# CHAPTER III
# RESEARCH METHODOLOGY

## 3.1 Research Design

This research is quantitative research with survey research design. In order to describe the views, opinions, behaviors, or features of the population, researchers may use survey research designs in quantitative research, which researchers conduct a survey to a sample or to the entire population of people. The data collected through questionnaires, interviews, and then analyze the data statistically to describe responses to the test research question. In contrast to experimental research, survey designs do not include the researcher giving a treatment to a participant. Due to the fact that survey researchers do not experimentally modify the setting (Creswell, 2012). This is aligned with the objective of this research, which to identify students' misconceptions of the content on the human excretory system. No circumstances are changed or manipulated; the only thing done is to present the findings as they are.

The type of survey design used in this research is the cross-sectional research design. In cross-sectional research, the researcher collects the data at one moment, even though collecting all of the data could take a day, a few weeks, or longer (Fraenkel et al., 2011). The cross-sectional can assess programs that give decision-makers meaningful data as well as estimate the community's needs for educational services (Creswell, 2012). The question used in this research consists of a four-tier test on the human excretory system topic that can diagnose misconceptions in-depth so the teacher can design better learning. The data were analyzed and interpreted to identify the misconceptions on the topic of human excretory system experienced by the sample.

## 3.2 Participant

The participants involved in this research were junior high school students. The criteria for the participants were that they were eighth-graders junior high school students who had studied the topic of human excretion in the 2013 National Curriculum. The populations of this research are all students of junior high school in Bogor, West Java, Indonesia. There were 127 eighth graders participated as sample in this research that came from a public school in Bogor. Convenience

sampling technique is used to determine the sample which is based on their convenience and availability (Creswell, 2012). This sampling technique is very suitable to use in this research because researchers have limited time and require many participants.

### 3.3 Research Instrument

The instrument required in order to collect the data from the participants for identifying students' conceptions and diagnosing students' misconceptions. The four-tier diagnostic test instrument in this research is developed from several researches (Aprilanti et al., 2016; Karomah et al., 2018; Dahlina et al., 2019; Luzyawati & Hidayah, 2019). The instrument was developed and followed some guidance steps for constructing the two-tier test by Treagust (1988) with certain modifications. The development of the instruments has been followed three phase as follows:

1) Determining the content of the study

   This is included the following steps:

   a. Identifying the key concept related to human excretion which are elaborated at the junior high school level.
   b. Generating theoritical knowledge statements relevant to the concept under investigation based on the science student book of the 2013 Curriculum for grade VIII semester 2.
   c. Validating the content to supervisors. In accordance with Treagust (1988), it is essential to be done to ensure that no questions are developed for multiple-choice tests that are not related to the concepts being taught.

2) Identifying students' misconceptions about human excretion

   This is done by identifying related literature and recent research for example are those reported by Aprilanti et al. (2016), Karomah et al. (2018), Dahlina et al. (2019), and Luzyawati & Hidayah (2019). By examining the explanations of students' misconceptions, it becomes feasible to gather a foundation of information for developing multiple-choice questions that address these misconceptions (Treagust, 1988).

3) Developing a diagnostic test of four-tier test about human excretion

   This process has several stages as follows:

a. Developing a draft of a four-tier test. A four-tier test is a multiple-tier diagnostic test that consists of four tiers developed by Caleon & Subramaniam (2010) with research subjects in physics on wave topics by adding a confidence tier following the content tier (Taban & Kiray, 2022). There are four tiers in each set of questions. The first tier of questions is made up of multiple choice with four possible answers that consist of three distractors that address specific misconceptions and one key response. The second tier presents two option answers which are confident and not confident in order to ask the confidence for the question in the first tier. The third tier asks for the scientific reason in the form of close-ended question for the question in the first tier, while the fourth tier asks the confidence for reasoning tier with two options which are confident and not confident. The distractors in the first and third tiers are the result of an analysis of misconceptions from several related literatures. This first draft consists of 29 sets of questions which are divided into three main subtopics. The distribution of the concept and the questions is shown in Table 3.1.
b. Validating the first draft by two lecturers that are experts in the biology field and one science junior high school teacher. The validators checked whether the questions are clear and in line with the scientific concepts. After the experts provided their feedback, a content validity test was carried out. The results can be seen in Table 3.3.
c. Preparing the instruments for a pilot test by revising the questions according to the suggestions from the validators in the previous stage.
d. Conducting a pilot test. In this research, a pilot test was conducted on a sample of 49 students of junior high school who had studied human excretion topic. Furthermore, the result is analyzed for validity and reliability using SPSS 25. The result can be seen in Table 3.4 and Table 3.6 respectively. Several tests were also carried out such as tests of false positive and false negative test, discriminating power, and level of difficulty. Only the questions that were declared valid from several previous validity tests would be used in the actual data collection. From the 29 sets of questions developed, there were 17 sets of questions were considered valid for both

the first and third tiers. The final instrument used for this research can be seen in the Appendix A.1.

