*Kyoto University,*
*Graduate School of Economics*
*Discussion Paper Series*

# Nudging by Beauty:
# Improving Women's Health Decisions and Well-Being in the Field

Hisaki Kono, Minhaj Mahmud, Yasuyuki Sawada, Nahoko Mitsuyama, Tomomi Tanaka

Discussion Paper No. E-23-009

*Graduate School of Economics*
*Kyoto University*
*Yoshida-Hommachi, Sakyo-ku*
*Kyoto City, 606-8501, Japan*

March, 2024

# Nudging by Beauty: Improving Women's Health Decisions and Well-Being in the Field

Hisaki Kono[1] Minhaj Mahmud[2] Nahoko Mitsuyama[3]
Yasuyuki Sawada[4] Tomomi Tanaka[5]

March 4, 2024

### Abstract

Health interventions often fail to influence behavior because they overlook the choice architecture. We assess a unique intervention targeting women in rural Bangladesh, which emphasized health, hygiene, and nutrition's role in skin beauty. This intervention aimed to attract the attention of women, who tend to be beauty-conscious. Using the high-dimensional covariate balancing propensity score method, we find significant impacts on beauty, health outcomes, social relationships, and subjective well-being. Our analysis suggests the intervention's effectiveness is unlikely due to omitted variable bias. Using meta-analysis, we highlight its effectiveness in leveraging beauty salience compared with existing health and hygiene programs.

**Keywords:** Hygiene, Health, Beauty.

**JEL:** I1, D9

[1] Graduate School of Economics, Kyoto University. E-mail: kono@econ.kyoto-u.ac.jp.

[2] Asian Development Bank. E-mail: mmahmud@adb.org

[3] General Education Center, International University of Health and Welfare. E-mail: n.mitsuyama@iuhw.ac.jp

[4] Graduate School of Economics, University of Tokyo. E-mail: sawada@e.u-tokyo.ac.jp

[5] Poverty and Equity Unit, World Bank. E-mail: tanami.911@gmail.com

**Acknowledgments** We thank Shiseido Co., Ltd. and Kaihatsu Management Consulting, Inc. for sharing their panel survey data; JITA for the excellent field assistance and data collection; Hikaru Kawarazaki for surveying the existing health intervention literature; and Hanko Oba and Tsuyoshi Kato for logistical support.

**Disclosure Statement** This research was financially supported by the JICA Research Institute and JSPS (26220502, 16K03623, and 18H00832). The authors declare no competing interests. The IRB at the University of Tokyo approved all aspects of the survey design (15-94).

**Data Replication Statement** The survey data in Tangail are available upon request, conditional on approval from Shiseido Co., Ltd. and Kaihatsu Management Consulting. The data for the meta-analysis and replication codes are available on `https://sites.google.com/site/hisakikono/research`.

# 1 Introduction

Existing health interventions through information provision, such as handwashing promotions, health and nutrition campaigns, and other hygiene or sanitation programs, have limited effectiveness in changing human behavior (Avitabile, 2012; Brauw et al., 2015; Dickinson et al., 2015; Fitzsimons et al., 2016; Galiani et al., 2016). Traditional interventions often fail to change behavior because they do not address the way people process information (Vlaev et al., 2016). This is the "last-mile" problem related to delivering effective health services, particularly in developing countries.

Behavioral science insights suggest that human behavior is strongly influenced by the context or environment in which choices are made; in other words, the choice architecture is important (Thaler and Sunstein, 2008). Given people's limited cognitive resources, they tend to be affected by anything that falls within the focus of their limited attention span (Vlaev et al., 2016). A growing body of evidence suggests that "changing contexts" can greatly influence human behavior by affecting people's choice architecture (Dolan et al., 2012). This underscores the importance of designing interventions for target populations based on a better understanding of human psychology and cognitive mechanisms to induce desired behavioral changes. A behavioral nudge, which is the choice architecture designed to change people's behavior toward a better option without excluding any options, can provide low-cost solutions to health problems (Loewenstein et al., 2012).

In this study, we evaluated a novel intervention to improve women's health and hygiene practices by emphasizing their beauty benefits. Many studies suggest that women are beauty-conscious. Evolutionary psychologists attribute this trait to adaptation to the long-term mating strategies of males, and it is believed that looking young and healthy and having good skin signals fertility (Sugiyama, 2015; Thornhill and Gangestad, 2008; Buss, 2016). The sociological literature argues that women receive a higher return than men on their "erotic capital," which is a combination of aesthetic, visual, physical, social, and sexual attractiveness (Hakim, 2010). The literature on the economics of beauty has shown that beauty improves an individual's economic and non-economic outcomes such as earnings (Hamermesh and Biddle, 1994; Neumark, 2018), labor market participation and occupational choice (Hamermesh and Biddle, 1994), and marital bargaining (Hamermesh, 2011; Zhang et al., 2023) and that being beautiful tends to make women happier than it does men (Hamermesh and Abrevaya, 2013).[1]

Given these significant advantages associated with possessing an attractive body and face, especially for women, considerable financial and time resources are invested globally on enhancing one's appearance (**?**). Indeed, existing surveys also show that women spend considerable time and resources on beauty enhancement, reading fashion magazines, using cosmetics and skincare products, and upgrading their clothes and hairstyles. For example, Japanese women spend 28% more time on personal care than men (Survey on Time Use and Leisure Activities, 2016), and women in the United States spend 48% more time on grooming than men (American Time Use Survey, 2018). Women also spend more money on personal care products and services annually. In Japan, women living alone spend an average of US$842 annually, while men living alone spend only US$222 annually on average (Family Income and Expenditure Survey, 2019). In the United States, single women spend an average of US$670 annually compared with just US$270 for single men (Consumer Expenditure Survey, 2018).

These observations suggest that beauty is psychologically and economically important for women and that beauty salience influences women's decision-making by affecting their choice architecture. To test this hypothesis, we investigate a health-related education program for women conducted by a Japanese cosmetics company in rural Bangladesh. This program aimed to empower women by improving their skin condition and self-confidence and emphasized the importance of health, hygiene, and nutrition for achieving beautiful skin. To the best of our knowledge, this is the first intervention to use beauty salience to improve women's decision-making.

Because we evaluated this program retrospectively, we face two empirical challenges in examining the importance of beauty salience in this intervention: (1) the non-random placement of the program and (2) the absence of a treatment arm without beauty salience. First, to address the issue of the non-random placement of the intervention and selection into the treatment, we meticulously selected the control villages to minimize differences in village-level characteristics from the treatment villages. We conducted new follow-up surveys in these

[1] Bursztyn et al. (2017) suggested that social image concerns affect the education and career decisions of single women.

villages and applied the high-dimensional covariate balancing propensity score (HDCBPS) method (Ning et al., 2020) to control for potentially large sets of confounding variables. While this method effectively handles observed confounders, the concern about bias stemming from unobserved confounders remains. To assess the robustness of our results to these unobserved confounders, we therefore perform coefficient stability analyses, as proposed by Oster (2019), which provide the bounds of the treatment effect, considering the potential impacts of unobserved confounders. The estimated bounds offer assurance that the positive impacts observed are not solely driven by unobserved confounders.

Second, to address the lack of a similar intervention without beauty salience, we compared our estimated impacts with the effect sizes of existing health and hygiene education programs. To account for heterogeneity in the program impacts across studies and differences in the outcome measures, we employ a three-level meta-analysis. Although we have to assume that these heterogeneous impacts are normally distributed, this approach yields concise summary measures of the impact of existing interventions. By comparing our estimated impacts with the summarized impact of existing education interventions lacking beauty salience, we aim to underscore the role of beauty salience in our intervention's effectiveness.

The estimation results show that the intervention increased respondents' interest in skincare and substantially improved their hygiene and dietary practices. The fact that the women in the treatment group were more likely to report applying good hygiene practices to improve their skin condition indicates that emphasizing respondents' beauty benefits improved the effectiveness of the education intervention. We also find that the intervention improved women's social relationships and subjective well-being. While the specific mechanisms through which the intervention influenced these outcomes remain unidentified, the results suggest that enhanced knowledge of health has empowered women. The coefficient stability analyses confirmed that these results are not driven by unobserved confounders. Moreover, the three-level meta-analysis revealed that our intervention was significantly and substantially more effective than the top 10th percentile of the effect sizes of existing interventions, suggesting that emphasizing beauty benefits can substantially increase the effectiveness of interventions targeting women.

We believe this study makes valuable contributions to the growing literature on changing health behaviors by utilizing insights from the recent developments in behavioral sciences.[2] A novel feature of our intervention is its emphasis on beauty. In the intervention, the workshop instructor focused on how to maintain good skin condition and presented desired hygiene and diet practices. Hence, the main focus of the content was on beauty, not on hygiene, diet, or women's empowerment. Providing information that matches participants' interests and presenting tips for achieving their goals could thus attract more attention and make the information more salient.

Moreover, this study further enriches the extensive literature on the economics of beauty, where numerous studies have demonstrated the presence of a beauty premium (Hamermesh and Biddle, 1994; Hamermesh, 2011; Hamermesh et al., 2023; Harper, 2000; King and Leigh, 2009) and some researchers have identified why beauty matters (Berggren et al., 2017; Mobius and Rosenblat, 2006). By contrast, we show the potential for leveraging people's interest in beauty to enhance policy effectiveness. Emphasizing beauty benefits may attract more attention and be more likely to induce behavioral changes. Nudging by beauty can be an effective way to address the last-mile problem.

The remainder of the paper is structured as follows. The next section describes the intervention, data, and empirical strategy. Section 3 presents our main results and Section 4 examines the robustness of the analyses. Section 5 presents the meta-analysis results to compare our results with existing interventions. Section 6 concludes the paper.

# 2 Data and methods

## 2.1 Intervention

The intervention was designed and implemented by Shiseido, a multinational personal care company, with the aim of improving the social status and lifestyle of rural women in Bangladesh. It targeted rural women in their late teens to 30s with a monthly household

[2] For example, Burger et al. (2010) investigated the extent to which information on salient descriptive norms affects food choices by female students. Dupas (2011) showed that sharing the relative risk of HIV infection with female students by the partner's age group substantially decreased the incidence of pregnancies with older, riskier partners. VanEpps et al. (2016) showed that presenting menus with traffic light calorie labels reduced calorie intake among employees.

income of over BDT 12,000 (approximately US$100), intending to empower them by improving their skin condition and self-confidence. The intervention included a series of workshops focused on the importance of health, hygiene, and nutrition for achieving beautiful skin, in addition to demonstrations of the company's skincare products. The workshops consisted of three parts: face washing, protecting skin from the sun, and moisturizing. Each of the three workshops took place on a different day and women from the targeted villages were invited to all three. During the workshops, Shiseido's local female staff demonstrated how to use the company's skincare products (facewash, sunscreen, and skin cream/gel), highlighting key points to apply the products effectively. In particular, they provided basic knowledge on handwashing, laundry, nutrition, cooking, and sleeping, emphasizing the importance of these skills to improve skin condition.

The workshops were conducted from January 2014 to January 2015 in 16 villages in the Tangail district, 80 km from the capital, Dhaka. Considering the relatively low literacy levels in the region, flip charts and handwashing materials were used to demonstrate hygiene practices to help the attendees understand the contents. The selection of the treatment villages was not random but was based on a rough assessment of the market potential for their products. In the following section, we describe the empirical strategy used to address this selection problem.

## 2.2 Research design and data collection

Evaluating the impact of this program requires controlling for two types of selection problems: the selection of the program villages and the selection of the respondents. An international development consulting company conducted a baseline survey in 2013 and an endline survey after the intervention, but the survey did not adopt any randomization designs.[3] It included five treatment villages and only one control village that was selected because of its proximity to the treatment villages. We found significant differences in village characteristics between the treatment and control villages, especially the low ratio of households involved in agricultural production in the control village. Hence, we conducted a follow-up survey in 2015

[3] The results of this original survey can be found in (Japan International Cooperation Agency and Shiseido Co., Ltd. and Kaihatsu Management Consulting, Inc., 2015).

with respondents from the five treatment villages and newly selected control villages sharing similar village characteristics.

The new control villages were selected based on the available statistics on village characteristics that could be correlated with female empowerment and beauty outcomes, namely, (1) the female population ratio, (2) average household income, (3) the share of the agriculture/forestry/livestock industry, and (4) the share of home businesses. We collected information on these variables for the surrounding 24 villages and calculated the Mahalanobis distance of these variables between the control villages and each of the five treatment villages. We then selected six control villages whose Mahalanobis distance was below 10 as the new control villages.[4] This procedure ensured the village characteristics were balanced between the treatment and new control villages. Since we lacked baseline data on the respondents in the new control villages, we collected recall data on the baseline characteristics. We also collected recall data on the baseline characteristics of the respondents in the treatment villages to check their reliability.

This balance in village characteristics, however, cannot control for selection bias because of selection into the survey. The original survey did not conduct random sampling, but rather interviewed 50 women in each village who expressed an interest in the program. Most of the respondents in the treatment villages attended at least one workshop: the participation rate was 93% for the first workshop, 67% for the second workshop, and 51% for the third workshop. As we could not know who in the new control villages would have shown interest in the program in 2013, we randomly sampled 50 women from each village aged between 16 and 40 years, which is consistent with the age range of the respondents in the treatment villages. Then, we use the propensity score method to control for the selection bias caused

[4] We computed the Mahalanobis distance of the control villages for each of the five treatment villages and then selected the villages with the minimum Mahalanobis distance below 10. The mean and median of the minimum Mahalanobis distance were 48.1 and 15.5, respectively. The minimum Mahalanobis distance of the original control village was 119.8, indicating a non-negligible difference in village characteristics. There was another village whose minimum Mahalanobis distance was below 10, but we decided to omit this village because its share of the agriculture/forestry/livestock industry was quite low (0.001) compared with the treatment villages (0.04–0.06) and working in agriculture is expected to affect skin condition and other outcome variables.

by selection into the survey as described below.[5]

Because many variables could affect both the outcomes and program participation, we applied the HDCBPS method (Ning et al., 2020), which allows for high-dimensional covariate spaces. This method uses a machine learning algorithm to select variables that affect the outcome from a large number of potential confounders and constructs the propensity score so that the selected variables are well balanced between the treatment and control groups. The HDCBPS estimator is $\sqrt{n}$ consistent and asymptotically normal if either the propensity score or the outcome model is correctly specified. Only balancing the variables that affect the outcome avoids the bias caused by balancing variables that do not affect the outcome (Wooldridge, 2016). Unlike standard propensity score methods, which compute the propensity score independently of the outcome model, the estimated propensity score differs depending on the outcome of interest.

