Report 0HM240 Assignment 2 Team 2

Introduction

In this assignment, the self-tracked data will be interpreted and reported, together with our conclusions. First, the data of all participants will be analyzed in Part A. In Part B, three individual participants will be analyzed. The analysis will take place in STATA 18 BE with the given dataset "Data_2024.dta". Commands in text will be **bold** to make a clear distinction between text and commands, variables in text will be encased in rectangular brackets, [variable], for instance.

Checking data

The dataset was inspected and optimized before the actual multilevel analysis. To take a first look, **summarize** was used. There are 72 variables in the dataset, of which 66 returned zero observations. After checking the dataset, it became clear that these 66 variables were of the type string, so the **summarize** command couldn't make computations with them. Table 1 displays a summary of the remaining 6 variables. Two variables, [MCTQ_Work_days] and [PSQI_02], have only 771 observations compared to 991 observations of the other variables because not all variables have been used for data collection in the three separate rounds of data collection. Lastly, [Participant_ID] has only one relevant value in Table 1, the number of observations. As it is a randomly assigned number per participant the calculation of mean, standard deviation, minimum value, and maximum value are meaningless.

Table 1

Results summarize command on the raw dataset.

Variable	Obs.	Mean	Std. dev.	Min	Max
Participant_ID	911	212840.2	77001.64	101082	325988
Energetic_sliderNegPos	911	5.532.272	1.920.128	0	100
Stress_sliderNegPos	911	6.112.184	1.759.831	4	100
Happy_smiley	911	6.361.581	1.705.793	8	100
MCTQ_Work_days	771	4.843.061	1.454.956	0	7
PSQI_02	771	1.980.545	1.336.407	5	70

To properly examine the data, the string variables were converted into continuous or categorical variables, when applicable: this can be seen in detail in the Do-File that is provided, with comments explaining the process. Furthermore, a selection was made of variables to keep for analysis. The decision was made because it would take a considerable amount of time to properly transform all variables, while in the end, we would only analyze a few for this assignment.

The variables kept, using **keep**, were firstly [Date_submitted] and [Participant_ID], so it is possible to distinguish between participants and between submission years, days, and moments. Next, we kept [Sleepiness], [Energetic_sliderNegPos], [Stress_sliderNegPos], [Happy_smiley], [Motivation], and [TimeOutdoors]. There is no rationale based on literature for these variables. Our group thought these variables were interesting and relevant for our understanding of using Experience Sampling as a method, and they were all used during all three data collections.

Some variables were renamed, with rename, to ease coding later in the process. Table 2 shows the old and new variable names.

Table 2

Old and new variable names

Original Variable Name	New Variable Name
Date_submitted*	Date_submitted*
Participant_ID	student
Sleepiness	sleepiness
Energetic_sliderNegPos	energylevel
Stress_sliderNegPos	stresslevel
Happy_smiley	happiness
Motivation	motivation
TimeOutdoors	timeoutdoors

**The Date_submitted variable did not change name*

Now that the variables have easier codable names they can be prepared for Multilevel Analysis (MLA). We ordered the data by student, day, and time of day, by separating the date and time from [Date_submitted] into two new variables, [Date] and [Time], and generating the variable [day] grouping each day. The data is sorted by [student] and [day], after which the variables [time] and [timeonday] are created, which contain the number for the n-th measurement per student and the n-th measurement per student on a specific day, respectively. Three students who did not submit data on all their respective submission dates or who only submitted one questionnaire per day(s) were dropped: these were participants 136569, 233885, and 325430.

There were no missing values in [Date_submitted], so the data should be sorted correctly.

Individual variables were inspected and edited to facilitate MLA. [motivation] is a categorical variable with a five-point Likert scale. Due to a capital letter difference, one category was split in two, and it was necessary to use separate commands to score all measurements correctly. The variable was transformed to display the categories in a numerical value ranging from 1, “not motivated at all”, to 5, “strongly motivated”.

[sleepiness] is a categorical variable with a six-point Likert scale. Due to typos and a difference in answers during the different collection years, there were ten different possible answers. The answers were transformed to display the categories in a numerical value ranging from 1, “Almost in reverie, sleep onset soon, lost struggle to remain awake”, to 6, “Feeling active, vital, alert or wide awake”.

The continuous variable [energylevel] was kept like it was, ranging from 0, no energy to 100, full of energy. The continuous variable [stresslevel], ranging from 0 to 100, was reversed to avoid confusion, since it first suggested that a higher score meant lower stress because of how the question was presented in the questionnaires. After recoding, the higher the value of [stresslevel] was, the higher the stress level. The categorical variable [happiness] was kept as is.

[timeoutdoors] should be a continuous variable, though, the answers were admitted as strings in the questionnaires. Because of this, there are four strange values that cannot be changed into a numerical value. One observation, without a number, was changed to a missing value. The other three observations were changed and cropped to keep the correct numerical value. Afterwards, [timeoutdoors] was destringed using **destring**.

Part A

Exploratory analysis

The data was now further explored to determine whether there are correlational relationships worth analyzing.

After transformation of the data, we first filtered the data by deleting the participants who have less than 10 data entries to ensure data quality. With the processed data above, we checked each dependent variable to find out whether there are any interesting points within them.

We first inspected whether there is any pattern within the data point distribution throughout days of the experiment, as shown in figure 1. Considering that our data comes from three years of experiment, and for each year the experiment lasts for 5 weekdays, the data points are mostly evenly distributed within each user. We see that each participant falls within one year, hence, why in figure 1 we see that there are no questionnaires on every single day (the first 5 days are in 2022, the next in 2023, the next in 2024).

