

## **Statistical Analysis Plan**

### **A Cluster Randomized Controlled Trial Evaluating the MEST Work Strategy to Improve Health Literacy Among Upper Secondary School Students**

# **A Cluster Randomized Controlled Trial Evaluating the MEST Work Strategy to Improve Health Literacy Among Upper Secondary School Students**

## **Statistical Analysis Plan**

### **1. Study design**

This is a cluster-randomized controlled trial with schools as the unit of randomization. Schools are allocated 1:1 to intervention or control. Outcomes are measured at baseline (t0), 4 months (t1), and 9 months/end of school year (t2).

### **2. Primary endpoint**

The primary endpoint is t2, measured at the end of the school year, approximately 9 months after baseline.

### **3. Outcomes**

The study includes three prespecified continuous outcomes:

#### **Primary Outcome Measure**

1. General Health Literacy [Time Frame: Baseline, 4 months, 9 months]

General health literacy assessed using the Health Literacy Survey 2019 - Young Persons 12-item scale (HLS19 YP12).

The instrument consists of 12 items; each rated on a 4-point scale. Total scores range from 12 to 48, with higher scores indicating higher (better) levels of health literacy.

#### **Secondary Outcome Measures**

1. Mental Health Literacy [Time Frame: Baseline, 4 months, 9 months]

Mental health literacy using the Mental Health Learning Scale (MHLS-9). The instrument consists of 9 items, each rated on a 5-point Likert scale, giving a total score from 9 to 45, with higher scores indicating higher (better) levels of mental health literacy.

2. Mental Well-Being [Time Frame: Baseline, 4 months, 9 months]

Mental well-being assessed using the Short Warwick-Edinburgh Mental Well-Being Scale (SWEMWBS). The instrument consists of 7 items; each rated on a 5-point scale. It ranges from 7 to 35 points, with higher scores indicating better mental well-being.

#### 4. Primary analysis

The primary analysis will estimate the effect of the intervention on each prespecified outcome at t2 using a prespecified covariate-adjusted linear mixed model with random intercept for school. The model will adjust for the baseline value of the same outcome and a fixed set of student-level baseline covariates selected a priori based on their expected association with the outcomes.

The model is specified as:

$$Y_{is2} = \alpha + \beta_0 Y_{is0} + \tau \text{Treat}_s + \gamma'X_{is0} + u_s + e_{is}$$

where  $Y_{is2}$  is the outcome at t2 for student  $i$  in school  $s$ ,  $Y_{is0}$  is the baseline value of the same outcome,  $\text{Treat}_s$  indicates whether school  $s$  was assigned to the intervention or control group,  $X_{is0}$  is the prespecified vector of student-level baseline covariates,  $u_s$  is the random effect for school, and  $e_{is}$  is the individual-level error term. The dependent variable is the outcome at t2. Fixed effects are treatment group, baseline value of the same outcome, and the prespecified baseline covariates. School is included as a random effect to account for clustering.

This covariate-adjusted model is the prespecified primary analysis model. The confirmatory primary hypothesis test concerns the primary outcome, General Health Literacy, at t2. The same model will be applied to the secondary outcomes at t2, and these analyses will be interpreted as supportive. The covariates are specified a priori and will be included simultaneously, with no data-driven selection or addition of covariates after inspection of outcome data.

#### 5. Estimand

The main estimand is the adjusted mean difference between intervention and control at t2.

Positive values will be interpreted as favoring the intervention because higher scores indicate better health literacy or well-being for all three outcomes.

#### 6. Analysis population

The primary analysis will follow the intention-to-treat principle at the cluster level. All randomized schools will be analyzed in their assigned group. At the individual level, all students with available outcome data at the relevant follow-up will be analyzed according to their school's randomized group.

For the primary covariate-adjusted model, students must have the relevant t2 outcome, the baseline value of the same outcome, and available data on the prespecified covariates included in the model. The handling of missing baseline, covariate, and outcome data is prespecified in Section 10.

## 7. Covariate adjustment

Baseline value of the same outcome will be included in all primary analyses. The following student-level baseline covariates are prespecified for the primary covariate-adjusted model and will be included together as fixed effects, with no data-driven selection: gender, age, socioeconomic status, self-reported grade average, self-reported health, leisure screen time, leisure physical activity, and minority/indigenous or immigrant background, as available in the baseline questionnaire.

If a prespecified covariate is not collected, is not available before outcome analysis, or has insufficient variation for model estimation, this will be reported and the covariate will be omitted from the primary covariate-adjusted model.

### Covariate coding rules

Covariates will be coded according to the baseline questionnaire before outcome analysis. Age and self-reported grade average will be included as continuous variables. Gender, self-reported health, leisure screen time, leisure physical activity, socioeconomic status, and minority/indigenous or immigrant background will be coded according to prespecified questionnaire categories. If sparse categories prevent model estimation, categories will only be collapsed according to conceptually adjacent categories, and the applied coding will be documented before outcome results are inspected. No coding decisions will be changed after inspection of outcome data.

