**Supplementary material for**

Vainik, U., Konstabel, K., Lätt, E., Mäestu, J., Purge, P., & Jürimäe, J. (2015). Diet misreporting can be corrected: confirmation of the energy intake - fat-free mass association in adolescents. *Submitted for publication*.

**Figure S1a**

A log-transformed version of the regression results

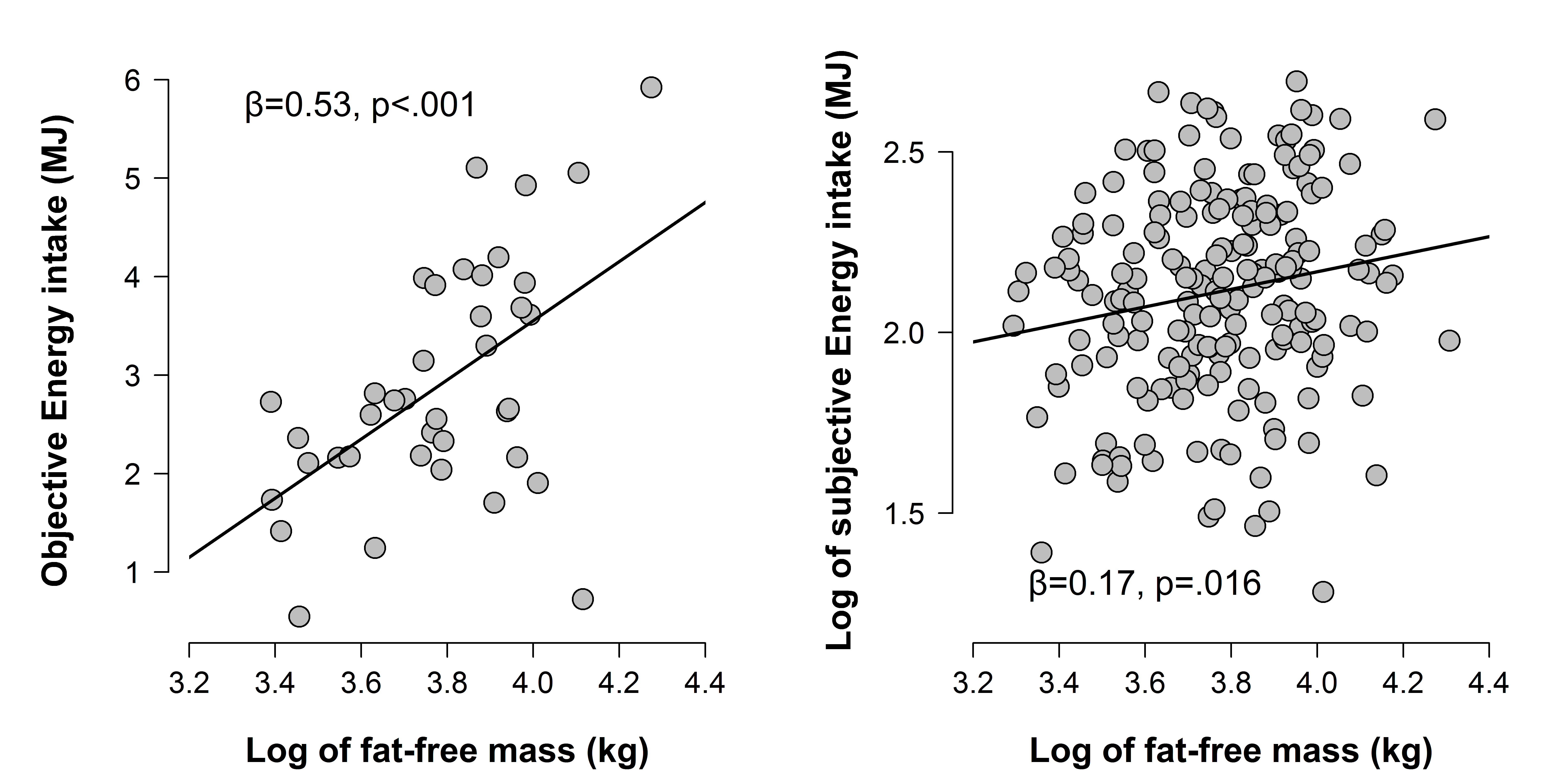


Figure S1a. Fat-free mass associations with objective (left) and subjective (right) energy intake.

