



TwinLife

TwinLife Technical Report Series

No. 16, June 2025

Documentation *TwinLife* Data:

Global Physical Activity Questionnaire (GPAQ)

F2F3 & F2F5

v2.0.0

by Elena T. T. Dang, Christoph H. Klatzka

christoph.klatzka@uni-saarland.de



Elena T. T. Dang, Christoph H. Klatzka

Documentation *TwinLife* Data: Global Physical Activity Questionnaire (GPAQ) F2F3 & F2F5 v2.0.0

TwinLife Technical Report Series No. 16

Project TwinLife “Genetic and social causes of life chances”

Bielefeld, June 2025

TwinLife Technical Report Series

General Editors: Christian Kandler, Bastian Mönkediek, and Frank M. Spinath

ISSN 2512-403X

Unless otherwise noted, this publication is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA). For more information see:

<https://creativecommons.org/licenses/by-nc-sa/4.0/> and <https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>

This publication has been funded by the German Research Foundation (DFG).

TwinLife Technical Reports are refereed scholarly papers. Submissions are reviewed by the general editors before a final decision on publication is made.

The Technical Report Series is a forum for presenting technical works (e.g., data documentation, field reports) in progress. Comments on the manuscript should be addressed directly to the author(s).

The papers can be downloaded from the project website:

<https://www.twin-life.de/twinlife-series>

TwinLife “Genetic and social causes of life chances”

University of Bielefeld

Faculty of Sociology

PO Box 100131

D-33501 Bielefeld

Germany

Phone: +49 (0)521 106 3563

Email: twinlife@uni-bielefeld.de

Web: <https://www.twin-life.de>

Table of content

Introduction	2
Measurement description	2
Adapted items in TwinLife study	2
Corrections of extreme / implausible values and handling missing data	3
TwinLife Dataset: generating variables	5
(1) time spent in physical activity in min per week (pac0400, pac0410)	5
(2) level of physical activity (pac0401, pac0411)	9
Limitations	9
References	11
Appendix A	13
Appendix B	14

Introduction

This technical report provides an overview of the physical activity assessment in the TwinLife study. The report begins with a description of the measurement and correction of the data, followed by an explanation of the procedure for generating variables that indicate physical activity in the TwinLife dataset. These generated variables are included in the data release starting with v8.0.0, and relevant information on how to use them are given in this report. Lastly, limitations and further aspects regarding the physical activity data are discussed. Further details can be found in the appendices.

Measurement description

The Global Physical Activity Questionnaire (GPAQ) developed by the World Health Organization (WHO) is a standardized tool to measure physical activity in face-to-face interviews (Armstrong & Bull, 2006). The final version of the GPAQ consists of 16 items (P1-P16) and collects information on physical activity participation in three domains: activity at work (P1-P6), travel to and from places (P7-P9), and recreational activities (P11-P15), as well as on sedentary behavior (P16) (WHO, n.d.). The assessment includes frequency and intensity of physical activity in different settings. To analyze the GPAQ data, the developers recommend calculating the Metabolic Equivalents (METs), a commonly used unit for expressing the physical activity intensity (WHO, n.d., p. 3). One MET is defined as the ratio of a person's working metabolic rate relative to the resting metabolic rate equivalent to a caloric consumption of 1 kcal/kg/hour (WHO, n.d., p. 3). However, the use of METs and its existing guidelines are debated (see Byrne et al., 2005; Lavie & Milani, 2007; deJong, 2010).

Studies on the GPAQ's reliability and validity reveal moderate to substantial reliability coefficients (Kappa .67 to .73; Spearman's rho .67 to .81) and a moderate relationship between the International Physical Activity Questionnaire (IPAQ) and GPAQ for concurrent validity (Spearman's rho .45 to .57) (Bull et al., 2009). A systematic review of 26 publications (Keating et al., 2019) found good reliability for the overall physical activity (Spearman's rho 0.58 to 0.89). The reviewed studies used accelerometers, pedometers, and physical activity log to examine the concurrent validity for work-related physical activity (Spearman's rho -0.03 to 0.50), transport-related physical activity (Spearman's rho 0.04 to 0.49), and leisure-related physical activity (Spearman's rho 0.02 to 0.41) (Keating et al., 2019). Physical activity measured with the GPAQ is associated with age (Wallmann-Sperlich & Froboese, 2014; Mogre et al., 2015), body mass index (BMI) (Liu et al., 2018), and depression (Rutherford et al., 2022).

Adapted items in TwinLife study

For the purposes of the TwinLife study, the GPAQ items were adapted to provide a rough measure of time spent with physical activity. The scales were modified as follows: In each of the three domains (work, commuting by foot/bike, and leisure), the questions were shortened

to two items capturing frequency and excluding intensity of physical activity (no distinction was made between lighter and heavier physical activity). The response format was adapted to assess the number of days per week, hours, and minutes spent on average in physical activity for these domains. The original additional item on sedentary behavior is not included. An overview of the adapted items is attached in [Appendix A](#).

Physical activity was assessed in data collection wave 3 of the face-to-face interviews (F2F3, 2018-2020) as part of the household interview (PAPI) and in wave 5 (F2F5, 2022-2024) in a hybrid format (PAPI or online questionnaire). Participants of ages 17 and older were asked about their physical activity in all three domains. If participants did not have an occupation at that time, they were instructed to state their non-working status and answer the remaining questions. Younger participants (aged 10 to 16) were asked about their activity only in the domains of transport and leisure. The F2F3 data includes GPAQ data from 6,796 cases, the F2F5 data includes data from 6,161 cases.

Corrections of extreme / implausible values and handling missing data

According to the WHO's GPAQ analysis guide regarding missing data (p. 9), the GPAQ data from F2F3 and F2F5 were filtered as follows:

- A common mistake is to indicate the minutes instead of the hours in the hour's variable. So, if the hours variables have a value of 15, 30, 45 or 60, then the value should be changed to the minutes variable if that one is 0 (counts as data recording error)
- Cases should be removed if:
 1. The combination of hours and minutes is higher than 16 hours (960 minutes) daily for one of the domains
 2. Total time spent with physical activity in total for the week is higher than 6720 minutes (16 hours per day on average)
 3. The case has implausible values (e.g. over 7 days per week)
 4. The answers are inconsistent (e.g. 0 days but specified hours)

Table 1.

