Milk or cowpea-containing peanut-based ready-to-use foods for school feeding of Ghanaian children 5-12 years of age to improve cognition

Statistical Analysis Plan (SAP)

Trial registration number: NCT04349007

Authors: Kevin Stephenson, Mark Manary

Contacts: Kevin Stephenson, email: k.stephenson@wustl.edu; Mark Manary, email: manarymj@wustl.edu

SAP history

Version	Date	Author	Reason
1	August 24, 2022	Kevin Stephenson	SAP for primary and secondary efficacy analyses
2	September 14, 2022	Kevin Stephenson	• Change to procedure for calculating DCCS, FICA accuracy scores, and PCPS scores • Addition of sensitivity analysis comparing trial procedure vs. NIH Toolbox procedure for calculating DCCS and FICA accuracy scores
3	November 30, 2022	Kevin Stephenson	Addition of post-hoc analysis

SAP Approval

Version	Date	Approver	Signature
1	August 24, 2022	Prof Mark Manary	Mark J. Manary
2	September 14, 2022	Prof Mark Manary	Mark J. Manary
3	November 30, 2022	Prof Mark Manary	Mark J. Manary

Table of Contents

1. Administrative Information

1.1 SAP version

This is version 3 of the Statistical Analysis Plan (SAP), dated 30 November 2022. The SAP describes the final analyses of primary and secondary endpoints, as well as post-hoc analyses, at the study end.

1.2 SAP revision history

Version 2: In a pre-specified data visualization step that took place between September 12-14, the team members responsible for statistical analysis, Kevin Stephenson and Mark Manary, first generated histograms, probability density plots, and mean/SD median (IQR) for the tablet-generated scores for the 4 primary outcomes irrespective of study food group (i.e., results were visualized without stratification by masked group identification). These 4 tests were dimensional change card sort (DCCS), flanker inhibitory control and attention (FICA), pattern comparison processing speed (PCPS), and list sorting working memory (LSWM). For DCCS and FICA, which have accuracy and speed sub-scores automatically calculated by NIH Toolbox program, accuracy scores were also evaluated. This descriptive evaluation was done prior to unmasking and was pre-specified because of the lack of data available prior to this trial regarding how children in rural Ghana would perform on the NIH Toolbox tests of fluid cognition. It was unclear, for instance, if certain assumptions underlying the Toolbox scoring guide would hold, including age-based assumptions on test results for DCCS and FICA.

After this evaluation, no changes were made to the analysis plan divergent from the NIH Toolbox scoring guide for the LSWM test or for the speed sub-scores of the DCCS and FICA tests.

Upon evaluation of the accuracy components of DCCS and FICA, however, an issue was identified. In the case of DCCS, the NIH Toolbox guide and program for the computed score (primary outcome) automatically adds 10 accuracy points (out of 40) for all children ≥ 8 years of age for the 10 pre- and post-switch trials based on the assumption that all children ≥ 8 years of age would obtain perfect scores therein. In the guide and program, all correct trials are multiplied by 0.125 to scale the accuracy score to 5 points; thus, all children ≥ 8 years of age would be given 1.25 / 5 (25%) accuracy points automatically. This assumption had been validated in the US testing cohort, where nearly all individuals scored at the ceiling, but after review of the present trial data, this assumption was deemed untenable in this study cohort. Frequent low accuracy scores were identified on DCCS (e.g., 45% of participants 8-11 years of age scored below 80% accuracy), as well as across the FICA, PCPS, and LSWM tests. Thus, the assumption underlying 10 free points for accuracy for participants ≥ 8 y was unlikely to reflect the abilities/scores of children in this cohort. Therefore, rather than following the NIH Toolbox accuracy scoring guide and automatically providing these 10 points, the decision was made instead to compute the accuracy score as the proportion of correct answers out of all trials taken by the participant. This was done because it was unknowable how children ≥ 8 years of age might have done on the 10 trials that were automatically skipped by the program on the basis of their age. This method required fewer assumptions about how these rural northern Ghanaian children would perform on these tests. As recommended in the NIH Toolbox guide, this proportion-correct accuracy score was then scaled to 5 accuracy points which, when added to the tablet-calculated speed score, composed the "computed score" used as the primary outcome.

