Building framework for assessing students' statistical reasoning in solving real-life medical problems

Hien Tran Thuy[1], Son Le Phuoc[2]

[1]University of Medicine and Pharmacy, Hue University, Hue City, Viet Nam

[2]University of Law, Hue University, Hue City, Viet Nam

tthienyd@hueuni.edu.vn, sonlp@hueuni.edu.vn

Abstract: The goal of this research was to provide a general assessment framework of students' statistical reasoning in medicine, then build three scales to assess students' statistical reasoning ability in solving practical medical problems, including Description, Interpretation and Prediction. On that basis, the study designed a set of tools to assess students' statistical reasoning ability in solving practical medical problems. Through the analysis of the students' performance, assessments of the students' statistical reasoning in medicine were done. These research results were suitable for student assessment in medical statistics courses, useful for faculty and students in teaching and learning medical statistics.

INTRODUCTION

In quality assessment, we think of low-quality and high-quality learning. According to Vui (2018), it can be said that high-quality learning is characterized by nurturing individual ability to gain knowledge and to understand, and then to apply them to real-life situations to make well-informed decisions and also enhance the individuals' ability, actively share ideas with others. To measure and assess the quality of learning, it is necessary to classify educational goals, classify students' thinking or perceptions in a unified way so that they can be exchanged and discussed among educators. There are many different classifications in terms of thinking, understanding, and goals, but in general, the classifications are hierarchical from low to high and provide educators with a basis for lesson design and appropriate learning quality assessment tools. Bloom's revised Taxonomy is presented in "A Taxonomy for Learning, Teaching, and Assessment: A Revision of the Bloom's Taxonomy of Educational Objectives" (Anderson et al., 2001). Here, we are interested in the cognitive process dimension structure: The number of original categories is still six, but there are important changes, from six levels of Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation to Remember, Understand, Apply, Analyze, Evaluate and Create. Three categories were renamed, the order of two categories was switched, and the names of the main categories were changed from noun to verb to match the way they are used in the objective: The Knowledge Category is changed to Remember, Comprehension is changed to Understand. This is explained by the fact that because taxonomy reflects different forms of thinking and thinking is an active process, the use of verbs is more appropriate. The math assessment task hierarchy abbreviated as MATH (referred to as MATH classification of thinking or MATH taxonomy) is designed and proposed by Smith et al. (1996). The MATH taxonomy is designed to assist in the development and construction of advanced math assessments to ensure that learners are assessed on a variety of knowledge and skills (Darlington, 2013). The Program for International Student Assessment (PISA) of the Organization for Economic Cooperation and Development (OECD) assesses students from participating countries on their ability to apply their learned knowledge and skills in real life, including four areas of Reading, Math, Science, and Problem Solving. PISA arranges cognitive activities to classify math ability in 3 clusters: Reproduction, Connections, Reflection. Each such cluster comprises 2 levels. Reproduction level 1, 2 is considered low competency; Connections level 3, 4 indicates average competency, and Reflection level 5, 6 is assessed as high competency. The combination of low-to-high competencies including Reproduction, Connections and Reflection in the PISA classification is understood as "basic math skills" in the sense of understanding specific math knowledge and applying it to real-world situations (OECD, 2009). We studied Bloom's revised taxonomy, and MATH taxonomy together with the classification of Mathematical Literacy of PISA to develop scales to evaluate students' statistical reasoning when solving real-life medical problems.

THEORETICAL FRAMEWORK

Statistical Reasoning in Medicine

"Statistical reasoning may be defined as the way people reason with statistical ideas and make sense of statistical information. This involves making interpretations based on sets of data, representations of data, or statistical summaries of data. Statistical reasoning may involve connecting one concept to another (e.g., center and spread), or it may combine ideas about data and chance. Reasoning means understanding and being able to explain statistical processes and being able to fully interpret statistical results" (Dani Ben-Zvi, 2006). So, statistical reasoning is a type of cognitive activity and can be found at the stages in the human thinking process where learners are required to make sense of, explain, or justify a conclusion from the statistical results. Garfield & Gal (1999), then Garfield (2002) identified six types of Statistical Reasoning that direct cognitive activities to statistical tasks, including: Reasoning about data, Reasoning about representations of data, Reasoning about statistical measures, Reasoning about uncertainty, Reasoning about samples, Reasoning about associations. According to delMas (2002), to develop statistical reasoning, the types of statistical tasks may be required to perform such as explaining why or how results are produced, explaining why a conclusion is reasonable, well-founded. Regarding Statistical Reasoning in Medicine (SRiM), Moyé (2006) describes it as follows: "The essence of statistical inference in medicine is the process of determining whether research findings based on a sample can be extended, generalized for the whole or not". Thus, statistical inference in Medicine can be viewed as statistical inference in the field of medicine, which focuses on the fundamentals of statistical inference in medicine to decide when and how to derive results from a sample and apply them to the whole health care research field.

Competency is a rather abstract concept, OECD (2002) has defined "competency as the ability of individuals to meet complex demands and successfully perform the task in a particular context". Accordingly, when it comes to competence, it means the ability of an individual to be associated with a particular task and the conditions to accomplish that task effectively. To evaluate the students' statistical reasoning competency in medicine, it is necessary to clarify specific statistical reasoning abilities in medicine, first, we consider the model of activities that demonstrate that competency. When studying the statistical reasoning development model of high school students, Jones et al. (2001), Mooney (2002) introduced four key statistical processes that characterize statistical reasoning, including data description, data organization, data representation, data analysis, and interpretation. However, we believe that medical students, must be equipped with the thinking methods of scientists, so the requirements for statistical processes must be expanded, and the level of statistical reasoning needs to be advanced. According to Riffenburgh (2012) “A scientist usually seeks to develop knowledge in three stages: The first stage is to describe a class of scientific events, the second stage is to explain these events and the third stage is to predict the occurrence of events”. Thus, the accumulation of knowledge in medicine also follows these three stages and in each stage, statistical reasoning in medicine is required. We propose three key medical statistical processes that demonstrate the medical students’ statistical reasoning ability including *Description*, *Interpretation*, and *Prediction,* and these processes are described as follows:

(1) *Description:* the stage in which we seek to describe the data generating process in cases for which we have data from that process. For example, in the description phase, the cause of the disease or the health status of the community is recorded. This stage generates data and generates scientific hypotheses to be tested. Thus, the process of Description also includes the process of describing data, organizing and representing data.

(2) *Interpretation:* the phase in which we seek to reason characteristics of the (overall) data generation process when we have only part (usually a small part) of the obtainable data. In the explanation phase, data are evaluated to explain disease-related problems. At this stage, scientific hypotheses are tested. The explanation often takes the form of statistical hypothesis testing. Thus, the process of Explanation also includes the process of analyzing and interpreting data.

(3) *Prediction:* The stage in which we seek to make predictions about a characteristic of the data-generating process and establish a mathematical model based on newly taken related observations is newly performed. In which, there is the integration of test results and disease progression models are formed. Predictions are used in the diagnosis and prevention of diseases and even in the evaluation of the effectiveness of a treatment or disease prevention or control measure. This allows decisions to be made regarding a disease treatment or prevention or control

to change the chances of an event. Prediction involves establishing a mathematical model (called a regression model) of the correlation between the predicted (dependent) variable and the predictor (independent) variable.