Objective energy intake was measured on the same day, whereas subjective energy intake was assessed from dietary interview from an earlier period of 3 days. Data not corrected for the effects of age. Fat-free mass, and subjective energy intake have been log-transformed.

**Figure S1b**

A log-transformed version of the regression results that includes only people from the oEI group, as requested by a reviewer.

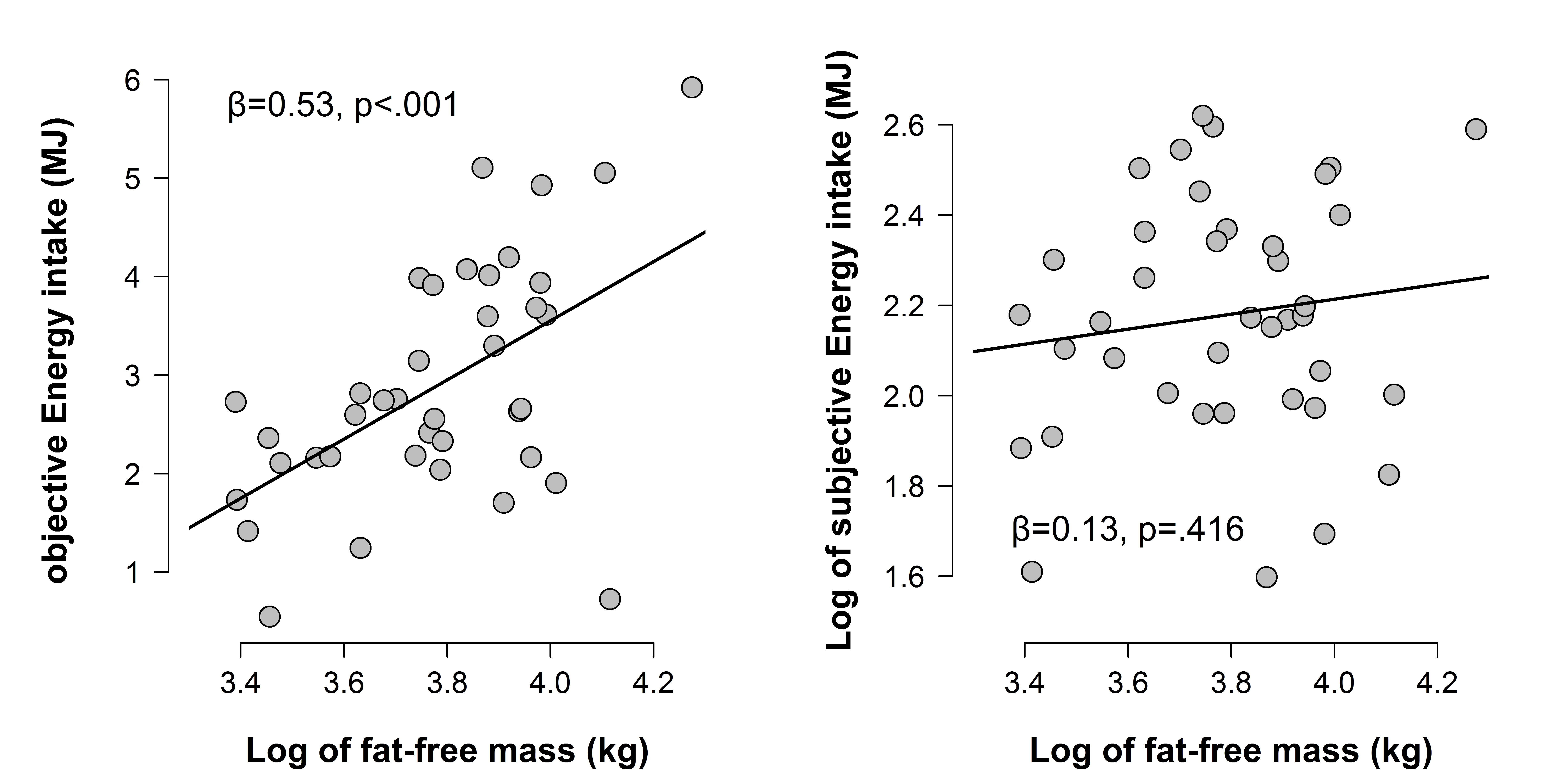


Figure S1b. Fat-free mass associations with objective (left) and subjective (right) energy intake.