An analogous issue was identified for the FICA test, wherein participants ≥ 8 y automatically received 20 accuracy points (out of 40), again resting on the assumption that they would test near the ceiling based on validation in the US cohort. As above, this assumption did not appear tenable in the trial cohort and seemed unlikely to be an accurate reflection of participant performance. As was done for DCCS, the proportion of correct answers on trials actually taken was used in place of this assumption, again scaled to 5 as recommended in the NIH Toolbox scoring guide, and added to the tablet-calculated speed score to compute the "computed score" used as the primary outcome.

In a sensitivity analysis, DCCS and FICA scores will also be computed using the NIH Toolbox scoring guide method, and ordinal logistic regression will be used to estimate odds ratios with 95% CIs for PM-RUF vs. FP and PC-RUF vs. FP. These will be displayed side-by-side with the method used in the primary analysis of this study. As in the primary analysis, baseline cognitive test scores and participant age will be used as covariates in the regressions.

Finally, the NIH Toolbox scoring guide recommends that the PCPS score be computed by adding the number of correct responses across all trials. Upon descriptive review of the trial data, again before unmasking and without stratification by study food group, it was noticed that multiple children's scores were consistent with guessing of the answer – that is, achieving a 50% score, as would be expected in the case of guessing on this 2-option test. These children were able to proceed through many trials by guessing quickly, and thus had high scores when the score was computed by simply adding together correct trials. As this did not seem to be an accurate measure of the child's performance, the decision was made to compute the PCPS score by subtracting the number of incorrect responses from the number of correct responses, such that a child guessing at random would be expected to have a score of 0. As an example, one child had 39 / 85 correct answers, which would be scored as 39 using the NIH Toolbox scoring guide. A different child had 38 / 39 correct answers, which would be scored as 38 using the NIH Toolbox guide. Thus, the first child would have had a higher score despite clearly inferior performance. In the scoring system chosen for this trial, the first child was given a score of 0, while the second received a score of 37. The score floor was 0.

<u>Version 3:</u> Post-hoc analysis was added to investigate whether degree of school attendance was an effect modifier in the 4 primary outcomes or the composite median ranking for PM-RUF vs. FP or PC-RUF vs. FP.

2. Introduction

2.1 Background and rationale

This study is an investigator-blinded, individually randomized, parallel group clinical trial aiming to determine whether peanut paste-based ready-to-use foods (RUFs) with either non-fat milk powder or cowpea would improve measures of fluid cognition from the NIH Cognitive Toolbox after a school year of feeding compared with a micronutrient-fortified millet porridge.

2.2 Objectives

The primary objective is to determine whether two peanut paste-based ready-to-use foods containing either added non-fat milk powder (peanut/milk ready-to-use food, PM-RUF) or cowpea (peanut/cowpea ready-to-use food, PC-RUF) given during the school day for a school year among 5-12 year old Ghanaian children improves cognition when compared with a micronutrient fortified millet porridge (FP) as measured by 4 tests from the NIH Toolbox: the dimensional change card sort test (DCCS), the flanker inhibitory control and attention test (FICA), the pattern comparison processing speed test (PCPS), and the list sorting working memory test (LSWM).

Secondary objectives are to determine whether PM-RUF or PC-RUF improves height-for-age z-score (HAZ), body mass index-for-age z-score (BAZ), fat-free mass, mid-upper arm circumference, a composite median ranking of the 4 cognitive tests, and sub-scores for speed and accuracy on DCCS and FICA compared with FP, and whether these effects might vary by age or gender. PM-RUF vs. PC-RUF comparisons are also to be done for all 4 primary outcomes and the composite median ranking.

A post-hoc analysis was done to determine whether degree of attendance (≥ 50%, ≥ 75%, and ≥ 90%) would modify the relative effect of PM-RUF or PC-RUF vs. FP.

3. Study Methods

3.1 Trial framework

The superiority framework is used for this trial. All comparisons will be presented on the basis of establishing superiority of PM-RUF or PC-RUF vs. FP.

3.2 Trial design

This is an investigator-blinded, 1:1:1 individually randomized, parallel group clinical trial designed to determine whether offer of school feeding with PM-RUF or PC-RUF vs. FP for a school year improves 4 NIH Toolbox measures of fluid cognition: DCCS, FICA, PCPS, or LSWM. Inclusion criteria were age 5-12 years and attendance at one of 6 participating schools within the first 6 weeks of school. Exclusion criteria were diagnosis of severe acute malnutrition, presence of a chronic debilitating illness, peanut or milk allergy, or caregiver intention to move out of the school district in the following year.