Accordingly, specific SRiM competencies including *Description, Interpretation, Prediction* can be regarded as the ability to efficiently and scientifically perform the tasks in the medical statistical processes, respectively, the process of *Description, Interpretation, Prediction.*

The Bloom's taxonomy, **Math Assessment Task Hierarchy and Classifying Mathematical Literacy according to PISA for Statistical Reasoning in Medicine**

Research on math assessment (Brown, 2010; Tan, 2011) has shown that because learning math is often seen as mastering a set of skills, procedures, and formulas, traditional math essay assessments generally focus on skills to be mastered, by assessing students' computational skills or their ability to recall information about procedures and formulas in memory. The questions that appear in the written test mainly test the individual skills to solve a problem while not testing whether the student understands the math concept or can integrate math knowledge to solve the problem or whether or not they can communicate using the language of math. Because the purpose of Math education changes and aims at a broader goal such as developing students' ability to think mathematically, to apply their knowledge to solve real-life problems, it is no longer appropriate to assess students' knowledge by simply asking them apply formulas and calculate answers, but solve real-world problems and make mathematical reasoning. In order to develop the objectives of the medical statistics module to meet the output standards of the medical doctor training program and in the direction of SRiM-focused innovation, and to build a medical statistical reasoning competency rating framework corresponding to the goals, we study the applications of the Bloom's revised taxonomy in the assessment of Mathematics. Bloom's taxonomy represents a hierarchy from low to high thinking, this idea is applied in the design of tools to assess learners' learning level for each specific subject, including Mathematics. Question compilation at the lower levels of the Bloom taxonomy: Remember, Understand, Apply is often appropriate to assess preparation and understanding; diagnose strengths, weaknesses and failures; review or summarize the content. Higher-level questions: Analyze, Evaluate, Create are often appropriate to encourage more critical and deeper thinking; problem-solving, encourage discussion and communication; Motivate learners to find information on their own. In researching math assessment, Smith et al. (1996) showed that Bloom's taxonomy is good for structuring program objectives and assessment tasks, but there are limitations when used for mathematics. The thing research group interested in is the skills required to complete a particular math task and is geared toward developing advanced math assessments across a wide range of knowledge and skills. The MATH taxonomy identifies 8 descriptors of skills and knowledge, arranging them into 3 groups A, B, C. These descriptors are arranged according to the nature of the activity required to complete the task well. Groups A, B, C corresponding to 3 levels are arranged in the order from low to high appropriately according to the context.

Level A: Reproduction Familiar processes			**Level B: Connection** Use your existing math knowledge in new ways		**Level C: Reason** Apply conceptual knowledge to construct mathematical arguments		
A1	A2	A3	B1	B2	C1	C2	C3
Factual knowledge	Comprehension	Routine use of procedures	Information Transfer	Applications in new situations	Justifying and interpreting	Application, conjectures and comparisons	Evaluation

Table 1: MATH Taxonomy (Smith et al., 1996)

Level A (Reproduction) involves recalling events, formulas and recognizing familiar situations and calculations and applying given algorithms. Level B (Connection) continues with the classification of a mathematical object, application in a situation or an answer, and the ability to design a plan or select features to perform an independent assignment. Level C (Reason) involves reasoning, justifying, counterexamples, arguing or proving, stating or discovering patterns, constructing an example, or extending a concept. The test questions for tasks in group A are mainly about recalling formulas to solve familiar problems. The math skills associated with group C are "those that we associate with practicing mathematicians and problem solvers" (Pountney et al., 2002). Some of the limitations of the MATH taxonomy are as popular as most other taxonomy, that is, some tasks may involve the use of more than one type of knowledge or activity. Even more advanced skills have procedural parts. In the assessment of mathematics literacy, PISA selects cognitive activities to classify math skills into three clusters: *Reproduction, Connections, Reflection*, each of which corresponds to two levels: Reproduction level 1, 2 is considered low capacity; Connections level 3, 4 indicates the average capacity and Reflection level 5, 6 is assessed high capacity. The combination of low-to-high competencies including Reproduction, Connections and Reflection in the PISA classification are presented in Table 2.

Competencies	**Clusters**	**Levels**	**Description**
High	*Reflection*	5, 6	*Be proficient and apply math literacy in all situations:* Set and solve complex problems. Reflective and insightful. Access to original math. Complex methods. Generalization.
Average	*Connections*	3, 4	*Know math knowledge and can apply it:* Modeling. Displacement and interpretation,

			standard problem-solving. The methods are well defined in various respects.
Low	*Reproduction*	1, 2	*Not proficient in math knowledge:* Standard presentation and definitions. Computations, procedures, familiar problem-solving.

Table 2: Classification of Math Literacy in PISA (OECD, 2009)

Developing a general framework for assessing Statistical Reasoning in Medicine

In our research on statistical literacy in medicine, SRiM and statistical thinking in medicine, statistical literacy in medicine is considered as the development of basic skills and knowledge and provides the necessary foundation for the development of SRiM and statistical thinking in medicine. With that in mind, when compared with Bloom's revised taxonomy of thinking levels, with 6 levels from low to high, we assume that SRiM may be suitable with the level of Understand and some aspects of the higher levels than Apply and Analyze. The mathematics literacy in classification PISA can also be used to develop an SRiM competency assessment. According to the description of the clusters of reproduction, connections, and reflection capability of PISA, these clusters have a relatively consistent correspondence with MATH's three-level cognitive classification. The overall framework for assessing SRiM competency that we propose as shown in Table 3 consists of six levels from low to high, corresponding to three clusters of competencies *Reproduction, Connections and Reflection*. Similar to the use of "verbs" to describe in Bloom's taxonomy, each task level is described in detail and begins with verbs such as recall, recognize, identify, describe, explain, implement, apply, use, and verify. Corresponding to the levels of the MATH taxonomy, PISA's mathematics literacy classification, in the rating scale we propose, level 1, 2 correspond to the *Reproduction* cluster, level 3, 4 correspond to the *Connections* cluster and level 5, 6 correspond to the *Reflection* cluster.

Clusters	Levels	Task Description
Reflection	6	*Generate new results through mathematical modeling to use medical statistical reasoning to solve real-world problems:* Explain statistical processes and can fully explain statistical results. Use a higher level of thinking and reasoning skills in a statistical context to create mathematical representations of real medical situations. Use insight, reflection, and reasoning to accurately communicate results.