Table 3.1
First Draft Question Distribution of Human Excretion Topics

| Sub-Concept | Indicator | Concept | Question |
|---|---|---|---|
| Structure and Function of Human Excretory System | Describing the functions of the excretory system | The function of excretion | Q1, Q2 |
| | Describing the organs that make up the excretory system in humans | The organs of excretory system | Q3, Q4 |
| | Analyzing the relationship between the structure and function of the kidneys | The structure and function of kidneys | Q5, Q6, Q7 |
| | | Urine formation | Q8, Q9, Q10, Q11 |
| | Analyzing the relationship between the structure and function of the liver | The structure and function of liver | Q12, Q13 |
| | Analyzing the relationship between the structure and function of the skin | The structure and function of skin | Q14, Q16 |
| | | Sweating for body regulation | Q15 |
| | Analyzing the relationship between the structure and function of the lungs | The structure and function of lungs | Q18, Q19 |
| Diseases and Disorders of the Human Excretory System | Identifying disorders and diseases that occur in the excretory system | Skin diseases | Q17 |
| | | Urin content test | Q20, Q21, Q24 |
| | | Diseases and disorders of the kidneys | Q22, Q23, Q25, Q26, Q27 |
| Lifestyle to Maintain The Health of The Excretory System | Identifying lifestyle patterns to maintain the health of the excretory system | The healthy life style for maintain kidneys health | Q28 |
| | | The healthy life style for maintain skin health | Q29 |

### 3.3.1 Validity Test

Validity can be defined as the agreement between a test score and the quality it is intended to measure (Kaplan & Saccuzzo, 2005). The research draw the conclusions based on the information they collected using the instruments, that is the reason behind the quality of the instruments used in research is very important (Fraenkel et al., 2011). According to Sireci (2007), validity tests used in educational tests should involve analysis of test content and empirical analysis. This is

supported by Lissitz & Samuelsen (2007) who stated that this is necessary to make the concept of validity more accessible, comprehendible, beneficial, and supportable. Analysis of the content of the test is related to content validity or internal validity which further requires empirical analysis to determine construct validity (Retnawati, 2016).

The content validity of an instrument is an analysis of the representation of questions with the ability to be measured that can be determine by using the expert agreement. To find out this agreement, the validity index introduced by Aiken (1985) can be used as follows:

$$V = \frac{\Sigma s}{n(c-1)}$$

where V is the rater agreement index regarding item validity; s is the score determined by each rater minus the lowest score in the category used ($s = r - l_o$, where r = score in the rater's chosen category; $l_o$ = the lowest score in the scoring category); n is number of raters; and c is the number of categories the rater can choose). From the calculation of the V index result, an item can be categorized based on its index. The categorization of the content validity of the instrument is based on Table 3.2.

Table 3.2
Validity Index Rating Scale

| Aiken's Index (V) | Validity Criteria |
|---|---|
| $0 \leq V \leq 0.3$ | Low |
| $0.3 < V \leq 0.7$ | Moderate |
| $0.7 < V \leq 1.0$ | High |

(Hsu et al., 2015)

When the V value is greater than 0.70, it indicates a strong agreement among experts, indicating that the item test is acceptable and appropriate for use. On the other hand, if Aiken's V value falls below 0.3, the experts consider the test item inadequate and unsuitable for use. Items with moderate V values, ranging from 0.3 to 0.7, indicate poor adequacy and require revision before they can be used effectively. Table 3.3 shows the recapitulation of Aiken's validity test results.