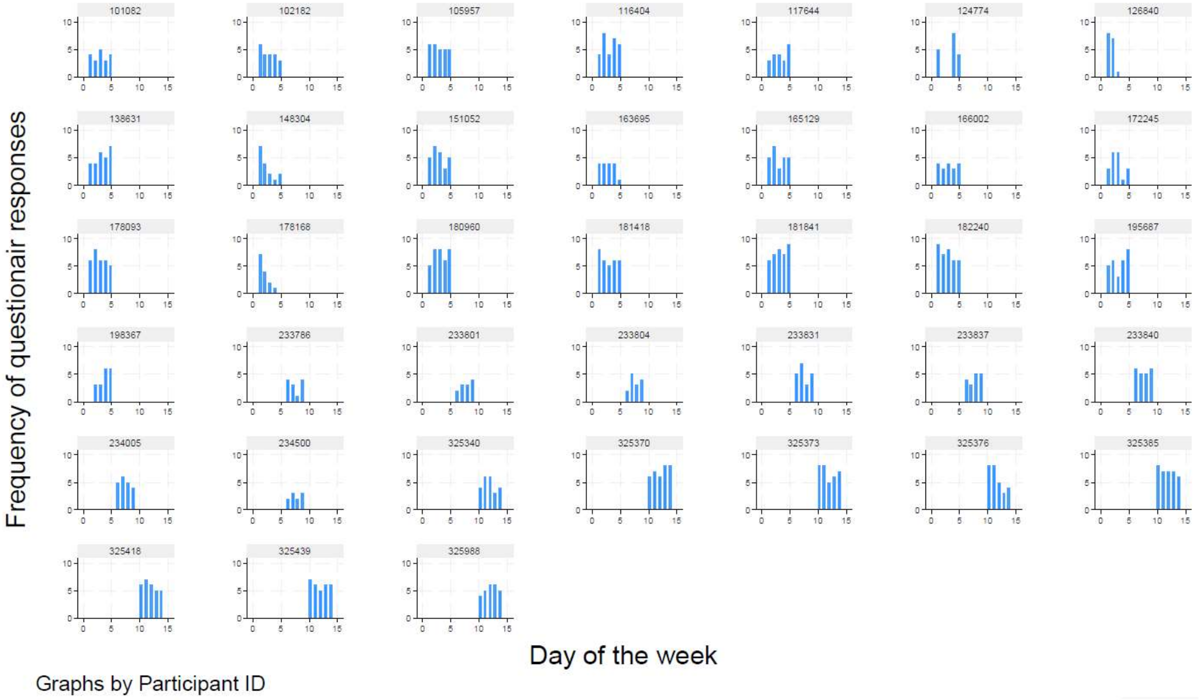


Figure 1

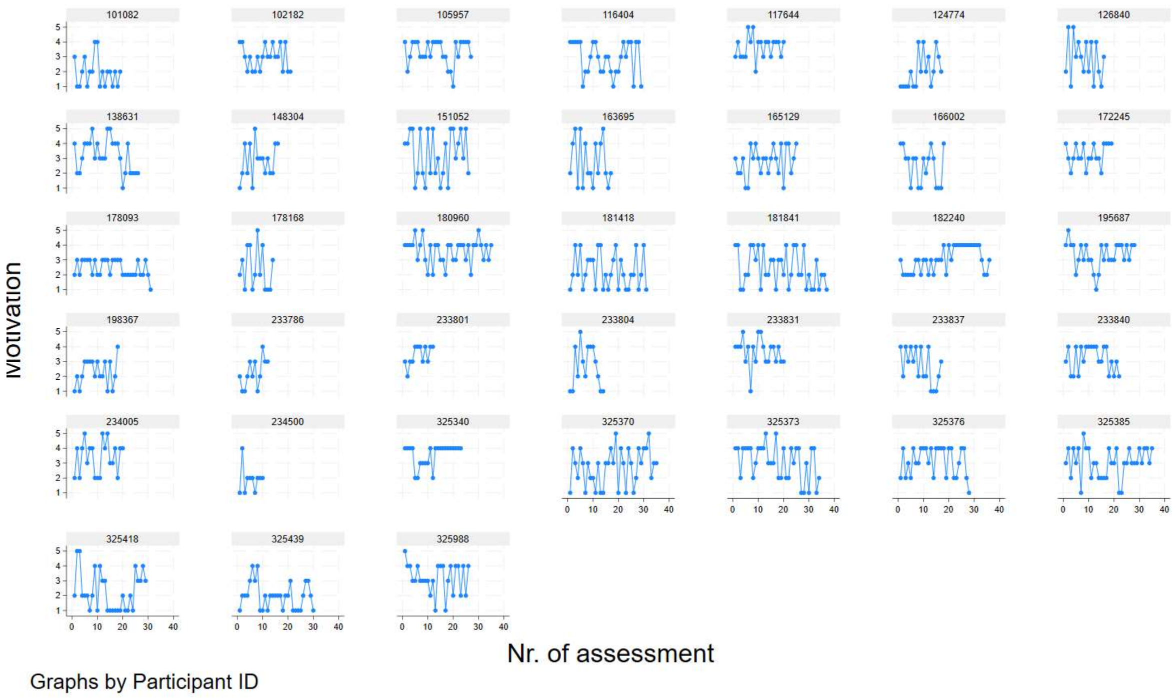
Frequency of compliance with questionnaires per participant
** There are three distinct clusters of data that correspond to the year of data collection. This is why each participant only answered questionnaires on a specific portion of days.*

After ensuring that there were no major distribution problems with the data that might influence further processing, we inspected the changes of each participant's data throughout the days and assessment moments.

For the relationship between motivation and assessment moment as shown by figure 2, the pattern varied between participants and seems not to have a clear overall relationship within it.

The results for vitality and energy level show a more interesting result, see figure 3. The distribution largely varied between participants while there is still a large proportion of the data could remain at a rather static level throughout data entries, which could lead to further analysis about whether it is connected to with any other data.