Objective energy intake was measured on the same day, whereas subjective energy intake was assessed from dietary interview from an earlier period of 3 days. Data not corrected for the effects of age. Fat-free mass, and subjective energy intake have been log-transformed. Data limited only to participants with objective energy intake data (n=42).

**Figure S2**

For a graphical representation of the findings, Models 1 and 2 were also plotted. The association tested in Model 1 is present in actual data (Figure S2, left; see also Figure 2, right), but is absent in the resampled data (Figure S2 right). However, Model 2 is largely unaffected by the data resampling procedure (Figure S2).



Figure S2. Graphical representation of the association between fat-free mass and subjective energy intake (left) and resampled subjective energy intake (right). Data was resampled by assigning each participant an energy intake value of another participant. Solid line depicts Model 1 in Table 4, where the fat-free mass predicts subjective intake (left) but not resampled subjective intake (right). Dashed line depicts Model 2 in Table 4, where the same association is tested only within plausible reporters. Intriguingly, the association on the left is reproduced also in resampled data (right), suggesting a method artefact.

Objective energy intake was measured on the same day, whereas subjective energy intake was assessed from dietary interview from an earlier period of 3 days. Data has not been log transformed or corrected for the effects of age

Figure S3

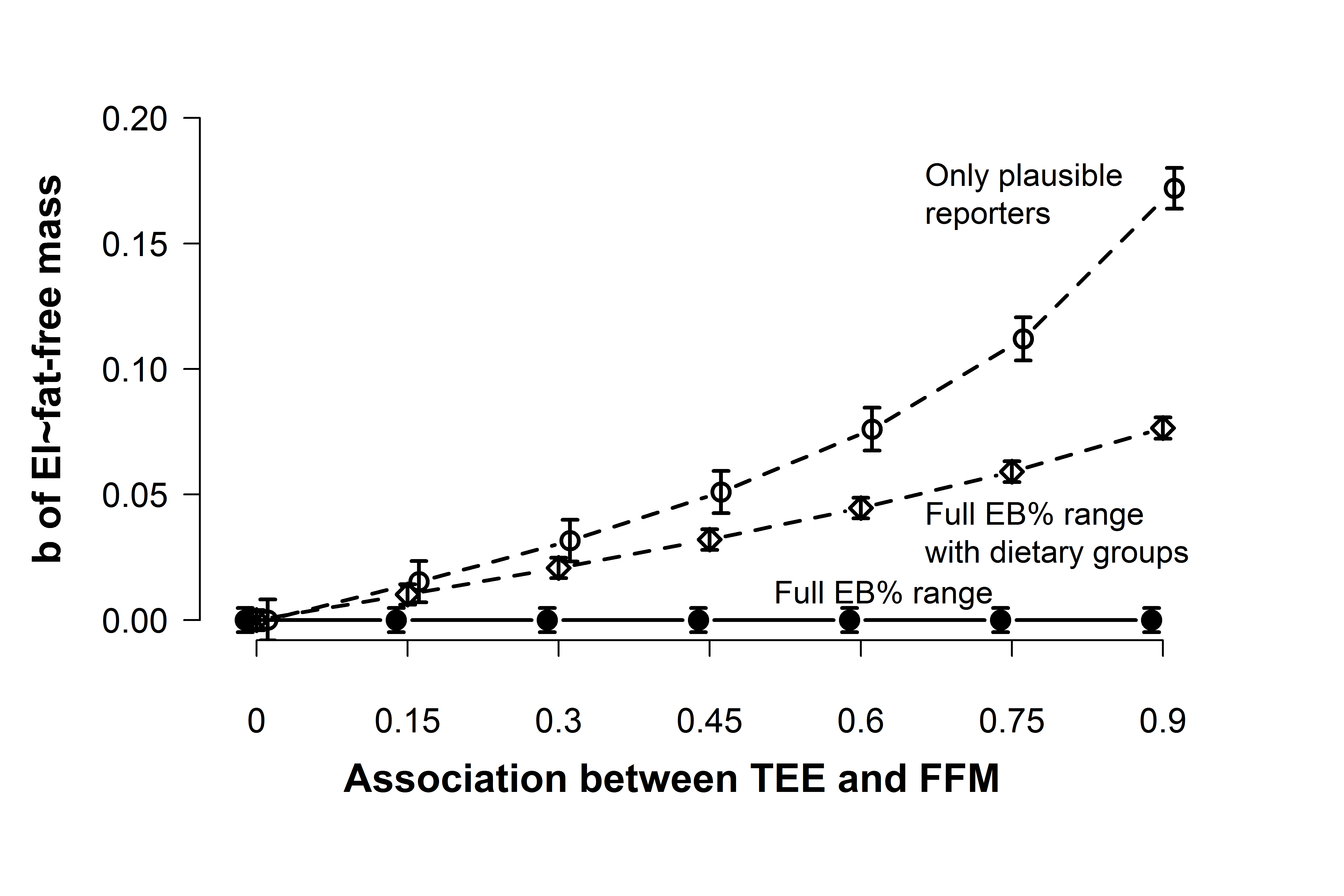


Figure S3. The effect of restricting variance based on a partly related variable on unstandardised regression using simulated data. The expected association (b) between fat-free mass (FFM) and energy intake (EI) is zero, and full sample data shows this (solid line with full circles). Dashed lines show the same analysis in case the analysis focused only on plausible reporters (dashed line with empty circles, like Model 2) or in case dietary groups information was added as a covariate (dashed line with diamons, like Model 3). In the latter two cases, the artificial association varies as a function of the association strength between TEE and fat-free mass.

Data simulated on 10,000 participants, 10,000 times. Variables have similar properties like actual data in terms of distribution. For precise parameters, see supplementary material. Error bars denote 95% confidende intervals (standard errors multiplied by 1.96). See Figure S3 for non-standardised regressions.

