**Supplemental information**

*Example of data analysis for one question set*

To illustrate how data analysis was structured, we will show an example of how the question set of one student was analysed. This student formulated a question set as part of the post-test assignment. This example shows an analysis performed only by the main researcher, interrater reliability was determined beforehand with agreement rates of 81-90% (κ = 0.41-0.81).

In the post-test, the students were tasked with formulating questions that could help in an investigation into why starling populations in the Netherlands had seen a steep decline in the previous decades. This particular student formulated eight research questions for this problem. These eight questions together form the ‘question set’ for this student for the post-test. As instructed, she also categorized each question with the *organism*, *population* or *ecosystem* question agenda by writing one of these words behind each question. The questions, with their categorization written after them, were:

1. *What is the most important source of food for the starling?* - *Organism*
2. *What are enemies for the starling? - Population*
3. *Has there been a disease that has affected the starlings? - Organism*
4. *Are there also starlings in other countries? - Ecosystem*
5. *Where do starlings incubate their eggs? In what kind of area? - Population*
6. *What is the most optimal temperature for starlings? - Ecosystem*
7. *Could there be a reason to shoot starlings down? - Ecosystem*

As the first step for the data analysis, each of these questions was provided a score for ‘Specificity’ and ‘Relevance’ with the rubric by the main researcher (table S1). All questions except question 4 and 7 scored 2 points for specificity and 2 points for relevance. For these questions, there was no ambiguity as to what the student meant with the question (hence the maximum 2 points for specificity) and it was clear that answering these questions could contribute to the investigation (hence the maximum 2 points for relevance). Question 4 was formulated unambiguously (score of 2 for specificity) but it was unclear how answering the question was relevant to the investigation (score of 0 for relevance). Question 7 was formulated such that there was ambiguity as to what the student meant (score of 0 for specificity), it was also unclear how possible answers could contribute to the problem (score of 0 for relevance). This resulted in a ‘mean specificity score’ of 0.86 and a ‘mean relevance score’ of 0.71 for this question set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Question | Specificity | Relevance | Categorization  (student) | Categorization (researcher) |
| 1 | 2 | 2 | Organism | Organism |
| 2 | 2 | 2 | Population | Organism |
| 3 | 2 | 2 | Organism | Population |
| 4 | 2 | 0 | Ecosystem | - |
| 5 | 2 | 2 | Population | Population |
| 6 | 2 | 2 | Ecosystem | Organism |
| 7 | 0 | 0 | Ecosystem | - |

Next, each question was categorized with the *organism*, *population* or *ecosystem* question agenda by the main researcher (table S1). Question 4 and 7 could not be categorized with a specific question agenda as it was unclear how the student found them relevant to the central problem.

Now, a score can be calculated for each question agenda by adding the points for specificity and relevance for each question that was categorized with a certain question agenda *by the researcher*. This results in a score of 12 for the *organism* question agenda (question 1, 2 and 6 each scored the max of 4 points), a score of 8 points for the *population* question agenda (question 3 and 5 scored the max of 4 points) and 0 points for the *ecosystem* question agenda. A ‘question set score’ can then be calculated by adding these scores together: 12 + 8 + 0 = 20 points. This question set score was used as our measure for the extent to which a student succeeded at *problem reduction*.

To differentiate between *deep* and *broad* questioning, we then looked at the maximum and minimum question agenda scores. The maximum question agenda score, indicating *deep questioning*, was 12 for this question set, the score for the *organism* question agenda. The minimum question agenda score, indicating broad questioning, was 0 for this question set, the score for the *ecosystem* question agenda.

Finally, in order to measure *problem abstraction* skills, the categorization of the questions was compared between the student and the main researcher and the similarity in percentage was calculated. In this case, the student had the same categorization as the researcher for two of the five questions that were categorized both by the researcher and the student, a score of 40%.

The scores that resulted from the analysis above formed the input for the analysis on the quality of student question sets, indicative of *problem reduction*, and the analysis on the students’ ability to correctly categorize their questions with the ecological question agendas, indicative of *problem abstraction*.