Supplementary Material

A Community-Based, Multi-Level, Multi-Setting, Multi-Component Intervention to Reduce Weight Gain among Low Socioeconomic Status Latinx Children with Overweight or Obesity: The Stanford GOALS Randomized Controlled Trial

Thomas N. Robinson, MD, MPH, Donna Matheson, PhD, Darrell M. Wilson, MD, Dana L. Weintraub, MD, Jorge A. Banda, PhD, Arianna McClain, PhD, Lee M. Sanders, MD, MPH, William L. Haskell, PhD, K. Farish Haydel, BA, Kristopher I. Kapphahn, MS, Charlotte Pratt, PhD, MS, RD, Kimberly P. Truesdale, PhD, June Stevens, PhD, Manisha Desai, PhD

Trial Protocol

For further details about the design and analysis of the study please refer to the protocol.[1] The trial protocol was published as:

Robinson TN, Matheson D, Desai M, Wilson DM, Weintraub DL, Haskell WL, McClain A, McClure S, Banda J, Sanders LM, Haydel KF, Killen JD. Family, community and clinic collaboration to treat overweight and obese children: Stanford GOALS – a randomized controlled trial of a three-year, multi-component, multi-level, multi-setting intervention. ***Contemporary Clinical Trials***, 2013;36:421-435.

Recruitment, Randomization and Follow-up Timeline

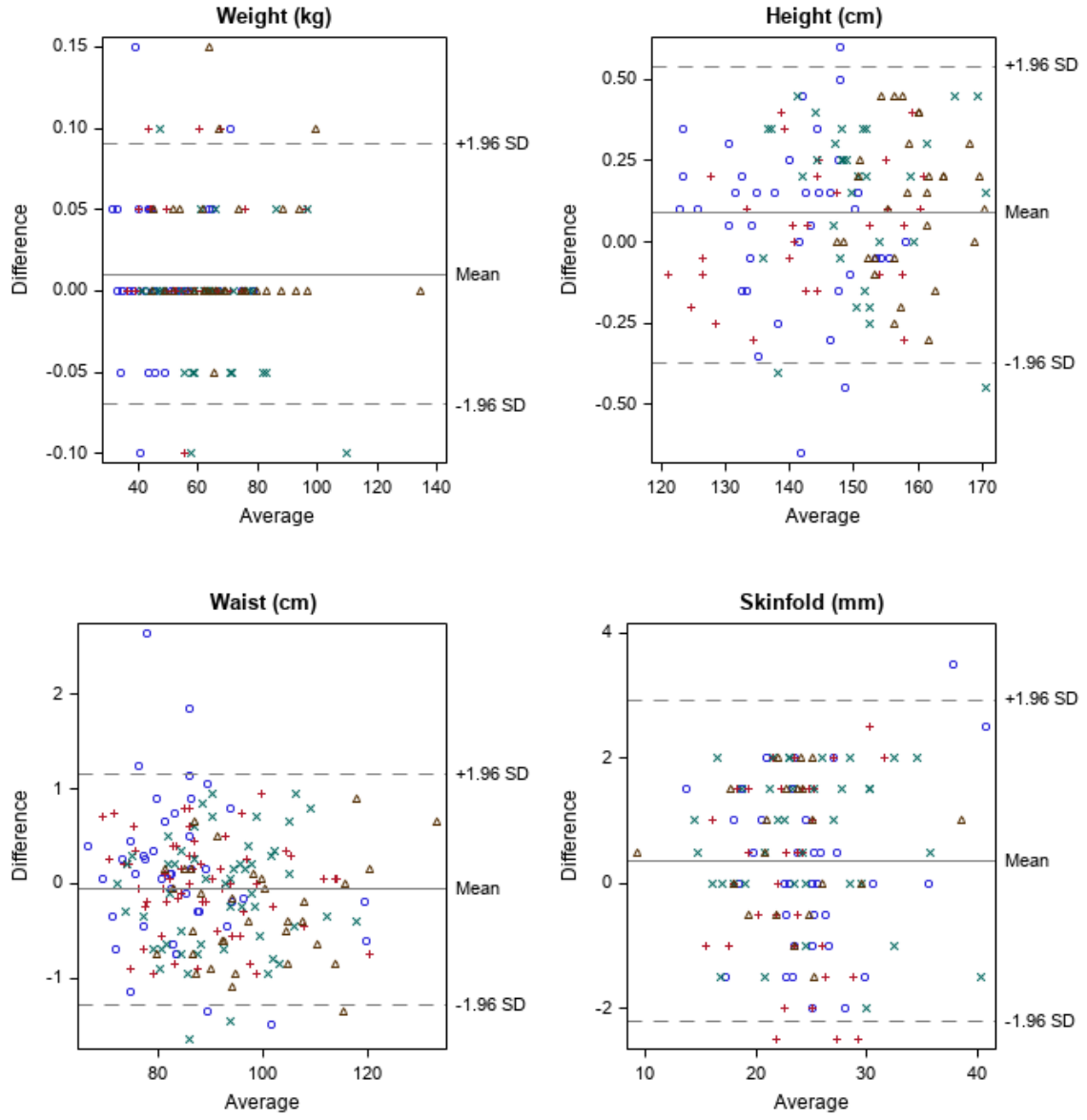
Recruitment start: July 13, 2012
First randomization: September 4, 2012
Last randomization: October 3, 2013
Final assessment of primary outcome: December 19, 2016
Completed data cleaning: January 18, 2018
Completed data analysis: September 11, 2020

Randomization

Children were randomized to treatment or control conditions after completing all baseline measures. All eligible children within a household were assessed for inclusion in the study. Prior to randomization, the database manager (who has no contact with participants) confirmed all baseline data were complete and verified that date of birth and sex match for each child, BMI was coded correctly, and randomization was occurring within 30 days of the BMI measurement.

For households that contributed multiple eligible children, one index child was randomly selected for randomization and inclusion in the analysis. Only the database manager and statistician were aware of which child in a multi-child household was in the analysis sample and used for the stratified randomization. A customized SAS program performed the randomization and verified all requirements were met (SAS 9.3). Strata for randomization were defined by the index child's baseline BMI percentile (BMI $\geq 85^{th}$ percentile and $< 95^{th}$ percentile for age and sex and BMI $\geq 95^{th}$ percentile for age and sex). Efron's biased coin randomization[2] was used to promote a balanced randomization within strata. When an imbalance occurs, the under-represented group is assigned with an allocation probability of 2/3, otherwise the allocation is made with an equal probability.

Figure S1. Bland-Altman plots of inter-rater agreement for measures of weight, height, waist circumference and triceps skinfold thickness.



Bias (mean difference of measurement - QC) is shown as a solid black line. Limits of Agreement (LOA) (mean bias ± 1.96 SD) are shown as dotted black lines. Bias (LOA range) for measures as follows: Weight: 0.01 kg (-0.07 to 0.09), height: 0.09 cm (-0.37 to 0.54), waist circumference: -0.06 cm (-1.28 to 1.16), triceps skinfold thickness: 0.40 mm (-2.20 to 2.93).