## 8. Secondary analyses

Secondary analyses will include analyses of the same outcomes at t1, using analogous covariate-adjusted linear mixed models with random intercept for school. These models will adjust for the baseline value of the same outcome and the prespecified student-level baseline covariates listed in Section 7. These analyses will be interpreted as secondary and supportive rather than confirmatory.

## 9. Prespecified subgroup / moderator analyses

Interaction terms will not be included in the primary confirmatory analysis. The following subgroup/moderator analyses are prespecified and will be conducted in separate mixed models by adding a treatment-by-subgroup interaction term to the primary covariate-adjusted model:

- Socioeconomic status: low SES versus not low SES. Low SES will be defined as the lowest tertile of the prespecified SES variable/index measured at baseline. Students in the

middle and highest tertiles will be classified as not low SES. This definition will not be changed after inspection of outcome data.

- **Minority/indigenous or immigrant background:** students reporting minority/indigenous or immigrant background at baseline versus students not reporting such background. Responses indicating “prefer not to answer” or missing information will not be included in this interaction analysis, but these students will remain included in the primary analysis if they otherwise meet the inclusion criteria.
- **Gender:** boys versus girls, as reported in the baseline questionnaire. Students reporting another gender category or not reporting gender will not be included in the gender interaction analysis but will remain included in the primary analysis if they otherwise meet the inclusion criteria for that analysis.

Subgroup analyses will estimate differential intervention effects by including the subgroup main effect and the treatment-by-subgroup interaction term in addition to treatment group, baseline outcome, the prespecified baseline covariates listed in Section 7, and school random intercept. These analyses are prespecified but secondary; they will be interpreted cautiously with emphasis on effect estimates and uncertainty rather than definitive subgroup claims. No additional subgroups will be introduced after inspection of outcome data.

## 10. Missing data

The primary analysis will use a complete/available-case approach for each outcome-specific covariate-adjusted model: students with the relevant follow-up outcome, the baseline value of the same outcome, and available data on the prespecified covariates included in the model will be included. This primary approach assumes that missingness is conditionally ignorable given randomized group, school, baseline outcome, and observed baseline covariates.

Before outcome analysis, the amount and pattern of missing data will be described by randomized group, school, time point, outcome, and prespecified baseline covariates. Differential follow-up by arm and baseline predictors of missingness will be summarized.

Additional exploratory missing-data sensitivity analyses, such as multiple imputation, may be conducted outside the prespecified confirmatory analysis if missingness is substantial or differs meaningfully between randomized groups. Such analyses will be clearly labelled as exploratory and will not replace the prespecified complete/available-case primary analysis.

## 11. Statistical inference

Results will be reported as adjusted mean differences with 95% confidence intervals. The confirmatory primary hypothesis test for the intervention effect at t2 will be one-sided with  $\alpha = 0.05$ , testing whether the intervention improves the outcome relative to control. The direction of improvement is prespecified as higher scores on the outcome scale. Two-sided

95% confidence intervals will be reported for estimation and comparability, but the primary decision rule is based on the one-sided  $\alpha = 0.05$  test.

The three outcomes will be reported separately. The primary outcome is General Health Literacy; Mental Health Literacy and Mental Well-Being are secondary outcomes. Findings for secondary outcomes will be interpreted as supportive and will not override the primary-outcome conclusion.

## 12. Sensitivity analyses

Planned sensitivity analyses will include:

**Baseline-adjusted model without additional covariates:** the primary and secondary t1 models will be repeated including treatment group, baseline value of the same outcome, and school random intercept, but without the additional prespecified student-level covariates.

**Alternative clustering specification:** models with cluster-robust standard errors or alternative random-effect specifications will be used if convergence or distributional diagnostics indicate concerns with the primary mixed model.

**Exploratory missing-data analyses:** additional missing-data analyses, such as multiple imputation, may be conducted if missingness is substantial or differs meaningfully between randomized groups. These analyses will be clearly labelled as exploratory and will not replace the prespecified complete/available-case primary analysis.

**School-level aggregate before-after analysis:** for each school and outcome, the mean change from t0 to t2 will be calculated using students with available data at the relevant time points. School-level mean changes will then be compared between intervention and control schools using a linear model or weighted linear model, with weights based on the number of contributing students per school. This ecological analysis will be interpreted as a robustness check and not as the primary estimand.

## 13. Interpretation

Primary emphasis will be placed on the size and precision of the estimated intervention effect at t2. Analyses at t1 will be interpreted as secondary analyses of shorter-term effects.

Because subgroup analyses, baseline-adjusted models without additional covariates, alternative clustering specifications, exploratory missing-data analyses, and school-level aggregate analyses are secondary or sensitivity analyses, consistency with the primary covariate-adjusted t2 analysis will be emphasized over isolated statistical significance.