The course of motivation level **Figure 2**



The course of motivation over the assessment moments per participant

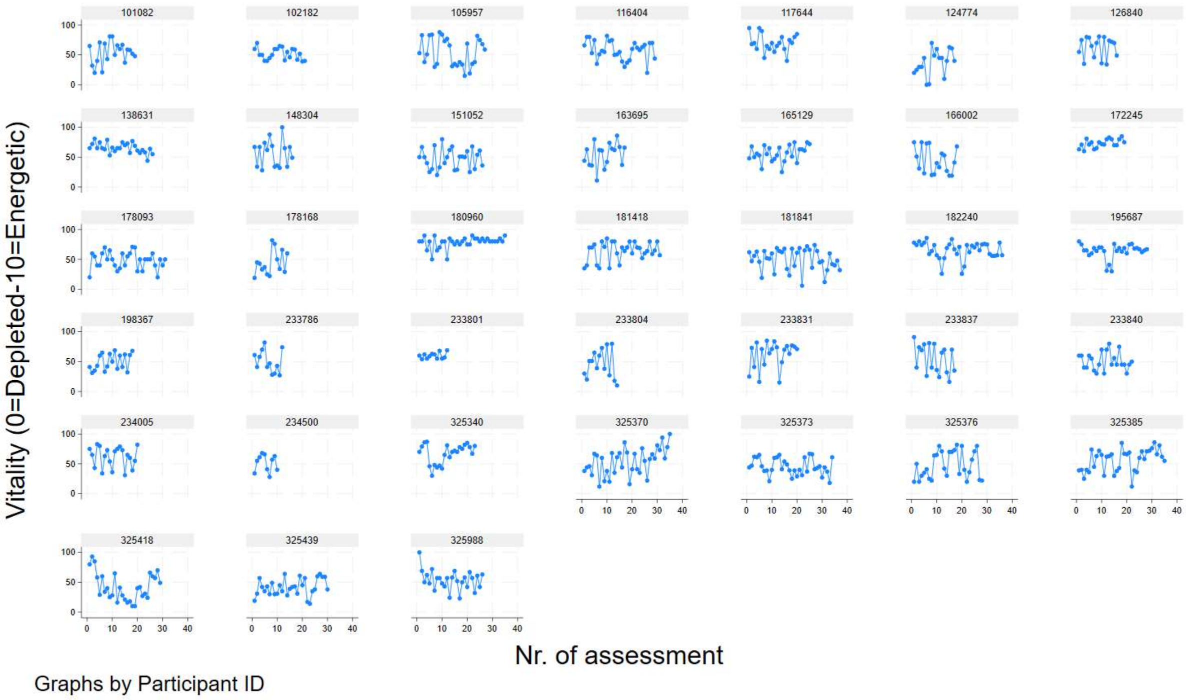


Figure 3

The course of energy levels (vitality) over the assessment moments per participant

For the stress level (see figure 4), we noticed that there seems to be a different tendency within the change of data for different participants. For some participants, the average stress level would gradually increase with higher numbers of assessment while for some other participants the stress level would gradually decrease, and for a third category of participants it seems relatively stable.

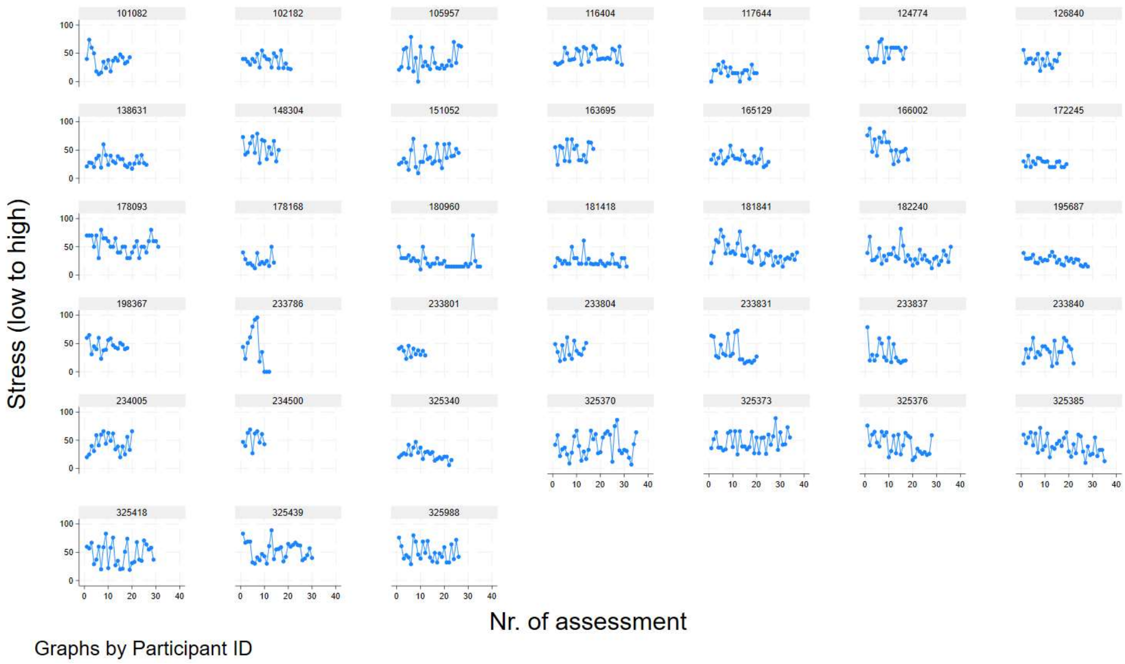


Figure 4

The course of stress level over the assessment moments per participant.

Finally, for happiness level, it shows a similarity with energy level in that some participants could remain at a more static level with an increasing number of assessments while others would vary a lot in their happiness levels and could be considered to be more unstable (see figure 5).

After we visualized the correlation between all pairs of standardized variables. In order to make it possible to easily go back to the original variable if required, we used small caps for standardized variables and kept all variables rather than recoding or discarding some variables.