	5	*Apply medical statistical reasoning to analyze and interpret data and draw overall-related conclusions from the data:* Applying statistical knowledge in complex medical situations, which in a sense somewhat structured and mathematically represented clearly explained. Use deductive insight to explain given information. Verify or explain a given result in a medically factual context. Deduct an application or hypothesis for a real-world medical situation. Reflect on their activities, establish and communicate their explanations and reasoning.
Connections	4	*Routine use of procedures and conceptual understanding in solving unfamiliar medical problems:* Selecting suitable statistical formulas and methods for a new situation for a specific context. Apply a concept to a real-life situation or other concepts. Apply basic statistical and probability concepts combined with logical reasoning in less familiar situations. Use reasoning based on the explanation of data, data representation, and statistical summaries. Constructing and communicating explanations and arguments: Explaining text, transferring written descriptions into statistical problems.
	3	*Understand statistical concepts, formulas, and procedures:* Interpret information and data. Explain information from the representation. Link different sources of information. Convert information from one form to another, from one concept to another. Explain familiar statistical concepts, formulas, and procedures in a process sense. Use basic reasoning with simple statistical and probability concepts. Forming an argument from a specific problem with a medical context.
Reproduction	2	*Recall formulas, use familiar procedures:* Practice some basic statistics skills such as organizing data, building, and representing tables, and working with different representations of data. Retrieve simple formulas, algorithms, and step-by-step procedures. Explain or read the results of a simple statistical procedure. Implement a statistical procedure or algorithm in a familiar, similar medical context.
	1	*Retrieve knowledge:* Recall information, facts, formulas and recognize situations, perform familiar calculations.

		Identify statistical information presented in familiar graphical formats. Identify and use basic probability ideas in simple and familiar medical-related experimental situations. Apply the given statistical formulas and algorithms.

Table 3: Overall framework for assessing statistical reasoning in medicine

For medical students, we define levels 1, 2, 3 as lower thinking levels and levels 4, 5, and 6 as higher thinking levels. Levels 1, 2, 3 in the new assessing scale only show the students' performance in remembering formulas, procedures, statistical calculations and basic statistical understanding, level 4 and higher clearly show the student's performance on SRiM.

Model for assessing Students' Statistical Reasoning in solving real-life medical problems

The objective of the PISA assessment is to examine students' ability to flexibly apply knowledge and skills in fundamental areas of expertise to new contexts and problems encountered in real life. PISA focuses on mastering processes, understanding concepts, and being able to handle different situations in each field. We find that this point of view of PISA fits the goal of assessing medical students' statistical reasoning in solving real-life problems. Learning Math for life is a PISA perspective. According to the OECD (PISA, 2010), assessment in the field of mathematics literacy includes students' abilities to analyze, reason, and communicate mathematical ideas effectively as they place, formulate, solve, and interpret mathematical solutions to math problems in a variety of contexts. Assessment involves Mathematical Content (overarching ideas: space and shape, relationships, uncertainty, Mathematical Process (defined by mathematical competencies) and Context where math is used. Uncertainty is one of four overarching ideas, proposed in two related topics, data and chance, corresponding with the topics of probability and statistics, respectively. Thus, in a sense, mathematics literacy includes the ability to make statistical inferences. Applying PISA's assessment of mathematics literacy, we propose a model to assess students' SRiM competency when solving real-life problems as shown in Figure 1.

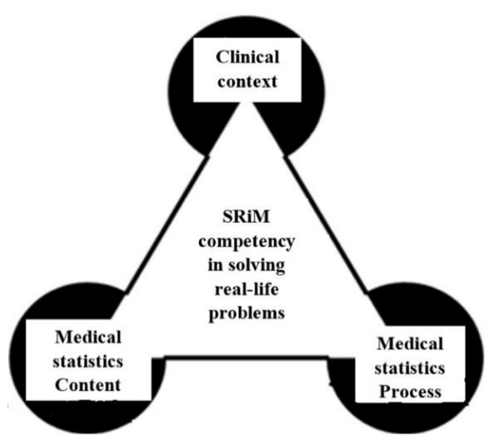

Figure 1: Model for assessing students' SRiM competency in solving real-life problems

In the diagram (Figure 1), medical statistics Content includes basic ideas used to solve problems, medical statistics Process is defined by three SRiM competencies, which are SRiM competence *Description, Interpretation, Prediction* and clinical context that uses SRiM. In a real-life context relevant to clinical medicine, it requires the ability to apply relevant knowledge and skills in medical statistics and medical statistical reasoning in a less structured context, students have to make decisions about what knowledge might be involved, what processes or procedures will lead to a possible solution, and have to use medical statistical reasoning in various degrees. Medical statistics content should focus on the big ideas of medical statistics, including Data; Distribution; Trend; Variability; Paradigm; Relationship; Samples and sampling; and Statistical Reasoning.

Framework for assessing students' Statistical Reasoning in solving medical real-life problems

The SRiM competencies are described according to three cognitive capability clusters as shown in Table 4. In the general assessment scale of SRiM competency we have proposed, six levels from low to high corresponding to the three clusters of reproduction, connections and reflection abilities. In which level 1, 2 corresponds to the reproduction cluster, level 3, 4 corresponds to the connections cluster and level 5, 6 corresponds to the reflection cluster. The cognitive capability levels of each SRiM competency are specifically described through three frameworks for assessing SRiM competency *Description, Interpretation, Prediction*.

Cognitive abilities cluster / SRiM competencies	*Reproduction*		*Connections*		*Reflection*	
	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Description	–	–	–	–	–	–
Interpretation	–	–	–	–	–	–
Prediction	–	–	–	–	–	–

Table 4: Two-dimensional matrix of three SRiM competencies and three ability clusters

We consider the SRiM competency *Description* as the ability to efficiently and scientifically perform the tasks in the medical statistical processes *Description*. Accordingly, the SRiM competency *Description* can be summarized by three key components: (1) Data description, organization and representation; (2) Description of the data generation process: description of the sampling process, sampling distributions; (3) Generating scientific hypotheses which need to be justified: forming preliminary overall predictions. As such, the SRiM competency *Description* can relate to various types of medical statistical reasoning including reasoning about data, reasoning about data representations, reasoning about statistical quantities. Accordingly, we build a framework for assessing SRiM competency *Description* which describes in detail each assessing level.

The SRiM competency *Interpretation* as the ability to efficiently and scientifically perform the tasks in the medical statistical processes *Interpretation*. Accordingly, SRiM competency *Interpretation* can be summarized by four key components: (1) Recognizing trends in data, using descriptive statistics from samples to make inferences and predictions about overall characteristics; (2) Justifying scientific hypotheses: generalizing conclusions about the population based on samples. Usually, first of all is the formulation of statistical hypotheses, often in a form different from the clinical hypothesis. Sample data are used to justify these hypotheses. The justification procedure is a special case of two-choice decision theory, where the decision strategy contains only one error, which is the risk (probability) of making the wrong decision. More specifically, the error is measurable, controlling for the probability of each of two possible types of error: choose hypothesis H_0 while hypothesis H_1 is true and choose hypothesis H_1 when hypothesis H_0 is true; (3) Reasoning for the plausibility of the results based on ideas about the correlation between the samples and the population, the representativeness of the sample with respect to its population; (4) Interpret the likelihood of outcomes using an understanding of the ideas of randomness, chance, and uncertainty. As such, SRiM competency *Interpretation* can involve various types of medical statistical reasoning including reasoning about statistical quantities, samples, uncertainty and correlation. Accordingly, we build a framework for assessing SRiM competency *Interpretation* which describes in detail each assessing level.