Table 3.3
Recapitulation of Content Validity Test Results

| Test Item | Tier | Aiken's Index (V) | Validity Criteria | Test Item | Tier | Aiken's Index (V) | Validity Criteria |
|---|---|---|---|---|---|---|---|
| Q1 | 1st | 0.5 | Moderate | Q16 | 1st | 1.0 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q2 | 1st | 0.5 | Moderate | Q17 | 1st | 1.0 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q3 | 1st | 1.0 | High | Q18 | 1st | 0.8 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q4 | 1st | 0.7 | Moderate | Q19 | 1st | 0.8 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q5 | 1st | 0.8 | High | Q20 | 1st | 0.8 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q6 | 1st | 0.8 | Moderate | Q21 | 1st | 0.8 | High |
| | 3rd | 0.8 | High | | 3rd | 1.0 | High |
| Q7 | 1st | 0.5 | Moderate | Q22 | 1st | 0.8 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q8 | 1st | 0.7 | Moderate | Q23 | 1st | 0.8 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q9 | 1st | 0.8 | High | Q24 | 1st | 1.0 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q10 | 1st | 1.0 | High | Q25 | 1st | 1.0 | High |
| | 3rd | 1.0 | High | | 3rd | 0.7 | Moderate |
| Q11 | 1st | 1.0 | High | Q26 | 1st | 1.0 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q12 | 1st | 0.8 | High | Q27 | 1st | 1.0 | High |
| | 3rd | 1.0 | High | | 3rd | 1.0 | High |
| Q13 | 1st | 0.5 | Moderate | Q28 | 1st | 1.0 | High |
| | 3rd | 0.7 | Moderate | | 3rd | 1.0 | High |
| Q14 | 1st | 0.8 | High | Q29 | 1st | 0.5 | Moderate |
| | 3rd | 1.0 | High | | 3rd | 0.7 | Moderate |
| Q15 | 1st | 1.0 | High | $V_{avg}$ | 1st | 0.8 | High |
| | 3rd | 0.8 | High | | 3rd | 0.9 | High |

In this research, content validity was carried out with the help of three expert judgments consist of two expert lecturers in the field of biology and one science teacher. According to the findings presented in Table 3.3, the results of the content validity for the 29 questions indicated that all of the questions both for the first and third tiers were determined to be valid with the average of Aiken's Index is 0.8 and 0.9 respectively which shows that the set of questions is in the category of high validity. Specifically, the analysis revealed that 81% of the questions were classified in the high category, while the remaining 19% were classified into the moderate category. The questions belonging to moderate category are revised according to the suggestions from the experts.

After the content validity was carried out, then the instrument was revised based on the suggestion from the experts and tested to determine the construct validity. Construct validity is the validity that indicates the extent to which the instrument reveals an ability to be measured (Nunnally, 1978; Fernandes, 1984; Rudner, 1994; Retnawati, 2016). It was determined from data collected from respondents throughout the pilot test (Sireci, 2007; Lissitz & Samuelsen, 2007). In this research, a pilot test was conducted on a sample of 49 students of junior high school who had studied excretory system topic.

Furthermore, the result was evaluated using Pearson Product Moment Correlation in SPSS 25 to determine the relationship between two variables (Cooksey, 2020). Items are considered to be valid if the Pearson correlation coefficient ($r$) calculation result is greater than the r table value ($r_{count} > r_{table}$) (Arikunto, 2005). The $r$ table for 49 respondents, 2 tailed test, in 0.05 significance level is $r(49) = 0.281$ (seen in Appendix A.2). From Table 3.4, it can be concluded that a total of 17 test are valid both of the first and third tier; Q3, Q7, Q8, Q9, Q10, Q11, Q15, Q16, Q17, Q20, Q21, Q24, Q25, Q26, Q27, Q28, and Q29. The invalid questions are not included in data collection.

Table 3.4
Recapitulation Result of Validity Construct Test

| Test Item | Tier | Pearson Correlation | Interpretation | Decision | Test Item | Tier | Pearson Correlation | Interpretation | Decision |
|---|---|---|---|---|---|---|---|---|---|
| Q1 | 1st | 0.257 | Not Valid | Not Used | Q2 | 1st | 0.191 | Not Valid | Not Used |
| | 3rd | 0.487** | Valid | | | 3rd | 0.504** | Valid | |