**Script for simulation**

## Simulation that shows how an artificial association

# between two variables can be created, after selecting

# a sub-sample based on a related variable.

#

### Part 1 - a walkthrough of the code.

# set path to where you want to save/load files.

path="C:/Users/Uku/Dropbox/Documents/teadusprojektid/kehakultuuri projekt/mar2015/"

setwd(path)

library(truncnorm) # a package for handy data simulation. use install.packages("truncnorm") if you do not have it

# function for plotting SE bars. From http://shinyapps.org/apps/RGraphCompendium/index.php

plotsegraph <- function(loc, value, sterr, wiskwidth, color = "grey", linewidth = 2) {

w <- wiskwidth/2

segments(x0 = loc, x1 = loc, y0 = value - sterr, y1 = value + sterr, col = color,

lwd = linewidth)

segments(x0 = loc - w, x1 = loc + w, y0 = value + sterr, y1 = value + sterr,

col = color, lwd = linewidth) # upper whiskers

segments(x0 = loc - w, x1 = loc + w, y0 = value - sterr, y1 = value - sterr,

col = color, lwd = linewidth) # lower whiskers

}

# A function to produce correlated variables with certain propoerties. # Supports,mean, sd, min, and max. Based on truncnorm and

# https://stackoverflow.com/questions/13291308/generate-numbers-with-specific-correlation

simcor2 <- function (n, xmean, xsd,xmin,xmax, ymean, ysd,ymin,ymax, correlation) {

x <- rtruncnorm(n=n, a=xmin, b=xmax, mean=xmean, sd=xsd)

y <- rtruncnorm(n=n, a=ymin, b=ymax, mean=ymean, sd=ysd)

z <- correlation \* scale(x)[,1] + sqrt(1 - correlation^2) \* scale(resid(lm(y ~ x)))[,1]

xresult <- xmean + xsd \* scale(x)[,1]

yresult <- ymean + ysd \* z

data.frame(x=xresult,y=yresult)

}

## Simulation parameters

n=10000 # sample size

set.seed(52) # set seed for reproducing current results

## Create variables. These could be any distributions,

# but I here I am setting the variables to have a similar

# distrubution to the ones seen in data

# Create TEE and FFM that are related at 0.75

vars=simcor2(n=n,xmean=11.8,xsd=2.7,xmin=7.3,xmax=26.1,ymean=44.37,ysd=9.08, ymin=26.7,ymax=74.3,correlation=0.75)

colnames(vars)=c("TEE", "FFM")

# Create EI that is unrelated to both

vars$EI=rtruncnorm(n=n, a=2.6, b=13.8, mean=7.6, sd=2.4)

# Verify associations

cor(vars)

# Calculate energy balance ratio

vars$EB=vars$EI/vars$TEE\*100

hist(vars$EB,breaks=100) # underreporting is present..

