For each description, indicate how much it is true for you." Possible answers: Strongly disagree; disagree; agree; strongly agree. Questions are: (i) do you think you have a certain number of good qualities?; (ii) all things considered, you are inclined to consider yourself a failure (rev.); (iii) you have a positive attitude towards yourself; (iv) overall, you are satisfied with yourself; (v) you wish you had more respect for yourself; (vi) sometimes you feel really useless (rev.); (vii) sometimes you think you're good for nothing (rev.). The index is constructed so that a higher value indicates *better* self esteem, i.e., more belief in oneself.

Menstruation-related knowledge *(sum of correct answers, standardized)*. Based on the CARE training curriculum, with the index constructed as the sum of correct answers to: (i) how often does a girl usually get her period? (Ans: once per month); (ii) how many days does the period last on average? (Ans: 3–7 days); (iii) at what age do women stop having their periods? (Ans: 40–60 years); (iv) what are the signs before or during your period? (Ans: ≥2 correct); (v) what are healthy behaviors to adopt during your period? (Ans: ≥3 correct); (vi–xii) seven true/false questions covering pregnancy risk, menstrual blood, pain, hygiene during menstruation, and menstruation as illness; (xiii) list different types of menstrual products (Ans: ≥3); (xiv) what are the most hygienic products? (Ans: ≥2 correct); (xv) how often should girls change menstrual products? (Ans: ≥3/day); (xvi) how often should girls change reusable pads? (Ans: ≥2/day); (xvii) how often should girls wash during their period? (Ans: ≥3/day); (xviii) what should you wash with? (Ans: soap); (xix) cost of one reusable sanitary napkin? (Ans: 5,000–10,000 Ary).

Hygiene knowledge *(sum of correct answers, standardized)*. Based on the CARE training curriculum, with the index constructed as the sum of correct answers to: (i) name 3 good habits to avoid getting sick; (ii) how to avoid contaminating water in a container (Ans: ≥2 correct); (iii) how to improve the hygiene of latrine use; (iv) how many times a day should we wash our hands? (Ans: ≥3); (v) when are the most important times to wash your hands? (Ans: ≥2 correct); (vi) name two methods for making water drinkable.

Menstruation-related behavior *(sum of recommended behaviors, standardized)*. Based on the CARE training curriculum, with the index constructed as the sum of "correct" answers to: (i) able to wash and change in private at home; (ii) able to wash and change in private at school; (iii) used reusable sanitary pads in the last 3 months; (iv) frequency of changing menstrual products (Ans: ≥3/day); (v) where sanitary pads were washed (Ans: not in the river); (vi) where sanitary pads were dried (Ans: outside in the sun); (vii) where underwear is washed (Ans: not in the river); (viii) where underwear is dried (Ans: outside in the sun); (ix–x) frequency of washing and washing with soap at home during last period (Ans: ≥3/day each); (xi–xii) frequency of washing and washing with soap at school during last period (Ans: ≥3/day each); (xiii) where sanitary products are dried (Ans: outside in the sun).

Hygiene behavior *(sum of recommended behaviors, standardized)*. Based on the CARE training curriculum, with the index constructed as the sum of "correct" answers to: (i) source of hand-washing water at home (Ans: clean source); (ii) what hands were washed with at home (Ans: soap); (iii) when hands were washed at home (Ans: ≥3 moments); (iv) where last urination/defecation at home (Ans: in a toilet); (v) frequency of hand-washing at home (Ans: ≥3/day); (vi) source of hand-washing water at school (Ans: clean source); (vii) what hands were washed with at school (Ans: soap); (viii) when hands were washed at school (Ans: ≥3 moments); (ix) where last urination/defecation at school (Ans: in a toilet); (x) frequency of hand-washing at school (Ans: ≥3/day); (xi) swimming location (Ans: not in the river).

[47] Also known as "yay-saying", acquiescence bias is when respondents tend to agree with a statement from the enumerator, even if doing so results in contradictory responses to positive and negative statements intended to measure the same trait. We correct for this bias, following the psychometric literature (Rammstedt & Farmer, 2013; Laajaj & Macours, 2019). For each girl, we calculate an acquiescence score by taking the average difference between the positively-coded questions and the reversed negatively-coded questions, and dividing it by two to retain centering. This score is then added to the reversed negatively-coded questions, and subtracted from the positively-coded questions.