| Test Item | Tier | Pearson Correlation | Interpretation | Decision | Test Item | Tier | Pearson Correlation | Interpretation | Decision |
|---|---|---|---|---|---|---|---|---|---|
| Q3 | 1st | .520** | Valid | Used | Q17 | 1st | 0.297* | Valid | Used |
| | 3rd | .474** | Valid | | | 3rd | 0.631** | Valid | |
| Q4 | 1st | 0.301* | Valid | Not Used | Q18 | 1st | 0.578** | Valid | Not Used |
| | 3rd | 0.175 | Not Valid | | | 3rd | 0.079 | Not Valid | |
| Q5 | 1st | 0.543** | Valid | Not Used | Q19 | 1st | 0.151 | Not Valid | Not Used |
| | 3rd | 0.195 | Not Valid | | | 3rd | 0.496** | Valid | |
| Q6 | 1st | 0.260 | Not Valid | Not Used | Q20 | 1st | 0.311* | Valid | Used |
| | 3rd | 0.439** | Valid | | | 3rd | 0.449** | Valid | |
| Q7 | 1st | .588** | Valid | Used | Q21 | 1st | 0.454** | Valid | Used |
| | 3rd | .601** | Valid | | | 3rd | 0.555** | Valid | |
| Q8 | 1st | .531** | Valid | Used | Q22 | 1st | 0.229 | Not Valid | Not Used |
| | 3rd | .492** | Valid | | | 3rd | 0.380** | Valid | |
| Q9 | 1st | 0.657** | Valid | Used | Q23 | 1st | 0.510** | Valid | Not Used |
| | 3rd | 0.721** | Valid | | | 3rd | 0.163 | Not Valid | |
| Q10 | 1st | 0.676** | Valid | Used | Q24 | 1st | 0.401** | Valid | Used |
| | 3rd | 0.570** | Valid | | | 3rd | 0.526** | Valid | |
| Q11 | 1st | 0.732** | Valid | Used | Q25 | 1st | 0.547** | Valid | Used |
| | 3rd | 0.673** | Valid | | | 3rd | 0.575** | Valid | |
| Q12 | 1st | 0.420** | Valid | Not Used | Q26 | 1st | 0.496** | Valid | Used |
| | 3rd | 0.194 | Not Valid | | | 3rd | 0.627** | Valid | |
| Q13 | 1st | 0.488** | Valid | Not Used | Q27 | 1st | 0.397** | Valid | Used |
| | 3rd | 0.267 | Not Valid | | | 3rd | 0.399** | Valid | |
| Q14 | 1st | 0.246 | Not Valid | Not Used | Q28 | 1st | 0.350* | Valid | Used |
| | 3rd | 0.208 | Not Valid | | | 3rd | 0.506** | Valid | |
| Q15 | 1st | 0.710** | Valid | Used | Q29 | 1st | 0.531** | Valid | Used |
| | 3rd | 0.639** | Valid | | | 3rd | 0.706** | Valid | |
| Q16 | 1st | 0.642** | Valid | Used | | | | | |
| | 3rd | 0.717** | Valid | | | | | | |

In addition, content validity also needs to be re-verified by the calculation of false negatives and false positives in percentages (Hasyim et al., 2018). According to the research conducted by Arslan et al. (2012), false positives and false negatives are the terms used to describe assessment errors in scientific research. Moreover, Hestenes & Halloun (1995) recommended that the false positive and false negative be used to provide evidence for content validity. By minimizing both false negatives and false positives, the validity of multiple-choice tests can be improved. They suggested that the occurrence of false negatives should be less than 10% (Arslan et al., 2012; Hestenes & Halloun, 1995). Multiple-choice questions have a 20% chance of false positive, it is possible to happen due to students having the opportunity to provide random responses on multiple-choice tests. Furthermore, a strong distractor will give rise to false positives in students (Daumer et al., 2008; Hestenes & Halloun, 1995; Istiyani et al., 2018). From the questions that were considered as valid, then the percentage of false negatives and false positives was calculated.

Table 3.5
False Positive and False Negative Validity Analysis

| Previous Number | Current Number | False Positive | False Negative |
|---|---|---|---|
| Q3 | Q1 | 17 | 5 |
| Q7 | Q2 | 4 | 12 |
| Q8 | Q3 | 19 | 2 |
| Q9 | Q4 | 6 | 2 |
| Q10 | Q5 | 15 | 3 |
| Q11 | Q6 | 1 | 5 |
| Q15 | Q7 | 0 | 3 |
| Q16 | Q8 | 4 | 2 |
| Q17 | Q9 | 13 | 2 |
| Q20 | Q11 | 4 | 8 |
| Q21 | Q12 | 8 | 3 |
| Q24 | Q13 | 4 | 1 |
| Q25 | Q14 | 10 | 1 |
| Q26 | Q15 | 9 | 4 |
| Q27 | Q16 | 16 | 2 |
| Q28 | Q17 | 7 | 4 |
| Q29 | Q10 | 11 | 4 |
| Total | | 148 | 63 |
| (%) | | 17.8 | 7.6 |

Based on the data provided in Table 3.5, it is known that the percentage of false positives and false negatives respectively is 17.8% and 7.6%. These values are still in the range of recommended values indicating that this test is valid and can be used to identify students' misconceptions.