# test regression on full data

summary(lm(EI~FFM, data=vars)) # Association close to zero.

# Like in Model 2: use group variable as control

# generate a grouping/categorising variable

vars$URPROR=cut(vars$EB,breaks=c(0,85,115,200), c("UR","PR","OR"))

# set PR as reference group for regression

vars$URPROR=relevel(vars$URPROR, ref="PR")

# regression model having dietary groups as a covariate

summary(lm(EI~FFM+URPROR, data=vars))

# A significant association, when there should be none.

# Like in model3: # A significant association, when there should be none.

#keep only plausible reporters, +- 15%

vars\_PR=subset(vars,URPROR=="PR")

# regression model

summary(lm(EI~FFM, data=vars\_PR))

# Again a large association!

### Repeat 10,000 times!

# To demonstrate the robustness of the findings,

# we are repeating the simulation 10,000 times.

# We also use other correlations to see, whether the

# effect dissappears with smaller effects.

# NB! Running the whole simulatuon can take several hours. For a quick

# estimate, Reduce the number of repetitions and / or

# sample size. For instance, nsim=100 will also work reasonably well.

# NB! Graphical parameters, such as graph limits or text labels are optimised for 10000 sample size

n=10000 # sample size

nsim =10000 # number of repetitions

cors=seq(0,0.90,0.15)

# Preallocate arrays to collect the results on regression

b\_full=matrix(NA,nsim,length(cors))

se\_full=matrix(NA,nsim,length(cors))

b\_cov=matrix(NA,nsim,length(cors))

se\_cov=matrix(NA,nsim,length(cors))

b\_restr=matrix(NA,nsim,length(cors))

se\_restr=matrix(NA,nsim,length(cors))

# Regular regression

for (k in 1:length(cors)) {

for (i in 1:nsim) {

#Create variables

vars=simcor2(n,11.8,2.7,7.3,26.1,44.37,9.08, 26.7,74.3, cors[k])

colnames(vars)=c("TEE", "FFM")

vars$EI=rtruncnorm(n=n, a=2.6, b=13.8, mean=7.6, sd=2.4)

vars$EB=vars$EI/vars$TEE\*100

# extract the regression coefficient

fit=lm(EI~FFM, data=vars)

m=summary(fit)

b\_full[i,k]=m$coefficients[2,1]

se\_full[i,k]=m$coefficients[2,2]

# generate a grouping/categorising variable

vars$URPROR=cut(vars$EB,breaks=c(0,85,115,200), c("UR","PR","OR"))

# set PR as reference group for regression

vars$URPROR=relevel(vars$URPROR, ref="PR")

# regression model having dietary groups as a covariate

fit2=lm(EI~FFM+URPROR, data=vars)

m2=summary(fit2)

b\_cov[i,k]=m2$coefficients[2,1]

se\_cov[i,k]=m2$coefficients[2,2]

# Omit implausible reporters

vars\_PR=subset(vars,URPROR=="PR")

# extractthe regression coefficient

fit3=lm(EI~FFM, data=vars\_PR)

m3=summary(fit3)

b\_restr[i,k]=m3$coefficients[2,1]

se\_restr[i,k]=m3$coefficients[2,2]

# remove temp diles

rm(vars,fit,m,fit2,m2,fit3,m3,vars\_PR)

}

}

b\_fullm=colMeans(b\_full)

se\_fullm=colMeans(se\_full)

b\_covm=colMeans(b\_cov)

se\_covm=colMeans(se\_cov)

b\_restrm=colMeans(b\_restr)

se\_restrm=colMeans(se\_restr)

# standardized regression coefficients. The only difference is that certain variables are scaled first.

b\_fulls=matrix(NA,nsim,length(cors))

se\_fulls=matrix(NA,nsim,length(cors))

b\_covs=matrix(NA,nsim,length(cors))

se\_covs=matrix(NA,nsim,length(cors))

b\_restrs=matrix(NA,nsim,length(cors))

se\_restrs=matrix(NA,nsim,length(cors))