Cases excluded due to extreme values on a GPAQ variable.

PID	Extreme value in the ...	Data collection
146669300	... work domain with 1,200 minutes per day	F2F3
219414300	... work domain with 1,200 minutes per day	F2F3
248025300	... total time spent per week with 6,900 minutes	F2F3
261505400	... total time spent per week with 6,960 minutes	F2F3
262361002	... leisure domain with 1,020 minutes per day	F2F3
292148300	... work domain with 1,200 minutes per day	F2F3
295902400	... total time spent per week with 6,960 minutes	F2F3
317281001	... commuting domain with 1,200 minutes per day	F2F3
321911300	... total time spent per week with 7,110 minutes	F2F3
333166110	... total time spent per week with 6,930 minutes	F2F3
423696002	... total time spent per week with 7,980 minutes	F2F3
430661001	... total time spent per week with 6,960 minutes	F2F3
442910400	... work domain with 1,200 minutes per day	F2F3
463853300	... work domain with 1,080 minutes per day	F2F3
471553300	... work domain with 1,200 minutes per day	F2F3
473571400	... total time spent per week with 7140 minutes	F2F3
481810001	... work domain with 1,080 minutes per day	F2F3
164528400	... total time spent per week with 7,000 minutes	F2F5
195502001	... commuting domain with 1,190 minutes per day	F2F5
195502300	... total time spent per week with 6,741 minutes	F2F5
212787002	... leisure domain with 1,080 minutes per day	F2F5
213171002	... work domain with 1,449 minutes per day	F2F5
221286001	... commuting domain with 1,200 minutes per day	F2F5
222088300	... total time spent per week with 9,660 minutes	F2F5
234013001	... work domain with 1,020 minutes per day	F2F5
266649002	... total time spent per week with 7,920 minutes	F2F5
268651400	... work domain with 1,440 minutes per day	F2F5
332200001	... commuting domain with 1,200 minutes per day	F2F5
333657001	... commuting domain with 1,200 minutes per day	F2F5
351076001	... work domain with 1,320 minutes per day	F2F5
377590002	... total time spent per week with 8,280 minutes	F2F5
397380002	... work domain with 1,230 minutes per day	F2F5
423643002	... total time spent per week with 10,318 minutes	F2F5
426992001	... commuting domain with 1,200 minutes per day	F2F5
485944300	... work domain with 1,440 minutes per day	F2F5

1568 participants for F2F3 and 957 participants in F2F5 did not provide enough data for the generation of scores (e.g., the hours and variable were missing; the number of days was missing; days and hours/minutes combination made no sense) and were set as missing ("-82: information incomplete") in the final scores. 8 cases with a value exceeding 16 hours (960 minutes) per day in one domain (see Table 1) and another 8 cases due to exceeding 6720 minutes per week in total activity have been removed from the 6,796 cases and set as missing ("-83: implausible value") in F2F3. Similarly, 18 cases for F2F5 were handled this way. In total there is complete and valid data from 5,213 participants in F2F3 and valid data from 5,186 participants in F2F5. There are 13 participants in F2F3 and 34 participants in F2F5 with valid data on the GPAQ variables but with no information on the age variable, however other information clearly indicated that they were adults, so they were treated accordingly.

TwinLife Dataset: generating variables

There are three ways to generate an indicator for physical activity or inactivity, as recommended in the GPAQ analysis guide (WHO, n.d., p. 14):

1. estimate a population's mean or median physical activity with
 - a. the continuous indicator MET minutes per week. The MET (Metabolic Equivalent) measures a person's working metabolic rate in relation to their resting metabolic rate, with one MET representing the energy expenditure of sitting quietly, equivalent to 1 kcal/kg/hour. For GPAQ data analysis, it is estimated that caloric consumption is four times higher during moderate activity and eight times higher during vigorous activity compared to sitting quietly.
 - b. The time spent on physical activity regardless of the activity level (low, moderate, high)
2. setting up a cut-point for a specific amount of physical activity to classify a certain percentage of a population as 'inactive' or insufficiently active

Considering that no detailed information on the intensity of the activities is available, a variable with the total number of minutes spent on physical activities per week has been calculated (Option 1b).

(1) time spent in physical activity in min per week (pac0400, pac0410)

For generating the variable indicating the total amount of weekly physical activity, the following formula has been applied:

For participants aged 17 and older:

$$pac0400 = (pac0101 * 60 + pac0102) * pac0100 + (pac0201 * 60 + pac0202) * pac0200 + (pac0301 * 60 + pac0302) * pac0300.$$

For participants 11 - 16 years old:

$$pac0410 = (pac0201 * 60 + pac0202) * pac0200 + (pac0301 * 60 + pac0302) * pac0300.$$

The total time spent in physical activity per week in minutes is the sum of time spent in the three domains for individuals aged 17 or older (*pac0400*) and for participants in the age group of 11 to 16 years it is the sum of two domains (commuting, leisure; *pac0410*). Both variables are calculated by converting the hour's variables (*pac0101*, *pac0201*, *pac0301*) to minutes, adding up the minute's variables (*pac0102*, *pac0202*, *pac0302*) and multiplying with the day's variables (*pac0100*, *pac0200*, *pac0300*). The distributions of the resulting variables are depicted in Table 2, Figure 1 (adults) and Figure 2 (teenager).

Table 2.

Descriptive statistics for the *pac0400* (adults) and *pac0410* (teenagers) variables.

Variable	N	Mean	SD	Median	Range	Data collection
Pac0400	4280	835.48	939.75	480	0-6,600	F2F3
Pac0410	933	569.37	434.99	460	15-3,900	F2F3
Pac0400	4375	828.06	930.76	480	0-6660	F2F5
Pac0410	811	557.77	526.22	435	0-5750	F2F5

Figure 1

Frequencies of physical activity time per week for age group 17 and older.