### 3.3.2 Reliability Test

The valid questions from the validity test then underwent a reliability test. Reliability refers to the consistency of scores or answers from one administration of an instrument to another, and from one set of item to another. Another check on internal consistency to estimating reliability of an instrument is alpha coefficient or frequently called Cronbach alpha (Fraenkel et al., 2011). A research instrument can be said reliable if the Cronbach's Alpha value is > 0.60 (Ghozali, 2016). All question in the both of first and third tiers are reliable with 0.856 and 0.893 Cronbach's Alpha Score respectively as shown in the Table 3.6. With that score, the instrument is considered reliable and acceptable (Taber, 2017). In total, 17 sets of questions are feasible to diagnose students' misconceptions about human excretion topics.

Table 3.6
The Result of Reliability Test

| N of Items | Tier | Cronbach's Alpha |
|---|---|---|
| 17 | 1 | 0.856 |
| 17 | 3 | 0.893 |

### 3.3.3 Discriminating Power

According to Ahmann & Marvin (1967), discriminating power refers to the ability of a test item to differentiate between high-ability students (upper group) and low-ability students (lower group). It is expected that the upper group students are able to answer the question correctly rather than the students in the lower group (Marie & Sreekala, 2015). Test items that lack discriminating power will not provide an accurate overview of the students' abilities. Thus, to create a good test item the discriminating power needs to be measured. Discriminating power determined by a value known as a discrimination index which calculated using the formula proposed by Heaton (1988) as follows:

$$D = \frac{U - L}{N}$$

where D is the index of item discriminating power, U is the number of students in the upper group who correctly answered the item, L is the number of students in the lower group who correctly answered the item, and N is the number of students in each of the group; grouping calculated by 27% of the students with highest score as the upper group and 27% of students with lowest score as the lower group. Discrimination index ranges from -1.00 to +1.00 with the discrimination index criteria shown in Table 3.7.

Table 3.7
The Discriminating Power Classification

| Values of D | Interpretation | Recommendation |
|---|---|---|
| $< -0.01$ | Worst | Definitely discard |
| $0.00 - 0.20$ | Poor | Discard or review in-depth |
| $0.20 - 0.29$ | Mediocre | Need to check/review |
| $0.30 - 0.39$ | Good | Possibilities for improvement |
| $> 0.39$ | Excellent | Retain |

(Backhoff, 2000)

The results of the discriminating power analysis of the instrument are presented in Table 3.8. Data analysis was carried out using Microsoft Excel 2016.

Table 3.8
Discriminating Power Result Analysis

| Test Item | Tier | D Value | Criteria | Test Item | Tier | D Value | Criteria |
|---|---|---|---|---|---|---|---|
| Q1 | 1st | 0.54 | Excellent | Q7 | 1st | 0.92 | Excellent |
| | 3rd | 0.46 | Excellent | | 3rd | 0.92 | Excellent |
| Q2 | 1st | 0.69 | Excellent | Q8 | 1st | 0.69 | Excellent |
| | 3rd | 0.54 | Excellent | | 3rd | 0.77 | Excellent |
| Q3 | 1st | 0.38 | Good | Q9 | 1st | 0.46 | Excellent |
| | 3rd | 0.69 | Excellent | | 3rd | 0.77 | Excellent |
| Q4 | 1st | 0.69 | Excellent | Q10 | 1st | 0.54 | Excellent |
| | 3rd | 0.92 | Excellent | | 3rd | 0.92 | Excellent |
| Q5 | 1st | 0.69 | Excellent | Q11 | 1st | 0.38 | Good |
| | 3rd | 0.77 | Excellent | | 3rd | 0.62 | Excellent |
| Q6 | 1st | 0.92 | Excellent | Q12 | 1st | 0.62 | Excellent |
| | 3rd | 0.92 | Excellent | | 3rd | 0.77 | Excellent |

| Test Item | Tier | D Value | Criteria | Test Item | Tier | D Value | Criteria |
|---|---|---|---|---|---|---|---|
| Q13 | 1st | 0.46 | Excellent | Q16 | 1st | 0.62 | Excellent |
| | 3rd | 0.69 | Excellent | | 3rd | 0.31 | Good |
| Q14 | 1st | 0.54 | Excellent | Q17 | 1st | 0.46 | Excellent |
| | 3rd | 0.62 | Excellent | | 3rd | 0.62 | Excellent |
| Q15 | 1st | 0.69 | Excellent | | | | |
| | 3rd | 0.69 | Excellent | | | | |

Based on the results of discriminating power analysis, it can be said that this set of questions is a good test item because it has a discriminatory index of more than 0.29 which include in several criteria which are good, and excellent.