Following assessment of these criteria and deeming the participant eligible, participants were randomized when their parent or guardian selected a small opaque envelope enclosing a colored piece of paper from a larger opaque envelope. This larger envelope would contain 24 such identical small envelopes, 8 per study group, wherein one color corresponded to each study group. There was no stratification.

Participants then underwent baseline anthropometric and cognitive testing with the 4 primary outcomes tests. Parents or guardians completed demographic and household food insecurity queries. Following testing, the participants began to receive their study food each day they attended school. Receipt and attendance were documented daily.

Participants were not masked. A study coordinator responsible for study food production and delivery to participating schools was unmasked but did not take part in outcomes assessment or data analysis. Two volunteer school employees at each school were trained as feeding supervisors and were responsible for disbursement of study foods and tracking their receipt, and thus were not masked. All outcome assessors were masked. The allocation key linking colors to food groups was kept locked and inaccessible to outcome assessors and the investigator responsible for statistical analysis (KBS) until after analyses were completed.

The study continued until the end of the school year, at which point the 4 primary outcome cognitive tests and anthropometric measurements were repeated.

3.3 Follow up

Participants who remained in school and for whom parents or caregivers did not withdraw consent were followed for the entire school year.

3.4 Outcomes

3.4.1 Primary outcomes

Scores on the dimensional change card sort, flanker inhibitory control and attention, pattern comparison processing speed, and list sorting working memory tests at the end of the school year over which the trial proceeded.

1. DCCS: two target pictures are presented that vary along 2 dimensions: shape and color. The participant is asked to match a new picture (test picture) based on 1 of the 2 dimensions, and the test picture then appears; the participant then selects one of the two target pictures based on the test picture and dimension requested. The computed score (max of 10) is composed of accuracy and speed components (max of 5 points each). The accuracy score is calculated by scaling the proportion of correct answers to 5. For example, a participant who gets 32/40 trials correct would have an accuracy score of 4. The speed score is calculated automatically by the tablet using the NIH Toolbox software. Briefly, median reaction time values are computed using only correct trials with reaction times greater than 100 ms and no larger than 3 SDs from the participant's mean. A log base 10 transformation is applied and the scores are then re-scaled such that higher scores correspond to faster reaction times.

2. FICA: A series of fishes (trials 1-20) and then arrows (trials 21-40) are shown on a screen. The participant is asked to correctly identify the direction that the central fish/arrow is pointing regardless of which direction the flanking fishes/arrows are pointing. Sometimes, the flanking fishes/arrows are congruent in direction, and sometimes not. The computed score (max of 10) is composed of accuracy and speed components (max of 5 points each). The accuracy and speed components are calculated using the same methods as in DCCS.

3. PCPS: Two images are shown and the participant is asked to answer whether or not the two images are the same. Participants are given 85 seconds to answer as many questions correctly as possible. The score is obtained by subtracting the number of inaccurate responses from the number of accurate responses. The maximum score is 130.

4. LSWM: Pictures of animals and/or foods are displayed for 4 seconds and then withdrawn. The participant is asked to recall as many animals/foods shown as possible, ordered from smallest to largest size. There are two tests. In the first test, the participant is shown foods or animals and asked to recall along that single dimension. In the second test, both foods and animals are shown, and the participant is asked to recall them separately, still from smallest to largest. The NIH Toolbox tablet program version was not used because some of the foods and animals were not recognizable to participants on pilot testing. Thus, new, familiar

images of culturally identifiable foods and animals were loaded into the tablet and used, and the assessors manually tabulated participant score. The assessors manually tabulated the score because the subjects are answering verbally. The score was computed as the number of correct items recalled in order from smallest to largest across the 2 trials.

3.4.2 Secondary outcomes

1. Composite median ranking of 4 primary outcome scores. The composite ranking will be generated by ranking each participant's score within each test, calculating the median of each participant's ranks, and ranking these medians.