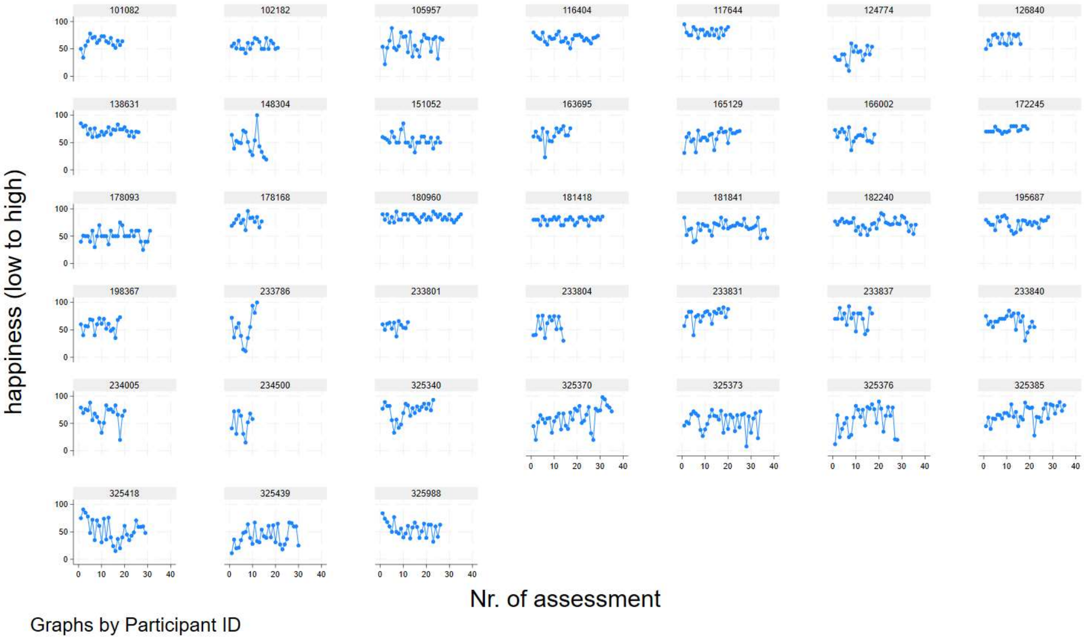


Figure 5

The course of happiness level over the assessment moments per participant

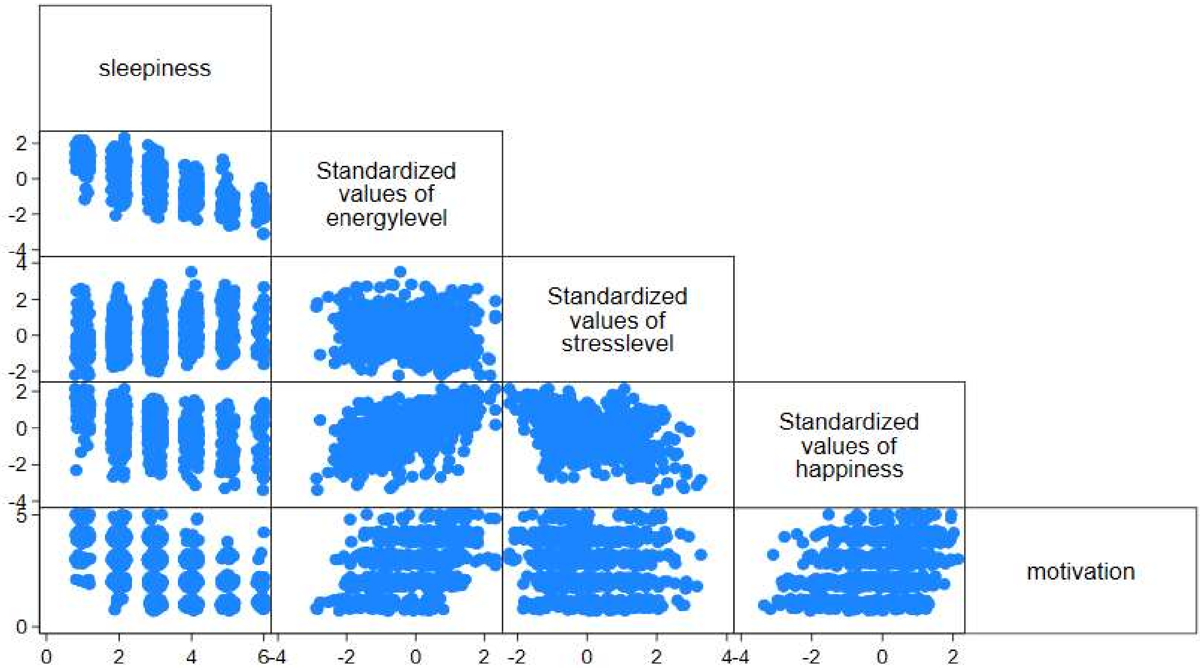


Figure 6

The correlations between all the pairs of variables

Analysis of Happiness and Energy

Our variables of interest are happiness and energy, as these two seem to be highly correlated; we would be interested to see if there are interindividual differences in the relationship between these two variables. After deleting some data from participants who had very low compliance, we are left with 893 observations. The mean for happiness is 63.5 (a little better than neutral, but not extremely happy), and the standard deviation is 17.1. The mean for energy is 55.4 (neither energetic nor depleted) with a standard deviation of 19.2.

As a first step, we ran an empty mixed model using the **mixed happiness** command. Next, we added a random intercept for each individual student by adding **|| student:** to the command. This yielded a significant result ($p<0.001$), meaning that the model is improved by adding a random slope on the student level. Intuitively, this makes sense: everyone has a different baseline level of happiness, where some people are generally happier than others.

After running this first unconditional model, we used the **estat icc** command to find the intra-class correlation and next stored the predicted residuals to check them for normality using the Shapiro-Wilk test. Unfortunately, the residuals were not normally distributed ($p<0.05$). However, in our eyes it did not make much sense to transform the variables, because it would become more difficult to interpret the results later.

We were interested in the relationship between someone's energy levels and their happiness at a given moment, so next we added [energylevel] as a predictor variable in the mixed model, first without using a random intercept. We stored the estimates for comparison with the random intercept model; then we ran the model with a random intercept for each student. In figure 7, we can see the model applied to every student. In figure 8, we see the deviation from the overall intercept per student. To make the graph more readable, we used the variable [studentnr], a number between 1 and 40, rather than the [Participant_ID], since the latter resulted in overlap between different students because of the size of the numbers.

In figure 7 it indeed becomes clear that some students are indeed happier at a low energy level than others. We do see a positive relationship between energy and happiness, where the more energy people have, the happier they feel. The results of the mixed conditional model with random intercepts for students suggests that for every extra point someone scores on the energy scale, they get 0.46 extra points on the happiness scale.

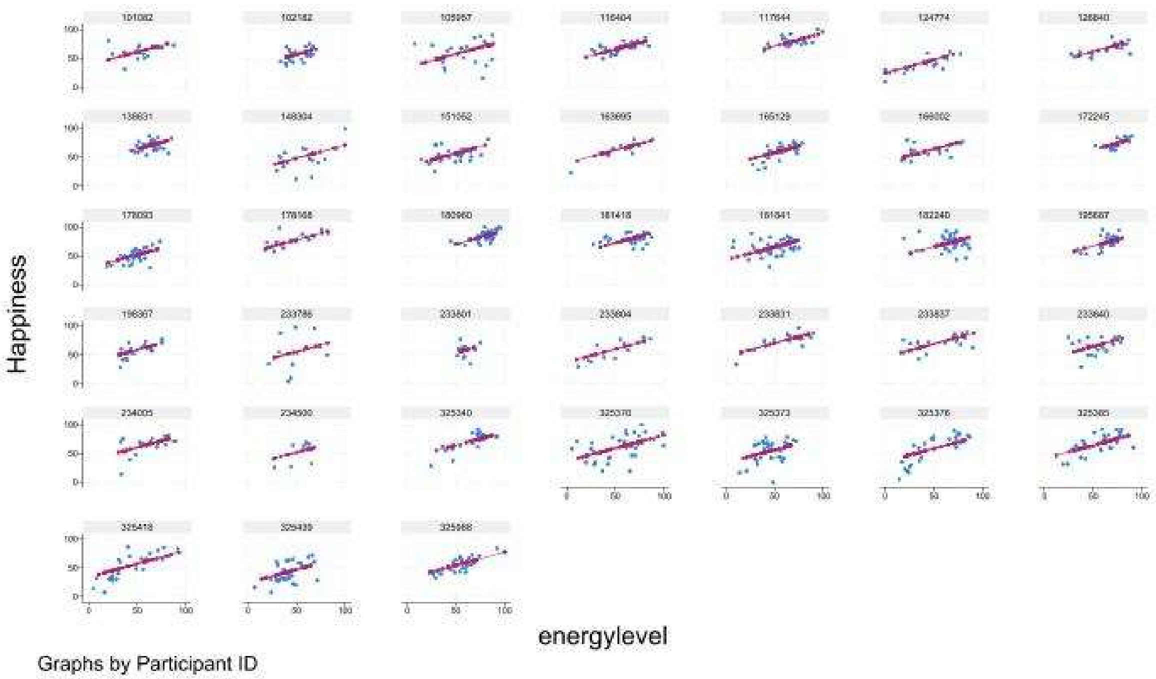


Figure 7

Mixed 2-level model with random intercepts fitted per student for the relationship between happiness and energy level