### 3.3.4 Difficulty Level

A good test item should have a certain level of difficulty, that is neither too easy nor too difficult. Questions that are too easy do not stimulate students to enhance students' efforts to solve them. Conversely, questions that are too difficult will cause students to become discouraged and have no motivation to solve them (Arikunto, 2006). Thus, the level of difficulty indicates how difficult or easy the question can be answered by the respondent. The higher number of correct answers, the easier the test item. Moreover, the lower the number of correct answers so the more difficult the test item. Difficulty level determined by a value known as difficulty index which calculated using the formula proposed by Gronlund (1993) as follows:

$$P = \frac{R}{N}$$

where P is difficulty index, R is the number of students who answered items correctly, N is the total number of students who conduct the test. The difficulty index interpreted based on Table 3.9.

Table 3.9
The Difficulty Level Classification

| Value of P | Interpretation |
|---|---|
| $P = 0.00$ | Very difficult |
| $0.00 < P \leq 0.30$ | Difficult |
| $0.30 < P \leq 0.70$ | Medium |
| $0.70 < P < 1.00$ | Easy |
| $P = 1.00$ | Very easy |

(Gronlund, 1993)

The results of the difficulty level analysis of the instrument are presented in Table 3.10. Data analysis was carried out using Miscrosoft Excel 2016.

Table 3.10
Difficulty Level Result Analysis

| Test Item | Tier | P Value | Interpretation | Test Item | Tier | P Value | Interpretation |
|---|---|---|---|---|---|---|---|
| Q1 | 1st | 0.63 | Medium | Q10 | 1st | 0.76 | Easy |
| | 3rd | 0.39 | Medium | | 3rd | 0.61 | Medium |
| Q2 | 1st | 0.51 | Medium | Q11 | 1st | 0.47 | Medium |
| | 3rd | 0.67 | Medium | | 3rd | 0.57 | Medium |
| Q3 | 1st | 0.90 | Easy | Q12 | 1st | 0.63 | Medium |
| | 3rd | 0.55 | Medium | | 3rd | 0.51 | Medium |
| Q4 | 1st | 0.67 | Medium | Q13 | 1st | 0.82 | Easy |
| | 3rd | 0.59 | Medium | | 3rd | 0.76 | Easy |
| Q5 | 1st | 0.71 | Easy | Q14 | 1st | 0.69 | Medium |
| | 3rd | 0.47 | Medium | | 3rd | 0.51 | Medium |
| Q6 | 1st | 0.33 | Medium | Q15 | 1st | 0.51 | Medium |
| | 3rd | 0.41 | Medium | | 3rd | 0.41 | Medium |
| Q7 | 1st | 0.45 | Medium | Q16 | 1st | 0.57 | Medium |
| | 3rd | 0.51 | Medium | | 3rd | 0.29 | Difficult |
| Q8 | 1st | 0.39 | Medium | Q17 | 1st | 0.76 | Easy |
| | 3rd | 0.35 | Medium | | 3rd | 0.69 | Medium |
| Q9 | 1st | 0.80 | Easy | | | | |
| | 3rd | 0.57 | Medium | | | | |

Followed by this analysis, this set of questions is said to be good apart from being supported by having a good discriminating power, the difficulty level in this set of questions is also dominated by the medium level which is good as said in research conducted by Amalia & Widayati (2012) that a good test item is the one that in the difficulty level category of moderate, that is not too easy or not to difficult.

### 3.3.5 Interviews

The interviews in this research were conducted using a semi-structured approach. This approach included a predefined set of questions and specific topics. The questions were posed to the interviewees in a structured and organized manner, but the interviewers were allowed to probe deeper. They were authorized to go beyond the scripted questions and explore more extensively during the interviews (Berg & Lune, 2012). Generally, the questions asked during the interview are like "Why are you confident with your answer?" "Why do you think of the answer in that way?". The interviews were conducted with teachers and students with the aim of delving deeper into the factors contributing to misconceptions about human excretion. The results of these interviews are used as supplementary data to the main data source, which is the diagnosis of misconceptions using a four-tier multiple-choice instrument.