The SRiM competency *Prediction* is the ability to efficiently and scientifically perform the tasks in the medical statistical processes *Prediction*. SRiM competency *Prediction* can be summarized by four key components: (1) Determining the existence of a correlation between variables, assessing the degree of association between variables; (2) Description of the correlation between variables: The description is based on a hypothetical model of the correlation, with at least one variable being an independent variable, i.e. hypothetical causal factor, predictor outcome and at least one dependent variable, the predictor outcome factor. This model can describe the fit, how good or bad the relationship is, how good the fit of the model is. If the model is incomplete, another suitable model can be found that better describes this correlation; (3) Testing the quality of describing the correlation between variables: using the test results to assess the nature of the correlation; (4) Predict the value of outcomes based on causal values: when the model is decided based on an understanding of the underlying physical or physiological factors that make the association, regression allows predict the value of outcomes arising from a specific clinical indicator of the independent variable. Thus, SRiM competency *Prediction* can involve inferring skills about uncertainty, reasoning about correlations. Within the scope of this article, we present in more detail frameworks for assessing SRiM competency *Prediction* in Table 5.

Clusters	Levels	Task Description
Reflecti on	6	Explain statistical processes and can fully and deeply interpret statistical results when considering correlations between variables, making predictions about variables.

		Use high-level thinking and reasoning skills in a statistical context to create mathematical representations of real medical situations. Use insight, reflection, and reasoning to accurately communicate results.
	5	Apply a deep understanding of statistical knowledge related to correlation analysis, univariate and multivariable linear regression analysis to complex medical research situations that are in some sense structured and somewhat clearly mathematically represented. Using reasoning insights to interpret given information, test, and interpret a given outcome about the correlation between variables in a real medical context. Create or work effectively with hypotheses for a real medical situation, with complex models. Reflect on their activities, form and communicate interpretations and reasoning.
Connections	4	Apply understanding of key concepts such as correlation coefficients, intercepts and regression coefficients, coefficients of determination, univariate and multivariate regression models to describe and interpret the correlations between variables through the data set. Use reasoning to explain statistical processes and fully interpret statistical results: interpret the reasonableness of the results about the correlation between variables; interpret the reasonableness of the established regression equation; make predictions about the outcome variable based on the predictor variable. Construct and communicate interpretations and reasoning: Interpreting descriptions included in an unfamiliar scientific situation; transferring written descriptions into statistical problems.
	3	Interpret information and data. Link different sources of information. Transfer information from one form to another, from one concept to another. Explain the meaning of concepts, formulas about correlation coefficient, intercept and regression coefficient, coefficient of determination, univariate and multivariable linear regression model, sample regression function. Explain the difference between correlation and regression. Use basic inference with simple statistical concepts in describing correlations between variables. Form an argument from a specific problem in a medical context. Select and apply options to solve simple problems by correlation analysis methods, univariate or multivariable linear regression.

Reproduction	2	Recall the formulas, algorithms, rules and conventions in linear regression analysis, correlation analysis: the formula for determining the Pearson correlation coefficient; univariate and multivariable linear regression models; estimation of the parameters in the regression model, the sample regression function; assumptions are reasonable. Interpret or read the results of a simple statistical procedure applied in situations where no more than a direct conclusion is required: read the results of the Pearson, Spearman correlation test; read the results of the conformity testing procedure of the regression model; read the results of the test procedure for the assumptions in the correlation analysis; Read the scatter plots in residual analysis to test assumptions in regression analysis. Implement simple procedures and algorithms in linear regression analysis, correlation analysis in the familiar, similar medical context.
	1	Directly apply the given formulas and algorithms to determine the estimated parameters in the regression model (intercepts and slopes or regression coefficients, coefficients of determination, correlation coefficients for simple experiment) Recall the meaning of correlation coefficients, parameters of the regression model, and coefficients of determination to evaluate the correlation between variables through the data set. Realize the correlation between variables through data tables, scatter plots.

Table 5: Framework for assessing SRiM competency *Prediction*

Build Test questions for assessing students' Statistical Reasoning in solving medical real-life problems

The general competency assessment scale and the detailed described SRiM competency scale *Description*, *Interpretation, Prediction* are an important basis for us to formulate assessment questions. Another basis for us to formulate assessment questions is the Course Alignment Map, with 5 course objectives and 29 objectives corresponding to 7 lectures, of which 21 are related to medical statistics. According to the SRiM competency assessment scale, each objective is identified corresponding to an SRiM competency cluster. Research on PISA assessment structure shows that PISA'S mathematical literacy assessment is done through a combination of the following question types: multiple-choice question (MCq), close-ended question (CEq) and open-ended question (OEq). According to lessons learnt in developing and using assessment questions for PISA exams, MCq is generally considered the most suitable for evaluating reproduction and connections capacity clusters. For some higher purposes and more complex processes, such questions as CEq or OEq will often be preferred. The main feature of OEq is that it allows students to prove their abilities by providing different solutions at various levels of mathematical complexity. To ensure a certain preset weight (such as a score weight) the percentage (%) of the questions in the *Reproduction: Connections: Reflection* of approximately 25: 50: 25 is appropriate.

Example 1. Matrix of test questions with a combination of both forms of objective multiple-choice questions (MCqs) and essay questions (CEq and OEq). The percentage (%) of the score weight of the questions in the cluster *Reproduction: Connections: Reflection* was ensured to be approximately 25: 50: 25, respectively. Table 6 is a matrix of SRiM competency assessment tests after students finish the module medical statistics (Matrix 1).

Course Objectives	*Reproduction*		*Connections*		*Reflection*		**Sum**		
	MCq	**Essay**	**MCq**	**Essay**	**MCq**	**Essay**	**MCq**	**Essay**	**Scores**
Objective 3	3		1	1 *(1)*			4 *(1.6)*	1 *(1)*	*(2.6)*
Objective 4	3		2	1 *(1.5)*		1 *(1.5)*	5 *(2)*	2 *(3)*	*(5)*
Objective 5			1	1 *(1)*		1 *(1)*	1 *(0.4)*	2 *(2)*	*(2.4)*
Sum	6 *(2.4)*		4 *(1.6)*	3 *(3.5)*		2 *(2.5)*	10 *(4)*	5 *(6)*	*(10)*
	(2.4)		*(5.1)*		*(2.5)*				

Table 6: Matrix of test questions for assessing SRiM competency (Matrix 1)

The ratio between MCqs and essay questions is 4: 6, the total test time is 60 minutes so the corresponding time is 24 minutes for MCqs (4 points) and 36 minutes for Essay questions (6 points). The set of questions includes 10 MCqs, 4 CEqs and 1 OEq. Each MCq has the same score weight of 0.4 and the average time to complete is about 2.4 minutes. The Essay question is numbered in accordance with the intended time of completion and the complexity of the reasoning intended to be evaluated, where the Essay question has a weighted score of 1 corresponding to an average time of completion of about 6 minutes. Weighted points corresponding to each course objective from 3 to 5 is 2.6: 5: 2.4. Weighted score corresponding to each level of reasoning *Reproduction: Connections: Reflection* is 2.4: 5.1: 2.5 which means that the percentage of 75% at the level of reproduction and connections, 25% at the reflection level, is in line with the ratio of the set goals. In Table 6, in each cell: the digit outside the bracket is the number of questions; the digit inside the bracket is the corresponding point weight.