Inter-rater agreement by Spearman ICCs: Weight Spearman ICC = .99, Height Spearman ICC = .99, Waist circumference Spearman ICC = .99, Triceps Skinfold Thickness Spearman ICC = .95

Multiple Imputation

The BMI slopes for the three participants with no post-baseline measurements were imputed in SAS (version 9.4) using PROC MI. The imputation model used was a function of a broad selection of available auxiliary variables. The MCMC option was specified, which tells SAS to treat all included variables as members of a joint multivariate normal distribution. This approach assumes missingness is no more severe than MAR. The number of auxiliary variables included was varied and regardless of which set of auxiliary variables were used results remained consistent in direction, magnitude, and significance. The number of imputed data sets was varied in increments of 5 from 5 to 25 and did not have appreciable effects on model estimates. Five imputed data sets were used in the analysis.[3]

Detectable Difference, Sample Size, and Power

Note: an abbreviated version of this discussion occurs in the design and protocol paper.[1]

Assumptions Regarding Effect Sizes and Standard Errors
The statistical literature recommends against the use of pilot studies to estimate effect sizes for clinical trials.[4] Instead, estimated sample size requirements should be based on the *a priori* minimum acceptable difference between groups to be considered of clinical or public health significance, from the experience and judgment of the investigators.[4-7] In this case, the effects of the MMM intervention compared to the enhanced standard Health Education control condition. Based on our judgment and experience, we estimate this minimum acceptable difference to be an effect size (Cohen's d) = 0.4. This is the equivalent of about 27% non-overlap of two normal distributions, or 50% of one group's distribution being greater than about 66% of the other group's distribution,[8] a Number Needed to Treat for one additional success (NNT) of 4.49, a Standardized Risk Difference (SRD) of .223, and an Area Under the ROC Curve (AUC) of .611.[5,6]

We can also use the changes observed in our past studies to better estimate the effects we expect to achieve in the proposed trial. The 12-week Dance for Health intervention resulted in a Cohen's d effect size = 0.43,[9] the 7-month school-based screen time reduction intervention resulted in a Cohen's d = 0.67,[10] and the 12-week Stanford GEMS Phase 1 Pilot Study resulted in a Cohen's d = 0.42.[11] Therefore, achieving an effect size of 0.4 or greater is realistic, and the MMM intervention strategy, with an increased intervention length and intensity compared to past studies, and starting with an overweight sample, may be expected to result in an even greater effect size.

To further aid in interpretation of meaningful effect sizes, we include clinically relevant scenarios that correspond to a Cohen's d statistic of 0.4. This statistic is a function of the difference in average slope for each group and the corresponding pooled standard deviation. Examples of clinically relevant scenarios that corresponds to a Cohen's d statistic of 0.4 include the following:

- Average decrease in treated children is 0.1 BMI units per year while controls increase at a rate of 0.4 BMI units per year with a standard deviation of 1.2.
- Treated children decrease by almost half a BMI unit per year (0.4) while controls have no change in BMI per year with a standard deviation of 1.0.
- Both groups increase in BMI each year where treated children increase by 0.2 BMI units per year and controls increase by 0.6 BMI units per year with a standard deviation of 0.9.

Power Calculations. For a two-tailed 5% alpha level test, the planned sample size of 120 children per group would provide approximately 90% power to detect intervention effects of that magnitude or greater.[7,8] Based on simulation studies (1000 simulations per scenario) we have assessed power for detecting meaningful treatment effects in the presence of an interaction between treatment and baseline BMI. The table below presents our simulation study and demonstrates that we have excellent power for detecting clinically relevant differences between treatment arms. Previous studies investigating rate of change in BMI give standard deviation estimates ranging from 0.8 to 1. We consider a wider and more conservative range of estimates from 0.9 to 1.8. For example, in Scenario 1 where children in the intervention group do not increase their BMI, while controls increase by 1.3 BMI units on average,

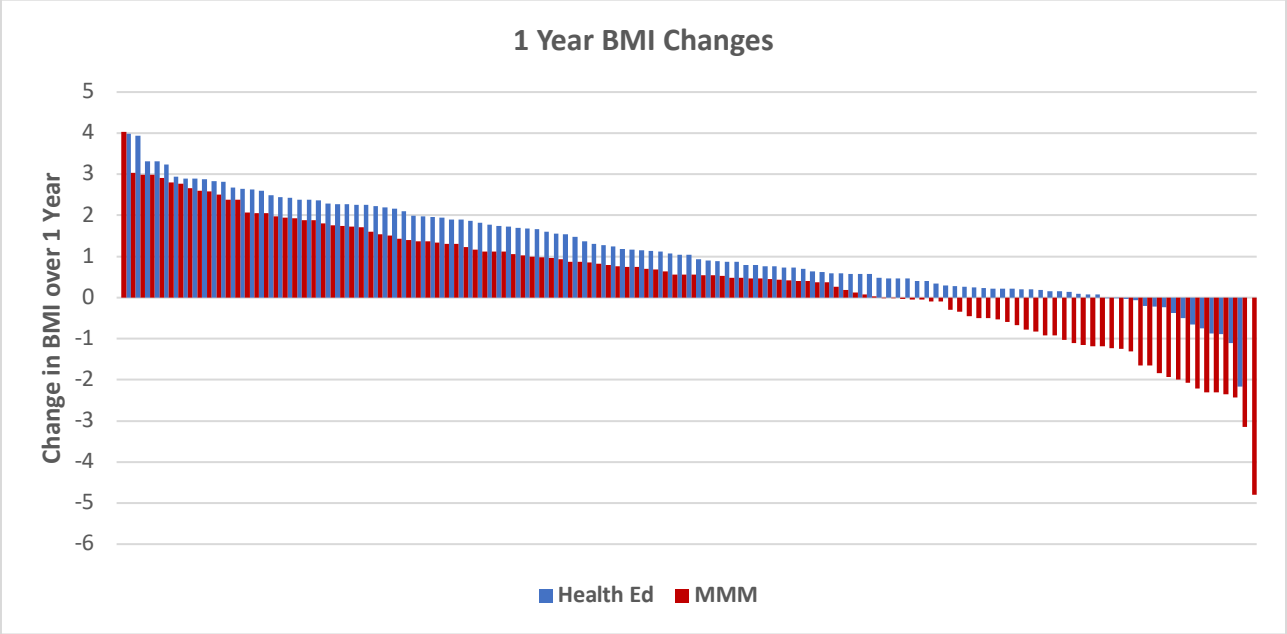
we have more than 95% power to detect an overall treatment effect. Scenarios 10 and 11 demonstrate we have sufficient power (94% and 89%) to detect a treatment effect if children in the intervention group have no change in BMI on average and children in the control group increase their BMI by about a half unit per year. Finally, we have 83% power to detect a main treatment effect if both arms increase in BMI with the treatment group increasing at an attenuated rate relative to controls (0.95 BMI units per year versus 1.3 units on average) (Scenario 12).