We considered the linear outcome model $E[y_i|\mathbf{X}_i] = \mathbf{X}_i\boldsymbol{\alpha}$, where $y_i$ is the outcome for individual $i$ $(i = 1, \dots, n)$ and $\mathbf{X}_i = (\mathbf{x}_{1i}, \dots, \mathbf{x}_{Ki})$ is a vector of the pre-treatment covariates. Let $T_i$ be the treatment indicator, $\pi(\mathbf{X}_i\boldsymbol{\beta})$ the propensity score, and $\lambda_1$ and $\lambda_2$ tuning parameters. First, we estimated the propensity score coefficients by penalized $M$-estimation:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[ \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{\mathbf{X}_i\boldsymbol{\beta}} \left\{ \frac{T_i}{\pi(u)} - 1 \right\} du + \lambda_1 ||\boldsymbol{\beta}||_1 \right]$$

where $||\cdot||_1$ is the standard $\ell^1$ norm. Using $\pi(\mathbf{X}_i\hat{\boldsymbol{\beta}})$, we constructed the weight $w(\cdot)$ by $w(\cdot) = \pi'(\cdot)/\pi^2(\cdot)$ and fit the outcome model using penalized weighted least squares estimation:

$$\tilde{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left[ \frac{1}{n} \sum_{i=1}^{n} T_i w(\mathbf{X}_i\hat{\boldsymbol{\beta}})(Y_i - \mathbf{X}_i\boldsymbol{\alpha})^2 + \lambda_2 ||\boldsymbol{\alpha}||_1 \right]$$

to select the variables that predict the outcome, $\mathbf{X}_{\tilde{S}} = \{x_k : \|\alpha_k\| > 0\}$. Then, we updated the propensity score coefficients of $\mathbf{X}_{\tilde{S}}$ so that the new propensity score $\tilde{\pi}_i$ satisfied the *covariate balancing property* for $\mathbf{X}_{\tilde{S}}$:

$$\sum_{i=1}^{n} \left( \frac{T_i}{\tilde{\pi}_i} - 1 \right) \mathbf{X}_{\tilde{S}i} = 0,$$

which requires the weighted mean of the treatment group to be equal to the control mean. The variables that hardly explained the propensity and outcome were excluded when com-

[5] As shown in Appendix Table S1, there was a significant difference in some of the baseline characteristics between the treatment and control groups.

puting $\tilde{\pi}_i$. Finally, the average treatment effect $\tau$ was estimated by inverse probability weighting:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\tilde{\pi}_i}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\tilde{\pi}_i}Y_i.$$

Given the low proportion of non-participation in the workshops and possible spillover effects,[6] we estimated the intention-to-treat effect by comparing similar respondents in the treatment and control villages.

We included the following variables as covariates $\mathbf{X}_i$: respondent's age; marital status; total number of people living in the same house; number of respondent's children; and categorical variables for household income, individual income, and individual education level, as well as the baseline values of all the outcome variables. We also included the interaction of the baseline outcome variables and the other covariates to allow for flexible forms of the dependence of the treatment status on these variables. The identifying assumption is that conditional on these covariates, the difference in the outcome variables between the treatment and control groups is solely attributed to the difference in program assignment. Because we carefully selected the control villages to be similar to the treatment villages and randomly selected respondents of a similar age, it is likely that we had a sufficient donor pool in the control villages whose potential outcomes were similar to those of the women in the treatment group. The estimated propensity score helps locate similar respondents in the treatment and control villages. Including a large number of the baseline variables is expected to capture the difference in the potential outcome. Since the assumption of the selection on the observables may be too strict even with the high-dimensional covariates, we checked the robustness of the results to unobserved confounders, as explained in Section 4.

## 2.3 Construction of the outcome variables

We focused on evaluating the impact of the intervention on health and women's empowerment. Because the survey included many outcome variables, we grouped similar outcome variables to reduce the number of hypothesis tests and avoid cherry-picking (Kling et al., 2007; Anderson, 2008; Casey et al., 2012). Of the 35 original variables, we computed nine aggregate outcome scores in three categories: beauty, health, and empowerment. For each

[6] Only 4% of the respondents in the treatment villages did not attend any workshop.

item, we first computed the $z$ scores and then aggregated them in the same group with equal weight (Kling et al., 2007). To check robustness, we constructed another aggregate score based on principal component analysis using the same variables (PC score). To facilitate comparison with other studies, we re-standardized the aggregate scores so that they had zero mean and unit standard deviation (SD) for the actual treatment baseline. Here, we briefly explain how we constructed the outcome measures and leave the detailed description of constructing the aggregate scores in Appendix A.1.

Before investigating the impact on the health and social outcomes, we examined whether our beauty-focused intervention actually changed participants' knowledge of and behavior toward beauty. For this purpose, we calculated the skincare product use score, skincare score, skincare behavior score, and skin condition score. The skincare product use score captures how frequently skincare products such as facewash, sunscreen, and skin cream/gel are used. The skincare score reflects the respondents' skincare practices and awareness of potential causes of skin damage, including the frequency of face washing and types of sun damage prevention. To focus exclusively on behavioral change, we calculated the skincare behavior score using only the outcomes related to skincare practices. Finally, the skin condition score captures the participants' subjective evaluation of their own skin.

For the health category, we computed the hygiene control score, diet score, and health behavior score. The hygiene control score reflects handwashing, laundry practices, and knowledge on the potential health risks caused by inappropriate handwashing and laundry practices. The diet score is based on knowledge on nutrition; the risks of consuming too much salt, oil, and sugar; methods of cooking vegetables; and well-balanced meals. The health behavior score is computed using data on health-related practices such as handwashing, laundry, and cooking.

For women's empowerment, we computed the decision-making score and social relationship score. The decision-making score captures the household decision-making power of female respondents regarding healthcare practices and the purchase of skincare products and other major household commodities. The social relationship score reflects women's willingness to go out and make friends, the variety of information sources available to them, information diffusion, and goal setting.

In addition, we used standard subjective well-being measures such as Rosenberg's self-esteem score (Rosenberg, 1965), a happiness score ranging from 0 to 10, and the mental health scale known as the K6 (Kessler et al., 2002).

## 2.4 Baseline characteristics

As we only had recall data for the baseline for the newly added control villages, in the main analysis, we used these data for the baseline variables, both for the treatment and for the control groups. The issues regarding recall bias and recall errors are addressed in Section 4.

Table 1 reports the summary statistics of our outcome variables and some demographic variables. The scatterplots of the aggregate $z$ scores and factor analysis scores show that these are highly correlated (Appendix Figure S1).

The average age of the respondents in the treatment village was 25.2 at baseline and 92.3% of them were aged 18 to 35. Women married quite young, with the average age of marriage among married women being 17.1 years. About 95% of the respondents aged above 20 years were married at baseline. The education level was quite low—46% of the respondents in the treatment villages reported their education level at or below the primary level and only 26% of them had completed secondary education. About a quarter of the respondents were still in school, 30% did not work, 20% engaged in farming, and 29% worked non-farm jobs. The majority of the women had no own income source (64%) or an individual monthly income below BDT 2,000 (approximately US$17). While 17.2% of the respondents reported a household income above BDT 20,000, 48% of them had a household income below BDT 10,000.

Although most of the respondents were married women with a relatively low income and low education living in rural areas, they frequently used skincare products. Nearly 90% of the respondents in the treatment villages reported using facewash twice or more a day, with only 2% not using facewash. For skin cream, 87% of them used it more than once a day, while only 6% had never used the product. On the contrary, the use of sunscreen was less common, with only 4% of the respondents regularly using the product. Seventy-three percent of the respondents in the treatment villages reported some skin problems, including acne (27%), eczema (31%), and oily skin (22%).

Handwashing was quite common among the respondents. Most of the respondents in the

Table 1: Summary statistics

| | Recall Baseline | | | | |
|---|---|---|---|---|---|
| | count | mean | sd | min | max |
| Age | 543 | 26.09 | 5.15 | 12.00 | 36.00 |
| Married | 543 | 0.84 | 0.36 | 0.00 | 1.00 |
| Age of marriage | 483 | 17.21 | 2.73 | 3.00 | 30.00 |
| Education | 540 | 3.79 | 1.48 | 1.00 | 6.00 |
| Household income | 540 | 3.84 | 1.38 | 1.00 | 6.00 |
| Individual income | 550 | 0.26 | 0.87 | 0.00 | 5.00 |
| Total number of people living in the same house | 543 | 4.70 | 1.72 | 1.00 | 12.00 |
| Number of your own children | 550 | 1.41 | 1.05 | 0.00 | 4.00 |
| Skincare products usage | 523 | -0.55 | 1.17 | -3.46 | 3.96 |
| Skincare | 543 | -0.81 | 1.02 | -5.45 | 2.99 |
| Skincare behavior | 523 | -0.88 | 1.24 | -4.89 | 4.85 |
| Skin condition | 542 | 0.30 | 0.80 | -2.79 | 1.55 |
| Hygiene control | 532 | -1.29 | 1.39 | -6.89 | 2.75 |
| Diet | 542 | -0.46 | 1.31 | -3.37 | 2.83 |
| Health behavior | 531 | -1.54 | 1.32 | -5.71 | 2.67 |
| Decision making | 536 | 0.70 | 1.17 | -1.89 | 2.15 |
| Social relationship | 541 | -2.41 | 2.00 | -6.02 | 1.00 |
| Happiness | 541 | 5.96 | 1.38 | 3.00 | 10.00 |
| Self Esteem | 543 | 17.17 | 2.25 | 11.00 | 25.00 |
| K6 | 541 | 16.58 | 4.26 | 5.00 | 24.00 |
| Hygiene for skin | 550 | 0.04 | 0.13 | 0.00 | 1.00 |
| Diet for skin | 550 | 0.12 | 0.21 | 0.00 | 1.00 |
| Skincare products usage(FA) | 523 | -0.49 | 1.18 | -3.67 | 1.62 |
| Skincare(FA) | 543 | -0.64 | 0.92 | -4.90 | 2.92 |
| Skincare behavior(FA) | 523 | -0.80 | 1.23 | -4.54 | 1.63 |
| Skin condition(FA) | 542 | 0.28 | 0.80 | -2.83 | 1.55 |
| Hygiene control(FA) | 532 | -0.74 | 0.63 | -2.33 | 3.88 |
| Diet(FA) | 542 | -0.63 | 0.85 | -2.54 | 1.77 |
| Health behavior(FA) | 531 | -1.09 | 0.68 | -2.17 | 3.62 |
| Decision making(FA) | 536 | 0.71 | 1.18 | -1.89 | 2.17 |
| Social relationship(FA) | 541 | -1.67 | 1.19 | -4.32 | 1.05 |

treatment villages (99%) used soap and water to wash their hands and a towel to wipe their hands after handwashing. However, some respondents tended to use the same towels and bed linen repeatedly without washing, resulting in poor sanitation. For example, only 36% of them changed towels every day and 23% changed towels only once a week. The majority of the respondents washed bed linen once a week (53%), but 15% of them washed it less than once a month. Very few respondents washed their hands with the intention to prevent skin problems (8%) or washed bed linen and dried the pillows in the sun to prevent acne (2%).

Most of the respondents (96%) in the treatment villages cooked food for their family members. While 89% of them answered that they are usually concerned about nutrition when they cook, they had limited knowledge of the risk of consuming too much salt, oil, and sugar. Only a third of the respondents were aware that consuming too much salt increases the risk of high blood pressure. Furthermore, only a few were aware that consuming too much oil increases the risk of diseases related to the heart (1.6%) and blood vessels (0.4%). They also did not know that consuming too much oil would cause skin trouble (4.4%). While 92% of the respondents knew that consuming too much sugar would cause diabetes, only 2% of them recognized that consuming too much sugar would make them fat. It was also quite rare that the respondents listed skin trouble as a result of consuming too much salt and oil and not enough vegetables.

In most cases, household decisions were jointly made by the husband and wife. For major household purchases, half of the respondents answered that they and their husbands jointly made decisions, while 40% answered that their parents and parents-in-law were also involved in the decision-making. For their own healthcare, 32% of the respondents reported that they themselves made the decisions, 51% reported they jointly made decisions with their husbands, and 10% answered that their husbands were the primary decision-makers. When purchasing skincare products, 47% of the respondents answered that they made decisions by themselves, while joint decisions with husbands (39%) and decisions by their husbands (11%) were also reported widely.

Table 2 reports the balance between the treatment and control groups after weighting by the propensity score estimated by the HDCBPS procedure. The reported $p$-values are for the null hypothesis that there is no difference between the weighted means of the treatment and control groups. Since the propensity scores differ depending on the outcome models, we only report the balance for the baseline outcome variables. As the propensity score is computed to balance the subset of the variables selected for the outcome model and the baseline outcome is a good predictor for the follow-up outcome, it is not surprising that most of the baseline outcomes were well balanced. For most of the baseline outcomes, there was no difference between the treatment and control means. Some of the baseline outcomes were not well-balanced (skincare score, diet score, and PC score of hygiene control), in which case the baseline outcomes tended to have less predictive power for the follow-up outcomes.