We develop Test 1 corresponding to matrix 1. A set of 15 questions is the SRiM competency assessment level *Description, Interpretation, Prediction* and one score weight (Table 7).

	PROBLEM	**Question**	**Question type**	**Clusters**	**SriM levels**	**Scores**
1	Coronary vessels	Question 1	MCq	*Reproduction*	2	0.4
2		Question 1	MCq	*Reproduction*	2	0.4

3	Protein concentration	Question 2	MCq	*Connections*	3	0.4
4	Hemoglobin concentration	Question 1	MCq	*Reproduction*	2	0.4
5	ALT concentration	Question 1	CEq	*Connections*	4	1.0
6	Cholesterol concentration	Question 1	MCq	*Reproduction*	2	0.4
7	Beta-Crosslaps concentration	Question 1	MCq	*Reproduction*	2	0.4
8		Question 2	CEq	*Connections*	4	1.5
9	Drug addiction treatment	Question 1	MCq	*Connections*	3	0.4
10	Sanspeed group	Question 1	MCq	*Reproduction*	2	0.4
11		Question 2	MCq	*Connections*	3	0.4
12	Prostate cancer	Question 1	CEq	*Reflection*	5	1.5
13	Carotid intima-media thickness	Question 1	CEq	*Connections*	4	1.0
14	Relationship between Glucose and Hormone	Question 1	MCq	*Connections*	3	0.4
15		Question 2	OEq	*Reflection*	5	1.0

Table 7: Describe the test questions corresponding to matrix 1 (Test 1)

Example 2. Considering the question 5 in Test 1. In terms of describing the problem, this is the Question 1 of the problem "Alt concentration".

PROBLEM: ALT CONCENTRATION

The ALT levels (U/L) of people with chronic hepatitis B who visit a hospital do not follow normal distribution. The distribution of ALT yeast concentrations is asymmetrical and deviates right (skewed towards higher values) with the mean of 301.8 and the standard deviation of 363.9.

Question 1. (Level 4) One research team generates random samples of 60 chronic hepatitis B patients each. Describe and interpret the distribution of the sample average ALT yeast concentrations.

Analysis. Question type: CEq. Medical statistic content: Description of the sampling distribution of means. SRiM competence: *Description*.

Clinical context: The question builds in the context of a study describing the ALT index (U/L), which is produced by the liver and usually has a fixed content in blood, when the liver is damaged, the ALT enzyme content will increase. The ALT enzyme index helps doctors in diagnosing liver disease, as well as for monitoring the treatment of patients with liver disease, tracking the progression of the disease with treatment.

Cluster of competence: *Connections* – level 4. The above question is aimed to assess the SRiM capacity of students in relation to the description of the sampling distribution, namely, description of the sampling distribution of means. A sampling distribution is a frequency distribution or frequency of a statistical sample based on random samples drawn from the population. This question is consistent with level 4 of our SRiM competency scale. Students have to solve a problem in an unfamiliar situation but involves familiar elements. Students need to think and reason to understand the problem statistically, make arguments and present arguments appropriately. To solve the problem, students need to determine that this is a descriptive problem of the sample mean distribution in the case of a random sample of which sample size $n = 60$ drawn from the whole that does not satisfy the standard distribution, which has finite expectations and variances. Then, knowledge of Central limit theorem is applied to describe the whole variability of the sample mean distribution including shape, center, and dispersion. The central limit theorem ensures that when samples are large enough, the probability distribution of the sample mean $\overline{X}$ will approximate the normal distribution even if the samples are generated from a whole that does not satisfy the normal distribution and $E\overline{X} = \mu$, $Var\overline{X} = \sigma^2/n$.

The question requires students to: Read and interpret a more complex context including overall identification, random variables, random samples, sample sizes, and distribution characteristics of the whole (non-normal distribution, expectation $\mu = 301.8$ và standard deviation $\sigma = 363.9$); To apply the central limit theorem to determine the sample mean distribution characteristic; Infer and present the results of this process.

For this question, the successful student is the student who fully describes the distribution characteristics of the sample average ALT yeast concentrations: the approximately normal distribution shape, with the sample mean of approximately 301.8, the sample standard deviation of approximately 46.98, and provide appropriate interpretations based on the central limit theorem.

Example 3. Considering the question 15 in Test 1. In terms of describing the problem, this is the Question 2 of the problem "Relationship between Glucose and Hormone".

PROBLEM: RELATIONSHIP BETWEEN GLUCOSE AND HORMONE

Study of the dependence of the content of Glucose (blood glucose) and Hormone. The blood glucose (mmol/L) and Hormone (IU/mL) levels of a sample of 35 patients are identified and stored in the file Data.sav. The obtained result of data processing and the linear regression model includes: Hormone = –0.44. Glucose + 5.77; $R^2 = 0.56$; *t*-test of the coefficient of regression (slope) has Sig. = 10^{-4}.

Question 2. (Level 5) From the obtained results of the data sample, write a report that makes well-founded statements for the above study. From there, provide an example of predicting a patient's hormone levels when knowing that patient's blood glucose index.

Analysis. Question type: OEq. Medical statistic content: Simple linear regression analysis. SRiM competence: *Prediction*

Clinical Context: The question builds on the context of studying the association between the levels of Glucose and Hormone. Blood glucose is an important indicator that helps doctors assess patients' body ability to control blood glucose, so that he can determine whether the patients have glucose-related diseases (diabetes), and can also assess whether the patients with diabetes are responding to the ongoing treatments or not. Hormones are substances produced by the endocrine glands that have a tremendous effect on the body's processes. They affect growth and development, mood, sexual function, reproduction and metabolism.

Cluster of competence: *Reflection* – Level 5. The above question was asked to assess the *Prediction* SRiM capacity of student relation to simple linear regression analysis. This question is consistent with level 5 described in the SRiM competency assessment scale we developed (the general assessment scale and the reflection SRiM competency assessment scale). Based on the provided information, writing a report for this study means a fully explained description of the association between glucose and hormone levels, students need to have a deep understanding of the statistical concepts in univariate linear regression analysis, must know how to integrate concepts and apply inferential insight to verify and interpret information and make well informed statements.

The question requires students to: Understand the question in detail. Write reports for the study includes: Read and interpret in a context to determine whether or not there exists an association between glucose and hormone, how close the association is; Whether the univariate linear regression model is suitable for describing the association between glucose and hormone, whether changes in glucose significantly account for hormone levels; A sample regression function is an estimate of the overall regression function from a sample of data, which is meant to interpret and predict the correlation between two variables. After the suitability of the sample regression model through the results of the *t*-test coefficient of regression (slope) and determination coefficient R^2, a specific example of the significance of the regression model is provided which allows to predict the value of the outcome arising from a specific clinical indicator of the variable glucose. Being able to infer flexibly, establish interpretation and effectively present the results of this process.