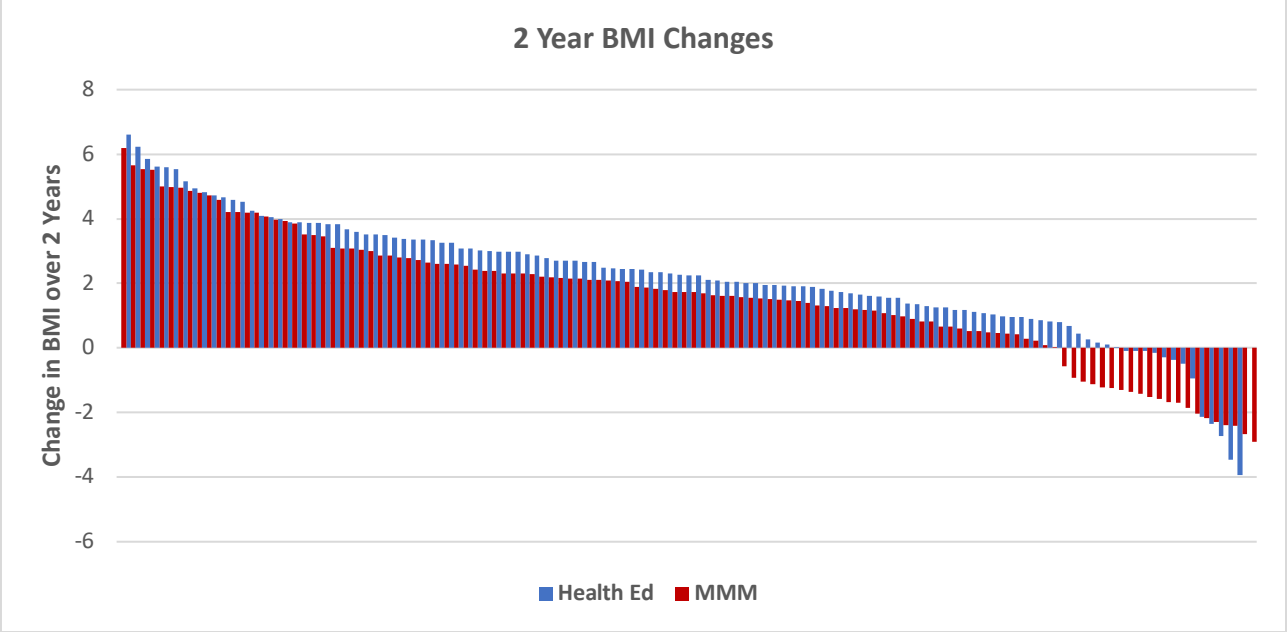
Power simulation scenarios

	Average Slope (SD) by Group					
Scenario	Treated OW	Treated Obese	Control OW	Control Obese	Effect Size (Cohen's d)	Power
1	0.0 (1.8)	0.0 (1.8)	1.2 (1.8)	1.4 (1.8)	0.7	>95%
2	0.5 (1.8)	-0.5 (1.8)	1.2 (1.8)	1.4 (1.8)	0.7	>95%
3	0.5 (1.8)	-0.9 (1.8)	1.0 (1.8)	1.0 (1.8)	0.6	>95%
4	0 (0.9)	-0.9 (0.9)	0 (0.9)	0 (0.9)	0.5	>95%
5	0.2 (1.0)	-1.0 (1.0)	0 (1.0)	0 (1.0)	0.4	87%
6	0.3 (1.2)	-0.5 (1.2)	0.4 (1.2)	0.4 (1.2)	0.4	90%
7	0.5 (0.9)	-0.2 (0.9)	0.5 (0.9)	0.5 (0.9)	0.4	86%
8	0.5 (0.9)	-0.1 (0.9)	0.5 (0.9)	0.6 (0.9)	0.4	87%
9	-0.14 (0.9)	-0.6 (0.9)	-0.1 (0.9)	0.1 (0.9)	0.4	89%
10	0.1 (0.9)	-0.1 (0.9)	0.4 (0.9)	0.4 (0.9)	0.4	94%
11	0.1 (1.2)	-0.1 (1.2)	0.4 (1.2)	0.6 (1.2)	0.4	89%
12	1.1 (0.9)	0.8 (0.9)	1.2 (0.9)	1.4 (0.9)	0.4	83%

Figure S2. Overlapping distributions of individual BMI changes in the MMM and Health Education intervention groups over 1, 2, and 3 years.

Overlapping waterfall plots showing each individual participant's change in BMI over 1, 2, and 3 years ordered within intervention group (MMM = red bars, Health Education = blue bars) from largest increase in BMI on the left to largest decrease in BMI on the right. Each bar represents an individual participant.





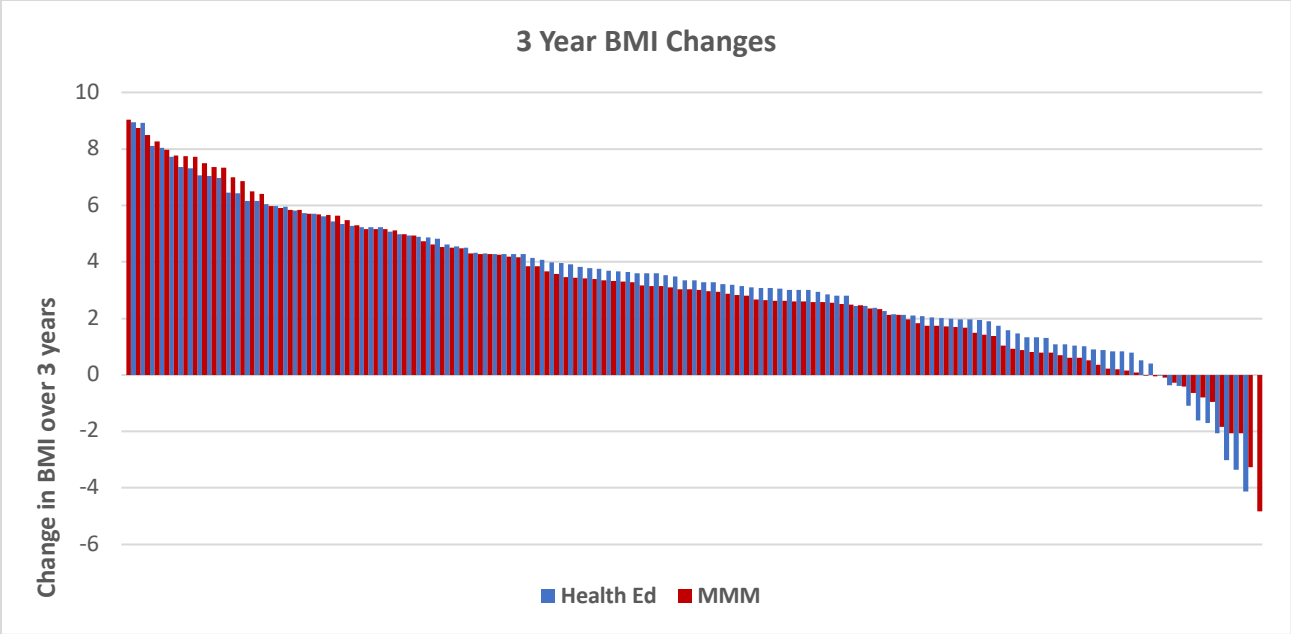

3 Year BMI Changes
Change in BMI over 3 years
10
8
6
4
2
0
-2
-4
-6
Health Ed
MMM

Figure S3: Plot of linear BMI trajectories for HE and MMM groups calculated from the linear model.

This figure illustrates linear BMI trajectories produced by the primary model. Both groups have increasing BMI, and MMM has a slightly smaller slope.

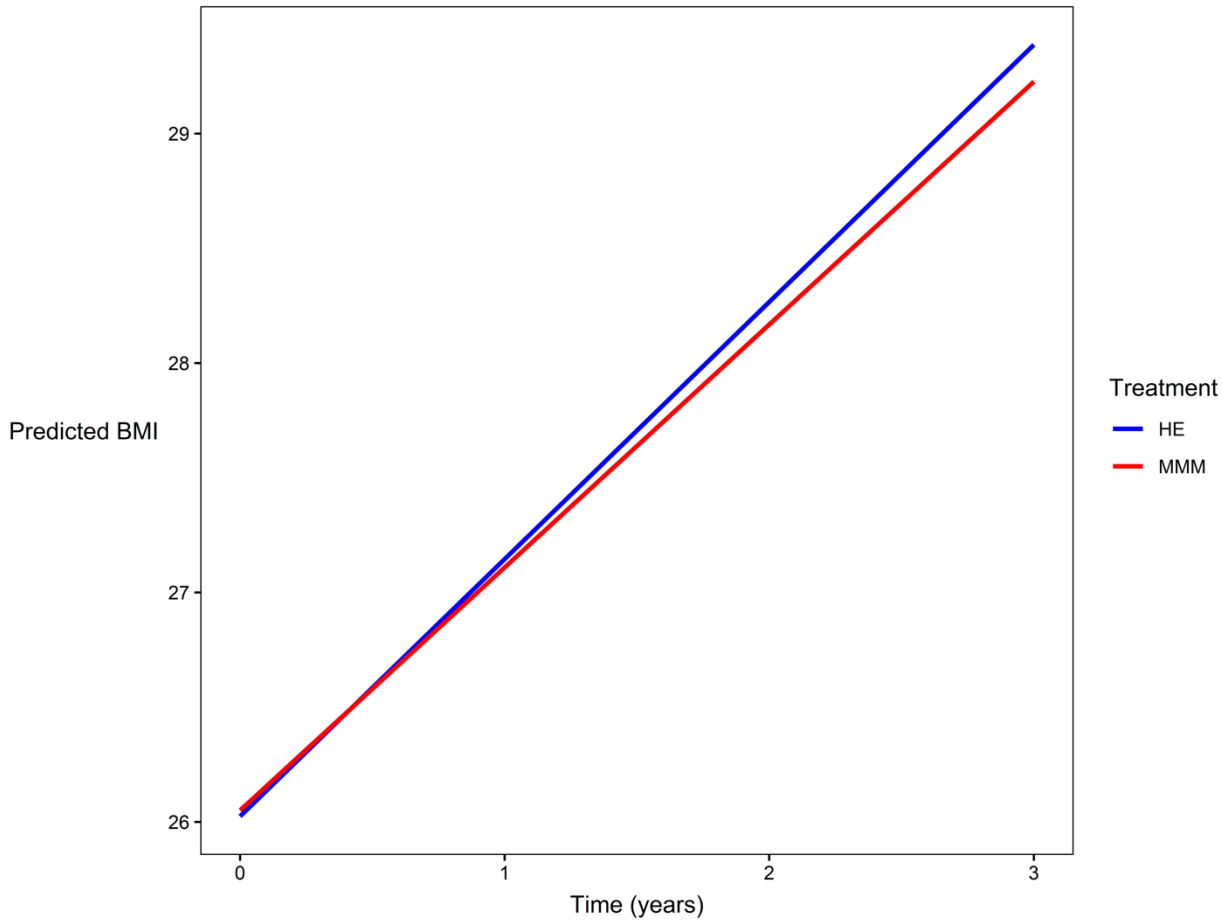


Figure S4. Forest plot of adjusted MMM-HE differences in BMI changes and 95% Confidence Intervals within baseline randomization strata of BMI (obesity and overweight) over 1, 2 and 3 years.