# Regular regression

for (k in 1:length(cors)) {

for (i in 1:nsim) {

#Create variables

vars=simcor2(n,11.8,2.7,7.3,26.1,44.37,9.08, 26.7,74.3, cors[k])

colnames(vars)=c("TEE", "FFM")

vars$EI=rtruncnorm(n=n, a=2.6, b=13.8, mean=7.6, sd=2.4)

vars$EB=vars$EI/vars$TEE\*100

# standardize variables so we get a standardized result for regression

vars$EIs=scale(vars$EI)

vars$FFMs=scale(vars$FFM)

# extract the regression coefficient

fit=lm(EIs~FFMs, data=vars)

m=summary(fit)

b\_fulls[i,k]=m$coefficients[2,1]

se\_fulls[i,k]=m$coefficients[2,2]

# generate a grouping/categorising variable

vars$URPROR=cut(vars$EB,breaks=c(0,85,115,200), c("UR","PR","OR"))

# set PR as reference group for regression

vars$URPROR=relevel(vars$URPROR, ref="PR")

# regression model having dietary groups as a covariate

fit2=lm(EIs~FFMs+URPROR, data=vars)

m2=summary(fit2)

b\_covs[i,k]=m2$coefficients[2,1]

se\_covs[i,k]=m2$coefficients[2,2]

# Omit implausible reporters

vars\_PR=subset(vars,URPROR=="PR")

# extractthe regression coefficient

fit3=lm(EIs~FFMs, data=vars\_PR)

m3=summary(fit3)

b\_restrs[i,k]=m3$coefficients[2,1]

se\_restrs[i,k]=m3$coefficients[2,2]

# remove temp diles

rm(vars,fit,m,fit2,m2,fit3,m3,vars\_PR)

}

}

b\_fullms=colMeans(b\_fulls)

se\_fullms=colMeans(se\_fulls)

b\_covms=colMeans(b\_covs)

se\_covms=colMeans(se\_covs)

b\_restrms=colMeans(b\_restrs)

se\_restrms=colMeans(se\_restrs)

# save/load results to/from "sim" folder in path.

# if the folder does not exist, you have to create it

save.image(file="sim/sim\_results10000x10000lines\_3M.Rdata")

#load(file="sim/sim\_results.Rdata")

#load(file="sim/sim\_results10000x10000lines.Rdata")

# plot the regular regression

tiff("Fig S3. plot sim regression.tiff", width = 18, height = 12, units = "cm", res = 800, pointsize = 10,compression="lzw")

par(cex.main = 1.5, mar = c(8, 4, 3, 4) + 0.1, mgp = c(3.5, 1, 0), cex.lab = 1.5, font.lab = 2, cex.axis = 1.3, bty = "n", las = 1)

x=seq(1.5,length(cors)+0.5,1)

jitter=0.15

xj=x+jitter

xj.5=x+jitter/2

# find max and min values for y axis. round them to nearest 0.05

ymax=round((max(b\_fullm,b\_covm,b\_restrm)+max(se\_fullm,se\_covm,se\_restrm))+0.01,2)

ymin=round((min(b\_fullm,b\_covm,b\_restrm)-max(se\_fullm,se\_covm,se\_restrm))-0.01,2)

ymax=round(ymax/0.05)\*0.05

ymin=round(ymin/0.05)\*0.05

plot(x, rep(-10, length(x)), type = "p", ylab = "", xlab = " ", cex = 1.5, ylim = c(ymin, ymax), xlim = c(1, max(xj)), lwd = 2, pch = 5, axes = F, main = " ")

axis(1,at=x+jitter/2, labels=F)

text(x=x+jitter/2, y=par()$usr[3]-0.1\*(par()$usr[4]-par()$usr[3]), labels=cors,xpd=TRUE, cex=1.3, pos=1) # adds labels

mtext("Association between TEE and FFM", side = 1, line = 4, cex = 1.5, font = 2)

axis(2, pos = 1.2 )

par(las = 0)

mtext("b of EI~fat-free mass", side = 2, line = 2, cex = 1.5, font = 2)

points(x, b\_fullm, cex = 1.5, lwd = 2, pch = 19)