Observed hygiene (girls) *(factor analysis)*. For the first 5 questions, the enumerator observes whether the girl is (i) very clean; (ii) quite clean; (iii) slightly dirty; or (iv) very dirty for the following: (i) hands, (ii) face, (iii) neck, (iv) clothes, (v) hair. The 6th question is whether the girl is wearing shoes. The index is constructed so that a higher value indicates *better* hygiene.

Observed hygiene (home) *(factor analysis)*. Enumerator observes the following about the girl's homestead, with (+) indicating a positive hygienic observation, (-) indicating a negative reverse coded observation: (i) water present at hand-washing area (+); (ii) soap, detergent, or ash at hand-washing area (+); (iii) human and animal feces in/around the house (−); (iv) waste accumulated around the house (−); (v) stagnant water around the house (−); (vi) stables, pigs, or kennels below or beside the house (−); (vii) lack of ventilation inside the house (−); (viii) separate spaces for cooking, sleeping, and socializing (+); (ix) house is neat (+); (x) kitchen area is neat (+); (xi) animals have access to food preparation areas (−); (xii) water served in the kitchen (+); (xiii) dishes in the kitchen are clean (+). The index is constructed so that a higher value indicates *better* hygiene.

Observed hygiene (school) *(factor analysis)*. Enumerators observe the following about the school environment, all of which indicate more hygienic environment: (i) soap available at the water point; (ii) garbage holes in the ground; (iii) garbage separated; (iv) trash collections in classrooms. The index is constructed so that a higher value indicates *better* hygiene.

Combined observed cleanliness *(equal-weighted average of Z-scores)*. We combine observed cleanliness from all three sources (girls, home, and school) into a single school-level index. Individual-level observations (girls and home) are first aggregated to the school level by taking means within each school. We then standardize each of the 22 individual items to have mean zero and standard deviation one in the control group. The combined index is the simple average of these 22 Z-scores. We use equal weighting rather than factor analysis or inverse covariance weighting because the 3 school-level items have extremely low control group means and standard deviations, which would otherwise give them outsized weight relative to the 19 items from girls and home surveys. One school-level item ("Is the garbage separated?") is excluded because it has zero variance in the control group, making Z-score standardization undefined.

Received information about menstrual hygiene *(factor analysis)*. Girls are asked the following questions: (i) how many times heard a classmate talk about menstrual hygiene; (ii) how many times heard a family member talk about menstrual hygiene; (iii) how many times heard a teacher talk about menstrual hygiene; (iv) asked anyone a question about menstrual hygiene; (v) anyone given you information about menstrual hygiene; (vi) has your mother given you information about menstrual hygiene. All questions refer to the past 3 months.

Willingness to speak about menstruation *(factor analysis)*. Answer options: (i) strongly disagree; (ii) disagree; (iii) neither agree nor disagree; (iv) agree; (v) strongly agree. Girls are asked whether they would be willing to discuss... (i) in a casual chat with friends, if only girls were present; (ii) in a casual chat with friends, if both boys and girls were present; (iii) in front of my class, if only girls were present; (iv) in front of my class, if boys and girls were present.

Attitudes towards menstruation *(factor analysis)*. Each question is a Likert scale, with possible answers: (i) strongly disagree; (ii) disagree; (iii) neither agree nor disagree; (iv) agree; (v) strongly agree. (i) "Girls should be allowed to go to school when they are on their period"; (ii) "Girls should be required to hide the fact that they have their period while at school" (rev.); (iii) "People should discuss menstruation openly with others"; (iv) "Mothers should broach the subject of menstruation with daughters only after their first period" (rev.); (v) "It is prohibited to dry washable towels in a public place" (rev.); (vi) "Girls should be punished in case their brothers discover its intimate effects during the period of menstruation" (rev.); (vii) "Girls should feel embarrassed or ashamed when they have their period in public" (rev.); (viii) "Boys should know about menstruation"; (ix) "Boys should be allowed to laugh at topics related to menstruation" (rev.); (x) "Periods are natural phenomena." The index is constructed so that a higher value indicates *more progressive* attitudes towards menstruation (e.g., being more likely to say that girls should be allowed to go to school during menstruation).

Second-order beliefs *(factor analysis)*. Girls were asked how many other girls and then other boys in their school would agree with the following three statements: (i) girls should be obliged to hide the fact that they are on their periods when they are at school; (ii) people should be able to openly discuss menstruation with others; (iii) boys should be allowed to laugh about things related to menstruation.