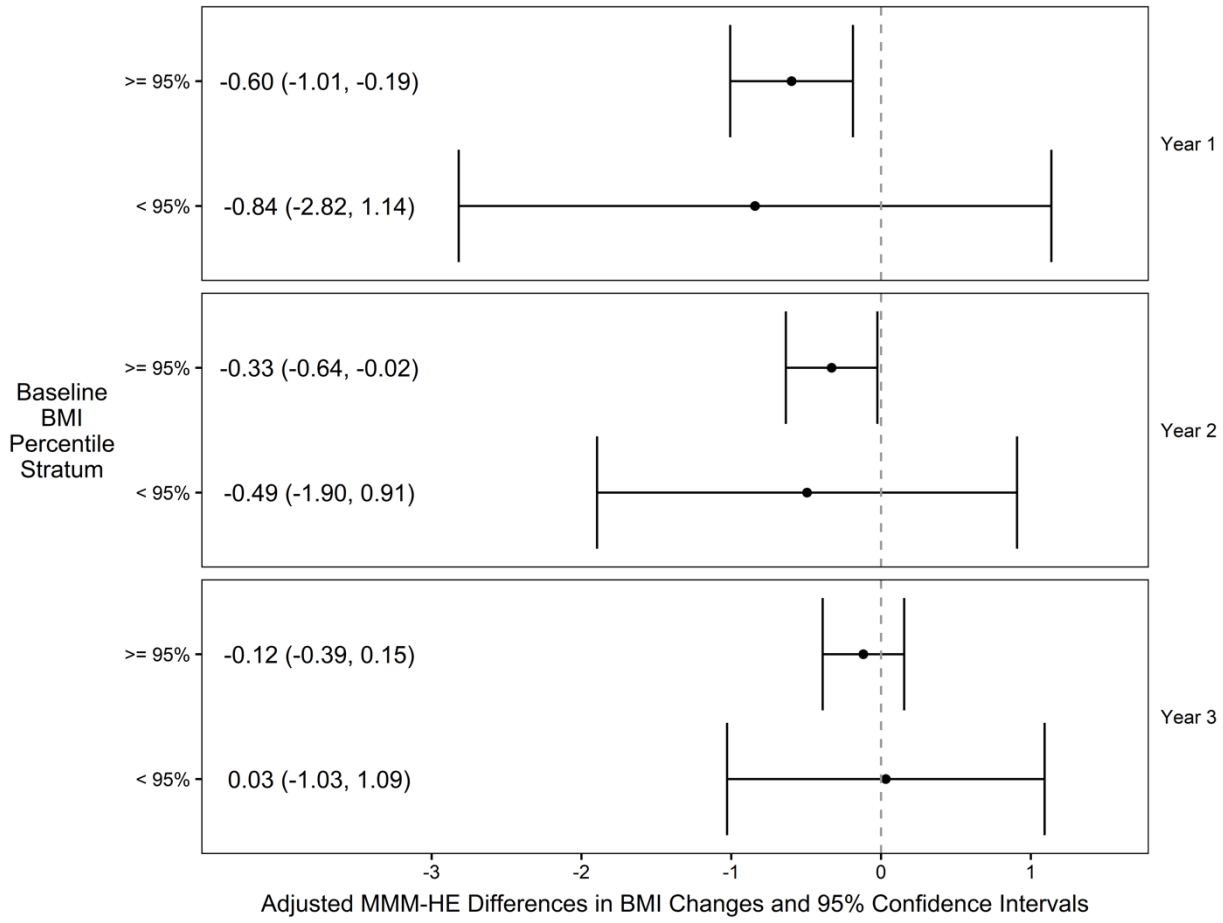


Table S1. Baseline prevalences and changes in categorical clinical classifications and relative differences between groups (relative risk) and number needed to treat (NNT).

	Baseline[a]		1 year[b]				2 years[c]				3 years[d]			
	N (%)		N (%)		MMM to HE Relative Risk over 1 year (95% CI)	NNT[e]	N (%)		MMM to HE Relative Risk over 2 years (95% CI)	NNT[e]	N (%)		MMM to HE Relative Risk over 3 years (95% CI)	NNT[e]
	MMM	HE	MMM	HE			MMM	HE			MMM	HE		
Prevalence of Obesity, BMI ≥ 95th Percentile	91 (75.8%)	92 (76.0%)	83 (69.2%)	90 (76.3%)	0.91 (0.8, 1.1)	14	79 (67.5%)	81 (69.8%)	0.97 (0.8, 1.2)	43	75 (66.4%)	74 (64.9%)	1.02 (0.8, 1.2)	69
Remission of Obesity[f]	NA	NA	10 (11.0%)	5 (5.6%)	1.96 (0.7, 5.5)	19	12 (13.6%)	12 (13.6%)	1.00 (0.5, 2.1)	NA	13 (15.5%)	17 (19.8%)	0.78 (0.4, 1.5)	23
Prevalence of Severe Obesity, ≥120% of the 95th Percentile BMI	46 (38.3%)	32 (26.4%)	35 (29.2%)	36 (30.5%)	0.96 (0.6, 1.4)	75	34 (29.1%)	37 (31.9%)	0.91 (0.6, 1.3)	35	40 (35.4%)	39 (34.2%)	1.03 (0.7, 1.5)	84
Remission of Severe Obesity[g]	NA	NA	12 (26.1%)	2 (6.5%)	4.04 (1.0, 16.8)	5	14 (31.1%)	3 (10.0%)	3.11 (1.0, 9.9)	5	10 (23.3%)	4 (13.3%)	1.74 (0.6, 5.0)	10
Systolic BP ≥ 90th percentile for age, sex & height	7 (5.8%)	5 (4.1%)	4 (3.3%)	5 (4.3%)	0.78 (0.2, 2.8)	106	4 (3.7%)	5 (4.5%)	0.81 (0.2, 2.9)	114	5 (4.6%)	4 (3.6%)	1.26 (0.3, 4.6)	105
Diastolic BP ≥ 90th percentile for age, sex & height	0	0	0	0	NA	NA	0	0	NA	NA	1 (0.9%)	0	NA	NA
Fasting Total Cholesterol ≥ 170 mg/dL	36 (30.0%)	33 (27.3%)	26 (22.8%)	33 (28.4%)	0.80 (0.5, 1.3)	18	28 (26.2%)	27 (25.0%)	1.05 (0.7, 1.7)	86	24 (23.1%)	23 (20.9%)	1.10 (0.7, 1.8)	46
Fasting LDL-Cholesterol ≥ 110 mg/dL	25 (20.8%)	25 (20.7%	13 (11.4%)	24 (20.7%)	0.55 (0.3, 1.0)	11	17 (15.9%)	16 (14.8%)	1.07 (0.6, 2.0)	93	13 (12.5%)	16 (14.5%)	0.86 (0.4, 1.7)	49
Fasting HDL-Cholesterol < 40 mg/dL	35 (29.2%)	35 (28.9%)	41 (36.0%)	41 (35.3%)	1.02 (0.7, 1.4)	161	37 (34.6%)	45 (41.7%)	0.83 (0.6, 1.2)	14	38 (36.5%)	38 (34.5%)	1.06 (0.7, 1.5)	50
Fasting Triglycerides ≥ 100 mg/dL	50 (41.7%)	45 (37.2%)	48 (42.1%)	47 (40.5%)	1.04 (0.8, 1.4)	63	54 (50.5%)	45 (41.7%)	1.21 (0.9, 1.6)	11	50 (48.1%)	52 (47.3%)	1.02 (0.8, 1.3)	124
Fasting Glucose ≥ 100 mg/dL	13 (10.8%)	12 (9.9%)	19 (16.7%)	13 (11.2%)	1.49 (0.8, 2.9)	18	7 (6.5%)	10 (9.3%)	0.71 (0.3, 1.8)	37	7 (6.7%)	7 (6.4%)	1.06 (0.4, 2.9)	272
Fasting Insulin ≥ 20 uIU/ml	32 (26.7%)	36 (29.8%)	33 (28.9%)	35 (30.2%)	0.96 (0.6, 1.4)	82	41 (38.3%)	37 (34.6%)	1.11 (0.8, 1.6)	27	43 (41.3%)	47 (42.7%)	0.97 (0.7, 1.3)	72
Hemoglobin A1c ≥ 5.7%	3 (2.5%)	2 (1.7%)	2 (1.8%)	3 (2.6%)	0.68 (0.1, 4.0)	120	1 (0.9%)	2 (1.9%)	0.50 (0.0, 5.5)	109	3 (2.9%)	4 (3.6%)	0.79 (0.2, 3.5)	133