This is an open-ended question, to which there can be various answers. Successful students are students who know how to fully exploit 3 given information: With the given information "$R^2 = 0.56$", this is the fixed coefficient of the regression model, determining the relevance of the sample regression function to the data, indicating that in 100% of the entire deviations of the data from the mean 56% is due to glucose, 44% is due to random errors and other factors (if any) that we don't include in this model to consider. Given that "*t*-test coefficient of regression (slope) has Sig. = 10^{-4}", slope b other than 0 is statistically significant ($p = 10^{-4} < 0,01$), there is an association between glucose and hormone, a linear regression model is appropriate to describe this association; Given the information "Hormone = –0.44. Glucose + 5,77" is a sample regression function, which is an estimate of the overall regression function, where an estimate is $b = -0.44$ và $a = 5.77$, showing an inverse correlation between glucose and hormone, within a consistent limit of glucose content, when glucose increase by 1 (mmol/L) the hormone decreases by 0.44 (IU/mL). The sample regression function allows the prediction of hormone levels in the blood for patients with specific glucose levels, for example, for patient with a blood glucose index of 6.5 (mmol/L), predicted blood hormone levels is then 2.91 (IU/mL).

The test scoring chart is developed in detail right from the compilation of the test matrix. Basing on the matrix of test questions to determine the score scale corresponding to SRiM levels and the test score (Table 8).

Clusters	SriM levels	Scores
Reflection	5, 6	(7.5; 10]
Connections	4	(4; 7.5]
	3	(2.4; 4]
Reproduction	1, 2	(0; 2.4]

Table 8: The score scale corresponding to SRiM levels of matrix 1

METHOD

Researches on math education focus on high school students, but in comparison to recognizing the shortcomings of assessment, new trends in assessment goals, and basic principles of math assessment, we think that there is a consistency with the assessment in teaching medical statistics to medical students. Therefore, we apply the theories in the assessment of math education as a reference for the research of assessing the medical statistical reasoning competency of medical students. Firstly, we apply a combination of Bloom's taxonomy, MATH taxonomy and PISA's mathematics literacy classification to propose an overall framework for assessing SRiM competencies. After that, applying PISA's assessment of mathematics literacy, we built a model to assess students' SRiM competency when solving real-life problems. From that, we built a framework for assessing students' SRiM competency in solving real-life problems. The framework is an important basis for designing a set of evaluation questions, we built Test 1 of MCqs, CEqs and OEqs. We conducted experiments with the self-developed assessment toolkit, analyzed the results and drew conclusions. We experimented on a classroom of $N = 103$ first-year students of the university medicine and pharmacy in Viet Nam. After completing the medical statistics course, the students took Test 1, in 60 minutes with a pen and paper. We collected student exam answers for Test 1 and use quantitative analysis methods combined with qualitative to analyze experimental results.

RESULTS

Test 1 score processing results help us determine the percentage (%) of students who reach the SRiM levels shown in Table 9. As can be seen in Table 9, among the surveyed students, the percentage of students who only achieve a low level of competence (reproduction) occupies only 7.8%, a majority of students who reach the competence levels 1, 2, 3 occupies 57.3%. Most students achieve competency levels at the levels of 3, 4 (connections) which occupies 79.6%, at higher levels of 5, 6 (reflections) is still low, occupying 12.6%. This proves that, SRiM level of students is rather alike, most of them are well equipped with basic statistical skills. Achieving the reproduction capacity level is not difficult for most students, according to the descriptive rating scale for the level of reproduction, that is, the task requires the export of knowledge, recall the formula, using familiar processes is not difficult for most students. Nevertheless, students meet difficulty in achieving a reflection level of competence, high levels of competence require students to know how to creatively apply known knowledge to produce new things, to deal with unfamiliar problems using unknown forms or apply SRiM to solve practical medical problems.

Cluster	SRiM level	Scores	Percentage (%) (N = 103)
Reflection	5, 6	(7.5; 10]	12.6
Connections	4	(4; 7.5]	30.1
	3	(2.4; 4]	49.5
Reproduction	1, 2	(0; 2.4]	7.8

Table 9: Percentage (%) of students achieving the levels of SRiM for Test 1

To know more about the students' SRiM capacity, we analyze students' answers to each question and found that a majority of students have good answers to questions at a reproduction level, for 6 questions requiring higher competency at the reproduction level (Question 1, 2, 4, 6, 7, 10) in Test 1, statistical data show that the percentage (%) of students who answers each question correctly at this level reaches 87.4% or more. Nevertheless, regarding the two questions at reflection level, for the percentage of students who achieves maximum score is 8.7% regarding Question 12 while only 1.9% regarding Question 15.

Example 4. Considering the students' answers to Question 5 in Test 1. Statistical data of students' responses to this question are shown in Table 10.

As can be seen from Table 10, among the participants, 62.1% of students gave the correct answers on the distribution characteristics of the mean ALT enzyme concentration of samples.

$N = 103$	Scores		
	0	0.5	1.0
Percentage (%)	38.8	23.3	37.9

Table 10: Results of answers corresponding to question 5 of Test 1

However, only 37.9% of students are successful with the question at this *Connections* level. Figure 2, Figure 3, Figure 4 are the work of three students (S1, S2 and S3) to this question.

The students' works show that, S1 can summarize the problem by converting information into symbols, but fails to answer the raised question;

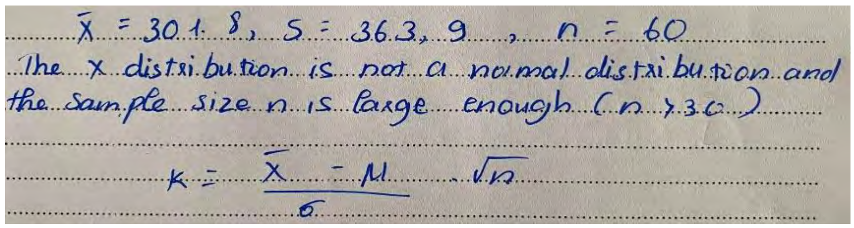
$\bar{X} = 301.8$, $S = 363.9$, $n = 60$
The X distribution is not a normal distribution and the sample size n is large enough ($n > 30$)
$K = \frac{\bar{X} - \mu}{\sigma} \sqrt{n}$

Figure 2: Student S1' answers to Question 5 of Test 1

S2 can identify the distribution shape of the sample mean ALT yeast concentration but can't provide an interpretation;

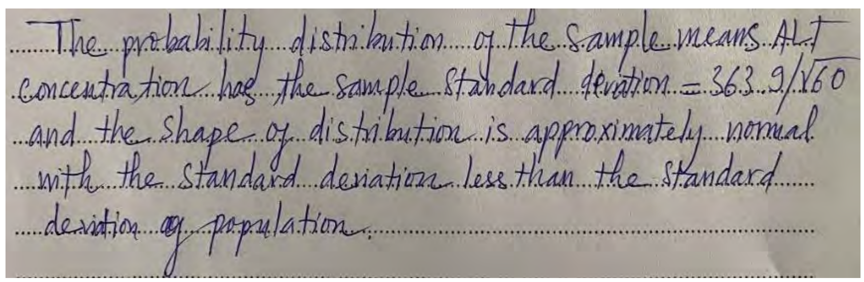
The probability distribution of the sample means ALT concentration has the sample standard deviation = 363.9/√60 and the shape of distribution is approximately normal with the standard deviation less than the standard devition of population.