To aid with the questions, we asked how many girls (or boys) out of 10 would agree with the statement, using 10 small stones and asking girls to move stones to the right to represent people agreeing and to the left to represent people disagreeing.

Girls were then asked about their teachers' opinions of the same statements (answer options: all teachers; most teachers; half of teachers; only a few teachers; no teachers), and then their mother's opinion for the same statements (answer options: strongly disagree; disagree; neither disagree nor agree; agree; strongly agree). For teachers and mothers, we also asked perceived agreement with a fourth statement: "Girls should be allowed to come to school when they have their periods."

The index is constructed so that a higher value indicates that girls believe that others have *more progressive* attitudes towards menstruation (e.g., being more likely to say that girls should be allowed to go to school during menstruation).

Shame in response to vignettes *(factor analysis)*. Enumerators describe 2 vignettes to girls: (i) Imagine that you went to school and you were speaking about periods with a friend. You turn around and realise that someone you don't know heard the whole conversation. (ii) Imagine that you go to school and that someone who you don't know realises that you have started your periods, because he has seen a reusable sanitary pad in your bag. After each vignette, girls are asked how much they feel the following emotions in response, with answer options: (i) I feel it strongly, (ii) I feel it a bit, (iii) I don't feel it at all (i) ashamed; (ii) indifferent (rev.); (iii) humiliated; (iv) guilty; (v) that it's funny (rev.); (vi) ridiculous; (vii) that it's not a problem (rev.).

When used in the paper, the score is reversed so that a higher value indicates *less* shame in response to vignettes.

Observed shame *(factor analysis)*. Enumerator is asked (after asking questions about menstruation) about the girl's behavior over the last 5 minutes of the survey: (i) looking at the floor when speaking; (ii) speaking quietly or very quietly; (iii) laughing while answering or listening to the questions; (iv) hesitating when answering; (v) seeming uneasy or slightly uneasy; (vi) seeming more uneasy when menstruation was mentioned than other subjects. When used in the paper, the index is constructed so that a higher value indicates *fewer* observable markers of shame.

Network *(factor analysis)*. Girls are asked the following questions: (i) number of friends (girls and boys); (ii) change in number of friends compared to last school year; (iii) who they play with at school; (iv) who they share the way to school with; (v) who they talk to during recess; (vi) who they organize cleaning/gardening with at school; (vii) whether and how often they share things with friends; (viii) whether and how often they do homework or school projects together; (ix) whether classmates help each other with school work; (x) who they talk to about menstruation. For most questions, counts of girls and boys are elicited separately.

Teasing and bullying *(factor analysis)*. Girls are asked about a random subset of 2-3 other girls in the same grade level as them in their school. If the respondent doesn't know the other girl being asked about, we ask about additional girls (up to a maximum of 5). Only girls whom the respondent knows are included in the results. Girls are also asked about teasing and bullying towards themselves. Teachers are asked about a random subset of 3 girls in their class.

The indexes are constructed so that higher values indicate *more* or more severe teasing/bullying. Separate factor analyses are conducted for light teasing, severe bullying, and menstruation-related teasing.

Teasing questions include: (i) how often has a classmate laughed at [NAME] in the past month; (ii) how often has a classmate teased [NAME]; (iii) how many classmates tease [NAME]; (iv) severity of teasing from class members. Severe bullying questions include: (i) how often has a classmate intimidated/harassed [NAME] in the past month; (ii) how many classmates intimidate/harass [NAME]; (iii) severity of the intimidation/harassment inflicted by class members.

Note that the teasing, bullying, and network questions were asked to all girls in the endline sample, including girls who had dropped out of school or changed schools since baseline. While the questions reference "classmates" and activities "at school," the survey did not include skip logic to exclude girls no longer enrolled. Of the 2,167 girls with teasing/bullying/network data, 372 (17%) had dropped out of school entirely by endline. Restricting to girls still enrolled also raises selection concerns, since these outcomes are most relevant for girls remaining in school. However, results are similar when excluding dropouts (Table A20), suggesting neither concern substantially affects our findings.