To further characterize the results in terms of standard clinical definitions, Table S1 presents the baseline prevalences of BMI and select physiological measures defined by standard clinical thresholds, and changes in those prevalences over one, two, and three years. Changes in prevalences within the MMM and Health Education intervention groups are compared with Relative Risks and 95% confidence intervals, and the number needed to treat (NNT) to produce one beneficial (harmful) change in the MMM group versus the Health Education group. Prevalence changes generally followed similar patterns to those seen with scaled versions of the outcomes but only remission of severe obesity resulted in relative risks with 95% confidence intervals excluding 1.Children who started the study with severe obesity (BMI ≥ 120% of the 95th percentile BMI for their age and sex) in the MMM intervention group were about four times more likely to experience a remission from severe obesity after one year and three times more likely after two years compared to children who started the study with severe obesity in the Health Education intervention group. The NNT= 5 suggests that for an average of every five children with severe obesity assigned to the MMM intervention, there was one fewer child with severe obesity after one and two years than those assigned to the Health Education intervention. Also of note, the overall prevalences of elevated blood pressure (≥90th percentile for age, sex, and height) and hemoglobin A1c ≥ 5.7% were low and remained below 5% throughout the three years of the study.

NA = not applicable

[a] Baseline data from all participants, N = 120 for MMM and N= 121 for Health Education for all variables at baseline

[b] All participants with 1 year follow-up data, N= 120 for MMM and N = 118 for Health Education for BMI measures, N= 120 for MMM and N = 117 for Health Education for systolic and diastolic blood pressures, N= 114 for MMM and N = 116 for Health Education for all blood measures.

[C] All participants with 2 year follow-up data, N= 117 for MMM and N = 116 for Health Education for BMI measures, N= 109 for MMM and N = 110 for Health Education for systolic and diastolic blood pressures, N= 107 for MMM and N = 108 for Health Education for blood measures except fasting insulin N =107 for MMM and N = 107 for Health Education.

[d] All participants with 1 year follow-up data, N= 114 for MMM and N = 113 for Health Education for BMI measures, N= 109 for MMM and N = 110 for Health Education for systolic and diastolic blood pressure, N= 104 for MMM and N = 110 for Health Education for all blood measures.

[e] NNT = Number Needed to Treat, the number of participants assigned to one group that need to be treated for one of them to benefit compared with the participants assigned to the other group.

[f] Remission of obesity calculated as the number no longer with obesity after 1, 2 and 3 years from among those who started the study with obesity at baseline.

[g] Remission of severe obesity calculated as the number no longer with severe obesity after 1, 2 and 3 years from among those who started the study with severe obesity at baseline.

Table S2. Baseline values, changes, and group differences in parent/guardian outcome measures.

			Changes over 1 year				Changes over 2 years				Changes over 3 years			
	Baseline, Mean (SD)[a]		Slope, Mean (SD)[b]		Adjusted MMM-HE difference (95% CI)[c]	Standardized Effect Size, (Cohen's d)	Slope, Mean (SD)[b]		Adjusted MMM-HE difference (95% CI)[c]	Standardized Effect Size (Cohen's d)	Slope, Mean (SD)[b]		Adjusted MMM-HE difference (95% CI)[c]	Standardized Effect Size (Cohen's d)
	MMM	HE	MMM	HE			MMM	HE			MMM	HE		
Weight (kg)	80.47 (19.38)	76.98 (18.43)	-0.49 (4.72)	0.46 (3.09)	-0.79 (-1.91, 0.33)	*d*=0.19	-0.20 (2.98)	0.26 (1.88)	-1.09 (-2.49, 0.31)	*d*=0.22	0.20 (2.29)	0.40 (1.54)	-0.56 (-2.08, 0.96)	*d*=0.10
Body Mass Index, BMI (kg/m^2)	32.30 (6.21)	31.28 (6.65)	-0.22 (1.90)	0.21 (1.25)	-0.34 (-0.78, 0.11)	*d*=0.21	-0.08 (1.16)	0.12 (0.79)	-0.45 (-1.01, 0.10)	*d*=0.23	0.08 (0.90)	0.16 (0.64)	-0.23 (-0.84, 0.37)	*d*=0.10
Waist Circumference (cm)	105.81 (13.73)	103.42 (14.24)	0.15 (5.42)	1.44 (4.35)	-1.12 (-2.48, 0.24)	*d*=0.22	0.32 (3.18)	0.50 (2.24)	-0.57 (-2.13, 0.99)	*d*=0.10	0.66 (2.24)	0.44 (1.78)	0.68 (-0.88, 2.23)	*d*=0.11
Index Parent/Guardian Physical Activity (minutes per week)	125.08 (121.79)	139.83 (123.67)	20.45 (135.32)	9.91 (134.72)	2.40 (-26.43, 31.23)	*d*=0.02	18.38 (78.75)	-1.68 (55.92)	33.48 (5.05, 61.91)	*d*=0.30	11.27 (55.67)	1.32 (50.84)	20.03 (-13.57, 53.64)	*d*=0.15
Other Parent/Guardian Physical Activity (minutes per week)	130.83 (130.80)	132.39 (121.82)	2.01 (148.59)	14.30 (152.24)	-5.39 (-42.30, 31.53)	*d*=0.04	14.05 (71.85)	-1.84 (85.54)	31.58 (-11.71, 74.87)	*d*=0.22	4.33 (60.93)	-3.31 (57.41)	26.95 (-15.29, 69.19)	*d*=0.18
Health Literacy (0 low – 6 high)	1.72 (1.46)	1.72 (1.43)	0.52 (1.41)	0.25 (1.18)	0.26 (-0.05, 0.56)	*d*=0.22	0.33 (0.65)	0.15 (0.66)	0.37 (0.05, 0.68)	*d*=0.29	NA	NA	NA	NA

NA = not applicable, parent/guardian Health Literacy not assessed after 3 years.

[a] Baseline data from participating parents/guardians. Weight, BMI and waist circumference measures were excluded from analysis when a parent reported being pregnant or within three months after childbirth or miscarriage. At baseline, N= 114 for MMM and N = 115 for Health Education for weight, BMI and waist circumference measures, N= 120 for MMM and N = 121 for Health Education for index parent/guardian physical activity and health literacy, and N= 96 for MMM and N = 109 for Health Education for other parent/guardian physical activity.

[b] Unadjusted slopes calculated from all participants with follow-up data. Weight, BMI and waist circumference measures were excluded from analysis when a parent reported being pregnant or within three months after childbirth or miscarriage. Over one year, N= 112 for MMM and N = 106 for Health Education for weight, BMI and waist circumference measures, N= 118 for MMM and N = 116 for Health Education for index parent physical activity and health literacy, and N= 91 for MMM and N = 99 for Health Education for other parent/guardian physical activity. Over two years, N= 117 for MMM and N = 113 for Health Education for weight, BMI and waist circumference measures, N= 119 for MMM and N = 116 for Health Education for index parent physical activity and health literacy, and N= 95 for MMM and N = 102 for Health Education for other parent/guardian physical activity. Over three years, N= 119 for MMM and N = 115 for Health Education for weight, BMI and waist circumference measures, N= 119 for MMM and N = 117 for Health Education for index parent physical activity, and N= 97 for MMM and N = 104 for Health Education for other parent/guardian physical activity.