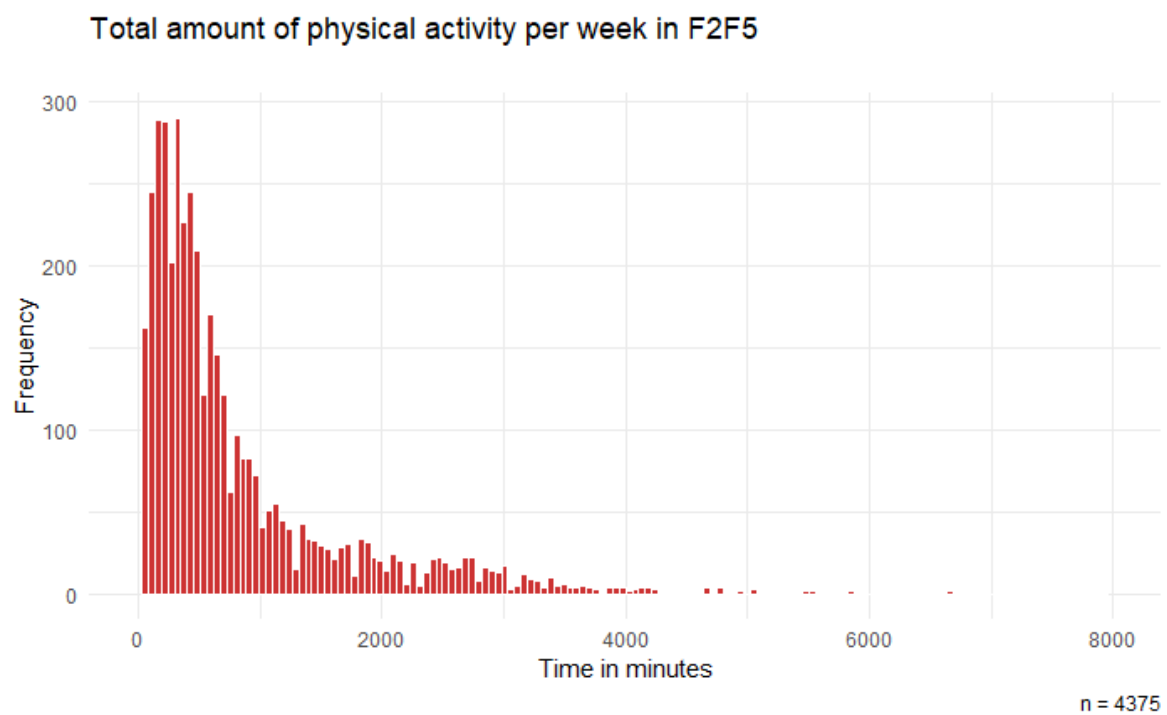
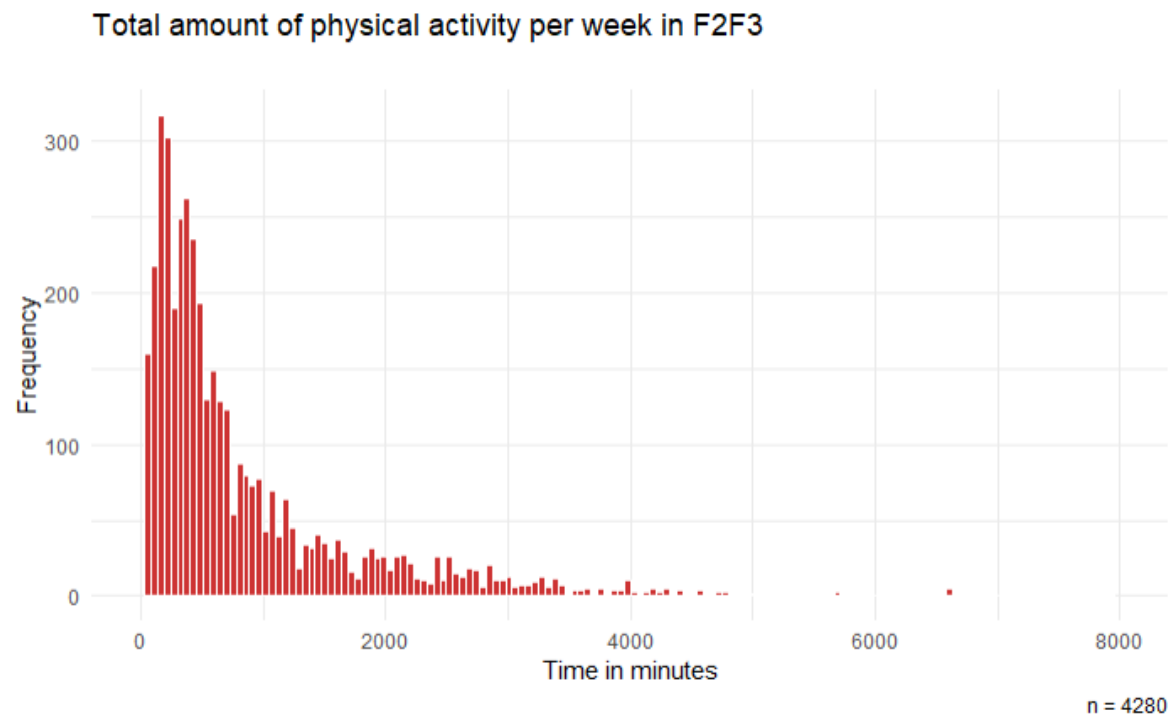