Figure 3: Student S2' answers to Question 5 of Test 1

S3 can define distribution shape, distribution characteristics, and provided an interpretation using the central limit theorem, however the interpretation is still incomplete.

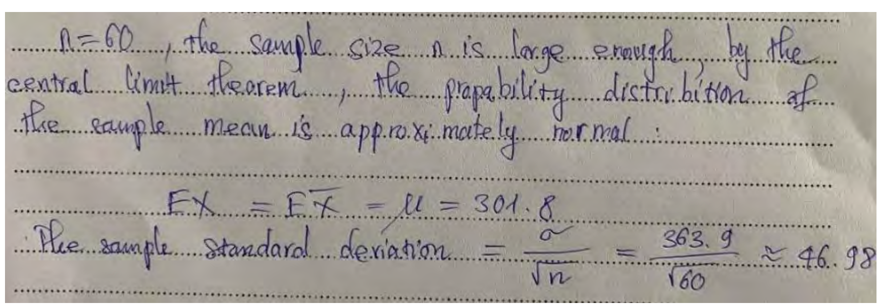
n=60, the sample size n is large enough, by the central limit theorem, the prapability distribition of the sample mean is approximately normal.
$E\bar{X} = EX = \mu = 301.8$
The sample standard deviation $= \frac{\sigma}{\sqrt{n}} = \frac{363.9}{\sqrt{60}} \approx 46.98$

Figure 4: Student S3' answers to Question 5 of Test 1

This question allows us to assess students' SRiM capability *Description* at the level of *Connections*. Through our analysis of the students' work on this question, we found that although students are able to accurately predict the characteristics of the sampling distribution, it is clear that they have difficulty in fully integrating all information to make arguments.

Example 5. Considering the students' answers to Question 15 in Test 1.

Statistical data of students' responses to this question are shown in Table 11.

$N = 103$	Scores		
	0	0.5	1.0
Percentage (%)	77.7	19.4	2.9

Table 11: Statistical results of answers to Question 15 of Test 1

The results of the students' work show that three students successfully complete accounting for only 2.9%, most students fail this question, 77.7% of students leave the answer blank, the rest either answers but don't have any correct answer to any part of the question, or can give an example of predicting hormone levels with patients with specific blood level index but can't explain why linear regression model is appropriate to describe the association between glucose and hormone. Except the three students who have maximum score for this question, most students only stop at the calculation level, can provide a specific glucose value and input into the sample regression function to identify hormone levels; however, none of the students can come to a conclusion and fully explain given the information "*t*-test of the coefficient of regression (slope) has Sig. = 10^{-4}". For example, Figure 5 and Figure 6 are the work of two students (S4, S5) to this question.

The students' works show that, student S4 can give an example of predicting hormone levels with patients with specific blood glucose index, but has not previously tested whether the linear regression model is appropriate to describe the association between glucose and hormone. S4 just stop at calculation, have not done SRiM, have not achieved SRiM levels.

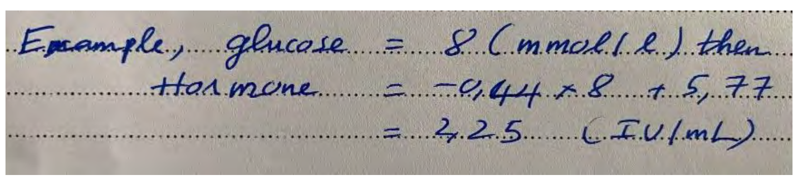

Figure 5: Student S4' answers to Question 15 of Test 1

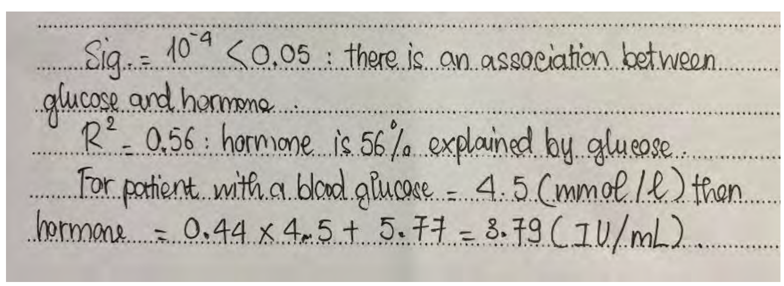

Figure 6: Student S5' answers to Question 15 of Test 1

Student S5 gives an example of predicting hormone levels for patients with specific blood glucose index, before considering and evaluating the relationship between glucose and hormone based on the coefficient of determination $R^2 = 0.56$, however, fully explain given the information "t-test of the coefficient of regression (slope) has Sig. = 10^{-4}", not fully explained to conclude that the linear regression model is suitable to describe the relationship between glucose and hormone. S5 achieved competency at the level 3 (*Connections*) in the cluster *Prediction.*

We conducted experiments and analyzed empirical results based on the assessment scale and assessment questions set (Test 1). The assessment scale is both the basis for building assessment questions and the basis for analyzing students' responses, from which assessments of students' SRiM competency when solving real-life problems. Analysis of initial empirical results reveals the effectiveness of the assessment scale and assessment questions set, allowing an overview of students' SRiM competency, through analyzing students' responses to each question, as well as analyzing each student's responses, we can assess this student's SRiM capacity, realize the difficulties students face in the process of reasoning, as well as the current situation of medical statistics teaching. Through the analysis of the results of the surveyed students' work, we obtained an overview of the student's SRiM capacity when solving practical problems: Most students achieved competencies at the levels 1, 2, 3, the number of students at the levels 4, 5, 6 are still very low. That is, most of them have achieved good results in basic statistical skills and understanding (shown at the levels 1, 2, 3), which is the foundation for the development of SRiM, medical statistical thinking, but it is clear that medical statistics teaching has not fully promoted that potential of students. Students can only solve problems that are partially or fully mathematically modeled. When facing essay problems in real world medical contexts, they can't successfully apply mathematical models to solve, especially meet difficulty with open-ended questions, have limited ability to create effective solutions to solve unknown problems, lack of flexibility in planning to solve unfamiliar practical medical problems. The limitation of our experiment is that we have only conducted on 103 first year students of 1 university, not yet diverse target study participants from many other medical universities in the country to have a comparison of SRiM competencies of different target groups, more empirical evidence to support the effectiveness of the assessment scale and assessment question set.