plot.errbars <- plotsegraph(x, b\_fullm, se\_fullm\*1.96, 0.1, color = "black") #0.1 = wiskwidth

lines(x,b\_fullm, lwd = 2, type = "c")

points(xj.5, b\_covm, cex = 1.5, lwd = 2, pch = 23)

plot.errbars <- plotsegraph(xj.5, b\_covm,se\_covm\*1.96, 0.1, color = "black") #0.1 = wiskwidth

lines(xj.5,b\_covm, lwd = 2, type = "c", lty="dashed")

points(xj, b\_restrm, cex = 1.5, lwd = 2, pch = 21)

plot.errbars <- plotsegraph(xj, b\_restrm,se\_restrm\*1.96, 0.1, color = "black") #0.1 = wiskwidth

lines(xj,b\_restrm, lwd = 2, type = "c", lty="dashed")

# add text labels

text(5, 0.01, "Full EB% range", cex = 1.2, font = 1, adj = 0)

text(6, 0.035, "Full EB% range\nwith dietary groups", cex = 1.2, font = 1, adj = 0)

text(6, 0.17, "Only plausible\nreporters", cex = 1.2, font = 1, adj = 0)

dev.off()

## plot standardized regression lines

tiff("Fig 5. plot sim regression std.tiff", width = 18, height = 12, units = "cm", res = 800, pointsize = 10,compression="lzw")

par(cex.main = 1.5, mar = c(8, 4, 3, 4) + 0.1, mgp = c(3.5, 1, 0), cex.lab = 1.5, font.lab = 2, cex.axis = 1.3, bty = "n", las = 1)

x=seq(1.5,length(cors)+0.5,1)

jitter=0.15

xj=x+jitter

xj.5=x+jitter/2

# find max and min values for y axis. round them to nearest 0.05

ymax=round((max(b\_fullms,b\_covms,b\_restrms)+max(se\_fullms,se\_covms,se\_restrms))+0.08,2)

ymin=round((min(b\_fullms,b\_covms,b\_restrms)-max(se\_fullms,se\_covms,se\_restrms))-0.01,2)

ymax=round(ymax/0.05)\*0.05

ymin=round(ymin/0.05)\*0.05

plot(x, rep(-10, length(x)), type = "p", ylab = "", xlab = " ", cex = 1.5, ylim = c(ymin, ymax), xlim = c(1, max(xj)), lwd = 2, pch = 5, axes = F, main = " ")

axis(1,at=x+jitter/2, labels=F)

text(x=x+jitter/2, y=par()$usr[3]-0.1\*(par()$usr[4]-par()$usr[3]), labels=cors,xpd=TRUE, cex=1.3, pos=1) # paneb sildid

mtext("Association between TEE and FFM", side = 1, line = 4, cex = 1.5, font = 2)

axis(2, pos = 1.2 )

par(las = 0)

mtext("β of EI~fat-free mass", side = 2, line = 2, cex = 1.5, font = 2)

points(x, b\_fullms, cex = 1.5, lwd = 2, pch = 19)

plot.errbars <- plotsegraph(x, b\_fullms, se\_fullms\*1.96, 0.1, color = "black") #0.1 = wiskwidth

lines(x,b\_fullms, lwd = 2, type = "c")

points(xj.5, b\_covms, cex = 1.5, lwd = 2, pch = 23)

plot.errbars <- plotsegraph(xj.5, b\_covms,se\_covms\*1.96, 0.1, color = "black") #0.1 = wiskwidth

lines(xj.5,b\_covms, lwd = 2, type = "c", lty="dashed")

points(xj, b\_restrms, cex = 1.5, lwd = 2, pch = 21)

plot.errbars <- plotsegraph(xj, b\_restrms,se\_restrms\*1.96, 0.1, color = "black") #0.1 = wiskwidth

lines(xj,b\_restrms, lwd = 2, type = "c", lty="dashed")

# add text labels

text(5, 0.05, "Full EB% range", cex = 1.2, font = 1, adj = 0)

text(6, 0.15, "Full EB% range\nwith dietary groups", cex = 1.2, font = 1, adj = 0)

text(6, 0.65, "Only plausible\nreporters", cex = 1.2, font = 1, adj = 0)

dev.off()