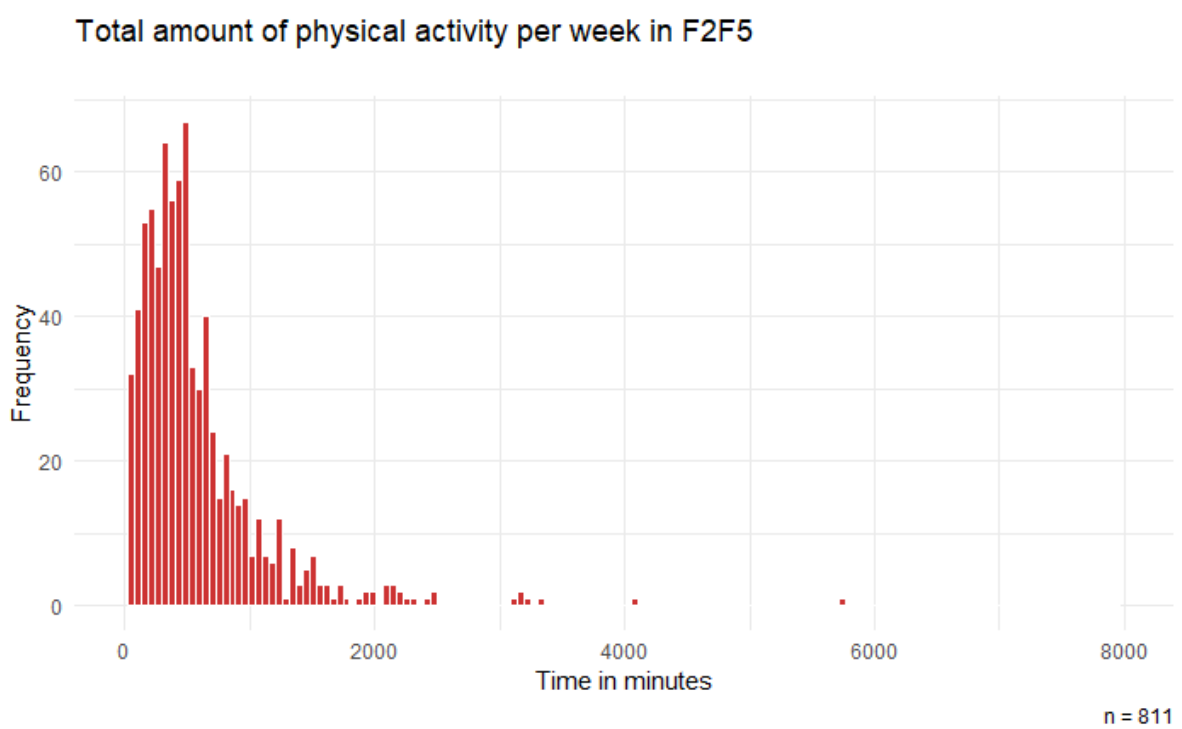
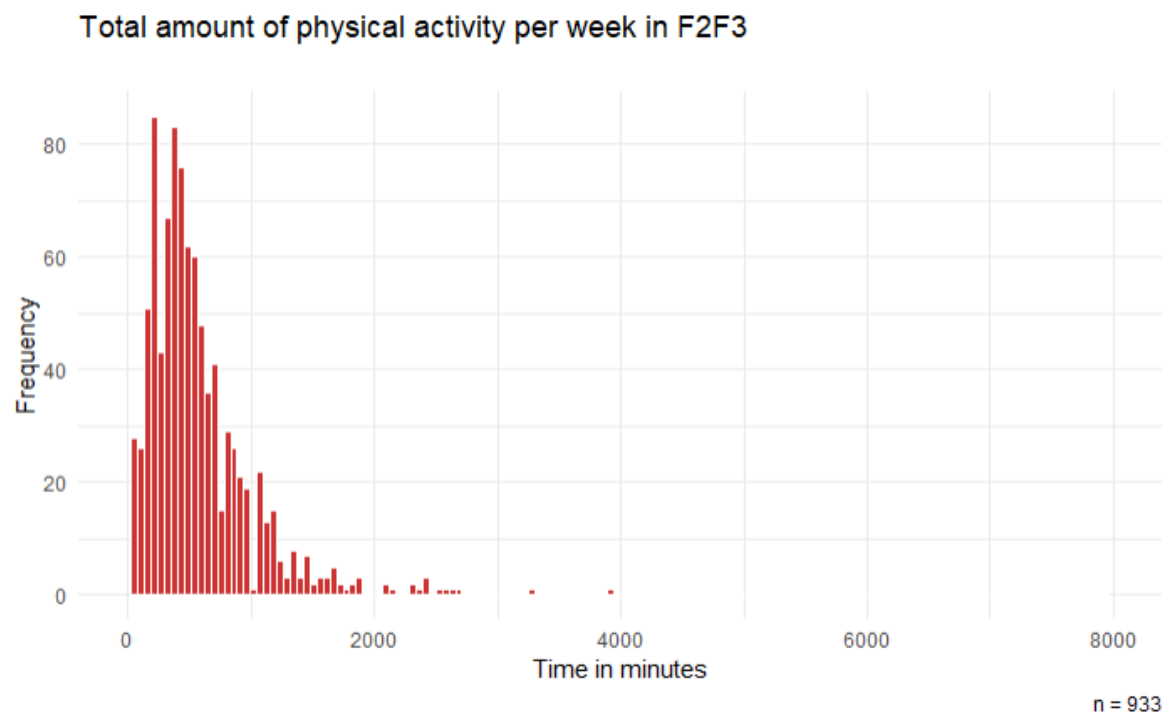