   This rank- and median-based composite was chosen because the primary outcomes are not interval-scaled and were unlikely to distribute normally. While the NIH Toolbox does auto-generate age-adjusted z-scores which can be combined, we *a priori* did not think comparison to US-derived normative scores would be valuable, in part because of the novelty of computer-based testing in the study cohort.
2. Change in height-for-age z-score (HAZ) from baseline to endline, calculated by subtracting the baseline HAZ from the endline HAZ. Z-scores will be computed using WHO Anthro 3.2.
3. Change in body mass index z-score (BAZ) from baseline to endline, calculated by subtracting the baseline BAZ from the endline BAZ.
4. Change in fat-free mass (FFM), as determined by bioelectrical impedance, calculated by subtracting baseline FFM from the endline FFM.
5. Change in mid-upper arm circumference (MUAC), calculated by subtracting baseline MUAC from endline MUAC.
6. Speed sub-scores on DCCS and FICA. The NIH Toolbox program automatically computes a speed score for these 2 tests, while PCPS and LSWM do not have such scores. Speed sub-scores are scaled to a maximum of 5.
7. Accuracy sub-scores on DCCS and FICA. This is proportion of prompts answered correctly, scaled to 5 as per NIH Toolbox protocol.
8. Sub-group analyses based on age (< 9 or ≥ 9 years) and sex (male or female) across the 4 primary outcomes and the composite median ranking
9. Rate of attendance over the final 10 weeks of trial, calculated as percent of days attended / total possible days.
10. Days enrolled, calculated by subtracting date of enrollment from date of final testing
11. Possible days of attendance / intervention receipt during trial (school days), from date of randomization to date of endline testing
12. Sensitivity analysis wherein DCCS and FICA accuracy and composite scores are computed using the NIH Toolbox scoring guideline for accuracy scoring and compared to results of the method used in this protocol

3.4.3 Post-hoc outcomes

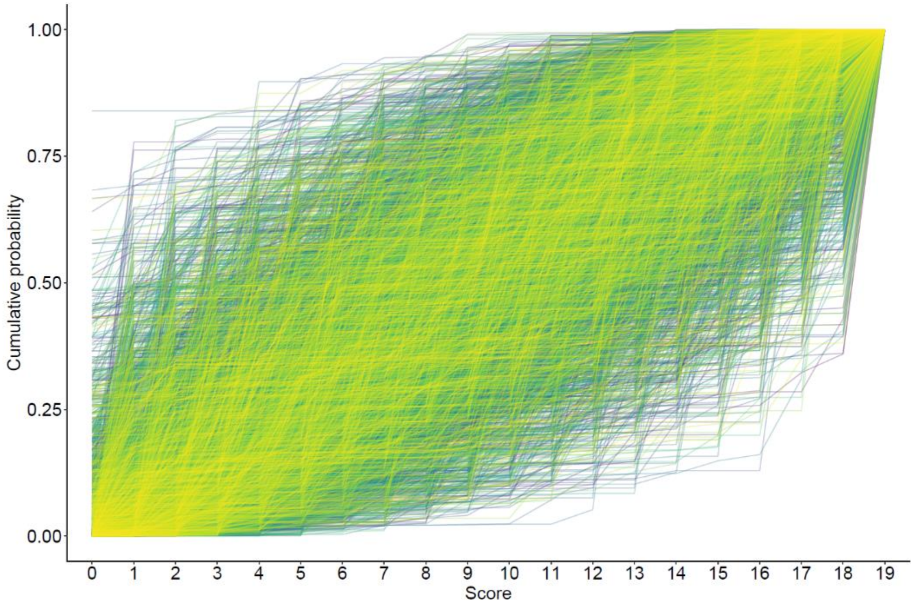
Per-protocol analysis based on attendance rates. DCCS, FICA, PCPS, LSWM, and composite median ranking compared between PM-RUF or PC-RUF vs. FP among those with attendance rates greater than 50%, 75%, and 90%.