[c] Adjusted MMM minus Health Education differences from linear trajectories (slopes) for each individual regressed on intervention group assignment (centered), with the baseline value (centered at its mean) and the Intervention x baseline interaction as covariates, using standard maximum likelihood linear regression techniques. Consistent with intent-to-treat principles, trajectories for parent/guardians with only a single measure were imputed using a joint modeling multiple imputation approach as implemented in SAS using PROC MI (version 9.4). Thus, N= 120 for MMM and 121 for Health Education, for all outcomes.

Sensitivity Analyses of the Primary Outcome

We conducted six sensitivity analyses to determine whether the relaxation or modification of the assumptions underlying our primary analyses would materially affect our results. Slope estimates produced by these analyses are shown in Tables S3 and S4. Three of these analyses involved modifications to how or whether data were imputed and they yielded results consistent with those of our primary analysis. The other three analyses explored the sensitivity of our primary results to how data were formatted and modelled. These analyses both confirm our primary, and also shed light on how the BMI trajectories of treatment groups differed in the intervening years of the study. Specifically, when we relaxed the linearity assumption regarding BMI changes over time, the BMI trajectory for the MMM group was significantly attenuated over three years compared to that of the Health Education group, in contrast to our primary analysis that constrained BMI changes to be linear over time.

Sensitivity to Missingness
The first three analyses varied their approaches to handling missing data. For the first, the primary model was fit to complete case data (SM1), that is, data where subjects with no post-baseline follow-up were excluded. For the second, (SM2), data were imputed on a longitudinally formatted version of the data set then collapsed into individual-level slope values, then primary models were refit. In the third (SM3), imputation was performed on missing slope values, but the imputed data were forced to take extreme values. This allowed for the assessment of whether our primary results hold under various degrees of NMAR missingness. None of these analyses produced results that differed in significance or direction from those produced by our primary analyses.

Sensitivity to Parameterization
For three remaining analyses, we fit models to longitudinally formatted data rather than data collapsed into a single observation. Each used a mixed effect regression model with subject-specific random intercepts to assess the relationship between BMI and treatment over time. The effect of treatment over time was estimated by including a time x treatment interaction in the models. The first of these (SM4) was a complete case analysis using the same participants as in the primary analysis. The second (SM5) used this same subset of the GOALS cohort with missing BMI values imputed using standard MI procedures. The slope estimates from SM4 and SM5 did not differ from the primary.

Table S3. Slope estimates and p values from sensitivity analyses.

Effects shown are MMM relative to HE. For models SM1-SM5.

Model	Coefficient (95% CI)	p
SM1	-0.08 (-0.26, 0.14)	0.47
SM2	-0.06 (-0.26, 0.15)	0.58
SM3	-0.12 (-0.34, 0.10)	0.27
SM4	-0.03 (-0.19, 0.12)	0.65
SM5	-0.04 (-0.21, 0.12)	0.61

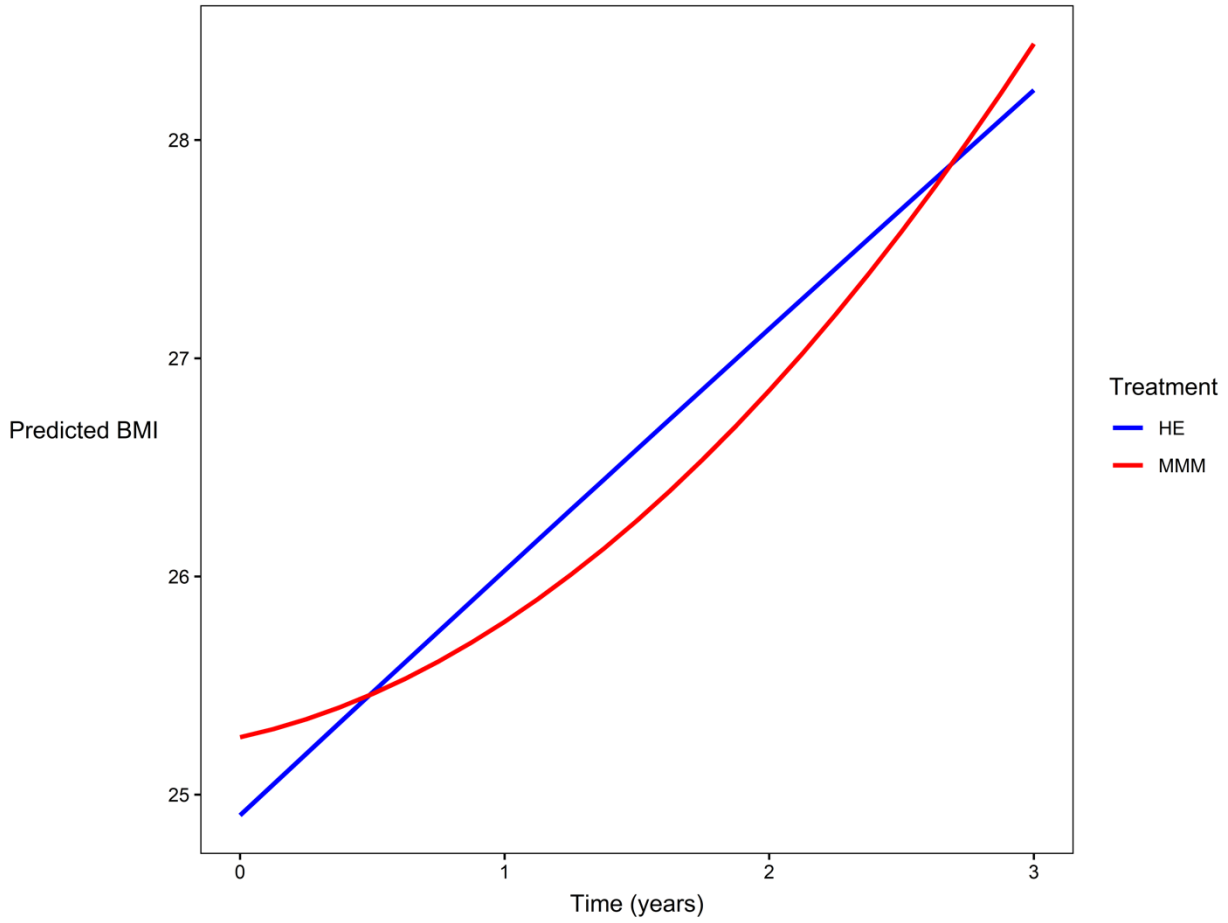
Sensitivity to Linearity
The third (SM6) fit complete case models with BMI as a function of quadratic and cubic functions of time and their interactions with treatment. For SM6, we found a significant quadratic relationship between BMI and treatment over time using a likelihood ratio test. The addition of cubic terms did not show a statistically significantly improvement over the quadratic model. A plot of BMI trajectories produced by the SM5 model serves as a useful guide in interpreting these results. Both the HE and MMM groups start in similar places. The significant quadratic term allows the BMI trajectory for the MMM group to bend away from the HE group for the intervening years of the study and return at the end. These results provide meaningful context to our primary results in that they show that BMI trajectories were significantly different for portions of the study, but ultimately both groups' BMIs ended up near the same place.

Table S4. Slope estimates and p values from sensitivity analysis.

Effects shown are MMM relative to HE. P value is from a likelihood ratio chi squared test.

Model	Linear Coefficient (95% CI)	Quadratic Coefficient (95% CI)	p
SM6	-0.86 (-1.37, -0.36)	0.27 (0.11, 0.43)	0.003

Figure S5. Model-based BMI curves for HE and MMM groups using SM6 quadratic models.