Figure 2

Frequency of physical activity time per week for age group 10 to 16.



(2) level of physical activity (*pac0401*, *pac0411*)

A categorial indicator can be generated through categorizing the total time spent on physical activity per week according to the WHO recommendations: throughout a week adults should do at least 150 minutes of moderate-intensity physical activity or 75 minutes of vigorous-intensity physical activity (WHO, n.d., p. 14). For children and adolescents aged 5 to 17 years it is recommended to do physical activity at least an average of 60 minutes per day of moderate-to-vigorous intensity, that means at least 420 minutes throughout the week. In this dataset the cut-point for classifying physical activity is set at 150 minutes per week for individuals aged 17 or older (*pac0401*) as there is no information on the intensity of physical activity available and this cut-off would not overestimate the inactive group. The cut-point for classifying physical activity for individuals under 17 years (*pac0411*) is set at 420 minutes per week. Descriptive statistics for these generated variables can be seen in Table 3.

Table 3.

Descriptive statistics for the *pac0401* (adults) and *pac0411* (teenagers) variables.

Variable	N	Sufficient Activity (1)	Percent	Data collection
Pac0401	4280	3736	87.3%	F2F3
Pac0411	933	539	57.8%	F2F3
Pac0401	4375	3769	86.1%	F2F5
Pac0411	811	428	52.8%	F2F5

Limitations

Some of the GPAQ's limitations are addressed as follows: Only self-reported data are assessed with the GPAQ; therefore, the accuracy of the data cannot be verified without objective data which is not available in the TwinLife dataset. General restrictions of self-reports (e.g. biased response due to social desirability) need to be considered when interpreting the data. The GPAQ measures the physical activity of a typical week, although other factors like adaptations to seasonal changes can impact the structure of a typical week leading to different data (Keating et al., 2019, p. 24). Keating et al. (2019, p. 24) also mentioned that the measurement of physical activity in the three domains can only be accurate if the participants clearly distinguish their activity related to work, transport and leisure. They further noted the

gap between the reports on work-related physical activity from working and non-working individuals, which limits tracking of physical activity among children and adolescents (Keating et al., 2019, p.24).

Please note that the items have been adapted to fit the economic needs of the TwinLife study. The physical activity data assessed in TwinLife can be used only as rough estimates of frequency of physical activity. However, the data should not be used to calculate METs because the intensity of physical activity has not been measured.

Furthermore, the time spent on physical education (PE) for individuals still enrolled in school was not assessed. However, studies with on average 11-year-old students indicated that only 19 minutes of an average PE session are used for physical activity (Kühnis et al., 2017). Nevertheless, there might be a slight bias in the classification of sufficient physical activity for participants aged 16 or younger, as time spent in PE is not explicitly included in the overall time measure.

References

- Armstrong, T., & Bull, F. (2006). Development of the world health organization global physical activity questionnaire (GPAQ). *Journal of Public Health*, 14(2), 66-70
<https://doi.org/10.1007/s10389-006-0024-x>
- Bull, F., Maslin, T. S., & Armstrong, T. (2009). Global Physical Activity Questionnaire (GPAQ): Nine Country Reliability and Validity Study. *Journal of Physical Activity and Health*, 6(6), 790–804. <https://doi.org/10.1123/jpah.6.6.790>
- Byrne, N. M., Hills, A. P., Hunter, G. R., Weinsier, R. L., & Schutz, Y. (2005). Metabolic equivalent: one size does not fit all. *Journal of Applied physiology*.
<https://doi.org/10.1152/japplphysiol.00023.2004>
- Chu, A. H. Y., Ng, S. Y., Koh, D., & Müller-Riemenschneider, F. (2015). Reliability and validity of the self- and Interviewer-Administered versions of the Global Physical Activity Questionnaire (GPAQ). *PLOS ONE*, 10(9), e0136944.
<https://doi.org/10.1371/journal.pone.0136944>
- deJong, A. (2010). The metabolic equivalent. *Acsm's Health & Fitness Journal*, 14(4), 43–46.
<https://doi.org/10.1249/fit.0b013e3181e438f9>
- Keating, X. D., Zhou, K., Liu, X., Hodges, M., Li, J., Guan, J., Phelps, A., & Castro-Piñero, J. (2019). Reliability and Concurrent Validity of Global Physical Activity Questionnaire (GPAQ): A Systematic review. *International Journal of Environmental Research and Public Health*, 16(21), 4128. <https://doi.org/10.3390/ijerph16214128>
- Kühnis, J., Eckert, N., Mandel, D., Imholz, P., Egli, S., Steffan, M., Arquint, L., & Schürpf, B. (2017). Zeitnutzung und Anstrengung im Sportunterricht: Befunde einer Querschnittsstudie auf der Primarstufe im Kanton Schwyz. *Swiss Sports & Exercise Medicine*, 65(3), 54–59.
- Lavie, C. J., & Milani, R. V. (2007). Metabolic Equivalent (MET) Inflation-Not the MET we used to know. *Journal of Cardiopulmonary Rehabilitation and Prevention*, 27(3), 149–150.
<https://doi.org/10.1097/01.hcr.0000270692.09258.6a>
- Liu, F., Wang, W., Ma, J., Sa, R., & Zhuang, G. (2018). Different associations of sufficient and vigorous physical activity with BMI in Northwest China. *Scientific Reports*, 8(1).
<https://doi.org/10.1038/s41598-018-31227-6>
- Mogre, V., Nyaba, R., Aleyira, S., & Sam, N. B. (2015). Demographic, dietary and physical activity predictors of general and abdominal obesity among university students: a cross-sectional study. *SpringerPlus*, 4(1). <https://doi.org/10.1186/s40064-015-0999-2>

- Rutherford, E. R., Vandelanotte, C., Chapman, J., & To, Q. G. (2022). Associations between depression, domain-specific physical activity, and BMI among US adults: NHANES 2011-2014 cross-sectional data. *BMC Public Health*, 22(1). <https://doi.org/10.1186/s12889-022-14037-4>
- TwinLife. (2023). *Codebook TwinLife Face-to-face survey of wave 3 (Version 7.1.0, Scientific Use File ZA6701_person_wid5)*. TwinLife. https://www.twinlife.de/documentation/images/TwinLife/Downloads/ZA6701_cod_wid5_v7-1-0.pdf
- Wallmann-Sperlich, B., & Froboese, I. (2014). Physical Activity during Work, Transport and Leisure in Germany - Prevalence and Socio-Demographic Correlates. *PLOS ONE*, 9(11), e112333. <https://doi.org/10.1371/journal.pone.0112333>
- World Health Organization. (n.d.). *Global Physical Activity Questionnaire (GPAQ) Analysis Guide*. Prevention of Noncommunicable Diseases Department. https://cdn.who.int/media/docs/default-source/ncds/ncd-surveillance/gpaq-analysis-guide.pdf?sfvrsn=1e83d571_2
- World Health Organization. (2022, October 5). *Physical activity*. <https://www.who.int/news-room/fact-sheets/detail/physical-activity>

Appendix A

Adapted GPAQ items in *TwinLife*

For further information (e.g. coding, filtering), please refer to the codebook of F2F3 (v.7-1-0, ZA6701_person_wid5).