3.4.4 Timing of outcome assessments

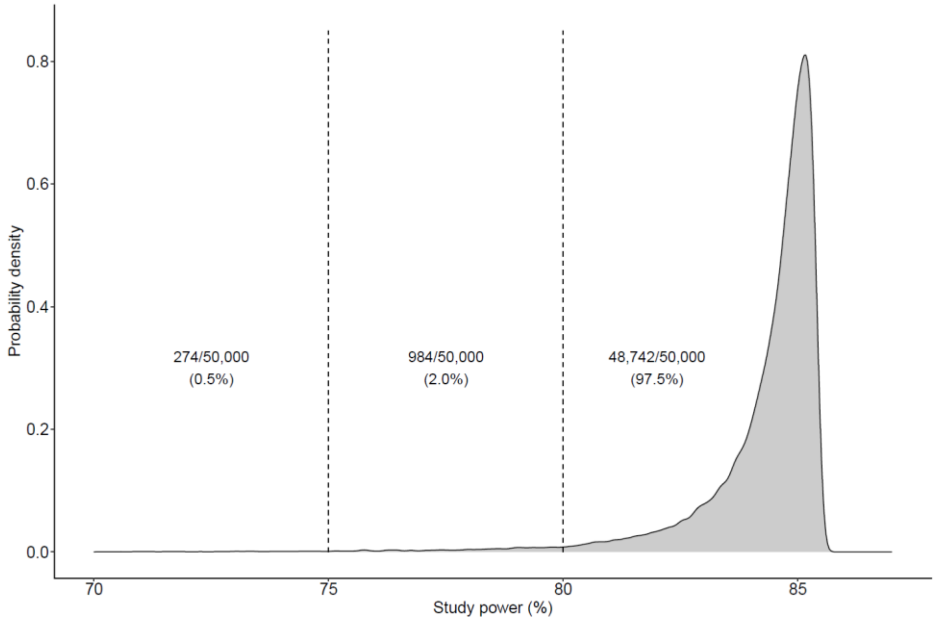
All primary, secondary, and post-hoc outcome comparisons will be done using outcomes data collected at the end of the school year; that is, after the year of school feeding offered. Cognitive and anthropometric testing will be done over the final 3 weeks of the school year. Attendance over the final 10 weeks of the trial will be compared between groups.

3.5 Sample size

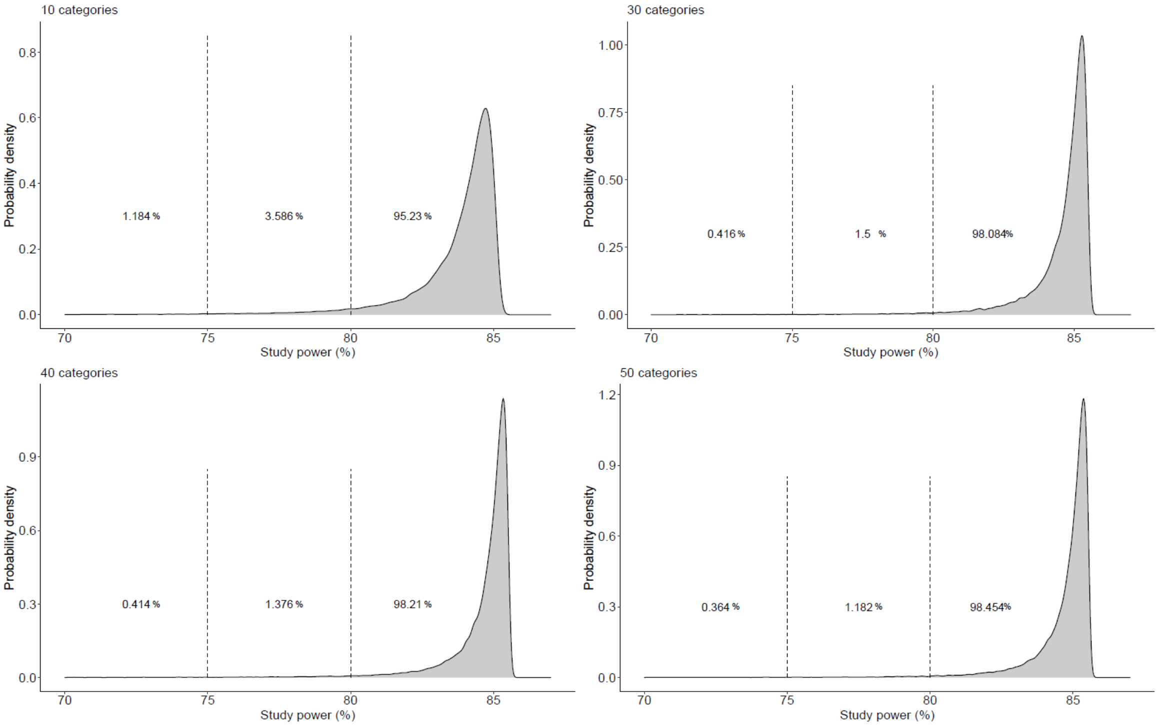
The study was planned to enroll 880 participants, anticipating that 15% would drop out during the school year, which would have provided >80% power to detect an odds ratio (OR) for a higher score of 1.6 at a two-sided α of 0.05. This calculation was based on the use of ordinal logistic regression applied to simulated datasets because (1) data were not available to estimate the primary outcome distributions in rural Ghanaian children, (2) the scores consisted of ordered numbers that were not interval-scaled, and (3) to allow for covariate adjustment and thereby increase power.[1] It was conservatively assumed that there would be at least 20 unique scores within each test (alternative choices were also tested, as shown at the end of this section). Using the Dirichlet distribution and a random vector generator, 50,000 distributions of scores were simulated that varied in degrees of skewness and kurtosis.[2] 10,000 such distributions are shown below in a cumulative probability plot, demonstrating the wide variety of degrees of skewness and kurtosis simulated. Each simulated distribution is shown in a slightly different color to improve visibility.