Table 2: Difference in the baseline outcomes between the treatment and control groups after the HDCBPS adjustment

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | Skincare product use | | Skincare | | Skincare behavior | | Skin condition | |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score |
| Treatment mean | -0.541 | -0.490 | -0.690 | -0.544 | -0.837 | -0.773 | 0.306 | 0.283 |
| Control mean | -0.541 | -0.490 | -0.750 | -0.600 | -0.837 | -0.773 | 0.306 | 0.283 |
| $p$ value | 1.000 | 1.000 | .486 | .459 | 1.000 | 1.000 | 1.000 | 1.000 |
| Observations | 488 | 492 | 490 | 494 | 488 | 492 | 489 | 493 |

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | Hygiene control | | Diet | | Health behavior | | Hygiene | Diet |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | for skin | for skin |
| Treatment mean | -1.017 | -0.720 | -0.101 | -0.459 | -1.468 | -1.090 | 0.044 | 0.123 |
| Control mean | -1.207 | -0.750 | -0.385 | -0.594 | -1.468 | -1.090 | 0.012 | 0.103 |
| $p$ value | 0.079 | .590 | 0.012 | .077 | 1.000 | 1.000 | .002 | .289 |
| Observations | 488 | 488 | 490 | 490 | 488 | 488 | 490 | 490 |

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | Decision-making | | Social relationships | | Subjective well-being measures | | |
| | $z$ score | PC score | $z$ score | PC score | Happiness | Self-esteem | K6 |
| Treatment mean | 0.748 | 0.769 | -2.400 | -1.616 | 5.967 | 17.253 | 16.589 |
| Control mean | 0.748 | 0.769 | -2.400 | -1.650 | 5.967 | 17.253 | 16.589 |
| $p$ value | 1.000 | 1.000 | 1.000 | .816 | 1.000 | 1.000 | 1.000 |
| Observations | 478 | 478 | 488 | 488 | 490 | 490 | 490 |

# 3 Results

## 3.1 Beauty-related outcomes

First, we examine whether our beauty-focused intervention actually changed women's knowledge and behaviors related to beauty. Table 3 reports the estimated impacts on our beauty outcome measures. To account for the fact that the intervention was assigned at the village level and the number of villages is small (11), we run Fisher's exact test clustered by village under the sharp null hypothesis of no treatment effect. The resulting $p$-values are reported in parentheses. To address the problem of multiple hypothesis testing, we apply the BKY procedure proposed by Benjamini et al. (2006) to control the false discovery rate (FDR).[7] We report the FDR $q$ values corresponding to the lowest $q$ at which the hypothesis is rejected (Anderson, 2008).

Table 3: Impacts on the beauty outcome measures: HDCBPS estimator

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | Skincare product use | | Skincare | | Skincare behavior | | Skin condition | |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score |
| Treatment | 0.676** | 0.283 | 0.935*** | 1.013*** | 0.618** | 0.329 | 0.006 | -0.014 |
| | (0.026) | (0.234) | (0.009) | (0.002) | (0.035) | (0.240) | (0.991) | (0.959) |
| | [0.055] | [0.191] | [0.032] | [0.018] | [0.055] | [0.191] | [0.472] | [0.472] |
| Observations | 488 | 492 | 490 | 494 | 488 | 492 | 489 | 493 |

Estimation results using the HDCBPS method are reported with $p$ values derived from the Fisher's exact test clustered by village in parentheses. Asterisks denote statistical significance: * $p < .1$, ** $p < .05$, and *** $p < .01$. The numbers in brackets indicate the FDR $q$ values computed by the BKY procedure.

Columns (1) and (2) show that the intervention significantly improved skincare product use by 0.678 SDs in the aggregate $z$ score and 0.268 SDs in the PC score, the latter of which was not significant. The difference in the results between the aggregate $z$ and PC scores was caused by the difference in weighting. While the aggregate $z$ score allocates greater weight to sunscreen use because it has a lower SD, the factor analysis allocates less weight to

[7] While Benjamini et al. (2006) derived this procedure by assuming the independence of the null hypotheses, their simulation exercises showed that it also works well for positively dependent $p$ values, as in the case here.

it. The intervention also improved skincare knowledge (Columns (3) and (4)) and practices (Columns (5) and (6)).

However, its impact on the skin condition scores was insignificant. Informal interviews with local staff indicated that the participants became more attentive to their skin condition and identified minor skin problems more frequently. This increased awareness might have decreased their confidence in their skin condition despite the lack of actual changes.

Overall, these empirical results imply that the beauty-focused intervention increased women's attention to their skin condition and improved their knowledge and behavior related to skincare. Controlling the FDR did not change the results, as most were significant.

## 3.2 Health-related outcomes

Now, we investigate if the beauty-focused intervention improved the health-related outcomes. To compute the FDR $q$ values, we considered multiple hypothesis testing for our six health-related outcome measures: the hygiene control $z$ score and PC score, diet $z$ score and PC score, and health behavior $z$ score and PC score.

Columns (1) and (2) of Table 4 demonstrate that the intervention led to a notable improvement in the hygiene control score, with an increase of 1.049 SDs in the aggregate $z$ score and 0.887 SDs in the PC score. The workshops emphasized the importance of hygiene in preventing skin issues and enhancing the efficacy of skincare products. This captured the participants' attention and raised their understanding of the importance of proper handwashing and laundry practices. This argument is supported by the findings in Column (7), where the higher value of the outcome variable indicates that the respondents cited better skin condition for applying good hygiene practices.[8] The result indicates that the intervention induced women to become more attentive to the consequences of their skin condition when making hygiene-related decisions.

The impact of the intervention on the diet scores was substantial, as evidenced by the significant 2.413 SD increase in the aggregate $z$ score and 1.618 SD increase in the PC score (see Columns (3) and (4)). The workshops emphasized the pivotal role of proper nutrition and diet in maintaining healthy skin, motivating the participants to enhance their dietary

[8] Appendix A.1 explains how we constructed the hygiene for skin score and diet for skin score.

Table 4: Impacts on the health-related outcome measures: HDCBPS estimator

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | Hygiene control | | Diet | | Health behavior | | Hygiene | Diet |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | for skin | for skin |
| Aggregate $z$ scores | | | | | | | | |
| Treatment | 1.049** | 0.887*** | 2.413*** | 1.618*** | 0.671** | 0.786*** | 0.462*** | 0.229* |
| | (0.032) | (0.002) | (0.002) | (0.002) | (0.032) | (0.002) | (0.002) | (0.002) |
| | [0.011] | [0.004] | [0.004] | [0.004] | [0.011] | [0.004] | | |
| Observations | 488 | 492 | 490 | 494 | 488 | 492 | 490 | 490 |

Estimation results using the HDCBPS method are reported with $p$ values derived from the Fisher's exact test clustered by village in parentheses. Asterisks denote statistical significance: * $p < .1$, ** $p < .05$, and *** $p < .01$. The numbers in brackets indicate the FDR $q$ values computed by the BKY procedure.

habits. The women in the treatment group exhibited a heightened awareness of nutrition concerns and a better understanding of the benefits and risks associated with their dietary choices. For instance, as shown in Column (8), the intervention increased women's awareness of the risks associated with skin problems resulting from excessive salt and oil consumption and insufficient vegetable intake when making dietary choices. Thus, this beauty-focused intervention also influenced women's dietary behavior by making the beauty-related benefits more salient, thereby making positive changes more prominent.

Moreover, as Columns (5) and (6) indicate, the intervention changed the participants' health practices, with an effect size ranging from 0.671 to 0.786 SDs. Altering behavior is typically more challenging than enhancing knowledge, and numerous existing interventions have struggled to bring about behavioral changes. Through a strategic combination of health information and the allure of beauty, our intervention not only heightened women's interest in the topic but also resulted in tangible behavioral changes, particularly in improving their dietary and hygiene practices.

## 3.3 Empowerment and well-being

Table 5 reports the impact on the empowerment category outcomes and subjective well-being measures, where we consider multiple hypothesis testing for four empowerment measures: the decision-making $z$ score and PC score as well as the social relationship $z$ score and PC score.

Columns (1) and (2) report the impact on the decision-making score. The point estimates are relatively low and not significantly different from zero. Hence, we see no positive impacts of the intervention on women's intrahousehold bargaining power.

By contrast, the intervention improved the social relationship score by 0.840 SDs in the aggregate $z$ score (Column (3)) and 0.675 SDs in the PC score (Column (4)). This result suggests that the participants became more inclined to socialize, make new friends, and discuss healthcare issues with their friends and family. As the intervention did not directly mention the role of social relationships, the observed improvement was attributed to the participants' interactions with the other women during the workshops and increased knowledge they gained, which they were eager to share with others.

Table 5: Empowerment and subjective well-being: HDCBPS estimator

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | Decision-making | | Social relationships | | Subjective well-being measures | | |
| | $z$ score | PC score | $z$ score | PC score | Happiness | Self-esteem | K6 |
| Treatment | 0.097 | 0.077 | 0.840*** | 0.675** | 1.582*** | -1.029** | -0.519 |
| | (0.132) | (0.290) | (0.008) | (0.017) | (0.002) | (0.015) | (0.140) |
| | [0.097] | [0.170] | [0.036] | [0.036] | | | |
| Observations | 478 | 490 | 488 | 492 | 490 | 490 | 490 |

Estimation results using the HDCBPS method are reported with $p$ values derived from the Fisher's exact test clustered by village in parentheses. Asterisks denote statistical significance: * $p < .1$, ** $p < .05$, and *** $p < .01$. The numbers in brackets indicate the FDR $q$ values computed by the BKY procedure.

In terms of the subjective well-being measures, our findings indicate that the intervention increased women's reported happiness (Column (5)), but concurrently decreased their self-confidence (Column (6)). No significant impact on K6 was observed (Column (7)). The decline in self-confidence could be attributed to women becoming more aware of their skin condition and noticing minor skin problems.

In summary, providing beauty-focused health information significantly enhanced the participants' knowledge and positively influenced their health-related behaviors and social relations. The observed effects were substantial across the various outcome measures.

# 4 Robustness checks

The validity of employing propensity score adjustment in our analysis relies on the assumption that, conditional on the covariates, there is no difference in the average potential outcomes between the treatment and control groups. Despite selecting control villages with similar characteristics to the treatment villages and constructing the control group by adjusting the propensity score based on the women in these control villages, concerns may arise about potential imbalances in the unobserved characteristics affecting the outcomes.

To assess the potential bias caused by unobserved characteristics, we employ the coefficient stability analysis proposed by Oster (2019). This method assesses the robustness to bias arising from omitting unobservables and calculates the bounds on the treatment effect. Specifically, it examines the changes in the coefficient and $R$-squared between a long regression (incorporating a full set of controls) and a short regression (using a restricted set of controls), along with the share of the treatment variance explained by the controls. Intuitively, a larger increase in the $R$-squared suggests that the added controls effectively capture the substantial variation in the outcome, diminishing the role of unobservables. A smaller change in the coefficient, despite the rise in $R$-squared, coupled with a small share of the treatment variance explained by the controls, indicates that the treatment assignment is not highly correlated with observables, making it less likely to be correlated with unobservables.

The coefficient stability analysis requires two additional inputs to bound the treatment effect: (1) the ratio of the correlation between the treatment and the unobservables to the correlation between the treatment and the included controls, denoted by $\delta$, and (2) the $R$-squared obtained from a hypothetical regression including all the controls and unobservables, denoted by $R_{max}$. If we knew the true values of $(\delta, R_{max})$, we could calculate the omitted variable bias and obtain the "bias-adjusted treatment effect," denoted by $\tau^*$. However, $(\delta, R_{max})$ are unknown. Since omitted variable bias increases in $\delta$ and $R_{max}$, Oster recommended setting these parameters sufficiently high to derive a conservative estimate of $\tau^*$, denoted by $\hat{\tau}^*$, and bounded the treatment effect by $\hat{\tau}^*$ and $\tilde{\tau}$, where $\tilde{\tau}$ is the estimate from the long regression. The identified set tends to be wider when the difference in the coefficient between the short and long regressions is large or when the long regression increases the $R$-squared only marginally. Oster suggested setting $\delta = 1$ and $R_{max} = 1.3\tilde{R}^2$ based on her re-analysis of previous studies, where $\tilde{R}^2$ is the $R$-squared of the long regression.[9] Following her suggestion, we set $\delta = 1$ and $R_{max} = \min(1.3\tilde{R}^2, 0.9)$.[10]

Oster's approach is based on linear regression and cannot be directly applied to propensity score weighting or matching methods. As an alternative estimation method that accommodates a high-dimensional covariate space, we consider the linear regression $y_i = \tau T_i + \mathbf{X}_i \boldsymbol{\alpha} + \epsilon_i$ with post-double-selection (PDS) lasso (Belloni et al., 2014). PDS lasso selects the control variables included in the regression, $\mathbf{X}_{\tilde{DS}}$, by applying the lasso to both a propensity score model $E(T_i|\mathbf{X}_i)$ and a reduced-form model $E(y_i|\mathbf{X}_i)$. The variables selected in either of these models are retained in $\mathbf{X}_{\tilde{DS}}$, and then ordinary least squares is performed using only $\mathbf{X}_{\tilde{DS}}$ as the controls. PDS lasso is consistent if at least one of the two models above is correctly specified, which aligns with the assumption of the HDCBPS method. In the following analyses, we include basic demographic variables (age, marital status, education) in both the short and the long regressions and add $\mathbf{X}_{\tilde{DS},i}$ into the long regression.

Table 6 shows the treatment effect estimated by the long regression, $\tilde{\tau}$, and the identified sets $[\hat{\tau}^*, \tilde{\tau}]$ obtained under the condition $(\delta, R_{max}) = \Big(1, \min(1.3\tilde{R}^2, 0.9)\Big)$. The long regression yields results akin to the HDCBPS method across all the outcome variables, providing support for the robustness of our main findings to the specification. The identified sets do not include zero for most of the outcome variables for which we found significant treatment effects, indicating the robustness of our main results to potential omitted variable bias.[11] Although the identified sets are wide for some of the outcome variables such as the scores for skincare, hygiene control, diet, health behavior, and social relationships, they still do

[9] $\delta = 1$ is the case in which the observables are as important as the unobservables in the omitted variable bias. Oster argued that carefully designed surveys would collect variables that explain the outcome, and therefore $\delta = 1$ is the appropriate upper bound. She also concluded that $R_{max} = 1.3\tilde{R}^2$ is appropriate based on the randomized data, where there should be no omitted variable bias.

[10] We set the maximum of $R_{max}$ be 0.9 given the aggregation and reporting errors (Oster, 2019).