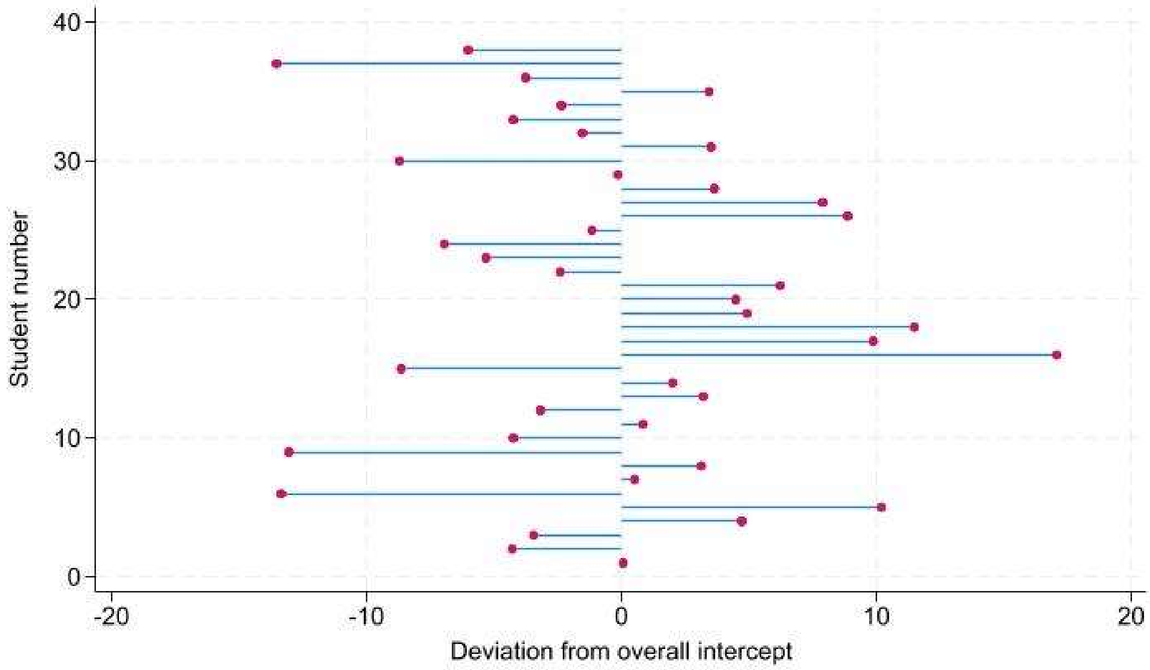


Figure 8

The deviation from the overall intercept per student

We see in figure 8 that there are quite some deviations from the overall intercept, so it makes sense to include random intercepts also from this point of view.

Next, we added a random slope to the model to see if this would be wise and logical. As $p<0.05$ it is probably wise to add random slopes for students as well, as some students' happiness can be more influenced by their energy level than others' - in other words, there might be different levels of stability for different participants. In figure 9 we can indeed see that there are some subtle differences in slopes between participants, where some slopes are steeper than others. However, in all cases, the relationship between happiness and energy level is a positive one.

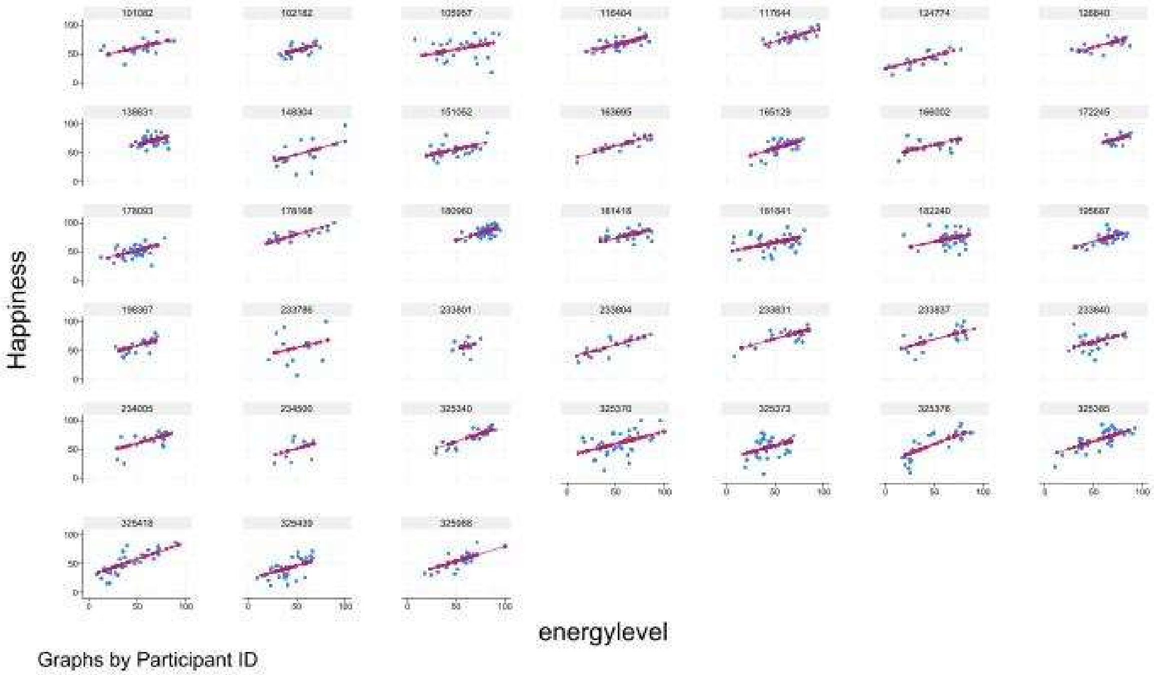


Figure 9

The mixed 2-level model with random intercepts and random slopes fitted per student for the relationship between happiness and energy level

In figure 10 we see there are different deviations from the slope, some of them quite large and some smaller, suggesting it is indeed wise to include both random intercepts and random slopes in our model.