Knowing that potential between-group differences would vary based on the data's underlying distribution, median scores in the control and intervention groups were simulated using the same 50,000 simulated score distributions at a variety of ORs. An OR = 1.6 yielded a median of median differences of 0.75 (IQR: 0.6, 0.9) along a 10-point scale, as in DCCS and FICA, corresponding to a 7.5% (IQR: 6%, 9%) relative improvement in scores. This was chosen as a modest target difference to detect. Using an OR = 1.6, 250 participants per group yielded power ≥ 80% in 97.5% of 50,000 simulations, and <75% in only 0.5%, and thus was considered sufficient.[3,4] The results of this power analysis are displayed visually below in a probability density plot.



Finally, alternative numbers of ordered outcome categories (10, 30, 40, 50) were tested to determine their effect on study power at OR = 1.6 across the 50,000 simulated baseline distributions. As expected, fewer categories slightly reduced study power, but >95% of simulations had >80% even at 10 categories.


10 categories
Probability density
0.0
0.2
0.4
0.6
0.8
1.184 %
3.586 %
95.23 %
70
75
80
85
Study power (%)
30 categories
Probability density
0.00
0.25
0.50
0.75
1.00
0.416 %
1.5 %
98.084%
70
75
80
85
Study power (%)
40 categories
Probability density
0.0
0.3
0.6
0.9
0.414 %
1.376 %
98.21 %
70
75
80
85
Study power (%)
50 categories
Probability density
0.0
0.3
0.6
0.9
1.2
0.364 %
1.182 %
98.454%
70
75
80
85
Study power (%)

4 Analysis time points

4.1 Significance level

α = 0.05 and 95% confidence intervals (CIs) will be applied in the interpretation of statistical significance tests. All reported *P* values will be two-sided. *P* values will be computed for primary outcomes and sub-group interaction effects; otherwise, 95% CIs with point estimates will be provided.

4.2 Correction for multiple comparisons

We will not correct for multiple comparisons, accepting an increased risk for type I error to avoid inflating the type II error in this trial of a low-risk intervention with 4 outcome measures that are expected to be highly correlated, which increases the chance of type II error with multiple-test corrections.[5,6] In the place of *P* value corrections, 95% CIs will be reported throughout, and a composite outcome metric chosen to rely on the fewest assumptions possible (median of ranks) was chosen to compare food groups.[7]

4.3 Timing of final analysis

Once all endline cognitive tests and anthropometric measurements are completed, all primary, secondary, and post-hoc analyses will be done.

4.4 Analysis sets

All primary and secondary outcomes will be analyzed in a modified intention-to-treat (mITT) population. The single modification (*contra* full intention-to-treat) is that participants must complete endline cognitive testing to be included in outcomes analysis, as the results of such testing are the study's primary outcomes. Thus, those who dropped out of school and did not return for endline testing are excluded from primary and secondary analyses, by necessity.

A post-hoc, per-protocol-type analysis wherein the 4 primary outcomes and the composite median ranking are compared across PM-RUF vs. FP and PC-RUF vs. FP groups will be undertaken based on degree of attendance, which will be a metric of adherence, as spot visits revealed that children reliably consumed all study food if they were present in school.

4.5 Study days and duration

The date of enrollment will be a participant's day 1, and the date of final testing will be the participant's final study day.

4.6 Participant flow chart

A CONSORT flow diagram will be presented displaying number of participants screened for eligibility, excluded prior to randomization, randomized to each study group, excluded due to dropping out during the school year (not returning for endline testing), and number who completed endline cognitive testing.