Sensitivity to Clustering

The published protocol included a plan to assess the sensitivity of the primary findings to assumptions of independent errors across participants, by accounting for potential clustering of responses.[1] An open question was whether this was appropriate in the setting where cluster membership changes over time and may be unknown, as in Stanford GOALS. To examine this question (for our study and more generally) we conducted simulation studies for a two-arm RCT where the number of clusters, the intra-cluster correlation (ICC) and the sample size per cluster varied.[12] We found that partial and complete misspecifications of cluster membership (where some and no knowledge of true membership were incorporated into assumptions) yielded inflated type I error rates.[12] Thus, trying to impose clusters on our analysis where cluster membership was dynamic and only partially defined was considered inappropriate and this analysis was not performed.

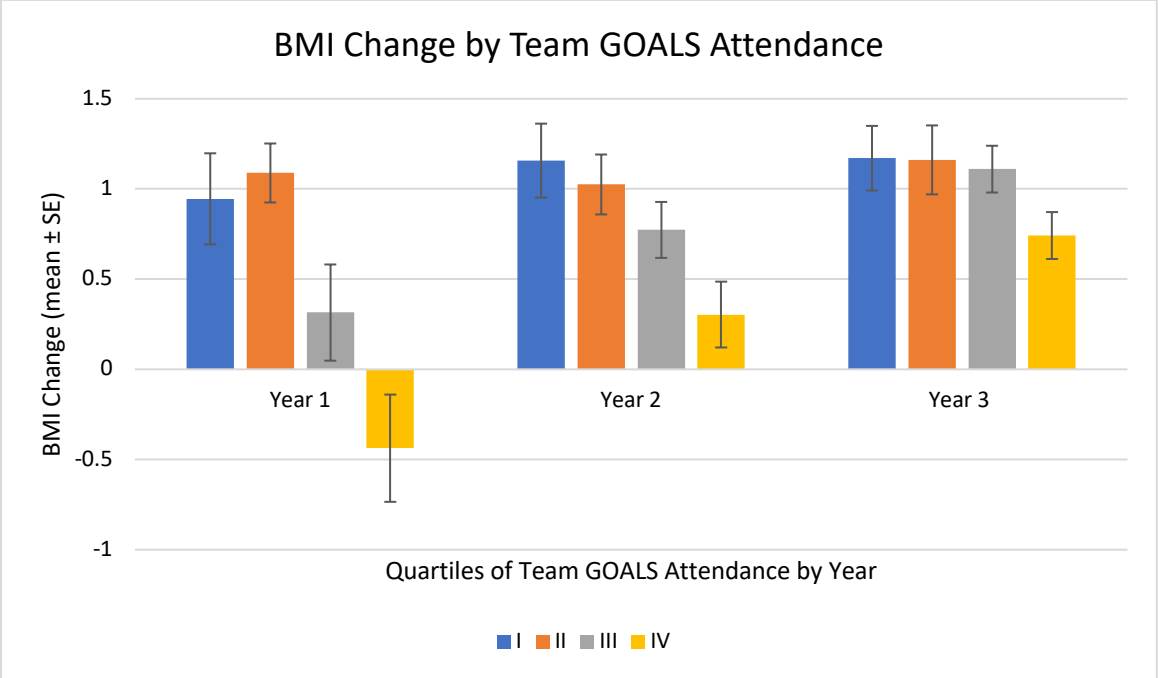
Moderator Analysis for BMI changes over 1 and 2 Years

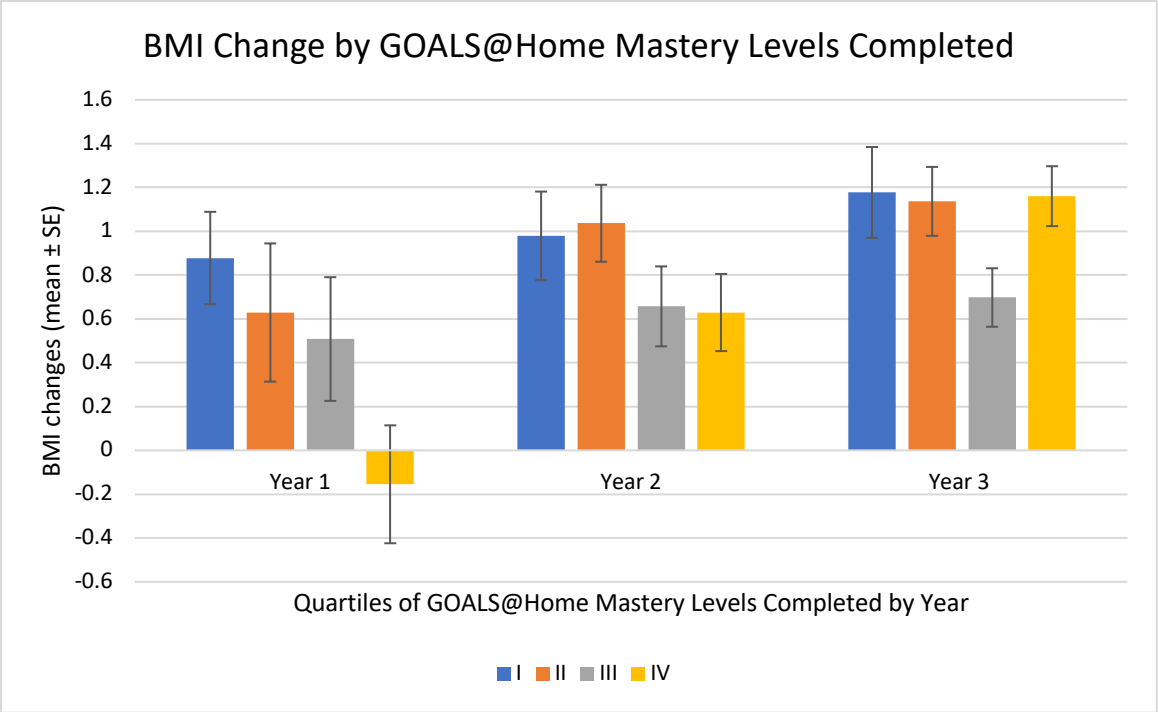
Over two years, the MMM intervention had greater effects than Health Education on BMI trajectories among children with lower baseline depressive symptoms, and over one year among children with a working mother, with private (versus government) health insurance, from households with fewer or greater than 3 adults, who reported less total screen time, had lower energy consumption with screens, and ate fewer breakfasts while watching TV at baseline.

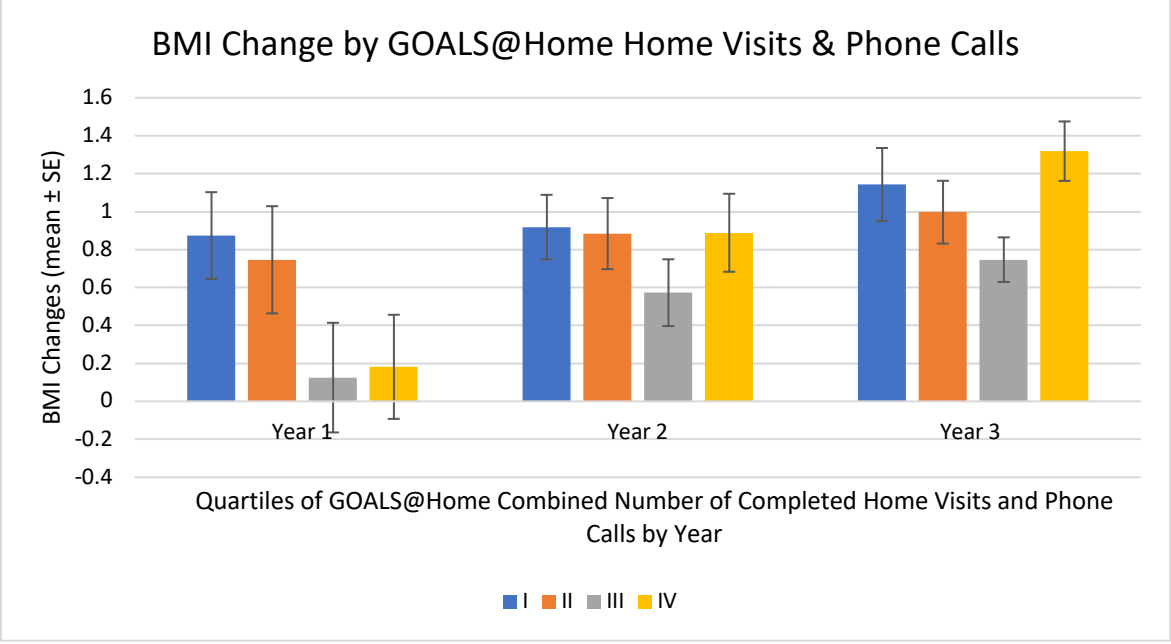
Dose Response

Change in BMI (slope) by Quartiles of participation (I low to IV high) in key intervention components in each year, among families randomized to the MMM intervention. In general, greater levels of participation were associated with less BMI gain. Because the GOALS@Home intervention was mastery based, greater BMI gain in Year 3 among children from families in the highest quartile (IV) of GOALS@Home mastery levels completed and combined home visits and phone calls may reflect attempts to "catch up" in levels and contacts among families who had lower levels of engagement and mastery during their first two years.