## 3.4 Research Procedure

As depicted in Figure 3.1, the research is conducted in three stages. Those are preparation stage, implementation stage, and completion stage. The steps are also described in greater detail as follow:

1) Preparation Stage
    a. Formulating research problem and develop it into several research questions.
    b. Analyzing some relevant research about four-tier test, misconceptions on topic of human excretory system, and human excretory system topic on the 2013 National Curriculum.
    c. Constructing research instrument from the idea of human excretory system topic that have been analyzed.

d. Conducting the expert judgment and revise the instrument, the expert judgement form provided in Appendix A.3.
e. Distributing the instrument to students who are not included in the research sample using Google Form (https://forms.gle/NLTRFGi8eZfKGve5A).
f. Conducting validity and reliability test to the result from pilot test for selecting the questions used in real test using SPSS. The result can be seen in Appendix A.4 and Appendix A.5.
g. Determining the research sample and the test date.

2) Implementation Stage

In this stage, the researcher has requested permission to conduct research at a public school in Bogor. The research instrument was distributed to the students using two different methods. Students with smartphones were provided with a Google Form (https://forms.gle/4MqkWffEUEwUP8g97), The documentation of the online instrument used can be seen in Appendix B.1 while the students who do not have smartphones used a paper-based test. The four-tier instrument test consists of 17 sets of questions. The data collection was conducted over a span of two days, specifically on the 22$^{nd}$-23$^{rd}$ of May 2023, carried out within each classroom under the direct supervision of the researchers, the documentation during the test provided in Appendix B.2. Furthermore, short interviews with students as sample were conducted. The permission letter to conduct research can be seen in Appendix C.1, and the completion letter from the school has been provided in Appendix C.2.

3) Completion Stage

a. Analyzing the data collection statistically using Microsoft Excel 2016 to investigate the profile of students' conceptions and common misconceptions about human excretion topic.
b. Constructing the research result, discussion, and conclusion.
c. Reporting the research. The approval from supervisor to follow thesis defense provided in Appendix C.3.
d. This research paper already submitted to journal that has been accredited SINTA 2 (International Journal of Education, Universitas Pendidikan Indonesia). The proof of submission can be seen in Appendix C.4.