4.7 Baseline characteristics

Baseline characteristics will be presented as number (%), mean ± SD, or median (IQR).

- Age
- Sex
- Mother's vital status
- Household's water source
- Household food insecurity score
- Household ownership of computer, radio, motorbike
- School attended during trial
- Height, cm
- Weight, kg
- Height-for-age z-score
- Body mass index
- Body mass index z-score
- Fat free body mass, kg

5 Analysis methods

5.1 General principles

The analysis and results reporting will follow CONSORT guidelines. Data for the primary and secondary outcomes will be analyzed using a modified intention-to-treat population. The modification from intention-to-treat is that only those with endline cognitive test results are included in primary and secondary analyses, which is a necessary modification.

Interpretation of the primary and secondary endpoints will be based on *P* values and 95% CIs. *P* values will be computed for the primary outcome comparisons, secondary outcome composite median ranking, and interaction terms in sub-group analyses. All *P* values will be two-tailed, and a 5% significance level will be used for all outcomes. No corrections for multiple comparisons will be used.

5.2 Modified intention-to-treat population

The mITT population consists of all randomized participants excepting those who dropped out of school prior to completion of endline cognitive testing. Participants will be analyzed according to their randomized group regardless of degree of attendance/receipt of study food (e.g., a child who is randomized, misses school frequently throughout the year, and then attends and undergoes endline cognitive testing, will be included in their randomized group for analysis).

5.3 Descriptive statistics

Categorical variables will be summarized as number (%), while continuous or multiple-ordered variables will be summarized as mean ± SD if their distribution approximates normality or median (IQR) if it is skewed or otherwise diverges from a normal distribution. Cognitive test results will be summarized using both mean ± SD and median (IQR) to provide a fuller picture of their distributions. Based on prior experience, it is expected that change in anthropometrics HAZ, BMI, MUAC, and FFM will be normally distributed and thus likely will be presented as mean ± SD.

5.4 Proportional odds assumption

The proportional odds assumption will be assessed using several methods. First, coefficients for study group will be compared using logistic regression across 10 outcome cut points that span possible scores for each test. For example, for DCCS, we will create 10 logistic regressions at DCCS ≥ 1, ≥2, ≥3, and so on. We will create plots with logit(outcome) along the x axis and food group plotted along the y axis. If the PO assumption holds, we would expect to constancy in the relationship between food groups at various cut-offs, e.g., the relationship between PM-RUF and FP coefficients will be the same at logit(DCCS score ≥ 7) and logit(DCCS score ≥ 8). Second, we will plot the logit of the cumulative distribution function of each outcome stratified by food group and evaluate for parallelism of distributions between groups. Third, we will use the plot.xmean.ordinaly function in the package “rms” to plot outcome-stratified means of predictor variables overlaid with $\hat{E}(X|Y = j)$ with *j* on the x-axis.[8]

If it appears that the proportional odds assumption may be violated, testing will be undertaken to assess the possible effects of disproportional odds on statistical inference, and the use of a partial proportional odds model may be considered.[9]

5.5 Analysis of primary outcomes

DCCS and FICA computed scores, PCPS scores, and LSWM scores will be compared between PM-RUF and FP and between PC-RUF and FP using ordinal logistic regression. The dependent variable will be endline cognitive test score. Independent variables will be randomized food group assignment and the 2 prespecified covariates: participant age and baseline cognitive test score. For example, the regression for DCCS will be:

$$\text{DCCS}_{\text{endline}} = \text{Food group} + \text{Baseline participant age} + \text{DCCS}_{\text{baseline}}$$

The fortified porridge group serves as the referent group for primary analyses. Because higher scores on cognitive tests indicate better performance, an odds ratio > 1.0 indicates more favorable results for the PM-RUF or PC-RUF group compared with FP. Results of the primary analysis of the primary outcomes will be presented as an adjusted common odds ratio and associated 95% CIs.

Cognitive scores will be visualized using ridgeline plots composed of kernel density-smoothed histograms with medians indicated.