[11] While Oster suggested using bootstrapping to calculate the standard error, bootstrap standard errors tend to overestimate dispersion due to computational errors. Oster's algorithm has multiple roots, and selecting the correct one is challenging. The risk of choosing a wrong root is greater when the difference in the population coefficient between the short and long regressions is smaller, as the sample estimate may have the opposite sign of the population estimate, particularly when the population estimates are closer to zero. Oster herself assessed robustness solely based on whether the identified set included zero, without computing standard errors.

not include zero because of the substantial magnitude of the estimated impact in the long regression. Hence, we believe our results are unlikely to be driven by the bias arising from unobserved characteristics.

Another concern is potential recall bias and measurement errors in the covariates. First, we address recall bias by comparing the actual baseline values from 2013 with the recalled baseline values collected in 2015 in the five treatment villages and one original control village (Appendix Table S2). While statistically significant differences are observed between the actual and recalled baseline values, there are no significant variations in the degree of recall bias between the treatment and control groups, except for individual income and the aggregate $z$ score of the diet score. Consequently, recall bias is not a significant concern when comparing the treatment and control groups.

However, even when the patterns of recall and measurement errors between the treatment and control groups are the same, these errors can still introduce bias when using propensity score methods (Battistin and Chesher, 2014). This is because errors in covariates can create discrepancies in the matched pairs with similar estimated propensity scores. Although the similarity in the estimation results based on PDS lasso to those derived using the HDCBPS method alleviates this concern to some extent, additional analyses are conducted to check the robustness of the results.

To mitigate concerns about recall bias, we adopt several strategies. First, we exclude from the treatment group those women whose reported characteristics at the endline were inconsistent with the baseline data. Specifically, we exclude those observations for which age differs by more than two years, the number of children differs by two or more, or education attainment differs by more than two years. Subsequently, we omit the baseline diet score (aggregate $z$) and baseline subjective well-being scores from the propensity score computation because we observe the substantial difference in mean recall bias for the diet score between the treatment and control groups as well as the potential for significant recall errors in subjective assessments of status from two years ago. Finally, we use endline education attainment and household income as covariates to reduce recall error. As long as education and household income were not influenced by the treatment, their use as covariates does not introduce bias. However, since individual income may be affected by the treatment, we continue to use baseline recall individual income.

Table 6: Oster's coefficient stability analysis exercises ($\delta = 1$, $R_{max} = 1.3\tilde{R}^2$)

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | Skincare product use | | Skincare | | Skincare behavior | | Skin condition | |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score |
| Treatment | 0.870** | 0.438* | 1.122*** | 1.153*** | 0.861** | 0.497* | 0.078 | 0.103 |
| | (0.311) | (0.224) | (0.294) | (0.315) | (0.327) | (0.256) | (0.148) | (0.150) |
| Identified set | [0.735,<br>0.870] | [0.429,<br>0.438] | [0.375,<br>1.122] | [0.236,<br>1.153] | [0.861,<br>0.867] | [0.497,<br>0.523] | [0.078,<br>0.125] | [0.103,<br>0.163] |
| Observations | 488 | 492 | 490 | 494 | 488 | 492 | 489 | 493 |
| $R^2$ | 0.562 | 0.589 | 0.552 | 0.555 | 0.539 | 0.582 | 0.444 | 0.435 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | Hygiene control | | Diet | | Health behavior | | Hygiene | Diet |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | for skin | for skin |
| Treatment | 1.176** | 0.967*** | 2.869*** | 1.858*** | 0.681* | 0.905*** | 0.428*** | 0.304*** |
| | (0.481) | (0.187) | (0.359) | (0.221) | (0.319) | (0.174) | (0.070) | (0.048) |
| Identified set | [0.389,<br>1.176] | [0.253,<br>0.967] | [1.777,<br>2.869] | [1.165,<br>1.858] | [0.267,<br>0.681] | [0.676,<br>0.905] | [0.428,<br>0.495] | [0.304,<br>0.443] |
| Observations | 488 | 492 | 490 | 494 | 488 | 492 | 490 | 490 |
| $R^2$ | 0.628 | 0.632 | 0.780 | 0.768 | 0.579 | 0.683 | 0.484 | 0.416 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | |
| | Decision-making | | Social relationships | | Subjective well-being measures | | | |
| | $z$ score | PC score | $z$ score | PC score | Happiness | Self-esteem | K6 | |
| Treatment | 0.111 | 0.108* | 0.903** | 0.802*** | 1.678*** | -0.396 | 0.108 | |
| | (0.063) | (0.051) | (0.389) | (0.246) | (0.294) | (0.267) | (0.297) | |
| Identified set | [0.111,<br>0.123] | [0.108,<br>0.126] | [0.631,<br>0.903] | [0.509,<br>0.802] | [1.678,<br>1.788] | [-0.554,<br>-0.396] | [0.108,<br>0.296] | |
| Observations | 478 | 490 | 488 | 492 | 490 | 490 | 490 | |
| $R^2$ | 0.793 | 0.799 | 0.765 | 0.761 | 0.691 | 0.296 | 0.682 | |

Estimation results of the long regression and Oster's coefficient stability analyses are reported. $p$ values were derived from the randomized inference clustered by village in parentheses. Asterisks denote statistical significance: * $p < .1$, ** $p < .05$, and *** $p < .01$.

Table 7 presents the estimation results following these adjustments, which closely resemble our main findings in Tables 2–4. This consistency in the results persists even when implementing one or two of the three procedures mentioned above. Consequently, any potential biases introduced by recall errors are expected to be minimal and should not impact our overall conclusions.[12]

# 5 Discussion: Comparison with existing interventions

We found that this intervention substantially improved the hygiene and nutrition knowledge and practices of the participants. While the emphasis on the beauty aspects of hygiene and nutrition practices is unique to the intervention, determining the specific impact of the beauty nudge was challenging since we lacked a treatment arm providing the same information without the beauty focus. To address this limitation, we inferred the effectiveness of beauty salience by comparing it with similar interventions in the existing literature, albeit sacrificing some credibility for comparability across studies.

To summarize the effectiveness of existing interventions, we conducted a meta-analysis of hygiene and nutrition intervention studies. First, we collected data from research published during 2000–2020 on EconLit using a combination of the following keywords: (1) handwashing, hygiene, sanitation, and nutrition, and (2) experimental, intervention, randomized control trial, treatment, and control. We excluded studies of interventions that did not include educational, information, or knowledge components as well as those that did not include knowledge improvement or behavioral change as outcome variables. To allow a comparison of the studies, we converted the outcome variables to the $z$ score. As a result,

[12] As an additional robustness check against recall and measurement errors, we conduct analyses using only variables less susceptible to these errors. These include age, marital status, household size, number of children, education, household income, and individual income at the endline survey; the interactions between age and education, household size and education, number of children and education, and age and individual income; and the use of skincare products at baseline. Given the relatively small set of covariates, we employ conventional propensity matching and Appendix Table S3 reports the results. We exclude observations with no matched counterparts within a bandwidth of 0.1. Appendix Figure S2 displays the distribution of the propensity score, indicating a good overlap between the treatment and control groups. The results remain consistent, underscoring the robustness of our findings.

Table 7: HDCBPS estimator: Mitigating the recall error problem

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | Skincare product use | | Skincare | | Skincare behavior | | Skin condition | |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score |
| Treatment | 0.689** | 0.313 | 0.942*** | 1.029*** | 0.632** | 0.364 | -0.027 | -0.034 |
| | (0.032) | (0.173) | (0.009) | (0.002) | (0.037) | (0.227) | (0.974) | (0.900) |
| | [0.059] | [0.161] | [0.032] | [0.018] | [0.059] | [0.179] | [0.435] | [0.435] |
| Observations | 449 | 449 | 451 | 451 | 449 | 449 | 450 | 450 |

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | Hygiene control | | Diet | | Health behavior | | Hygiene | Diet |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | for skin | for skin |
| Treatment | 1.066** | 0.872*** | 2.333*** | 1.566*** | 0.665*** | 0.787*** | 0.466*** | 0.221*** |
| | (0.011) | (0.002) | (0.002) | (0.002) | (0.006) | (0.002) | (0.002) | (0.006) |
| | [0.006] | [0.005] | [0.005] | [0.005] | [0.005] | [0.005] | | |
| Observations | 449 | 453 | 451 | 455 | 449 | 453 | 451 | 451 |

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | Decision-making | | Social relationships | | Subjective well-being measures | | |
| | $z$ score | PC score | $z$ score | PC score | Happiness | Self-esteem | K6 |
| Treatment | 0.121 | 0.108* | 0.850** | 0.698*** | 1.555*** | -1.225 | -0.617 |
| | (0.199) | (0.067) | (0.019) | (0.004) | (0.004) | (0.366) | (0.229) |
| | [0.099] | [0.047] | [0.031] | [0.018] | | | |
| Observations | 439 | 451 | 449 | 453 | 451 | 451 | 451 |

Estimation results are reported with $p$ values derived from the Fisher's exact test clustered by village in parentheses. Asterisks denote statistical significance: * $p < .1$, ** $p < .05$, and *** $p < .01$. The numbers in brackets indicate the FDR $q$ values computed by the BKY procedure.

we excluded papers that did not report the SD or information necessary to calculate it. Ultimately, we identified 12 papers (Briceno et al., 2015; Cameron et al., 2013; Chase and Do, 2012; Dickinson et al., 2015; Dillon et al., 2019; Elbers et al., 2012; Fitzsimons et al., 2016; Galiani et al., 2016; Guiteras and Mobarak, 2015; Levere et al., 2016; Patil et al., 2013; Zhao and Yu, 2020) with 115 outcomes across four types of interventions: sanitation training, sanitation and handwashing promotion, sanitation (latrine use) awareness campaigns, and nutrition knowledge provision. Appendix Figure S2 shows the distribution of the effect sizes.

To account for the differences in intervention types and outcome measurements as well as the presence of multiple outcomes in a study, we employed a three-level meta-analysis. This model accommodates sampling variation for each effect size (level one), variation across the outcomes within a study (level two), and variation across studies (level three), accounting for the correlation between the outcomes in the same study and allowing for heterogeneity in effect sizes. (Geeraert et al., 2004; Van den Noortgate et al., 2015). All three of these components were assumed to be normally distributed. If the impacts of interventions differ considerably because of differences in program designs and local contexts, large between-study variance results.

Since both hygiene and nutrition intervention studies were considered, we included an indicator variable for nutrition interventions as the moderator variable. Because our intervention did not include components related to latrine use or open defecation, we also included an indicator variable for interventions related to latrine use or open defecation. Appendix A.2 details the literature search procedure and model specification.

Table 8presents the results of the three-level model estimation, where Column (1) displays the results when we only included indicators for nutrition and latrine-related interventions. The point estimate for the overall mean impact of existing interventions was 0.063, with a 95% confidence interval (CI) of [-0.121, 0.248]. Notably, the between-outcome variance was smaller than the between-study variance, indicating greater heterogeneity across studies than within the outcomes of a study.

As reported in Table 4, the impact of our intervention on the hygiene control, diet, and health behavior scores ranged from 0.671 to 2.413. For comparison, we use the estimated impact of our intervention on the hygiene control PC score and health behavior PC score, both of which were relatively precisely estimated. The 95% CIs for the impact on these outcome variables were $[0.284, 1.490]$ and $[0.251, 1.321]$, respectively.[13] Importantly, even the lower bounds of these 95% CIs surpassed the upper bound of the 95% CI of the overall mean impact of existing interventions (0.248), underscoring the notable effectiveness of our intervention.

[13] We obtained the $p$-values using Fisher's exact tests and did not derive standard errors or CIs to avoid statistical inference concerns about the small number of clusters. Here, we introduced an additional assumption of the normality of the estimates to compute the CIs. With this assumption, the standard errors of the estimates $\hat{\tau}_k$ were derived using the formula $\hat{\sigma}_k = \frac{\hat{\tau}_k}{\Phi^{-1}\left(\frac{1-p_k}{2}\right)}$, where $\Phi^{-1}$ was the inverse cumulative distribution function of the standard normal distribution and $p_k$ were the $p$-values associated with the estimates $\hat{\tau}_k$. The CIs were then computed as $[\hat{\tau}_k - 1.96\hat{\sigma}_k, \hat{\tau}_k + 1.96\hat{\sigma}_k]$.

In our three-level meta-analysis, we assumed that the heterogeneity of the effect sizes across studies and outcome variables are independently and normally distributed. With this assumption, we can compute the $100p$th percentile effectiveness of existing interventions and test whether our impact is greater than this computed value, as detailed in Appendix A.2. However, caution should be exercised when interpreting our results, as the independent normal distribution may not accurately represent the true distribution of effect sizes. Since the approximation of the distribution will be poor at its tails, we focus on the 90th percentile.

Given the relatively small between-study and between-outcome variances in the three-level model, the 90th percentile effectiveness of existing interventions, $\tau_{90}$, was not large. As shown in Column (1) of Table 8, the 90th percentile effect size of the existing intervention was 0.167, and we can reject the null hypothesis that our impact, $\tau$, is no greater than the 90th percentile effect size at the 5% level ($p = 0.016$ for the impact measured by the hygiene control PC score and $p = 0.063$ for the impact measured by the health behavior PC score).

Next, we excluded the three studies related to latrine usage or open defecation, which left us with 109 outcomes. While the estimated coefficients and variance components changed only slightly, the CI shrank substantially to [-0.008, 0.135] (see Column (2)). This also shrank the CI for the 90th percentile effect size, resulting in a lower $p$-value for the null hypothesis that $\tau \leq \tau_{90}$. Further, we included the country-level factor score based on the 88 health and education indicators from the World Development Indicators of the World Bank to allow for possible differences in the effect size depending on a country's macro conditions.[14] As shown in Column (3), the CI for the predicted overall mean was computed with the factor score equal to that of Bangladesh in 2013 (-0.538). Including the country-level factor score reduced the between-study variance. When we additionally excluded the studies related to latrine usage or open defection, the between-study variance decreased to 0.0042 (see Column (4)).

Further, we included an indicator for the behavioral change outcomes; its negative coefficient indicated that changing behavior is more difficult than improving knowledge (Columns

[14] We also implemented the multilevel lasso to identify the variables affecting the effect size, but none of these variables had non-zero coefficients. Given the possible violation of the sparsity assumption of lasso, we included the first factor score in the analysis. Including the second or third factors did not affect the results and the coefficients of these two factors were not statistically significant.