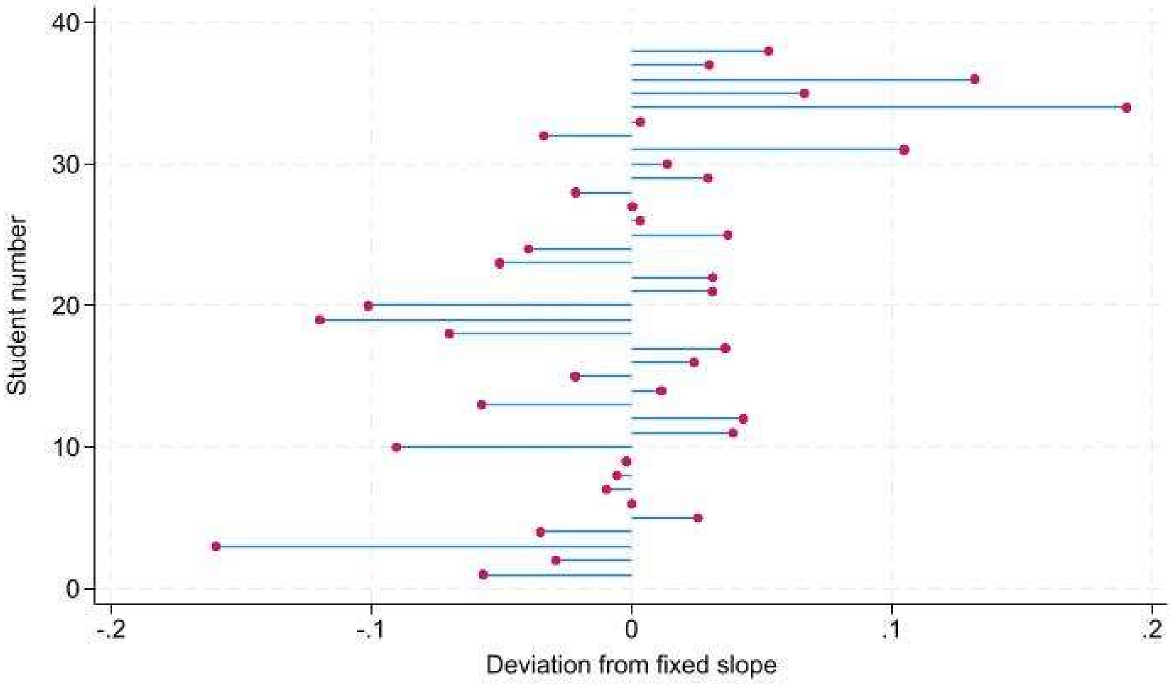


Figure 10

The deviations from the fixed slope per student

It would also be possible to use standardized scores for the analysis, but we chose not to do so because both our variables range between 0 and 100 and make most sense to interpret without transformation.

However, the analysis we discussed above is about 2-level model, while the data could also be described as a 3-level model, which we ignored so far. To find out the performance of 3-level model, we will use unconditional models (null models) and the results as shown below:

```
. mixed happiness || student: ||day:                              //3-level model

Performing EM optimization ...

Performing gradient-based optimization:
Iteration 0:   Log likelihood = -3614.4664
Iteration 1:   Log likelihood = -3614.4664

Computing standard errors ...

Mixed-effects ML regression                          Number of obs = 893

        Grouping information
        -------------------------------------------------------------
                        |     No. of       Observations per group
         Group variable |     groups    Minimum    Average    Maximum
        ----------------+--------------------------------------------
                student |         38         10       23.5         37
                    day |        176          1        5.1          9
        -------------------------------------------------------------

                                                     Wald chi2(0)    =     .
Log likelihood = -3614.4664                          Prob > chi2     =     .

------------------------------------------------------------------------------
   happiness | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _cons |   62.95182   1.716745    36.67   0.000     59.58706    66.31658
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects parameters  |   Estimate   Std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
student: Identity            |
                  var(_cons) |   92.84422   25.82313      53.82783    160.1411
-----------------------------+------------------------------------------------
day: Identity                |
                  var(_cons) |   53.84826   10.57265       36.6476     79.1221
-----------------------------+------------------------------------------------
               var(Residual) |   145.3241   7.694326      130.9996     161.215
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 370.74                Prob > chi2 = 0.0000
```

Figure 11

The results for the 3-level model

The results suggest that it is better to use a 3-level model since $p<0.05$. This makes sense intuitively speaking since happiness and energy levels can differ significantly between days (just based off our own experiences).

Finally, we tested the performance of centered score of the data. With the random intercepts for students and day (nested in student), we have the results of $p < 0.05$. Furthermore, the normality is also rejected by the swilk test. Due to the reasons above, we are not going to use the centered score and the results as shown below.

```
Mixed-effects ML regression                     Number of obs =    893

        Grouping information
        -------------------------------------------------------------
                        |     No. of       Observations per group
         Group variable |     groups    Minimum    Average    Maximum
        ----------------+--------------------------------------------
                student |         38         10       23.5         37
                    day |        176          1        5.1          9
        -------------------------------------------------------------

                                                Wald chi2(2)   = 176.30
Log likelihood = -3441.5615                     Prob > chi2    = 0.0000

------------------------------------------------------------------------------
              happiness | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
------------------------+-----------------------------------------------------
             Energy_CMC |   .4292749   .0389246    11.03   0.000     .3529841    .5055656
Energy_ClusterMean_centered |   .8748655   .1182105     7.40   0.000     .6431772    1.106554
                  _cons |   63.26132   1.081278    58.51   0.000     61.14206    65.38059
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects parameters  |   Estimate   Std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
student: Independent         |
              var(Energy_CMC) |   .0348386   .0122157      .0175226    .0692665
                   var(_cons) |   34.47678   10.71838      18.74559    63.40948
-----------------------------+------------------------------------------------
day: Identity                |
                   var(_cons) |   22.38771   5.996484      13.24396     37.8444
-----------------------------+------------------------------------------------
                var(Residual) |   101.3299   5.538295      91.03623    112.7875
------------------------------------------------------------------------------
LR test vs. linear model: chi2(3) = 186.71              Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

. predict residuals_G, res

. swilk residuals_G                       // normality is rejected

                    Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
 residuals_G |        893    0.97998     11.391     5.997    0.00000
```

Figure 11

The results for the centered-score model

Part B

Student A – participant 325370 – Robin Flaton

This participant answered the prompts 35 times, with a minimum of 6 times per day and a maximum of 8 – their compliance was high. Both energy level and happiness level cover a broad range of values: energy levels between 12 and 100 (mean = 54.1, standard deviation =

22.4) and happiness between 20 and 98 (mean = 60.9, standard deviation = 18.9). Compared to the group statistics, their values are similar, although a fraction lower on average on both variables. Their standard deviation is a fraction higher.

First, we created a scatterplot for this specific participant to get a feeling for what the relationship between their happiness and energy level looked like. As expected, we see a positive relationship, see figure 11.

