The Impact of the GQM Framework on Software Engineering Exam Outcomes

Reem Abdulaziz Alnanih
Department of Computer Science-Faculty of Commuting and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Assessment is crucial in educational systems, particularly in Software Engineering (SE) programs, where fair and effective evaluations drive continuous improvement. The shift to student-centric methodologies has evolved assessment strategies to focus on aligning educational processes with students' developmental needs rather than merely measuring academic outputs. This paper adapts the Goal-Question-Metric (GQM) framework to enhance learning in software engineering education by linking educational goals, learning activities, and assessment methods. This approach specifies expected learning outcomes and integrates mechanisms for continuous improvement, aligning teaching strategies with student performance metrics. A systematic framework for course assessment using the GQM framework is presented, aligning assessment methods with Intended Learning Outcomes (ILOs) and Student Learning Outcomes (SLOs) to ensure data-driven enhancements. To validate this approach, a template was introduced to assess the impact of a tailored GQM approach on the final exam outcomes of a software engineering course at King Abdulaziz University's Department of Computer Science. A controlled experiment was conducted over two semesters with students from the CPCS 351 course. The control group, in the first semester, completed their finals without applying GQM, while the experimental group in the following semester employed a customized GQM framework. Statistical analyses, including ANOVA and Mann-Whitney U tests, were utilized to compare exam performance between the groups. Results indicated a significant improvement in the exam scores of the experimental group, thereby validating the effectiveness of the GQM framework in boosting academic performance through structured exam preparation and execution.**

Keywords—Goal-question-metric (GQM); software engineering; education; learning process; learning outcomes; continuous improvement; statistical analysis

I. INTRODUCTION

Assessment remains a cornerstone of educational systems, particularly within Software Engineering (SE) programs, where the fairness and effectiveness of evaluations are essential for fostering continuous educational enhancement. The shift towards student-centric methodologies has marked a significant evolution in assessment strategies, emphasizing the importance of aligning educational processes with the developmental needs of students rather than merely measuring academic outputs.

Institutions of higher learning often grapple with the dual obligations of advancing knowledge and fulfilling societal service demands. This balancing act requires a commitment to academic and intellectual autonomy to avoid diluting scholarly standards under external pressures.

Well-designed assessment strategies can transform the educational landscape by prioritizing learning enhancement over simple output measurement. In the context of higher education, there is a growing consensus on the importance of developing 'knowledge workers' who are adept at critical thinking, effective communication, teamwork, and self-directed learning. These competencies are crucial, surpassing the traditional focus on rote memorization, and are vital for meeting various accreditation standards that emphasize outcome-based education.

Experienced educators are increasingly aware of the processes that contribute to effective education, supporting a more nuanced articulation of teaching methodologies alongside traditional performance metrics. In the specialized field of SE, the discipline holds a status of expertise, with ongoing global discussions about the requisite knowledge and skills. This focus draws attention from diverse sectors, including media, industry, and academia.

Many challenges persist in assessing and managing educational processes within SE. Innovations such as performance-based assessments, decision-support systems, peer reviews, automated grading, and flexible assessment strategies are being implemented; however, comprehensive studies on their adoption and effectiveness remain limited.

Recent systematic reviews indicate that the Goal Question Metric (GQM) method is predominantly used for evaluating SE processes and products in professional settings, but its application in educational settings is not well-documented [1-4]. GQM, originally designed for SE, is adaptable for assessing various product development processes beyond software, including hardware production or service development.

SE inherently requires robust measurement systems to provide feedback and evaluate processes. Such systems not only assist in course planning and improvement but also enable educators to respond dynamically to educational needs through ongoing assessments.

The GQM approach has significantly influenced both industry and academia by fostering systematic data collection and analysis aimed at enhancing the quality of software processes and products [5]. This method emphasizes the importance of defining clear goals, linking these goals to specific data, and establishing a structured framework for data interpretation. This ensures that measurements are purposeful and aligned with organizational objectives [6].

Over the past fifteen years, the GQM methodology has expanded to facilitate the collection and analysis of externally oriented quality metrics aligned with business objectives, such as end-user satisfaction, market share, and customer retention. This goal-centric approach is recognized for adding significant value to the industry by enhancing the design and maintenance of both process and product quality [7].

The GQM method operates on three distinct levels: the conceptual level (Goal), which identifies the subject of research and its rationale; the operational level (Question), where specific, measurable questions are defined; and the quantitative level (Metric), which outlines the measurements needed to address these questions. This structured framework assists organizations in determining which data to collect and how to interpret it, establishing a clear, goal-oriented approach to software measurement [6].