(5) and (6)). We also excluded studies related to latrine usage or open defecation. The variance of the study-specific component was much smaller at 0.0034 and the between-outcome variance was 0.0004 (see Column (6)). In all these specifications, we found that our intervention was more effective than the top 10th percentile of existing studies. Since the main difference between our and existing interventions is its focus on beauty, we inferred that the beauty nudge can substantially enhance the effectiveness of information interventions for women.

Finally, we included our results on the hygiene and health behavior scores in the database for the three-level meta-analysis as well as an indicator for our beauty nudge treatment as an additional predictor. As shown in Column (7), the coefficient of the beauty nudge was high, 0.762, significant at the 1% level. This supports our conclusion that the beauty nudge can improve the effectiveness of information programs targeted at women.

# 6 Conclusion

Many health information interventions have failed to generate the desired changes in human behavior. Behavioral science insights suggest the importance of considering the choice architecture of the target population when designing information interventions. Attracting their attention is crucial for changing their behavior, and tailoring the information to their specific interests can substantially improve the effectiveness of the intervention at a relatively low cost. In this study, we show that the effectiveness of information provision programs targeted at women can be enhanced by adding a beauty nudge because the beauty components of the workshops pique women's interest. Our beauty-focused hygiene and nutrition intervention improved the health knowledge and behaviors, social relationships, and happiness of the targeted women. Although our research design did not allow us to separately estimate the impact of the beauty nudge, our three-level meta-analysis of comparable studies indicates that the intervention with the beauty nudge was much more effective than the top 10th percentile of existing interventions. Thus, we infer that this difference is attributed to our beauty nudge program design.

Despite implementing various statistical strategies to address empirical challenges in non-experimental research designs, our results should be viewed as suggestive evidence of the effectiveness of beauty salience when targeting women. Although the derived bounds of the treatment effects exclude zero impacts, the conclusion could change if the underlying role of unobserved confounders was much larger than assumed. Moreover, although our three-level meta-analyses show the greater impacts of our intervention compared with existing interventions, such impacts could be partly attributed to favorable local contexts. To enhance the credibility and external validity of our findings, further randomized studies with and without beauty salience are necessary.

Finally, while our beauty-focused program demonstrated improvements in the social relationship and happiness measures, along with the positive impacts on health knowledge and behaviors, the mechanism for the former effects remains unclear. Specifically, we could not separately identify if the intervention improved women's social relationships and happiness through increased knowledge or directly affected these aspects by making women more beautiful. Although our findings of no positive impact on skin condition or self-esteem provide little support for the existence of a direct effect, exploring whether these products can empower women by enhancing their beauty and confidence could be an intriguing area for future research, especially given growing demand for skincare and cosmetics products in developing countries.

# References

**Anderson, Michael L**, "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*, 2008, *103* (484), 1481–1495.

**Avitabile, Ciro**, "Does Information Improve the Health Behavior of Adults Targeted by a Conditional Transfer Program?," *Journal of Human Resources*, 2012, *47* (3), 785–825.

**Battistin, Erich and Andrew Chesher**, "Treatment effect estimation with covariate measurement error," *Journal of Econometrics*, 2014, *178* (2), 707–715.

**Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen**, "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 2014, *81* (2), 608–650.

**Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli**, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, 2006, *93* (3), 491–507.

**Berggren, Niclas, Henrik Jordahl, and Panu Poutvaara**, "The right look: Conservative politicians look better and voters reward it," *Journal of Public Economics*, 2017, *146*, 79–86.

**Brauw, Alan, Patrick Eozenou, and Mourad Moursi**, "Programme Participation Intensity and Children's Nutritional Status: Evidence from a Randomised Control Trial in Mozambique," *Journal of Development Studies*, 2015, *51* (8), 996–1015.

**Briceno, Bertha, Aidan Coville, and Sebastian Martinez**, "Promoting Handwashing and Sanitation: Evidence from a Large-Scale Randomized Trial in Rural Tanzania," 2015. Policy Research Working Paper, World Bank.

**Burger, Jerry M, Heather Bell, Kristen Harvey, Jessica Johnson, Claire Stewart, Kelly Dorian, and Marni Swedroe**, "Nutritious or delicious? The effect of descriptive norm information on food choice," *Journal of Social and Clinical Psychology*, 2010, *29* (2), 228–242.

**Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais**, "'Acting Wife': Marriage Market Incentives and Labor Market Investments," *American Economic Review*, 2017, *107* (11), 3288–3319.

**Buss, David M.**, *The evolution of desire: Strategies of human mating*, New York: Basic Books, 2016.

**Cameron, Lisa, Manisha Shab, and Susan Olivia**, "Impact Evaluation of a Large-Scale Rural Sanitation Project in Indonesia," 2013. Policy Research Working Paper, World Bank.

**Casey, Katherine, Rachel Glennerster, and Edward Miguel**, "Reshaping institutions: Evidence on aid impacts using a preanalysis plan," *The Quarterly Journal of Economics*, 2012, *127* (4), 1755–1812.

**Chakraborti, S and J Li**, "Confidence interval estimation of a normal percentile," *The American Statistician*, 2007, *61* (4), 331–336.

**Chase, Claire and Quy-Toan Do**, "Handwashing Behavior Change at Scale: Evidence from a Randomized Evaluation in Vietnam," 2012. Policy Research Working Paper, World Bank.

**Cheung, Mike W-L**, *Meta-analysis: A structural equation modeling approach*, John Wiley & Sons, 2015.

**den Noortgate, Wim Van, José Antonio López-López, Fulgencio Marín-Martínez, and Julio Sánchez-Meca**, "Meta-analysis of multiple outcomes: a multilevel approach," *Behavior research methods*, 2015, *47* (4), 1274–1294.

**Dickinson, Katherine L, Sumeet R Patil, Subhrendu K Pattanayak, Christime Poulos, and Jui hen Yang**, "Nature's Call: Impacts of Sanitation Choices in Orissa, India," *Economic Development and Cultural Change*, 2015, *64* (1), 1–29.

**Dillon, Andrew, Joanne Arsenault, and Deanna Olney**, "Nutrient production and micronutrient gaps: evidence from an agriculture-nutrition randomized control trial," *American Journal of Agricultural Economics*, 2019, *101* (3), 732–752.

**Dolan, Paul, Michael Hallsworth, David Halpern, Dominic King, Robert Metcalfe, and Ivo Vlaev**, "Influencing behaviour: The mindspace way," *Journal of Economic Psychology*, 2012, *33* (1), 264–277.

**Dupas, Pascaline**, "Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya," *American Economic Journal: Applied Economics*, 2011, *3* (1), 1–34.

**Elbers, Chris, Samuel Godfrey, Jan W Gunning, Matteus van der Velden, and Melinda Vigh**, "Effectiveness of Large Scale Water and Sanitation Interventions: The One Million Initiative in Mozambique," 2012. Tinbergen Institute Discussion Paper, Tinbergen Institute.

**Fitzsimons, Emla, Bansi Malde, Alice Mesnard, and Marcos Vera-Hemasdez**, "Nutrition, Information and Household Behavior: Experimental Evidence from Malawi," *Journal of Development Economics*, 2016, *122*, 113–126.

**Galiani, Sebastian, Paul Gertler, Nicolas Ajzenman, and Alexandra Orsola-Vidal**, "Promoting Handwashing Behavior: The Effects of Large-Scale Community and School-Level Intervention," *Health Economics*, 2016, *25*, 1545–1559.

**Geeraert, Liesl, Wim Van den Noortgate, Hans Grietens, and Patrick Onghena**, "The effects of early prevention programs for families with young children at risk for physical child abuse and neglect: A meta-analysis," *Child maltreatment*, 2004, *9* (3), 277–291.

**Guiteras, Raymond P and Ahmed Mushfiq Mobarak**, "Does development aid undermine political accountability? Leader and constituent responses to a large-scale intervention," 2015. NBER Working Paper No.2144, National Bureau of Economic Research.

**Hakim, Catherine**, "Erotic capital," *European Sociological Review*, 2010, *26* (5), 499–518.

**Hamermesh, Daniel S**, *Beauty pays: Why attractive people are more successful*, Princeton University Press, 2011.

**Hamermesh, Daniel S. and Jason Abrevaya**, "Beauty is the promise of happiness?," *European Economic Review*, November 2013 2013, *64*, 351–368.

**Hamermesh, Daniel S and Jeff E Biddle**, "Beauty and the Labor Market," *American Economic Review*, 1994, *84* (5), 1174–1194.

**Hamermesh, Daniel S., Rachel A. Gordon, and Robert Crosnoe**, " **"** O Youth and Beauty: **"** Children's looks and children's cognitive development," *Journal of Economic Behavior & Organization*, 2023, *212*, 275–289.

**Harper, Barry**, "Beauty, Stature and the Labour Market: A British Cohort Study," *Oxford Bulletin of Economics and Statistics*, 2000, *62* (s1), 771–800.

**Japan International Cooperation Agency and Shiseido Co., Ltd. and Kaihatsu Management Consulting, Inc.**, "Preparatory research for a project to improve the living standard of rural women in Bangladesh through skincare products executive summary," 2015.

**Kessler, Ronald C, Gavin Andrews, Lisa J Colpe, Eva Hiripi, Daniel K Mroczek, S-LT Normand, Ellen E Walters, and Alan M Zaslavsky**, "Short screening scales to monitor population prevalences and trends in non-specific psychological distress," *Psychological medicine*, 2002, *32* (6), 959–976.

**King, Amy and Andrew Leigh**, "Beautiful Politicians," *Kyklos*, 2009, *62* (4), 579–593.

**Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz**, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

**Levere, Michael, Gayatri Acharya, and Prashant Bharadwaj**, "The Role of Information and Cash Transfers on Early Childhood Development: Evidence from Nepal," 2016. Policy Research Working Paper, World Bank.

**Loewenstein, George, David A Asch, Joelle Y Friedman, Lori A Melichar, and Kevin G Volpp**, "Can behavioural economics make us healthier?," *British Medical Journal*, 2012, *344*, 3482–3484.

**Mobius, Markus M. and Tanya S. Rosenblat**, "Why Beauty Matters," *American Economic Review*, March 2006, *96* (1), 222–235.

**Neumark, David**, "Experimental research on labor market discrimination," *Journal of Economic Literature*, 2018, *56* (3), 799–866.

**Ning, Yang, Sida Peng, and Kosuke Imai**, "Robust Estimation of Causal Effects via High-Dimensional Covariate Balancing Propensity Score," *Biometrika*, 2020, *107* (3), 533–554.

**Oster, Emily**, "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 2019, *37* (2), 187–204.

**Patil, Sumeet R, Benjamin F Arnold, Alicia Salvatore, Bertha Briceno, John M Colford Jr., and Paul J Gertler**, "A Randomized, Controlled Study of a Rural Sanitation Behavior Change Program in Madhya Pradesh, India," 2013. Policy Research Working Paper, World Bank.

**Rosenberg, Morris**, *Society and the adolescent self-image*, Vol. 11, Princeton, NJ: Princeton University Press, 1965.

**Sugiyama, Lawrence S.**, "Physical Attractiveness: An Adaptationist Perspective," *Handbook of Evolutionary Psychology*, 2015, pp. 1–68.

**Thaler, Richard H and Cass R Sunstein**, *Nudge: Improving decisions about health, wealth, and happiness*, New Haven, CT: Yale University Press, 2008.

**Thornhill, R. and S. W. Gangestad**, *The Evolutionary Biology of Human Female Sexuality*, New York: Oxford University Press, 2008.

**VanEpps, Eric M, Julie S Downs, and George Loewenstein**, "Calorie label formats: using numeric and traffic light calorie labels to reduce lunch calories," *Journal of Public Policy & Marketing*, 2016, *35* (1), 26–36.

**Vlaev, Ivo, Dominic King, Paul Dolan, and Ara Darzi**, "The theory and practice of " nudging " : changing health behaviors," *Public Administration Review*, 2016, *76* (4), 550–561.

**Wooldridge, Jeffrey M**, "Should instrumental variables be used as matching variables?," *Research in Economics*, 2016, *70* (2), 232–237.

**Zhang, Junsen, Shulan Fei, and Yanbing Wen**, "How Does the Beauty of Wives Affect Post-marriage Family Outcomes? Helen's Face in Chinese Households," *Journal of Economic Behavior & Organization*, 2023, *212*, 122–137.

**Zhao, Qiran and Xiaohua Yu**, "Parental Nutrition Knowledge, Iron Deficiency, and Child Anaemia in Rural China," *Journal of Development Studies*, 2020, *56* (3), 578–595.

Table 8: Meta-analysis: Three-level models

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | All | Ex.Latrine | All | Ex.Latrine | All | Ex.Latrine | Incl.Ours |
| Fixed coefficients | | | | | | | |
| Intercepts | -0.083 | -0.081 | -0.035 | -0.048 | 0.010 | 0.002 | 0.002 |
| | (0.135) | (0.086) | (0.130) | (0.080) | (0.128) | (0.079) | (0.079) |
| Nutrition | 0.147** | 0.144** | 0.121** | 0.119** | 0.112** | 0.110** | 0.110** |
| | (0.060) | (0.058) | (0.058) | (0.054) | (0.056) | (0.051) | (0.051) |
| Latrine | 0.000 | | -0.014 | | -0.010 | | |
| | (0.067) | | (0.064) | | (0.061) | | |
| First factor | | | 0.044* | 0.042* | 0.044* | 0.042* | 0.042* |
| | | | (0.026) | (0.025) | (0.025) | (0.023) | (0.023) |
| Behavioral change | | | | | -0.022 | -0.023* | -0.023* |
| | | | | | (0.014) | (0.013) | (0.013) |
| Beauty nudge | | | | | | | 0.762*** |
| | | | | | | | (0.187) |
| 95% CI for prediction | [-0.121, | [-0.008, | [-0.110, | [-0.017, | [-0.090, | [0.003, | |
| | 0.248] | 0.135] | 0.234] | 0.115] | 0.242] | 0.129] | |
| Variance components | | | | | | | |
| Between-study variance | 0.0059 | 0.0055 | 0.0050 | 0.0042 | 0.0045 | 0.0034 | 0.0027 |
| Between-outcome variance | 0.0006 | 0.0004 | 0.0006 | 0.0004 | 0.0007 | 0.0004 | 0.0009 |
| Test for heterogeneity ($p$ val.) | $< 0.001$ | $< 0.001$ | $< 0.001$ | 0.001 | $< 0.001$ | 0.006 | $< 0.001$ |
| $\tau_{90}$ | 0.167 | 0.161 | 0.129 | 0.119 | 0.138 | 0.128 | |
| Comparison with the impact of our intervention ($p$ value for $H_0 : \tau \leq \tau_{90}$) | | | | | | | |
| Hygiene control PC score | 0.016 | 0.010 | 0.014 | 0.008 | 0.015 | 0.009 | |
| Health behavior $z$ score | 0.020 | 0.012 | 0.018 | 0.009 | 0.019 | 0.010 | |

Estimation results of the three-level meta-analyses are reported with standard errors in parentheses. Asterisks indicate statistical significance: * $p < .10$, ** $p < .05$, and *** $p < .01$. $\tau_{90}$ represents the estimated 90th percentile effectiveness of the existing intervention (see Appendix A.2).