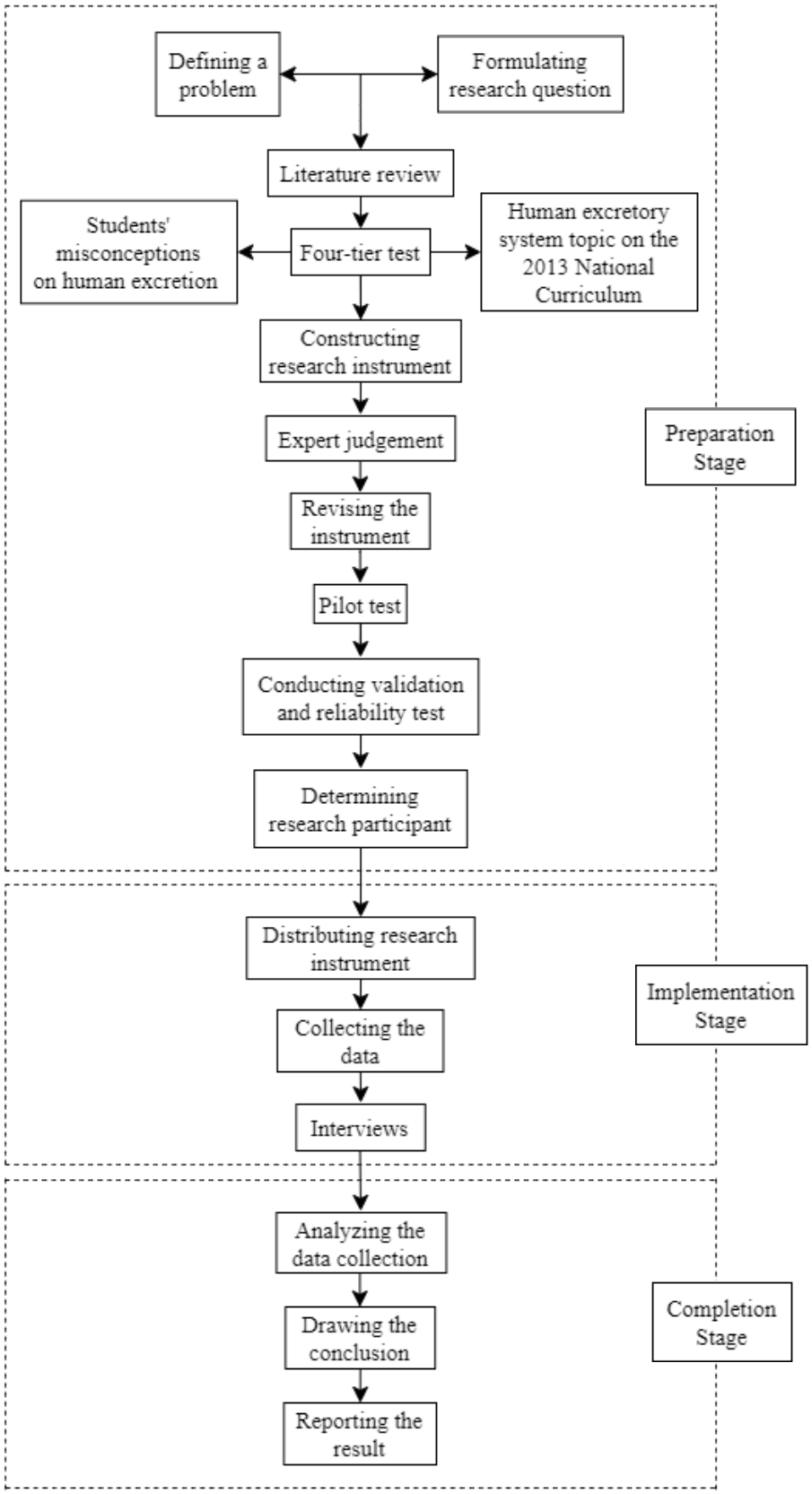


Figure 3.1 Research Procedure Flowchart

### 3.5 Data Analysis

As shown in Table 3.11, responses to each question will be grouped into the following categories: scientific knowledge (SK), lack of knowledge (LK), misconception (M), false negative (FN), and false positive (FP) (Kiray & Simsek, 2021). When students confidently answer the first and third tier correctly, they asserted scientific knowledge. False positive is when students confidently answer both tiers but give correct answer to the first tier and incorrect answer to the third tier. False negative occurs when students incorrectly answer the first tier while correctly answering the third tier and are confident about first and third tiers. When students incorrectly answer the first and third tiers but are confident about both tiers, this is referred to as misconception. Combinations different than those already described are classified as lack of knowledge.

Table 3.11
Combination Response and Decision of Four-Tier Test

| 1st Tier | 2nd Tier | 3rd Tier | 4th Tier | Decision of Four-Tier Test |
|---|---|---|---|---|
| True | Confident | True | Confident | SK |
| True | Confident | False | Confident | FP |
| False | Confident | True | Confident | FN |
| False | Confident | False | Confident | M |
| True | Confident | True | Not Confident | LK |
| True | Not Confident | True | Confident | LK |
| True | Not Confident | True | Not Confident | LK |
| True | Confident | False | Not Confident | LK |
| True | Not Confident | False | Confident | LK |
| True | Not Confident | False | Not Confident | LK |
| False | Confident | True | Not Confident | LK |
| False | Not Confident | True | Confident | LK |
| False | Not Confident | True | Not Confident | LK |
| False | Confident | False | Not Confident | LK |
| False | Not Confident | False | Confident | LK |
| False | Not Confident | False | Not Confident | LK |

(Kiray & Simsek, 2021)

The correct answer for the first and third tiers are coded as “1”, and the wrong answer for that tiers are coded as “0”. The answer of second and fourth tiers are confident will coded as “1”, and the not confident one will coded as “0”. For instance, the code of scientific knowledge is the sequence of score “1” given to the four tiers (1-1-1-1) when students’ answer of the first and third tiers are correct and the answer of second and fourth tiers have to be confident. False positive was code as (1-1-0-1) since answer of the first tier is correct but false answer in the third tier, and confidently answer for both tiers. The false negative was coded as (0-1-1-1) following the answer of the first tier is false, and correct answer the third tier, and the answer of second and fourth tiers are confident. Misconception was coded as (0-1-0-1) when students’ answer for the first and third tiers are false and confidently answer for both tiers. The responses were labeled as lack of knowledge if the pattern are not following the previous patterns given.

In the first stage of data processing, the level of students’ conceptions was classified into the categories in Table 3.11 and the percentages of each category was calculated. A simple formula was used to calculate the percentage of each category from the results of the categories. The calculation is as follows:

$$P = \frac{s}{N} \times 100\%$$

Description:

P : Percentage of each category

s : Number of the students for each group

N : Total number of students

The second stage, the percentages of misconceptions in each question were calculated and the misconceptions that arise in the questions examined are analyzed further.