The integration of the National Qualification Framework (NQF) in Saudi Arabia significantly enhances the structured approach to educational assessments [8]. Mandated for all Saudi universities, this framework is pivotal in defining the competencies that graduates should possess across various learning domains [9]. By focusing on outcome-based educational models, the NQF ensures that educational processes prioritize achieving specific, measurable outcomes rather than merely delivering content. This alignment supports both academic and professional expectations.

The NQF categorizes learning into five distinct domains, although the proposed assessment method in the discussed study focuses on the first three: Knowledge, Cognitive Skills, and Interpersonal Skills and Responsibility [10-11]. This strategic emphasis targets the core competencies that are most relevant and critical for SE students. These domains are essential for developing professionals who are not only proficient in technical skills but also capable of effective communication, teamwork, and ethical practice in their careers.

The exclusion of the last two domains from the primary focus of the assessment—Communication, IT, and Numerical Skills, and Psychomotor Skills—is justified within the context of SE education, as it typically does not require extensive psychomotor capabilities.

In this context, an essential aspect is content validity, which ensures that a representative sample of the intended learning outcomes (ILOs) is assessed. This is particularly relevant given the NQF's emphasis on the first three domains: Knowledge, Cognitive Skills, and Interpersonal Skills and Responsibility. Each assessment item must align with at least one ILO to ensure effective content validity. This approach considers the curriculum and the competencies expected of students, thereby evaluating the effectiveness of the assessment methods. Any systematic errors in the assessment may indicate that the course objectives or delivery methods are not adequately aligned with the intended content. This focus on alignment is crucial for ensuring that SE students develop both the technical and interpersonal skills necessary for their careers [12-13].

A. Motivations and Objectives

This study is motivated by the need to adapt and apply the GQM framework in educational settings, specifically within SE programs. The central motivation is to refine educational methodologies to align with industry demands, ultimately elevating the caliber of educational outcomes. By integrating GQM, we aim to diminish the divide between academic theories and industry practices, creating a more robust learning environment for aspiring software engineers.

To realize these aims, we propose the use of an assessment-based GQM as a pedagogical tool tailored for SE education. Our objectives are structured as follows:

1) Framework adaptation: Modify and refine the GQM framework to address the unique requirements of SE education effectively.

2) Assessment consistency: Implement a comprehensive assessment framework to ensure that evaluations of student performance are consistent and relevant.

3) Guideline provision: Offer practical guidelines for the effective deployment of this GQM-based assessment methodology in educational contexts.

4) Feasibility demonstration: Conduct initial stages of controlled experiments to validate the practicality of this approach and elucidate its foundational logic.

The innovative contribution of this research lies in the development of a novel pedagogical strategy for assessing educational processes and outcomes, firmly rooted in the GQM framework and tailored testing strategies. This study not only deepens the understanding of the GQM approach but also advocates for its broader adoption in academic circles. Specifically, it aims to enhance the application of GQM within educational frameworks, ensuring that these adaptations are appropriately customized to meet academic needs. Through this research, we seek to advance the field of SE education, aligning it more closely with industry standards and expectations.

The structure of this paper is outlined as follows: In Section II, a comprehensive literature review highlights significant research contributions related to the application of GQM in assessment contexts. Section III describes the adapted GQM model for the learning process, while Section IV presents the proposed framework for course assessment design and evaluation. Section V focuses on validating the proposed model, and Sections VI and VII present the results and discussion, respectively. Finally, Section VIII offers the conclusion of the paper.

II. Literature Review

Assessments are primarily conducted through written exams; however, this approach has several drawbacks, including an uneven distribution of questions across topics, restricted sampling, and ambiguous questions, all of which can undermine its validity. The GQM framework is recognized as a valuable tool for addressing these issues and promoting optimal educational practices. This section highlights significant recent contributions to the application of GQM in educational and assessment contexts.

In [14], Meng et al. present a method for recommending software process patterns using the GQM framework. They propose a systematic approach that aligns software process patterns with specific goals and questions, enabling organizations to select appropriate patterns based on their unique needs. The study highlights the effectiveness of the GQM-based method in enhancing decision-making in software process management, ultimately aiming to improve software development outcomes. Additionally, the paper includes a case study that demonstrates the practical application of the proposed recommendation approach.

Idahmash and Gravell in [15] explore the application of the GQM framework to assess success in agile software development projects. They identify key success factors and propose a structured methodology for measuring these factors using GQM. Through case studies, their research demonstrates how GQM can provide insights into project performance, helping teams align their goals with measurable outcomes. The findings emphasize the importance of tailored metrics in enhancing the effectiveness of agile practices.