Mothers' survey. To evaluate whether any changes in knowledge and behavior among girls are transmitted vertically to their mothers (e.g., through interactions in the household), we elicit from mothers at endline the same indexes of knowledge and behavior related to general hygiene and menstrual hygiene. We also ask mothers the same set of attitude questions about menstruation as girls. We further ask mothers a set of questions about their perception of the behavior of their daughter, using the standardized Strengths and Difficulties Questionnaire (SDQ). This was intended to evaluate girls' psychosocial well-being and also to understand if mothers had a negative or positive reaction to the changes in girls' behavior induced by the program. Some questions focused on emotional behavior (e.g., whether they are often worried, or often angry), while others focused on their relations with others (e.g., are they often alone, do they have at least one friend?). Finally, we ask all mothers about the fertility status of their daughters (specifically, were they pregnant in the last year or are they pregnant now) to examine potential impacts on fertility.

Boys' survey. We measure boys' hygiene knowledge and menstruation-related attitudes using the same survey modules as the girls.

K.2 Baseline variable definitions for LASSO

Following Belloni et al. (2014), we select controls by running a LASSO of the outcome on the candidate baseline variables and a LASSO of treatment status on the same candidates, and including the union of the variables selected in either step. The baseline controls available for LASSO selection include: *girl's age* (years); *girl has a brother* (=1); *distance to school* (km); *time to school* (min); *was absent last 3 months* (=1); *days absent last 3 months*; *years of school delay*; *father in household* (=1); *mother in household* (=1); *guardian can read and write* (=1); *guardian has no education* (=1); *guardian has secondary or higher education* (=1); *guardian occupation: agriculture* (=1); *household size*; *menstruation inhibiting school activity index* (Z), constructed as the sum of positive responses to whether the girl (over the last 3 months) missed school, left school during the day, did not fully participate in school, or missed a social activity because of her period; *menstrual health problems index* (Z), constructed as the sum of positive responses to having had menstrual cramps, a urinal infection, or other health problems linked to menstruation; and *wealth index* (Z), constructed from household asset ownership during the mother's interview. Other indexes are constructed analogously to the endline outcomes documented in Appendix K.1.

K.3 Other variable definitions

Social desirability bias. At baseline, we elicited a shortened version of the Crowne & Marlowe (1960) social desirability scale for the sample of girls, translated into the local language, Malagasy. The questions ask whether the respondent has a number of "too good to be true" traits (or the reverse), so that answering in the socially desirable way is likely to indicate a desire to please the interviewer. The questions used were: (i) it is sometimes difficult for me to continue my work if I am not encouraged (−); (ii) I sometimes feel resentful when I don't get what I want (−); (iii) on a few occasions, I have given up on doing something because I thought too little of my abilities (−); (iv) there were times when I wanted to rebel against the authorities, even though I knew they were right (−); (v) no matter who I talk to, I always know how to listen; (vi) there have been occasions when I have taken advantage of someone (−); (vii) I'm always willing to admit when I make a mistake; (viii) I sometimes try to take revenge rather than forgive and forget (−); (ix) I am always polite, even to unpleasant people; (x) I have never been irritated when people expressed ideas very different from mine; (xi) there were times when I was quite jealous of other people's good fortune (−); (xii) I sometimes get irritated by people who ask me for favors (−); (xiii) I have deliberately said something that hurt someone (−). We construct an index of each girl's social desirability score using factor loadings, and then construct *High SDB*, which indicates whether a girl has an above-median social desirability score. This is the variable we use as an interaction in Table A6.

YGL positive deviance index. During the YGL selection process, all girls nominated by teachers as potential candidates were asked a series of survey questions to evaluate their suitability for the role. The positive deviance index is constructed using a factor-analysis combination of the following questions:

(i) how often do you tell your friends about new things you have learned?; (ii) willingness to adopt new knowledge even if different from habits of those around you; (iii) comfort when someone talks to you about a sensitive subject (like menstruation or sexual health); (iv) whether you have discussed sensitive subjects in front of those around you in the previous year; (v) comfort level during those discussions; (vi) what actions you would take if people around you showed embarrassment or anger during such discussions; (vii) comfort explaining how to use a sanitary pad in front of all the girls in your class (with follow-up for smaller or larger groups). To construct a school-level indicator of YGLs' positive deviance, we take the average positive deviance of all YGLs selected in the school in the first year of the program. To construct the *relative* measure of YGL's positive deviance, we take this measure and subtract the average positive deviance of all the candidates who were in the potential pool of YGLs (whether they were selected or not).