Variable name	Variable label	Question text
pac0100	physical activity at work – number of days per week	On how many days in a usual week do you engage in physical activity at work (e.g., lifting or carrying loads, delivering on foot or bicycle, working on your knees, construction work, or digging)?
pac0101	physical activity at work - time/working day: hours	How much time do you usually spend on these activities during such a working day? [item text: hours]
pac0102	physical activity at work - time/working day: minutes	How much time do you usually spend on these activities during such a working day? [item text: minutes]
pac0200	longer distances on foot or by bike - number of days per week	On how many days in a usual week do you cover longer distances on foot or by bicycle (for example, to go grocery shopping or on the way from home to school or for a walk)?
pac0201	longer distances on foot or by bike - time / day: hours	How much time do you usually need to cover these distances on such a day? [item text: hours]
pac0202	longer distances on foot or by bike - time / day: minutes	How much time do you usually need to cover these distances on such a day? [item text: minutes]
pac0300	physical activity in leisure time - number of days per week	On how many days of a usual week do you engage in physical activity in your free time (e.g., through sports such as soccer, tennis, weight training, jogging, swimming, bicycling, or through other activities such as gardening)?
pac0301	physical activity in leisure time - time / day: hours	How much time do you usually spend on these activities on such a day? [item text: hours]
pac0302	physical activity in leisure time - time / day: minutes	How much time do you usually spend on these activities on such a day? [item text: minutes]

Appendix B

Sample Script for the variable generation of F2F3 data. R (Version 4.2.2) and RStudio (Version 2023.12.0 Build 369) have been used to generate the variables *pac0400*, *pac0401*, *pac0410* and *pac0411*.

```
library(haven)
library(dplyr)

library(tidyverse)

library(ggplot2)
library(psych)

search_var_function <- function(dataset, stamm, variables = "[0-9]{4}", suffix=" ", zeitpunkt="(f2f|cati|cov)[1-5](|_inv|_rec)?"){

  if (nchar(as.character(substitute(variables))) == 1){
    number_char <- paste0(as.character(substitute(variables)), "[0-9]{3}") # define search pattern here
  } else if (nchar(as.character(substitute(variables))) == 2){
    number_char <- paste0(as.character(substitute(variables)), "[0-9]{2}")
  } else if (nchar(as.character(substitute(variables))) == 3){
    number_char <- paste0(as.character(substitute(variables)), "[0-9]{1}")
  } else if (nchar(as.character(substitute(variables))) == 4){
    number_char <- paste0(as.character(substitute(variables)))
  } else {number_char <- paste0(as.character(substitute(variables)))}
  suchmuster <- paste0("^", as.character(substitute(stamm)), number_char, as.character(substitute(suffix)), "\\_", as.character(substitute(zeitpunkt)))
  print("This is the searching pattern: ")
  print(suchmuster)
  ergebnis_vector <- c()
  for (i in colnames(dataset)){
    if (grepl(suchmuster, i)){
      ergebnis_vector <- c(ergebnis_vector, i)
    }
  }
  if (length(ergebnis_vector) == 0) {
    print("Variables not found")
  }
  inverted <- ergebnis_vector[grep("_rec|_inv", ergebnis_vector)]
  inverted_first <- substr(inverted, 1, nchar(inverted)-4)
  ergebnis_v <- setdiff(ergebnis_vector, inverted_first)
  return(ergebnis_v)
}

# import relevant dataset (file should be in the same working directory)
f2f3_data <- read_dta("Y:/Release_7-1-0/TL_v7-1-0_Stata/ZA6701_person_wid5_v7-1-0.dta")

# add a suffix to f2f3 data
f2f3_data <- f2f3_data %>% rename_at(vars(everything()), ~ paste0(., "_f2f3"))
```



```

# rename the pid variable used for merging datasets
f2f3_data <- f2f3_data %>% rename(pid = pid_f2f3)

# copy of dataset
data <- f2f3_data

# creating a vector to indicate which variables to keep
pac_vector <- search_var_function(data, pac, "0[1-3]0[0-2]")

## [1] "This is the searching pattern: "
## [1] "^pac0[1-3]0[0-2]\\_(f2f|cati|cov)[1-5]( _inv|_rec)?"

var_keep_f2f3 <- c("pid", "cgr_f2f3", "ptyp_f2f3", "sex_f2f3", "zyg0102_f2f3",
, "age0100_f2f3")
var_keep <- c(var_keep_f2f3, pac_vector)

# subset data to relevant variables
sub_data <- subset(data, select = var_keep)
sub_data [] <- lapply(sub_data , as.numeric)

# check variables and distributions
describe(sub_data)

# view labels of a GPAQ variable (also described in the codebook of F2F3)
# print_labels(sub_data$pac0100_f2f3)

# create variable indicating if case is valid (1) or not (0)
# invalid cases: cases with too many missings that means coding lesser than
-81 in more than one domain on days variable
sub_data <- sub_data %>% mutate(valid_95 = case_when(pac0100_f2f3 >= -81 |
pac0200_f2f3 > -85 | pac0300_f2f3 > -85 |
pac0101_f2f3 >= -81 |
pac0201_f2f3 > -85 | pac0301_f2f3 > -85 |
pac0102_f2f3 >= -81 |
pac0202_f2f3 > -85 | pac0302_f2f3 > -85 ~ 1))

table(is.na(sub_data$valid_95))

##
## FALSE TRUE
## 6796 3707

# replace NA with zero
sub_data$valid_95[is.na(sub_data$valid_95)] <- 0
# 6796 cases with data
table(sub_data$valid_95)

##
## 0 1
## 3707 6796

# next filtering: invalid cases: missing data on both hours and minutes var
iables
# view dataset to check filtering
# View(sub_data[, c("pid", "age0100_f2f3", "pac0100_f2f3", "pac0101_f2f3", "p
ac0102_f2f3", "pac0200_f2f3", "pac0201_f2f3", "pac0202_f2f3", "pac0300_f2f3", "p