# A Appendix

## A.1 Construction of the Outcome Scores

We constructed each outcome score based on questions as stated below. We computed a $z$-score for each item by

$$z_i = \frac{y_i - \bar{y}_{BT}}{s_{y,BT}}$$

where $\bar{y}_{BT}$ and $s_{y,BT}$ represent the sample mean and standard deviation of the actual baseline outcome $y$ of the treatment group. While the existing study (Kling et al., 2007) subtracts the control group mean and divides by the control group standard deviation to compute the $z$-scores, we standardized the index using the baseline actual data because the intervention was not randomized and the simple average and standard deviation of the control group are not meaningful. We then aggregated these scores in the same category by taking the simple average.

For the PC scores, we first conducted a factor analysis for the actual baseline data for the treatment group to obtain the weight for each variable. Then, we applied these weights for the baseline recall data and follow-up actual data to compute the aggregate score.

**Skincare Product Usage score:**

The following three questions comprise the Skincare Product Usage score.

(a) How frequently do you use face wash?: "Once"= 1, "Twice"= 2, "Three times or more"= 3, "Fewer than once"= 0.5, "Never"= 0

(b) How frequently do you use skin cream/gel?: "Once"= 1, "Twice"= 2, "Three times or more"= 3, "Fewer than once"= 0.5, "Never"= 0

(c) How frequently do you use sun cream?: "Once a day"= 30, "A couple times a week"= 10, "A couple times a month"= 2, "Fewer than once a month"= 0.5, "Never"= 0

**Skincare score:**

The following four questions comprise the Skin Care score.

(a) How frequently do you wash your face?: “Once a day”= 1, “Twice a day”= 2, “Three times a day or more”= 3, “Fewer than once a day”= 0

(b) Why do you wash your face?: “Keep face clean/ can remove dirt and gems”= 1, “Prevent acne on face”= 1, “Prevent unwanted face color”= 1, “Prevent oily skin”= 1, “Make my face beautiful”= 1, “Other “= 0, “Do not know”= 0

(c) What do you usually do to take care of your skin when you are outside in the sunshine? “Stay in the shade”= 1, “Cover skin with clothes and others”= 1, “Put on sunscreen”= 1, “Stay at home”= 0, “Other”= 0, “Nothing”= 0

(d) What will happen if you do not take care of your skin at all when exposed to strong sun for a long time?: “Sunburn”= 1, “Skin trouble”= 1, “Get tired”= 1, “Other”= 0, “Nothing”= 0

**Skincare Behavior scores:**

The Skincare Behavior score is computed based on questions (a)–(c) in the Skincare Product Usage score, and questions (a) and (c) in the Skin Care score.

**Skin Condition score:**

The following three questions comprise the Skin Condition score.

(a) What kind of skin trouble do you have?: “Acne”= 1, “Eczema”= 1, “Birthmark”= 0, “Skin color”= 1, “Inflammation”= 1, “Itchiness”= 1, “Dry and rough”= 1, “Cracking”= 1, “Spot”= 1, “Wrinkles”= 1, “Shadows under eye”= 1, “Oily skin”= 1, “Other”= 1, “None”= 0, (*Equation = 1 − [(sum of the score)/12])

(b) How satisfied are you with the skin condition of your face?: “Very much”= 3, “Somewhat”= 2, “Not much”= 1, “Not at all”= 0

(c) How confident are you with your skin care?: “Very much”= 3, “Somewhat”= 2, “Not much”= 1, “Not at all”= 0

**Hygiene Control score:**

The following eight questions comprise the Hygiene Control score.

(a) What do you usually use when you wash your hands?: “Soap and water”= 1, “Cleansing agents and water”= 0, “Water only”= 0, “Other”= 0

(b) Why do you wash your hands?: “Keep hands clean/can remove dirt and germs”= 1, “Prevent diarrhea”= 1, “Prevent skin problems”= 0, “Maintain good health”= 1, “Other”= 0, “Do not know”= 0

(c) What do you usually do after you wash your hands?: “Nothing”= 0, “Wipe with towel”= 1, “Wipe with clothes”= 0, “Other”= 0

(d) How frequently do you wash or change the towel with which you wipe your hands after washing hands?: “Everyday”= 7, “ Once every two to three days”= 2.5, “Once a week”= 1, “Fewer than once a week”= 0.5

(e) What do you think will happen if you do not wash or change the towel with which you wipe your hands, at the right time?: “Get dirty and affect your sanitation”= 1, “Germs affect your health”= 1, “Other”= 0, “Nothing”= 0

(f) How frequently do you change your and your family’s bed linens?: “Daily”= 30, “ Once every two to three days”= 12.5, “Once a week”= 4, “ Once every two to four weeks”= 1.5, “Fewer than once a month”= 0.5, “None”= 0

(g) How frequently do you hang your and your family’s pillows in the sun?: “Daily”= 30, “ Once every two to three days”= 12.5, “Once a week”= 4, “Once for two to four weeks”= 1.5, “Fewer than once a month”= 0.5, “No”= 0

(h) Why do you wash your bed linens or hang your pillows in the sun regularly?: “Feel comfortable”= 1, “It is hygienic/ can remove dirt and gems”= 1, “Prevent infection”= 1, “Prevent acne”= 0, “Prevent bad smell”= 1, “Other”= 0, “Do not know”= 0

**Diet score:**

The following nine questions comprise the Diet score.

(a) What are you usually concerned about when you cook?: “Number of dishes”= 1, “Nutrition”= 1, “Taste”= 0, “How the meal looks”= 0, “Amount”= 0, “Other”= 0

(b) What do you understand by a “nutritious meal”?: “Good balance”= 1, “Good amount”= 0, “Fresh meal”= 1, “Other”= 0, “Do not know”= 0

(c) What do you think is the risk to your health if you eat too much salt?: “Get tired”= 0, “Get fat”= 0, “Skin trouble”= 0, “Increase in blood pressure”= 1, “Diabetes”= 0, “Heart-related diseases”=1, “Diseases related to blood vessels”=1, “Acidity”= 0, “Other”= 0, “Do not know”= 0

(d) What do you think is the risk to your health if you eat too much oil?: “Get tired”= 0, “Get fat”= 1, “Skin trouble”= 0, “Increase in blood pressure”= 1, “Diabetes”= 0, “Heart-related diseases”=1, “Diseases related to blood vessels”=1, “Acidity”= 1, “Other”= 0, “Do not know”= 0

(e) What do you think is the risk to your health if you eat too much sugar?: “Get tired”= 0, “Get fat”= 1, “Skin trouble”= 0, “Increase in blood pressure”= 0, “Diabetes”= 1, “Heart-related diseases”=1, “Diseases related to blood vessels”=1, “Acidity”= 0, “Other”= 0, “Do not know”= 0

(f) What kind of vegetables did you eat yesterday?: “Root vegetables”= 1, “Leafy green vegetables”= 1, “Alliums (onion, leek, garlic, shallot)”= 1, “Potatoes”= 1, “Legumes (peas and beans)”= 1, “Squash”= 1, “Tomatoes”= 1, “Salad crops”= 1, “Other vegetables”= 1

(g) When do you usually wash vegetables?: “Before cutting”= 1, “After cutting”= 0, “Do not wash”= 0, “Other”= 0

(h) What are you concerned the most about when you are cooking vegetables?: “Do not heat too much”= 1, “The amount is enough”= 1, “Ensure enough variety”= 1, “Cook leafy vegetables daily”= 1, “Choose fresh ones”= 1, “Other”= 0

(i) What is the risk to your health if you do not eat enough vegetables?: “Skin trouble”= 0, “Anemia”= 1, “Constipation”= 1, “Get sick (high blood pressure, high blood sugar)”= 1, “Other”= 0

Note that the risks of eating too much salt, oil, and sugar are manifold and difficult to exclude any potential health effects. The computation of the score above is ad hoc but our results are robust to various changes in the definition of correct answers to questions (c), (d), and (e).

**Health Behavior scores:**

The Health Behavior score is computed based on questions (a), (c), (d), (f), and (g) in the Hygiene Control score and questions (f) and (g) in the Diet score.

**Hygiene for the Skin score and Diet for the Skin score:**

These are not outcome variables but intend to capture the extent to which the respondents are conscious of the effect on their skin when taking hygiene and diet practices. The hygiene for the skin score is equal to 1 if the respondent chose the alternative "prevent skin problems" to question (b) and "prevent acne" to question (h). It is equal to 0.5 if she chose any of these alternatives to only one of these two questions, and is equal to 0 if she did not choose these alternatives at all. The diet for the Skin score is computed as the number of times the respondent chose the alternative "Skin trouble" to questions (c), (d), and (i), divided by 3.

**Decision-Making score:**

The following two questions comprise the Decision-Making score.

(a) Who usually makes decisions on health care for yourself?: "Respondent"= 2, "Husband"= 0, "Respondent and husband"= 1, "Mother/mother in law"= 0, "Father/ father in law"= 0, "Other"= 0

(b) Who usually makes decisions on major household purchases?: "Respondent"= 2, "Husband"= 0, "Respondent and husband"= 1, "Mother/mother in law"= 0, "Father/ father in law"= 0, "Other"= 0

**Social Relationship score:**

The following five questions comprise the Social Relationship score.

(a) How willing are you to go out?: "Very much"= 2, "If needed"= 1, "Want to stay home"= 0, "Other"= 0

(b) Are you willing to make new friends?: "Very much"= 3, "Somewhat"= 2, "Not much"= 1, "Not at all"= 0

(c) From whom/what do you usually get information related to health, sanitation, and/or nutrition?: "Family"= 0, "Friends"= 1, "TV"= 1, "Radio"= 1, "School"= 1, "Shop sellers"= 1, "AO/aparajita"= 0, "Other"= 1, "Do not get information"= 0

(d) To whom do you teach good information and/or practices related to health, sanitation and/or nutritional issues?: "Husband"= 1, "Children"= 1, "Your family"= 1, "Your husband's family"= 1, "Your friends"= 1, "Other"= 1, "No conversation"= 0

(e) Do you have goals you want to achieve in the future?: "Yes"= 1, "No"= 0

## A.2 Meta-analysis

### Search procedure

We searched for academic papers on EconLit using a combination of the following keywords (2000–2019): (1) hand washing, handwashing, hygiene, sanitation, and nutrition and (2) experiment, intervention, randomized controlled trial, treatment, and control. We excluded studies without treatment, including education, information, or knowledge components (e.g., cash transfer only) and also excluded studies that did not include knowledge improvement or behavioral change as outcome variables. We further excluded papers that did not report the standard deviation, since we cannot standardize the impact for comparison. Finally, we identified 14 papers, with 91 cases under the following 10 types of interventions: sanitation and hand washing promotion, sanitation training, hand washing promotion (information), health session (education), nutrition information, nutrition information and in-kind promotion, nutrition meeting (information), nutrition session (education), sanitation (latrine use) awareness campaign, and sanitation (latrine use) awareness campaign including subsidies.[15] Some of these studies investigate so-called Community-Led Total Sanitation (CLTS),

[15] Twelve of these papers are (Briceno et al., 2015; Cameron et al., 2013; Chase and Do, 2012; Dickinson et al., 2015; Dillon et al., 2019; Elbers et al., 2012; Fitzsimons et al., 2016; Galiani et al., 2016; Guiteras and

which focuses on stopping open defecation. More specifically, CLTS aims at raising collective awareness on the open defecation problem in the community through analysis and discussions of the community sanitation situation and fecal contamination. Total Sanitation and Sanitation Marketing (TSSM) combines CLTS with intensive sanitation marketing and promotion campaigns, including community-based events and mass media (print and radio). Brief explanations of the interventions are listed below. The outcome variables are listed in the data file of the meta analysis.

**Briceño et al. (2015 )** evaluate TSSM and handwashing with soap (HWWS) interventions in Tanzania. HWWS provides technical assistance in the building of handwashing stations with local materials and promotes handwashing, with a special focus on mothers.

**Cameron et al. (2013)** evaluate TSSM in Indonesia.

**Chase and Do (2012)** evaluate a communication campaign in Vietnam, which focused on changing the perceptions of and addressing other motivational barriers to handwashing with soap by using TV ads and interpersonal communication (IPC) activities. The control group only received TV ads. Meanwhile, the IPC activities included group meetings, household visits, market meetings, loudspeaker announcements, HWWS festivals, cooking contests, and the distribution of informational and promotional materials.

**Dickinson et al. (2015)** evaluate CLTS in India. Their strategy emphasizes the links between sanitation and health outcomes, as well as the non-health benefits of latrine use, including those benefits related to saving time, privacy, and dignity.

**Dillon et al. (2019)** evaluate a two-year enhanced-homestead food production (E-HFP) program in Burkina Faso, which focused on improving women's agricultural production of nutrient-rich foods, coordinated with nutrition, health, and hygiene behavior change communication (BCC). The nutrition and health BCC strategy focused on educating the public regarding essential nutrition actions and encouraging the adoption of optimal Infant and Young Child Feeding practices, including the consumption of nutrient-rich foods.