Figure S6. Change in BMI trajectory (slope) by quartiles of participation in key MMM intervention parameters (MMM sample only)




BMI Change by GOALS@Home Mastery Levels Completed
BMI changes (mean ± SE)
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
Year 1
Year 2
Year 3
Quartiles of GOALS@Home Mastery Levels Completed by Year
I II III IV


BMI Change by GOALS@Home Home Visits & Phone Calls
BMI Changes (mean ± SE)
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Year 1
Year 2
Year 3
Quartiles of GOALS@Home Combined Number of Completed Home Visits and Phone Calls by Year
I II III IV

Adverse Events

Serious Adverse Event - Any adverse event which:

a. Results in death
b. Is life-threatening
c. Requires in-patient hospitalization, or prolongation of an existing hospitalization
d. Results in persistent or significant disability and/or incapacity
e. Is a congenital anomaly / birth defect in the offspring of a participant.

Adverse Event - Any injury, illness or other medical problem requiring a visit to a health care provider (a doctor, an emergency room or urgent care center, a hospital, a primary care provider, or a public health clinic) related to participation in study.

Table S5. Adverse events and serious adverse events by relatedness, group and child and adult participants.

		Child Participant		Adult participant	
		MMM	**Health Ed**	**MMM**	**Health Ed**
Serious Adverse Events	**Related to study**	0	0	0	0
	Possibly related to study	0	1*	0	0
	Not related to study	4	3	22	15
Adverse Events	**Possibly or probably related to study**	3	0	0	0

* Hospitalization for poor food intake and an abnormal heart rhythm.

Systematic annual monitoring of adverse events did not find evidence of differential risks associated with the MMM and Health Education interventions in children or parents/guardians. One potentially serious adverse event, hospitalization for poor food intake and abnormal heart rhythm, was judged (blinded to group assignment) to be possibly related to study participation and was subsequently found to occur in a child randomized to Health Education. There were 3 injuries or other medical problems requiring a visit to a medical care provider judged (blinded to group assignment) to be possibly or probably related to study participation, two broken fingers in the same child in different years, and one child with rapid weight loss, both subsequently found to be randomized to MMM. There were no reported adverse events among parents/guardians judged to be possibly or probably related to study participation.

In addition, through our clinical monitoring of physiological measures, three children were diagnosed with diabetes during their participation in the study. One child in the MMM group was diagnosed with type 1 diabetes during year 1, judged not related to study participation, and two children, one in each group, were diagnosed with type 2 diabetes first identified from their blood tests at their year 3 study assessments, upon completing the study.

Parents/guardians were informed if children's measures indicated poor statural growth, hypertension, dyslipidemias, impaired fasting glucose, pre-diabetes, diabetes mellitus, or excessive weight loss, with an explanation of the result and referral to their primary care medical professional for further evaluation.

Data and Database Management and Data Quality Control

Physical data measures for both children and adults are entered directly into a customized FileMaker database system which is set to prompt the data collector to follow all MOP rules, including rules as to when a 3rd physical measure is needed or if the measurement needs to be checked as valid. The database is designed to prompt data collectors when to perform random test-retest measures and, unbeknownst to the data collectors, will additionally prompt a height re-measurement when a child participant is measured as shorter than at a previous visit or an adult is measured more than 0.5cm shorter than a previous visit.

Child survey measures will be directly entered into the same database, using similar systems to make sure all questions are answered and within expected ranges, to prevent transcription errors.

Once the baseline visit record is complete in the database it is locked to prevent any further manipulation and can only be unlocked by the database manager and data aide in order to be modified.

Parent surveys are completed on paper. To ensure data completion, paper surveys are reviewed once in the field and again in the office before being double-entry keypunched. Once in electronic form, data are built into SAS databases and at least 20% are reviewed for accuracy by a staff data aide.

The data team meets weekly to review visit progress and ensure that all visits are being completed as expected. Each overdue visit is reviewed and any visits not fully completed within 30 days are reviewed. Questions and concerns are raised with the Principal Investigator as needed.

Data reports, including completeness and range and frequency for categorical measures and range and mean (SD) for continuous measures are reviewed at least monthly and typically bi-weekly by the database manager and principal investigator.

Data will be uploaded quarterly to the RCU as directed by the MOP.

In addition, we include these general approaches for quality control in data collection and data management, prior to database management.

- All measures are made according to a detailed MOP
- Updated protocols and MOP are kept both online and in hard copy binders for easy access
- Training is conducted using step-by-step instructions and data collectors meet weekly throughout the study to share experiences and problem solve, if necessary
- Data collectors must pass certification prior to collecting data
- Protocols are followed for physical measures instrument validity/calibration checks
- Inter-rater reliability is assessed throughout the study on a random 10% of participants
- Ongoing (booster) trainings are provided throughout the study
- Data collectors use checklists for each visit to ensure completeness
- Any paper surveys are color coded
- ID labels are preprinted
- IDs include a last digit as a check digit
- Direct entry of most data (custom filemaker pro database) to reduce transcription errors and eliminate readability errors
- Automated real-time safeguards to prevent illogical data entry (e.g., range checks, longitudinal checks)
- Data entry software conducts all calculations in real-time (e.g., eligibility, outliers)
- Double-entry keypunching of paper survey data
- Standardized data cleaning rules
- Manual and automated checks for completion of all measures
- Tracking of all data in database
- 24-hour recalls – quality control checks on outliers
- Actigraph accelerometers – immediate download and review for completeness upon receipt
- Color-coded alerts in Filemaker Pro data management system
- Data backed-up on external USB drive at each visit
- Database backed-up daily on remote server

References cited in Supplementary Material

1. Robinson TN, Matheson D, Desai M, et al. Family, community and clinic collaboration to treat overweight and obese children: Stanford GOALS-A randomized controlled trial of a three-year, multi-component, multi-level, multi-setting intervention. Contemp Clin Trials 2013;36:421-35.
2. Efron B. Forcing a sequential experiment to be balanced. Biometrika 1971;58:403-17.
3. Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: John Wiley & Sons; 1987.
4. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch Gen Psychiatry 2006;63:484-9.
5. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. Biol Psychiatry 2006;59:990-6.
6. Kraemer HC, Morgan GA, Leech NL, Gliner JA, Vaske JJ, Harmon RJ. Measures of clinical significance. J Am Acad Child Adolesc Psychiatry 2003;42:1524-9.
7. Kraemer HC, Thiemann S. How many subjects? Statistical Power Analysis in Research. Newberry Park, CA: Sage Publications; 1987.
8. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Second ed. Hillsdale, NJ: Lawrence Erlbaum Associated, Publishers; 1988.
9. Flores R. Dance for health: improving fitness in African American and Hispanic adolescents. Public Health Rep 1995;110:189-93.
10. Robinson TN. Reducing children's television viewing to prevent obesity. JAMA 1999;282:1561-7.
11. Robinson TN, Killen JD, Kraemer HC, et al. Dance and reducing television viewing to prevent weight gain in African-American girls: The Stanford GEMS pilot study. Ethn Dis 2003;13:s65-s77.
12. Desai M, Bryson SW, Robinson T. On the use of robust estimators for standard errors in the presence of clustering when clustering membership is misspecified. Contemp Clin Trials 2013;34:248-56.