```

```

ac0301_f2f3", "pac0302_f2f3", "invalid"]])
# three variables for three domains indicating if there is data on the critical variables
# work
sub_data$valid_work <- ifelse(
  ((sub_data$pac0100_f2f3 == -81 | sub_data$pac0100_f2f3 == 0) & (sub_data$pac0101_f2f3 <= 0 & sub_data$pac0102_f2f3 <= 0)) |
  (sub_data$pac0100_f2f3 > 0 & (sub_data$pac0101_f2f3 > 0 | sub_data$pac0102_f2f3 > 0)), 1, 0 )
table(sub_data$valid_work)

##
##      0      1
## 5522 4981

# travel
sub_data$valid_commuting <- ifelse(
  ((sub_data$pac0200_f2f3 == 0) & (sub_data$pac0201_f2f3 <= 0 & sub_data$pac0202_f2f3 <= 0)) |
  (sub_data$pac0200_f2f3 > 0 & (sub_data$pac0201_f2f3 > 0 | sub_data$pac0202_f2f3 > 0)), 1, 0 )
table(sub_data$valid_commuting)

##
##      0      1
## 4525 5978

# leisure
sub_data$valid_leisure <- ifelse(
  ((sub_data$pac0300_f2f3 == 0) & (sub_data$pac0301_f2f3 <= 0 & sub_data$pac0302_f2f3 <= 0)) |
  (sub_data$pac0300_f2f3 > 0 & (sub_data$pac0301_f2f3 > 0 | sub_data$pac0302_f2f3 > 0)), 1, 0 )
table(sub_data$valid_leisure)

##
##      0      1
## 4334 6169

# variable indicating if all critical data exist for each case: missing code -82 for incomplete cases
# 1 = valid, 0 = invalid
# invalid missings: missing on days variable or missings on hours / minutes variables (critical info)
# 5274 invalid missings, 5229 valid missings
sub_data$valid_82 <- ifelse((sub_data$age0100_f2f3 >= 17 | sub_data$age0100_f2f3 == -99) & (sub_data$valid_work == 1 & sub_data$valid_commuting == 1 & sub_data$valid_leisure == 1), 1, 0)
sub_data$valid_82 <- ifelse((sub_data$age0100_f2f3 <= 16 & sub_data$age0100_f2f3 > 10) & (sub_data$valid_commuting == 1 & sub_data$valid_leisure == 1), 1, sub_data$valid_82)

table(sub_data$valid_82[sub_data$valid_95 == 1])

##      0      1
## 1568 5228

```

```

# 13 participants with missing age value
table(sub_data$age0100_f2f3[sub_data$valid_82 == 1])

# Subset with cases missing value on age variable
missing_age <- sub_data %>% filter(age0100_f2f3 == -99 & valid_82 == 1)
# check hours variable with values 15, 30, 45, 60 while 0 on minutes variable
mismatched_values <- sub_data %>% filter(valid_82 == 1 & (pac0101_f2f3 == 15 | pac0101_f2f3 == 30 | pac0101_f2f3 == 45 | pac0101_f2f3 == 60 | pac0201_f2f3 == 15 | pac0201_f2f3 == 30 | pac0201_f2f3 == 45 | pac0201_f2f3 == 60 | pac0301_f2f3 == 15 | pac0301_f2f3 == 30 | pac0301_f2f3 == 45 | pac0301_f2f3 == 60) )
# 3 cases with mismatched values on hours variable: correcting case-by-case
# pid: 230895400 262672300 488432110
print(mismatched_values$pid)

## [1] 230895400 262672300 488432110

# mismatched_values[, c("pid", "pac0101_f2f3", "pac0102_f2f3", "pac0201_f2f3", "pac0202_f2f3", "pac0301_f2f3", "pac0302_f2f3")]

# copy dataset to have the raw data as a reference
sub_data_copy <- sub_data
sub_data_copy <- sub_data_copy %>% rename_at(vars(everything()), ~ paste0(. , "_copy"))

# correcting case-by-case for mismatched values
# case 1: pid 230895400
sub_data$pac0301_f2f3[sub_data$pid == 230895400]

## [1] 15

sub_data$pac0302_f2f3[sub_data$pid == 230895400]

## [1] -99

sub_data$pac0301_f2f3[sub_data$pid == 230895400] <- -99
sub_data$pac0302_f2f3[sub_data$pid == 230895400] <- 15
sub_data$pac0301_f2f3[sub_data$pid == 230895400]

## [1] -99

sub_data$pac0302_f2f3[sub_data$pid == 230895400]

## [1] 15

# case 2: pid 262672300
sub_data$pac0301_f2f3[sub_data$pid == 262672300]

## [1] 15

sub_data$pac0302_f2f3[sub_data$pid == 262672300]

## [1] 0

sub_data$pac0301_f2f3[sub_data$pid == 262672300] <- 0
sub_data$pac0302_f2f3[sub_data$pid == 262672300] <- 15
sub_data$pac0301_f2f3[sub_data$pid == 262672300]