**Elbers et al. (2012)** evaluate the One Million Initiative in Mozambique, which provided sanitation and hygiene training and improved water availability by creating or rehabilitating water points. Notably, the program combined training with the CLTS approach.

Mobarak, 2015; Levere et al., 2016; Patil et al., 2013; Zhao and Yu, 2020).

**Fitzsimons et al. (2016)** evaluate an infant feeding counseling intervention in Malawi, which provided mothers with information and advice on feeding infants; encouraged exclusive breastfeeding; and provided information on locally available nutritious foods, the importance of a varied diet, and how to prepare nutritious and easy-to-digest foods.

**Guiteras and Mobarak (2015)** evaluate a community motivation campaign, called the Latrine Promotion Program, in Bangladesh, which sought to end open defecation and urged households to adopt hygienic latrines, rather than simply using any latrine.

**Levere at al. (2016)** evaluate the provision of information on best practices regarding nutrition and childcare for children in Nepal. The information includes breastfeeding, care when sick, and supplemental feeding when older.

**Patil et al. (2013)** evaluate the Total Sanitation Campaign in India, which included subsidies for and the promotion of individual household latrines, school sanitation and hygiene education, Anganwadi toilets, and community sanitation complexes. It also supported rural sanitary marts and production centers. Awards were given to the communities that became Open Defecation Free.

### Model Specification

Given that many extant studies reported multiple outcomes, we use a three-level approach (Geeraert et al., 2004; Van den Noortgate et al., 2015) that models the sampling variation for each effect size (level one), variation across studies (level two), and variation across outcomes within a study (level three). Specifically, the effect size of outcome $j$ in study $k$ is modeled as:

$$\tau_{jk} = \tau_{00} + \mu_{0k} + \xi_{jk} + \nu_{jk},$$

where $\tau_{00}$ is the overall mean population effect size, $\mu_{0k}$ is the deviation of the study mean from the overall mean population effect, $\xi_{jk}$ is the deviation of the population effect $j$ in study $k$ from the mean population effect in study $k$, and $\nu_{jk}$ is the variation of the effect size due to the sampling variation. Both $\mu_{0k}$ and $\xi_{jk}$ are normally distributed random effects, with variances $\sigma^2_\mu$ and $\sigma^2_\xi$, respectively. Variance $\sigma^2_\mu$ is between-studies and variance $\sigma^2_\xi$ refers to the within-study variance. The deviation due to the sampling variation, $\nu_{jk}$, is assumed to be normally distributed with zero mean and variance $\sigma^2_{\nu_{jk}}$, which can differ across outcomes and studies reflecting the size of the study.

This three-level model accounts for the dependence of outcomes in the same study sample as well as the heterogeneity of the program effect. To allow for the possibility that program characteristics $\mathbf{w}_{1k}$ and country characteristics $\mathbf{w}_{2k}$ affect the program effect, we also parameterize $\tau_{00}$ as:

$$\tau_{00} = \mathbf{w}_{1k}\boldsymbol{\delta}_1 + \mathbf{w}_{2k}\boldsymbol{\delta}_2.$$

For the country characteristics variables, $\mathbf{z}_k$, we used data one year before the start of the intervention for each country. The models are estimated by the restricted maximum likelihood estimation to correct for the downward bias of the maximum likelihood estimates of the variance components (Cheung, 2015).

In Table 8 in the main text, the between-study variance refers to the estimate of $\sigma_\mu^2$ and the between-outcome variance refers to the estimate of $\sigma_\xi^2$. Given the independence of random effects, the variance of the heterogeneous effect size is $\sigma_\tau^2 \equiv \sigma_\mu^2 + \sigma_\xi^2$.

We are interested in testing if the effectiveness of our treatment, $\hat{\tau}$, is much greater than the effectiveness of the existing interventions. Under the normality assumption, the maximum likelihood estimate of the 100$p$th percentile effectiveness of the existing intervention is obtained as

$$\hat{\tau}_p = \hat{\tau}_{00} + z_p\hat{\sigma}_\tau,$$

where $\hat{\tau}_{00}$ denotes the estimate of $\tau_{00}$, $z_p$ the 100$p$th percentile of the standard normal distribution, and $\hat{\sigma}_\tau$ the estimate of $\sigma_\tau$. Given $\hat{\tau}_p$, we test null hypotheses as $H_0 : \tau \leq \tau_p$ for $p = 0.75$ and $p = 0.9$. The $t$ statistics for this test is simply

$$t = \frac{\hat{\tau} - \hat{\tau}_p}{\sqrt{(s.e.(\hat{\tau}))^2 + (s.e.(\hat{\tau}_p))^2}},$$

where $s.e.(\hat{\tau})$ and $s.e.(\hat{\tau}_p)$ are the standard errors for $\hat{\tau}$ and $\hat{\tau}_p$.

To construct the standard errors for $\hat{\tau}_p$, we need to allow for the sampling error of the estimate $\hat{\sigma}_\tau$ in addition to that of $\hat{\tau}_{00}$, as shown in equation (A.2). Note that as $z_p = 0$ for the mean, the sampling error of $\hat{\sigma}_\tau$ need not be considered for computing the standard errors for the average treatment effects. Chakraborti and Li (2007) argue that under the normality assumption, the confidence intervals are well approximated by using the standard errors

$$s.e.(\hat{\tau}_p) = s.e.(\hat{\tau}_{00})\sqrt{1 + nz_p^2(C_n^2 - 1)}$$

where $n$ is the sample size and

$$C_n \equiv \sqrt{\frac{n-1}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} > 1,$$

in which $\Gamma(\cdot)$ is the gamma function. Compared to the case for the mean, the standard error for the percentile is inflated by $D_n \equiv \sqrt{1 + nz_p^2(C_n^2 - 1)}$. While $C_n$ converges to 1 as $n$ gets larger, the term $nz_p^2(C_n^2 - 1)$ will not converge to 1. Computation shows that $D_n$ gets close to 1.349 for the 90th percentile as $n$ gets larger.[16]

This inflation factor depends on the sample size $n$, but there is no obvious counterpart to $n$ in the multi-level meta-analysis in which we allow for heterogeneity and dependence in the effect sizes. To make the tests conservative, we set $n$ equal to the number of studies (14) minus the number of the parameters in the estimated model. That is, in Table 4, we set $n = 11, 9, 10, 8, 9$, and 7 for Columns (1), (2), (3), (4), (5), and (6), respectively.[17]

# References

**Briceno, Bertha, Aidan Coville, and Sebastian Martinez**, "Promoting Handwashing and Sanitation: Evidence from a Large-Scale Randomized Trial in Rural Tanzania," 2015. Policy Research Working Paper, World Bank.

**Cameron, Lisa, Manisha Shab, and Susan Olivia**, "Impact Evaluation of a Large-Scale Rural Sanitation Project in Indonesia," 2013. Policy Research Working Paper, World Bank.

**Chakraborti, S and J Li**, "Confidence interval estimation of a normal percentile," *The American Statistician*, 2007, *61* (4), 331–336.

**Chase, Claire and Quy-Toan Do**, "Handwashing Behavior Change at Scale: Evidence from a Randomized Evaluation in Vietnam," 2012. Policy Research Working Paper, World Bank.

[16] For the 90th percentile, $D_{10} = 1.3915$, $D_{20} = 1.3695$, $D_{50} = 1.3573$, $D_{100} = 1.3534$, and $D_{100000} = 1.3496$.

[17] For the 90th percentile, $D_7 = 1.4122$, $D_8 = 1.4034$, $D_9 = 1.3967$, $D_{10} = 1.3915$, and $D_{11} = 1.3873$.

**Cheung, Mike W-L**, *Meta-analysis: A structural equation modeling approach*, John Wiley & Sons, 2015.

**den Noortgate, Wim Van, José Antonio López-López, Fulgencio Marín-Martínez, and Julio Sánchez-Meca**, "Meta-analysis of multiple outcomes: a multilevel approach," *Behavior research methods*, 2015, *47* (4), 1274–1294.

**Dickinson, Katherine L, Sumeet R Patil, Subhrendu K Pattanayak, Christime Poulos, and Jui hen Yang**, "Nature's Call: Impacts of Sanitation Choices in Orissa, India," *Economic Development and Cultural Change*, 2015, *64* (1), 1–29.

**Dillon, Andrew, Joanne Arsenault, and Deanna Olney**, "Nutrient production and micronutrient gaps: evidence from an agriculture-nutrition randomized control trial," *American Journal of Agricultural Economics*, 2019, *101* (3), 732–752.

**Elbers, Chris, Samuel Godfrey, Jan W Gunning, Matteus van der Velden, and Melinda Vigh**, "Effectiveness of Large Scale Water and Sanitation Interventions: The One Million Initiative in Mozambique," 2012. Tinbergen Institute Discussion Paper, Tinbergen Institute.

**Fitzsimons, Emla, Bansi Malde, Alice Mesnard, and Marcos Vera-Hemasdez**, "Nutrition, Information and Household Behavior: Experimental Evidence from Malawi," *Journal of Development Economics*, 2016, *122*, 113–126.

**Galiani, Sebastian, Paul Gertler, Nicolas Ajzenman, and Alexandra Orsola-Vidal**, "Promoting Handwashing Behavior: The Effects of Large-Scale Community and School-Level Intervention," *Health Economics*, 2016, *25*, 1545–1559.

**Geeraert, Liesl, Wim Van den Noortgate, Hans Grietens, and Patrick Onghena**, "The effects of early prevention programs for families with young children at risk for physical child abuse and neglect: A meta-analysis," *Child maltreatment*, 2004, *9* (3), 277–291.

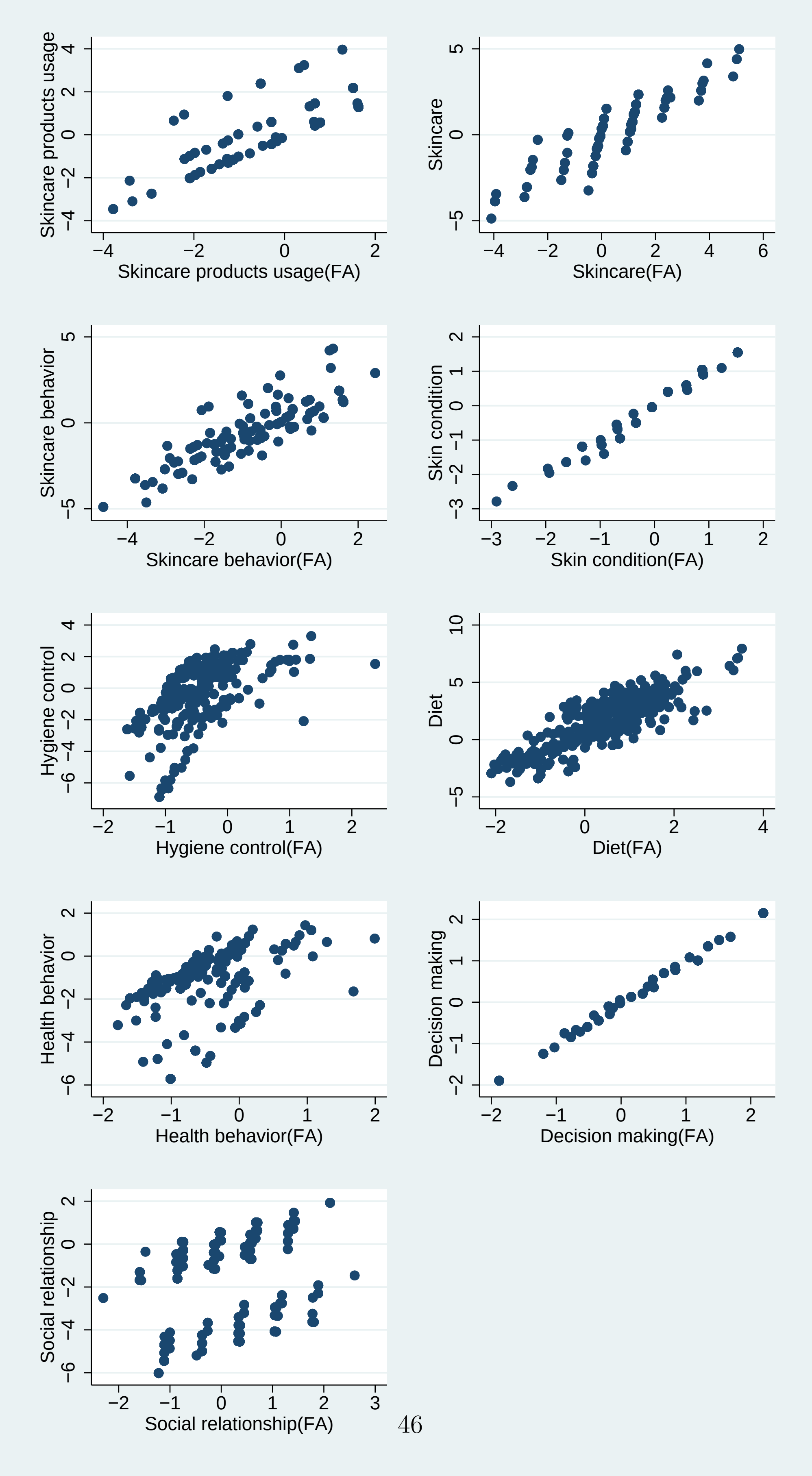
**Guiteras, Raymond P and Ahmed Mushfiq Mobarak**, "Does development aid undermine political accountability? Leader and constituent responses to a large-scale intervention," 2015. NBER Working Paper No.2144, National Bureau of Economic Research.

**Levere, Michael, Gayatri Acharya, and Prashant Bharadwaj**, "The Role of Information and Cash Transfers on Early Childhood Development: Evidence from Nepal," 2016. Policy Research Working Paper, World Bank.