5.6 Analysis of secondary outcomes

1. Changes in HAZ, BMI-Z, MUAC, and FFM will be compared PM-RUF vs. FP and PC-RUF vs. FP using linear regression, after confirming normality of residuals and homoscedasticity.
2. Composite median rankings will be compared PM-RUF vs. FP and PC-RUF vs. FP using ordinal logistic regression, with covariates of baseline participant age and baseline composite median ranking.
3. Speed sub-scores for DCCS and FICA will be compared PM-RUF vs. FP and PC-RUF vs. FP using ordinal logistic regression, with covariates of baseline participant age and baseline speed sub-score.
4. Accuracy sub-scores for DCCS and FICA will be compared PM-RUF vs. FP and PC-RUF vs. FP using ordinal logistic regression, with covariates of baseline participant age and baseline accuracy sub-score.
5. We will assess for possible heterogeneity in effect of PM-RUF or PC-RUF vs. FP on the 4 cognitive primary outcomes and the composite median ranking by pre-

specified sub-groups based on age (< 9 or ≥ 9 years) and sex (male or female). First, ordinal logistic regression will be applied to participants filtered by the sub-group variable in question, covariate-adjusted for baseline participant age and baseline cognitive test score, to generate 95% CIs for food group effect within each sub-group. Second, an interaction will be introduced between study group and the possible effect modifier, age (as a continuous variable) or sex. For instance, for DCCS with assessment for sex effect modification, the regression would be:

$$\text{DCCS}_{\text{endline}} = \text{Food group} * \text{Sex} + \text{Baseline participant age} + \text{DCCS}_{\text{baseline}}.$$

Presence of absence of effect modification will be assessed qualitatively by visualization of the OR / 95% CIs between the sub-groups, and quantitatively by evaluation of the *P* value for the interaction term, with < 0.05 indicating significant effect modification. It should be noted that the study was not powered to assess for interaction effects reliably.

6. Rate of attendance over the final 10 weeks of trial will be compared across the 3 groups using the Wilcoxon rank sum test. A continuity correction will be used to produce 95% CIs around the median of differences between groups.
7. DCCS and FICA accuracy and computed scores, as calculated by the NIH Toolbox tablet-based scoring guideline, will be compared between PM-RUF vs. FP and PC-RUF vs. FP using ordinal logistic regression, with covariates baseline cognitive score and participant age. ORs with 95% CIs will be shown.

5.7 Post-hoc analyses

The 4 primary outcomes and the composite median ranking will be compared PM-RUF vs. FP and PC-RUF vs. FP according to degree of school attendance over the study's final 10 weeks, as a type of per-protocol analysis. Attendance cut-offs will be ≥ 50%, ≥ 75%, and ≥ 90%. Attendance rate will be introduced as an interaction term with study group into ordinal logistic regressions to assess for effect modification, with covariates baseline participant age and baseline score.

5.8 Missing data

No missing data will be imputed.

5.9 Adverse evets

Rates of rash and gastrointestinal intolerance (gas, bloating, diarrhea, nausea, vomiting, poor oral intake) will be tracked periodically throughout the study. The number of episodes per participant will be compared across groups using Wilcoxon rank sum test.

5.10 Statistical software

Statistical analyses will be carried out using R version 4.2.1 (R Foundation for Statistical Computing).[10]

6 References

1. Lingsma H, Roozenbeek B, Steyerberg E, investigators I. Covariate adjustment increases statistical power in randomized controlled trials. *J Clin Epidemiol*. Dec 2010;63(12):1391; author reply 1392-3. doi:10.1016/j.jclinepi.2010.05.003
2. *simstudy: Illuminating research methods through data generation*. The Journal of Open Source Software; 2020. https://doi.org/10.21105/joss.02763
3. *Hmisc: Harrell Miscellaneous*. Version R package version 4.7-1. 2022. https://CRAN.R-project.org/package=rms
4. Whitehead J. Sample size calculations for ordered categorical data. *Stat Med*. Dec 30 1993;12(24):2257-71. doi:10.1002/sim.4780122404
5. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. Jan 1990;1(1):43-6.
6. Feise RJ. Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology*. 2002-12-01 2002;2(1)doi:10.1186/1471-2288-2-8
7. Althouse AD. Adjust for Multiple Comparisons? It’s Not That Simple. *The Annals of Thoracic Surgery*. 2016-05-01 2016;101(5):1644-1645. doi:10.1016/j.athoracsur.2015.11.024
8. *rms: Regression Modeling Strategies*. Version R Package version 6.3-0. 2022. https://CRAN.R-project.org/package=rms
9. Peterson B, Harrell, Frank. Partial Proportional Odds Models for Ordinal Response Variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 1990;39(2):205-217.
10. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* 2022. https://www.R-project.org/