We conducted experiments and analyzed empirical results based on the assessment scale and assessment questions set (Test 1). The assessment scale is both the basis for building assessment questions and the basis for analyzing students' responses, from which assessments of students' SRiM competency when solving real-life problems. Analysis of initial empirical results reveals the effectiveness of the assessment scale and assessment questions set, allowing an overview of students' SRiM competency, through analyzing students' responses to each question, as well as analyzing each student's responses, we can assess this student's SRiM capacity, realize the difficulties students face in the process of reasoning, as well as the current situation of medical statistics teaching. Through the analysis of the results of the surveyed students' work, we obtained an overview of the student's SRiM capacity when solving practical problems: Most students achieved competencies at the levels 1, 2, 3, the number of students at the levels 4, 5, 6 are still very low. That is, most of them have achieved good results in basic statistical skills and understanding

(shown at the levels 1, 2, 3), which is the foundation for the development of SRiM, medical statistical thinking, but it is clear that medical statistics teaching has not fully promoted that potential of students. Students can only solve problems that are partially or fully mathematically modeled. When facing essay problems in real world medical contexts, they can't successfully apply mathematical models to solve, especially meet difficulty with open-ended questions, have limited ability to create effective solutions to solve unknown problems, lack of flexibility in planning to solve unfamiliar practical medical problems. The limitation of our experiment is that we have only conducted on 103 first year students of one university, not yet diverse target study participants from many other medical universities in the country to have a comparison of SRiM competencies of different target groups, more empirical evidence to support the effectiveness of the assessment scale and assessment question set.

CONCLUSIONS

In this article, we have applied the theories of assessment in math education including Bloom's taxonomy, MATH taxonomy, and PISA's mathematics literary classification to build an overall framework and a model for assessing medical students' SRiM competency in solving real-life problems. An overall assessing framework of SRiM competency includes six levels from low to high, corresponding to three clusters of Reproduction, Connections and Reflection. The model is to evaluate medical students' SRiM competency of solving practical problems, in which the assessment is considered regarding three aspects including the content of medical statistics, the process of medical statistics demonstrating SRiM competency, occupational clinical context, the medical statistical process is defined by the three SRiM competencies Description, Interpretation, Prediction. On that basis, we built the framework for assessing SRiM competency Description, Interpretation, Prediction including 6 levels corresponding to 3 cognitive competencies clusters Reproduction, Connections and Reflection. The development of assessment models and frameworks provided us a basis to design toolkits to assess SRiM competency of medical students when solving real-world problems. These research results on assessing students' SRiM competence in solving real-life problems have a significant effect on the innovation of medical statistics teaching, contributing to the innovation of assessment methods in teaching and learning medical statistics in particular, contribute to improving the quality of teaching medical statistics for medical students in general. Medical students have a good foundation in basic math skills, good foundation in statistical literacy in medicine, high level skills related to the application of statistical rules and procedures, which are the foundation for the development of SRiM, statistical thinking in medicine. Students could do well in *Reconstruct*, 87.4 percent or more. They had more difficulty but still did well in *Connect* with 79.6 percent. In specific, they appeared to understand what formula-tests were needed in most exercises and could apply them quite well. The reflection stage was where the matters appeared, only 12.6 percent could successfully make the required deductions and fewer could justify or provide satisfactory answer to how their work reflected upon the patients' condition. Which means medical statistics teaching should not just be teaching students to recognize, understand concepts, or apply procedures, processes and perform statistical calculations, but more importantly, students must be able to apply medical statistics knowledge to solve real-life problems in professional practice. Medical statistics teaching needs innovation to maximize the great potential of the students' foundation of basic math skills to the fullest, to motivate students to use basic statistical skills in solving real-life medical problems that require a high level of SriM. Medical statistics teaching must undergo a real and synchronous innovation, from its objectives, curriculum, contents, teaching materials, teaching methods and assessment methods. Innovation in the direction of integration, further connection of medical statistics with basic medicine, clinical medicine and medical research.

REFERENCES

[1] Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. *New York: Longman*, page 352.

[2] Brown, T. (2010). Truth and the renewal of knowledge: The case of mathematics education. *Educational Studies in Mathematics*, 75 (3), (pages 329–343). https://doi.org/10.1007/s10649-010-9259-0

[3] Dani Ben-Zvi, J. G. (2006). The Challenge of Developing Statistical Literacy, Reasoning and Thinking. In *Kluwer Academic Publishers* (Vol. 1999, Issue December).

[4] Darlington, E. (2013). The Use of Bloom's Taxonomy in Advanced Mathematics Questions. *British Society for Research into Learning Mathematics*, (pages 7–12). www.bsrlm.org.uk/IPs/ip33-1/BSRLM-IP-33-1-02.pdf

[5] delMas, R. C. (2002). Statistical literacy, reasoning and learning: A commentary. *Journal of Statistics Education*, 10 (3). https://doi.org/10.1080/10691898.2002.11910679

[6] Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10*(3). https://doi.org/10.1080/10691898.2002.11910676

[7] Garfield, J., & Gal, I. (1999). Teaching and assessing Statistical reasoning. *Developing Mathematical Reasoning in Grades K-12*, 207–219.

[8] Jones, G. A., Langrall, C. W., Thornton, C. A., Mooney, E. S., Wares, A., Jones, M. R., Perry, B., Putt, I. J., & Nisbet, S. (2001). Using students' statistical thinking to inform instruction. *Journal of Mathematical Behavior*, 20 (1), (pages 109–144). https://doi.org/10.1016/S0732-3123(01)00064-5

[9] Mooney, E. S. (2002). A Framework for Characterizing Middle School Students' Statistical Thinking. *Mathematical Thinking and Learning*, 4 (1), (pages 23–63). https://doi.org/10.1207/s15327833mtl0401_2

[10] Moyé, L. A. (2006). Statistical Reasoning in Medicine. *Statistical Reasoning in Medicine*. https://doi.org/10.1007/978-0-387-46212-7

[11] OECD. (2002). Definition and Seclection of Competencies (DeSeCo): Theoretical and Conceptual Foundation. *OECD, Paris, France*.

[12] OECD. (2009). The PISA 2009 Assessment Framework–Key competencies in reading, mathematics and science. *OECD, Paris, France*.

[13] PISA. (2010). *Learning Mathematics for Life*. Programme for International Student Assessment.

[14] Pountney, D., Leinbach, C., & Etchells, T. (2002). The issue of appropriate assessment in the presence of a CAS. *International Journal of Mathematical Education in Science and Technology*, 33 (1), (pages 15–36). https://doi.org/10.1080/00207390110086616

[15] Riffenburgh, R. (2012). Statistics in Medicine. In *Statistics in Medicine* (Vol. 10, Issue 4). Academic Press. https://doi.org/10.1016/C2010-0-64822-X

[16] Smith, G., Wood, L., Coupland, M., Stephenson, B., Crawford, K., & Ball, G. (1996). Constructing mathematical examinations to assess a range of knowledge and skills. *International Journal of Mathematical Education in Science and Technology*, 27 (1), (pages 65–77). https://doi.org/10.1080/0020739960270109

[17] Tan, K. (2011). Assessment for learning in Singapore: Unpacking its meanings and identifying some areas for improvement. In *Educational Research for Policy and Practice* (Vol. 10, Issue 2). https://doi.org/10.1007/s10671-010-9096-z

[18] Vui, T. (2018). Bringing Mathematics Education into the global orbit to develop thinking, logic and creativity in solving realistic problems with closed-open approach. *Vietnam Journal of Education*, 5, (pages 28–33).