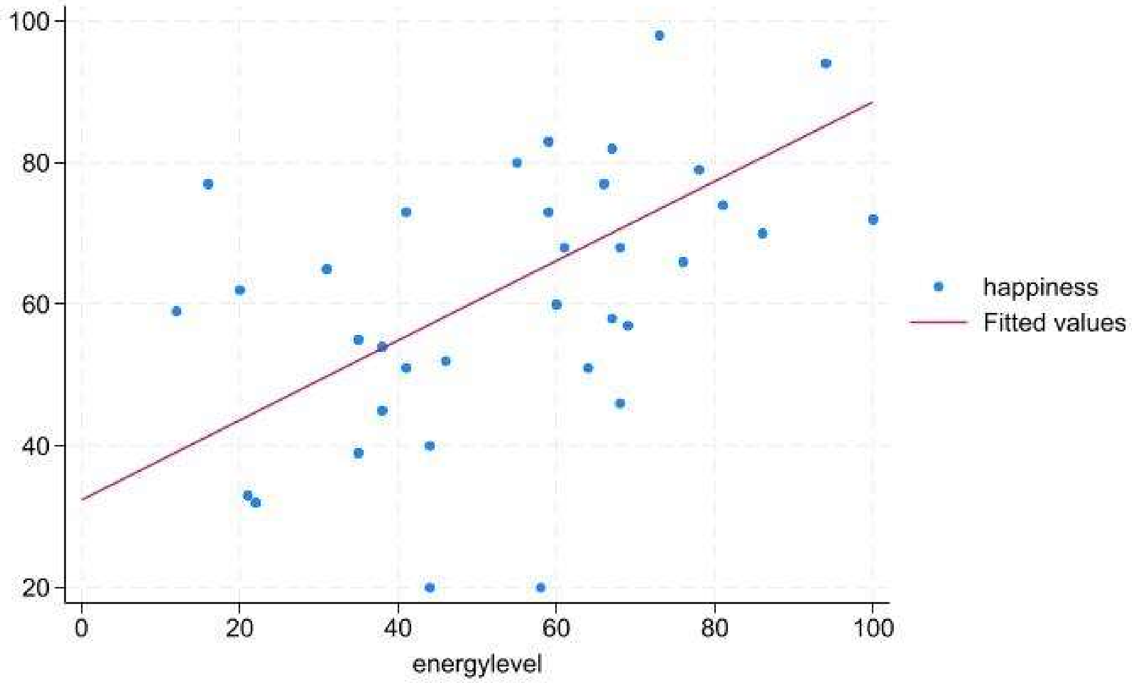


Figure 12: scatter plot for participant 325370

Then we ran a mixed model without random intercepts to see the parameters we are working with. For this student, the coefficient is 0.39, which is lower than the group level coefficient we found before (0.46), meaning that their energy level influences their happiness level a little less than for the average student.

Next, we added a random intercept for [day] to see if the relationship between [happiness] and [energylevel] differed over days – one could consider the data to be nested within days – and compared the results with the previous analysis. Adding a random intercept did not improve the model further (p=0.237), so it does not make sense to use a 2-level model for this individual student.

Student B – participant 325988 – Han Xiao

Within the experiment, participant 325988 answered the prompt questions 26 times. The answers from the participant distributed with a high-level of consistency across the five days of experiment, there are two days that participant answered 5 times, two days answered 6 times and one day answered 4 times.

For the distribution of happiness level and energy level, the energy level of the participant shows a higher level of variance as it distributed from 23 to 100 while the happiness level distributed from 32 to 84. That fact can also be reflected by the slandered division of these two factors. The std for happiness is 13.16 while the std for energy level is 16.13. When it comes to the mean value, the result for these two factors is closer, the mean value for happiness is 55.85 and the mean value for energy is 53.7.

To start with the analysis of individual data, we generated the scatterplot for the data of participant 325988 as shown below, which reveals the positive relationship between happiness and energy level of the participant.

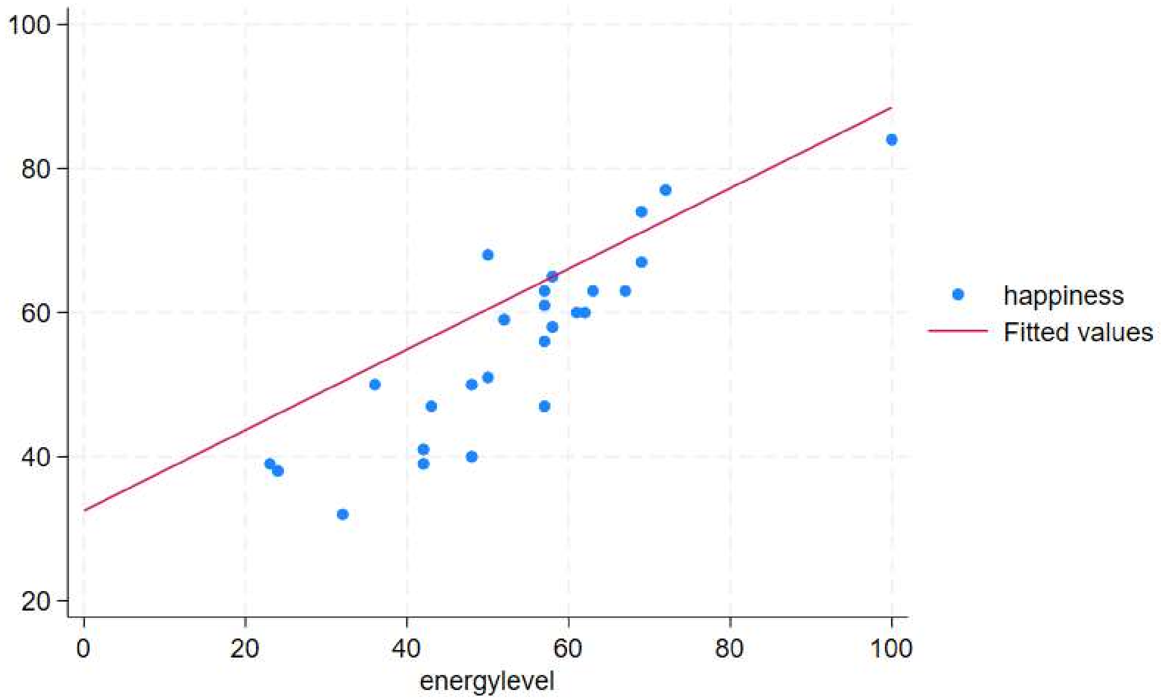


Figure 13: scatter plot for participant 325988

With the result above, we can run a naive regression model first for the data of the participant. With the results below, we can see that there are 76.8% of variation of happiness could be explained energy level as the R-squared is 0.768. The coefficient is 0.715, which means that happiness would increase 0.715 when energy level increases 1. As the coefficient is larger than the group level coefficient of 0.46, we can notice that participant is influenced more by the energy level compared to the normal condition.