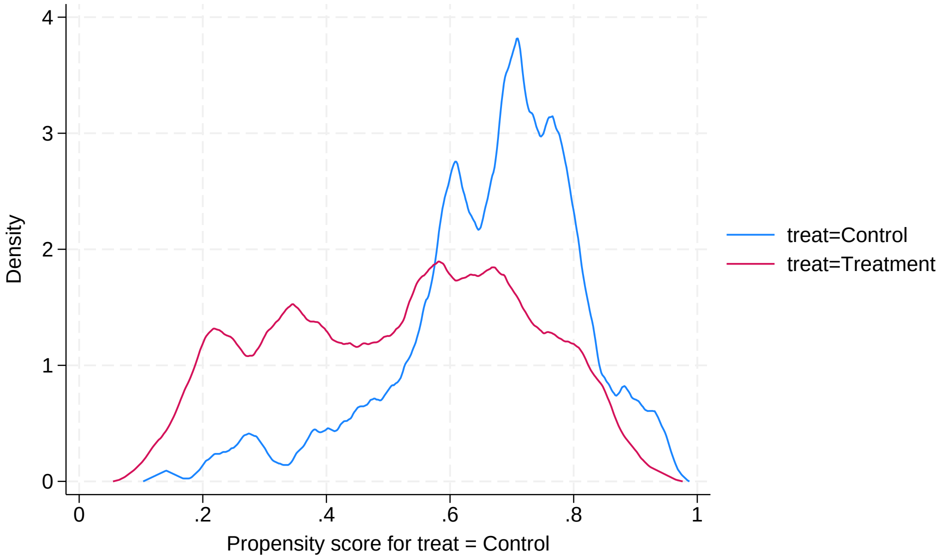
**Patil, Sumeet R, Benjamin F Arnold, Alicia Salvatore, Bertha Briceno, John M Colford Jr., and Paul J Gertler**, "A Randomized, Controlled Study of a Rural Sanitation Behavior Change Program in Madhya Pradesh, India," 2013. Policy Research Working Paper, World Bank.

**Zhao, Qiran and Xiaohua Yu**, "Parental Nutrition Knowledge, Iron Deficiency, and Child Anaemia in Rural China," *The Journal of Development Studies*, 2020, *56* (3), 578–595.

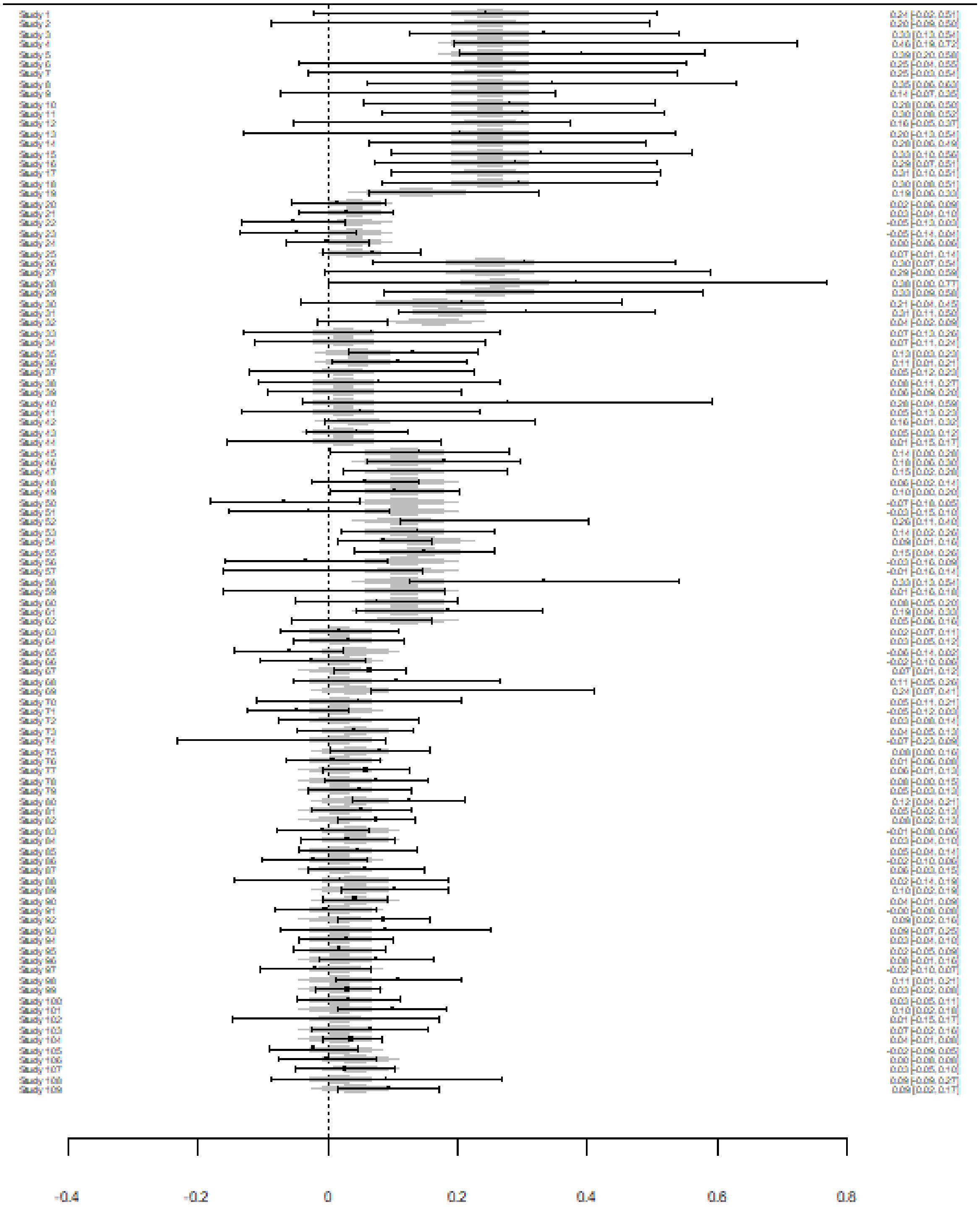
Appendix Figure S1: Comparison between the aggregate $z$ scores and the PC score



Appendix Figure S2: Overlap of the estimated propensity score



Appendix Figure S3: Forest plot of the effect size of the existing studies



The bar indicates the 95% confidence intervals of the effect size.

# Appendix Tables

Appendix Table S1: Difference in the baseline variables between the treatment and control groups

| | (1)<br>Treat | | (2)<br>Control | | (3)<br>Difference | |
|---|---|---|---|---|---|---|
| Age | 25.196 | (0.416) | 26.737 | (0.232) | -1.541* | (0.725) |
| Married | 0.763 | (0.028) | 0.930 | (0.016) | -0.166*** | (0.031) |
| Education | 4.112 | (0.084) | 3.593 | (0.095) | 0.519* | (0.257) |
| Household income | 3.982 | (0.092) | 3.711 | (0.084) | 0.271 | (0.280) |
| Individual income | 0.353 | (0.067) | 0.178 | (0.046) | 0.175 | (0.196) |
| Total number of people living in the same house | 4.746 | (0.116) | 4.600 | (0.104) | 0.146 | (0.281) |
| Number of your own children | 1.228 | (0.073) | 1.600 | (0.057) | -0.372*** | (0.095) |
| Skincare products usage | -0.287 | (0.086) | -0.742 | (0.061) | 0.455** | (0.144) |
| Skincare | -0.455 | (0.069) | -0.989 | (0.051) | 0.534* | (0.234) |
| Skincare behavior | -0.670 | (0.087) | -0.977 | (0.067) | 0.307 | (0.201) |
| Skin condition | 0.232 | (0.061) | 0.380 | (0.043) | -0.149 | (0.170) |
| Hygiene control | -0.770 | (0.076) | -1.584 | (0.083) | 0.814* | (0.320) |
| Diet | 0.376 | (0.069) | -1.012 | (0.069) | 1.388*** | (0.246) |
| Health behavior | -1.300 | (0.076) | -1.626 | (0.083) | 0.325 | (0.302) |
| Decision making | 0.574 | (0.086) | 0.786 | (0.066) | -0.213 | (0.154) |
| Social relationship | -1.834 | (0.124) | -2.856 | (0.123) | 1.022* | (0.521) |
| Happiness | 5.915 | (0.084) | 6.015 | (0.091) | -0.100 | (0.388) |
| Self Esteem | 17.381 | (0.147) | 17.163 | (0.138) | 0.218 | (0.487) |
| K6 | 16.464 | (0.307) | 16.687 | (0.243) | -0.222 | (1.037) |
| Hygiene for skin | 0.071 | (0.012) | 0.007 | (0.004) | 0.064** | (0.024) |
| Diet for skin | 0.222 | (0.017) | 0.037 | (0.006) | 0.185** | (0.060) |
| Observations | 224 | | 270 | | | |

Standard errors in parentheses. Asterisks indicate statistical significance: * $p < .10$, ** $p < .05$, and *** $p < .01$.

Appendix Table S2: Recall data bias in the treatment villages and the original control village

| | (1) | | (2) | | (3) | |
|---|---|---|---|---|---|---|
| | Treatment | | Control | | Difference | |
| Recall bias: Age | -0.181 | (0.095) | 0.000 | (.) | -0.181 | (0.215) |
| Recall bias: Married | 0.000 | (0.012) | 0.021 | (0.021) | -0.021 | (0.028) |
| Recall bias: Education | 0.665*** | (0.090) | 0.553* | (0.219) | 0.112 | (0.226) |
| Recall bias: Household income | 0.019 | (0.108) | 0.696** | (0.241) | -0.677** | (0.260) |
| Recall bias: Individual income | -0.347*** | (0.100) | 0.106 | (0.183) | -0.454 | (0.239) |
| Recall bias: Total number of people living in the same house | -0.102 | (0.083) | 0.000 | (0.135) | -0.102 | (0.198) |
| Recall bias: Number of your own children | -0.004 | (0.023) | 0.000 | (.) | -0.004 | (0.051) |
| Recall bias: Skincare products usage | -0.330** | (0.101) | -0.264 | (0.310) | -0.066 | (0.287) |
| Recall bias: Skincare | -0.490*** | (0.095) | -0.524** | (0.201) | 0.033 | (0.231) |
| Recall bias: Skincare behavior | -0.696*** | (0.105) | -0.757** | (0.257) | 0.061 | (0.288) |
| Recall bias: Skin condition | 0.207* | (0.082) | 0.179 | (0.178) | 0.027 | (0.201) |
| Recall bias: Hygiene control | -0.902*** | (0.104) | -0.865*** | (0.177) | -0.037 | (0.253) |
| Recall bias: Diet | 0.273** | (0.100) | -0.521** | (0.175) | 0.794*** | (0.232) |
| Recall bias: Health behavior | -1.412*** | (0.100) | -1.368*** | (0.153) | -0.044 | (0.234) |
| Recall bias: Decision making | 0.579*** | (0.101) | 0.426* | (0.193) | 0.153 | (0.242) |
| Recall bias: Social relationship | -1.816*** | (0.143) | -2.149*** | (0.248) | 0.333 | (0.338) |
| Recall bias: Skincare products usage(FA) | -0.336** | (0.103) | -0.237 | (0.318) | -0.099 | (0.293) |
| Recall bias: Skincare(FA) | -0.289** | (0.098) | -0.335 | (0.225) | 0.045 | (0.242) |
| Recall bias: Skincare behavior(FA) | -0.661*** | (0.108) | -0.687* | (0.289) | 0.026 | (0.299) |
| Recall bias: Skin condition(FA) | 0.192 | (0.101) | 0.111 | (0.212) | 0.081 | (0.245) |
| Recall bias: Hygiene control(FA) | -0.482*** | (0.100) | -0.760*** | (0.199) | 0.278 | (0.248) |
| Recall bias: Diet(FA) | -0.221* | (0.113) | -0.491* | (0.243) | 0.269 | (0.270) |
| Recall bias: Health behavior(FA) | -1.021*** | (0.103) | -1.059*** | (0.202) | 0.038 | (0.249) |
| Recall bias: Decision making(FA) | 0.573*** | (0.105) | 0.388* | (0.194) | 0.185 | (0.252) |
| Recall bias: Social relationship(FA) | -1.184*** | (0.110) | -1.682*** | (0.175) | 0.499 | (0.257) |

Standard errors are in parentheses. Asterisks indicate statistical significance: * $p < .10$, ** $p < .05$, and *** $p < .01$. The number of non-missing observations is 212 to 239 in the treatment group, and 44 to 47 in the control group, except for the skincare products usage score and skincare behavior score where the number of non-missing observations is 33 in the control group.

Appendix Table S3: Propensity score matching only based on variables less subject to recall errors

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | Skincare product use | | Skincare | | Skincare behavior | | Skin condition | |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score |
| Treatment | 0.985*** | 0.359** | 1.525*** | 1.596*** | 0.830*** | 0.366** | -0.072 | -0.055 |
| | (0.142) | (0.154) | (0.125) | (0.136) | (0.127) | (0.145) | (0.113) | (0.113) |
| Observations | 472 | 472 | 473 | 473 | 472 | 472 | 473 | 473 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | Hygiene control | | Diet | | Health behavior | | Hygiene | Diet |
| | $z$ score | PC score | $z$ score | PC score | $z$ score | PC score | for skin | for skin |
| Treatment | 1.784*** | 1.183*** | 3.616*** | 2.296*** | 0.998*** | 0.966*** | 0.451*** | 0.343*** |
| | (0.151) | (0.077) | (0.206) | (0.121) | (0.120) | (0.074) | (0.027) | (0.022) |
| Observations | 465 | 465 | 473 | 473 | 465 | 465 | 473 | 473 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | |
| | Decision-Making | | Social Relationship | | Subjective measures | | | |
| | $z$ score | PC score | $z$ score | PC score | Happiness | Self-Esteem | K6 | |
| Treatment | 0.137 | 0.136 | 1.420*** | 1.323*** | 1.712*** | -0.134 | 0.712 | |
| | (0.135) | (0.135) | (0.278) | (0.170) | (0.170) | (0.281) | (0.510) | |
| Observations | 455 | 455 | 471 | 471 | 471 | 473 | 471 | |

The estimation results using propensity score matching are reported with standard errors in parentheses. Asterisks indicate statistical significance: * $p < .10$, ** $p < .05$, and *** $p < .01$. The covariates include age, marital status, household size, number of children, education, household income, and individual income at endline survey; interactions between age and education, househld size and education, number of children and education, and age and individual income; and usage of the skincare product at baseline. We drop the obervations without any matched observations within bandwidth of 0.1.