Example of a hypothesis refinement

Hypothesis: "Students who do sports perform better in their university studies."

Underline keywords

Underline any words that could be more specific:
"Students who do sports perform better in their university studies."

Ask questions

For each of the underlined keywords, ask (write down) questions that could help specify this part. Even trivial things are welcome. You do not need to answer these questions right now, that comes in the next step.

Students

- Which students?
  - Which university/ies?
  - Which degree program(s)?
  - Which period/year(s)?
  - Full-time only or also part-time?
  - Any other requirements?
    - Age?
    - Gender?
    - Health?
    - Relationship?
    - ...

do sports

- Which sports? Does it include e.g. chess or esports? Does it include 1 hour biking to the university every day?
- How often?
  - n *times* per day/week/month/year?
  - x *hours* per day/week/month/year?
- When?
  - At any point during their studies?
  - Specifically during the periods where they are taking courses?

perform better

- What is 'better'?
  - Which one is better:
    - 15 obtained/25 registered ECTS, grades: 8, 8, 8, 4, 2 OR
      15 obtained/15 registered ECTS, grades: 6, 6, 6
    - 15 ECTS, grades: 8, 8, 8 OR
      20 ECTS, grades: 6, 6, 6, 6
  - More ECTS per year?
  - Fewer failed courses per year?
  - Less total time to complete degree?
  - Higher grade average? (how to deal with extra courses when computing averages?)
- Better than what/whom?
  - Students who do not do sports?
  - Correlation more sports $\leftrightarrow$ better performance?

university studies

- Any questions relevant here actually already appear with the other keywords, so we can skip this one.

Answer questions

Now start answering the questions from the previous step. Some will be very easy or just need a decision based on common sense. Some will be harder and might need some research and/or discussion. You may also want to check the available data for some of them.

Remember that this is just an example and other answers to the questions could have been chosen.

Students

- Which students?
  - Which university/ies?
    If we have data on all universities in NL, but only data on sports associations in Eindhoven, it doesn't make sense to include all universities.
  - Which degree program(s)?
    If we have sufficient data, we can do the analysis separately per program, and see if there are differences. Otherwise, we can disregard this factor completely.
  - Which period/year(s)?
    We will limit the study to only bachelor students in 3-year programs to get a more homogeneous group. This could eliminate some confounding variables.
  - Full-time only or also part-time?
    Full-time only.
  - Any other requirements?
    - Age?
      Let's limit this from 17-23 to include students that start a bit earlier and also those that finish 1-2 years late. We should check that this does not cost us too much data.
    - Gender?
      It could be a confounding variable. If we have that data, we can split the data on gender; otherwise we ignore gender but be aware that there may be an alternate reason for our observations.
    - Health?
      Not in our data, we cannot take it into account and should report that it could potentially be a confounding variable.
    - Relationship?
      Also not in our data.

do sports

We only have data on membership of sports associations.

We can do some research online to see if we should filter out esports and chess to focus on more traditional physical exercise. It is also possible to investigate differences between these groups.

We cannot include those who do sports outside of these associations. That means that some students labeled as 'not doing sports' actually do sports, and if there *is* a positive correlation between doing sport and study success, then we will have some better-performing sport-doing students mixed in with the worse-performing non-sport-doing students.

We will also have to assume that those who are member of a sports association actually do sports regularly. There may be inactive members. This means that some students labeled as 'doing sports' actually don't do sports, and if there *is* a positive correlation between doing sport

and study success, we will have some worse-performing non-sport-doing students mixed in with the better-performing sport-doing students.

We do have data on which years students were members, so we can take students per year and sort them into 2 categories: those who did sports that year, and those who did not.

Alternatively, we should get data from different sources, like student surveys. (Whole different story.)

If we proceed with the data from the university sport associations, we have to reformulate the hypothesis and replace "students who do sports" with "students who are members of a university sport association" and choose answers to the questions we formulated in the same way we did it for answering questions about students.

perform better

We can eliminate some options based on available data, for example if we don't have individual grades. We can also search online for common metrics of student performance and see if we have the data for those. The rest is left open to discussion and a decision needs to be made with the group. Let's assume an online search shows predominantly results in ECTS/year. "perform better" now becomes "earn more ECTS per year".

Write about the hypothesis refinement in your notebook

The original hypothesis was "Students who do sports perform better in their university studies." We have refined it as follows:

- Students who
  - are enrolled in a full-time 3-year bachelor program – to make the population more homogeneous, eliminating some potential confounding variables
  - at TU/e – because we only have data from student sports associations in Eindhoven
  - between 17 and 23 years of age – again homogenizing the population (provided we don't lose too much data)
- and who are members of a university sport association – due to lack of more detailed data
- earn more ECTS that year in their university studies – determined by available data and popular online metrics
- than similar students who are not members of a university sport association for that year. – something to compare against

The refined hypothesis thus becomes: "Students between 17 and 23 years of age enrolled in a full-time 3-year bachelor program at TU/e earn more ECTS in a year in their university studies if they are a member of a university sport association for that year than similar students who are not members of a university sport association that year."