```

```

## [1] 0
sub_data$pac0302_f2f3[sub_data$pid == 262672300]

## [1] 15
# case 3: pid 488432110
sub_data$pac0201_f2f3[sub_data$pid == 488432110]

## [1] 15
sub_data$pac0202_f2f3[sub_data$pid == 488432110]

## [1] -99
sub_data$pac0201_f2f3[sub_data$pid == 488432110] <- -99
sub_data$pac0202_f2f3[sub_data$pid == 488432110] <- 15
sub_data$pac0201_f2f3[sub_data$pid == 488432110]

## [1] -99
sub_data$pac0202_f2f3[sub_data$pid == 488432110]

## [1] 15
# recoding to 0
sub_data[sub_data <= -81] <- 0

# rename the pid variable used for merging datasets
sub_data_copy <- sub_data_copy %>% rename(pid = pid_copy)
sub_data <- left_join(sub_data, sub_data_copy, by = "pid")

# calculate total amount of time spent on physical activity weekly
sub_data <- sub_data %>% mutate(pac0400 = (pac0101_f2f3 * 60 + pac0102_f2f3
) * pac0100_f2f3 + (pac0201_f2f3 * 60 + pac0202_f2f3 ) * pac0200_f2f3 +
(pac0301_f2f3 * 60 + pac0302_f2f3 ) * pac0300_f2f3,
                                patime_work_daily = (pac0101_f2f3 *
60 + pac0102_f2f3 ) ,
                                patime_commute_daily = (pac0201_f2f
3 * 60 + pac0202_f2f3 ) ,
                                patime_leisure_daily = (pac0301_f2f
3 * 60 + pac0302_f2f3 ) )

# if one domain exceeds 16 hours (960 minutes) exclude the case
table(sub_data$patime_work_daily)
table(sub_data$patime_commute_daily )
table(sub_data$patime_leisure_daily)

extreme_values<- filter(sub_data, ((sub_data$valid_95 == 1 & sub_data$valid
_82 == 1) & (patime_work_daily > 960 | patime_commute_daily > 960 | patime_
leisure_daily > 960 | sub_data$pac0400 >6720)))
# View(extreme_values[, c("patime_work_daily", "patime_commute_daily", "pati
me_leisure_daily", "age0100_f2f3", "pid")])

# 1 = valid case, 0 = invalid case because of extreme value in at least one

```

```

domain
sub_data$valid_83 <- ifelse(sub_data$ptime_work_daily > 960 | sub_data$ptime_commute_daily > 960 | sub_data$ptime_leisure_daily > 960 | sub_data$pac0100_f2f3 > 7 | sub_data$pac0200_f2f3 > 7 | sub_data$pac0300_f2f3 > 7 | sub_data$pac0400 > 6720, 0, 1)
table(sub_data$valid_83[(sub_data$valid_95 == 1 & sub_data$valid_82 == 1)])

##
##      0      1
##      8 5220

# missing codes
table(sub_data$pac0400)

sub_data$pac0400 <- ifelse(sub_data$valid_83 == 0, -83, sub_data$pac0400)
sub_data$pac0400 <- ifelse(sub_data$valid_82 == 0, -82, sub_data$pac0400)
sub_data$pac0400 <- ifelse(sub_data$valid_95 == 0, -95, sub_data$pac0400)

table(sub_data$pac0400)

# calculate total amount of time spent on physical activity weekly
# age <= 16 --> pac0410
# copy from pac0400 allowed because working domain was set to 0 for age group 11-16
sub_data$pac0410 <- sub_data$pac0400
table(sub_data$pac0400)

# generating variable for level of physical activity
# pac0401 for group 1: aged 17 or older
# 1 = high, 0 = low
sub_data$pac0401 <- ifelse (sub_data$pac0400 >= 150, 1, 0)

sub_data$pac0401[sub_data$pac0400 == -95] <- -95
sub_data$pac0401[sub_data$pac0400 == -82] <- -82
sub_data$pac0401[sub_data$pac0400 == -83] <- -83

# View(sub_data[, c("pac0100_f2f3_copy", "pac0101_f2f3_copy", "pac0102_f2f3_copy", "pac0200_f2f3_copy", "pac0201_f2f3_copy", "pac0202_f2f3_copy", "pac0300_f2f3_copy", "pac0301_f2f3_copy", "pac0302_f2f3_copy", "pac0400", "pac0401")])

# pac0411 for group 2: aged between 11 to 16
# 1 = high, 0 = low
sub_data$pac0411 <- ifelse (sub_data$pac0410 >= 420, 1, 0)
sub_data$pac0411[sub_data$pac0410 == -95] <- -95
sub_data$pac0411[sub_data$pac0410 == -82] <- -82
sub_data$pac0411[sub_data$pac0410 == -83] <- -83

table(sub_data$pac0401)

##
##   -95  -83  -82    0    1
## 4922    7 1287  544 3743

table(sub_data$pac0411)

```

```
##
##   -95   -83   -82    0    1
## 9288    1  281  394  539

# group 1: aged 17 or older
group_1 <- sub_data %>% filter((age0100_f2f3 >= 17 | sub_data$age0100_f2f3_
copy == -99) & valid_95 == 1 & valid_82 == 1 & valid_83 == 1 )

describe(group_1$pac0400)

describe(group_1$pac0401)

ggplot(group_1, aes(x=pac0400)) +
  geom_histogram( bins = 150, colour = "white", fill = "brown3") +
  theme_minimal() +
  labs(
    title = "Total amount of physical activity per week in F2F3",
    subtitle = " ",
    caption = "n = 4287",
    x = "Time in minutes",
    y = "Frequency" )

# group 2: 16 years old or under
# over60 = total time spent on physical activity on daily basis
group_2 <- sub_data %>% filter(age0100_f2f3 <= 16 & age0100_f2f3 > 10 & val
id_95 == 1 & valid_82 == 1 & valid_83 == 1 ) %>% mutate(over60 = ifelse ((
patime_commute_daily > 60 | patime_leisure_daily > 60), 1, 0 ))

ggplot(group_2, aes(x=pac0410)) +
  geom_histogram( bins = 150, colour = "white", fill = "brown3") +
  theme_minimal() +
  labs(
    title = "Total amount of physical activity per week in F2F3",
    subtitle = " ",
    caption = "n = 933",
    x = "Time in minutes",
    y = "Frequency" )

table(group_2$over60)

describe(group_2$pac0410)

describe(group_2$pac0411)

# check coding of "not working" on work domain days variable only: 647 vali
d cases with -81
table(sub_data$pac0100_f2f3_copy[sub_data$pac0100_f2f3_copy == -81 & sub_da
ta$valid_95 == 1 & sub_data$valid_82 == 1 & sub_data$valid_83 == 1])

##
## -81
## 647

sub_data$wid <- 5
final_data <- select(sub_data, c(pid, wid, pac0400, pac0401, pac0410, pac04
```

```
11))  
# save final dataset as .rda-File  
write_dta(final_data, "final_data.dta")
```