Source	SS	df	MS			
				Number of obs	=	26
				F(1, 24)	=	79.48
Model	3325.2424	1	3325.2424	Prob > F	=	0.0000
Residual	1004.14222	24	41.8392591	R-squared	=	0.7681
				Adj R-squared	=	0.7584
Total	4329.38462	25	173.175385	Root MSE	=	6.4683

happiness	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
energylevel	.714964	.0801982	8.91	0.000	.5494431	.8804849
_cons	17.43059	4.491952	3.88	0.001	8.159658	26.70152

Figure 14: naive regression results for participant 325988

After that, we could run a mixed model with random intercepts of day to find out what is the of that relationship over days. However, the result shows that by adding a random intercept of day, it would not improve the mode as the p = 1.0. Due to the reason above, we would not apply 2-level model for this participant.

```
Performing EM optimization ...

Performing gradient-based optimization:
Iteration 0:  Log likelihood = -84.515571
Iteration 1:  Log likelihood = -84.393907
Iteration 2:  Log likelihood = -84.391703
Iteration 3:  Log likelihood = -84.391703

Computing standard errors ...

Mixed-effects ML regression                     Number of obs     =     26
Group variable: day                             Number of groups  =      5
                                                Obs per group:
                                                              min =      4
                                                              avg =    5.2
                                                              max =      6
                                                Wald chi2(1)      =  86.10
Log likelihood = -84.391703                     Prob > chi2       = 0.0000

------------------------------------------------------------------------------
   happiness | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
 energylevel |    .714964   .0770519     9.28   0.000      .563945    .8659829
       _cons |   17.43059   4.315727     4.04   0.000      8.97192    25.88926
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects parameters  |   Estimate   Std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
day: Identity                |
                  var(_cons) |   1.31e-08     .00003             0           .
-----------------------------+------------------------------------------------
               var(Residual) |   38.62085    10.7115      22.42544    66.51241
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 0.00          Prob >= chibar2 = 1.0000
```

Figure 15: mixed regression results with intercept for participant 325988

Student C – Participant 325373 – Tom Coenen

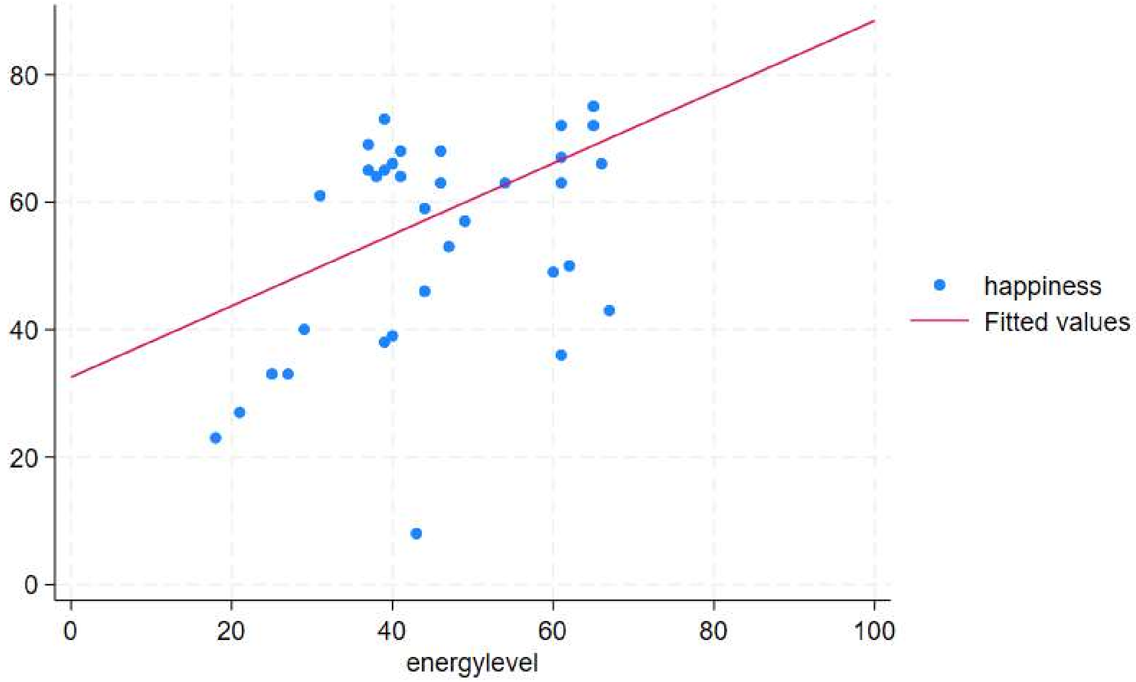
Participant 325737 answered the questionnaire 34 times, between 5 and 8 times a day. [happiness] and [energylevel] both have a broad range of values for this participant: [happiness] between 8 and 75 (*M* = 54.06, *SD* = 16.70), and [energylevel] between 18 and 67 (*M* = 45.41, *SD* = 13.72). Both variable means are lower than the group average of [happiness] (*M* = 63.51, *SD* = 17.06), and [energylevel] (*M* = 55.41, *SD* = 19.20).

To examine the correlational relationship between happiness and energy level for participant 325373, a scatterplot was created (figure 15). Figure 15 shows a positive relationship between happiness and energy level, which is expected.

Next, an empty mixed model without random intercepts was run to see the parameters of this participant’s model. The coefficient is 30.87, which is lower than the group-level coefficient, 0,45. This participant's energy level, therefore, influences the happiness less compared to the average student.

Figure 15

A Scatterplot showing the correlational relationship between happiness and energy level for participant 325373.



To continue the analysis, a random intercept was added for [day] and later for [timeonday] as the data could be nested in both categories. Both models with [day] or [timeonday] as clustering variables, using the **mixed** command, were not better than a normal regression based on the LR test (p = 1.00) for both models. It does not make sense to use a 2-level model for this individual student.

Conclusion

As can be deduced from the group analysis in comparison to some individual analyses, there are indeed different models that work for individuals vs. the whole group. Whereas a model that uses data nested within participants within days is ideal for the whole group analysis, adding the day level for participants 325370 and 325988 does not improve the model (p>0.05, for the latter even p=1). This underlines the importance of making a distinction between individuals and not blindly applying rules found for a group of participants to single persons, which we think is one of the main takeaways of this course.