In 2018, Tahir et al. [16] investigated the current state of software measurement practices within the Pakistani software industry. Through a comprehensive survey, they assessed the effectiveness and challenges of existing measurement processes. The study identifies gaps in the application of these practices and provides recommendations for improving software quality and project management. The findings highlight the need for better integration of measurement frameworks and tools to enhance overall software development practices in the region.

Shojaeshafiei in [17] introduces a novel approach for assessing web application vulnerabilities using the Analytic Hierarchy Process (AHP) integrated with fuzzy measurement techniques. This methodology aims to quantify and prioritize vulnerabilities in a structured manner, addressing the inherent uncertainties in risk assessment. By employing the GQM framework, the author establishes clear evaluation goals and metrics, facilitating effective decision-making in vulnerability remediation. The study underscores the importance of a systematic approach to enhance security measures in web applications, ultimately contributing to more robust risk management strategies in SE.

Calvo and Beltrán in [18] examine the use of the GQM framework to create tailored cyber risk metrics. They argue that conventional risk assessment methods are often inflexible and insufficiently adaptable to changing cyber environments. By utilizing the GQM approach, the study clearly defines cybersecurity goals, formulates relevant questions, and establishes measurable metrics aligned with organizational needs. This framework enhances real-time monitoring and evaluation of cyber risks, facilitating better decision-making and proactive risk management against evolving threats.

Philippou, Frey, and Rashid [19] present a methodology that utilizes the GQM framework to align security metrics with business objectives. They emphasize the importance of contextualizing security measures to ensure that they support organizational goals. The proposed methodology facilitates the identification of relevant security metrics by defining specific goals and questions that reflect business needs. The study highlights the benefits of this alignment in improving the effectiveness of security assessments and enhancing the overall organizational security strategy.

Falco and Robiolo [20] present the development of a comprehensive catalog that effectively maps ISO/IEC 25010 quality characteristics to specific measures used in industrial settings. Their aim is to bridge the gap between theoretical quality standards and practical implementation by identifying relevant metrics for each quality attribute. The study emphasizes the critical importance of these measures in assessing software quality and offers valuable insights into their application in real-world scenarios, ultimately contributing to enhanced quality assurance practices in the industry.

Hsueh, Wang, and Bilegjargal [21] discuss the development of a learning analysis system using the GQM methodology combined with the ELK Stack (Elasticsearch, Logstash, and Kibana). They outline how GQM helps define learning objectives and questions which guide data collection and analysis. The ELK Stack is utilized for real-time data processing and visualization, enabling educators to gain insights into student learning behaviors. The study demonstrates the effectiveness of this integrated approach in enhancing educational outcomes through data-driven decision-making.

Finally, Minhas et al. [22] examine the differences between research and practical approaches to regression testing through the lens of the GQM framework. They analyze existing literature and industry practices to identify gaps and alignments in the goals, questions, and metrics used in regression testing. The study finds that while research provides theoretical insights, practical applications often lack the structured measurement approaches advocated in academic literature. The findings suggest that adopting GQM can enhance collaboration between researchers and practitioners, leading to improved regression testing methodologies.

Table I summarizes the significant contributions of the studies discussed above regarding the application of the GQM methodology across various domains, including education and SE.

From the above table, it is evident that each study highlights specific challenges in measurement and assessment processes, demonstrating how a GQM-based approach can effectively address these issues. In academic settings, especially, adopting a test-based GQM framework is essential for enhancing the validity and reliability of assessments. By aligning educational objectives with measurable outcomes, GQM can improve teaching practices and student performance, fostering a culture of continuous quality improvement.

Moreover, this paper indicates the importance of integrating GQM into the design and development of tests. Specifically, there is a need to apply GQM within the Department of Computer Science at FCIT, KAU. None of the studies reviewed have tailored GQM to computer science courses or defined the weight, domain, or time requirements for different types of questions in these courses.

TABLE I. SUMMARY OF THE LITERATURE REVIEW

Ref.	Domain	Focus Area	GQM Methodology Application	Challenges Addressed
[14]	Software Process	Process Patterns	Recommending patterns with GQM.	Selecting patterns for unique needs.
[15]	Agile Development	Project Success	Assessing success with GQM.	Aligning goals with measurable outcomes.
[16]	Software Measurement	Industry Practices	Survey of measurement in Pakistan.	Gaps in current measurement processes.
[17]	Software Engineering	Risk Management	Application of GQM for defining goals related to security measures and evaluation metrics.	Addressing the ambiguity in risk assessment through fuzzy measurement techniques.
[18]	Cybersecurity	Risk Assessment	Application of the GQM framework to define specific cybersecurity goals and develop tailored metrics.	Lack of flexibility and adaptability in traditional risk assessment methods for dynamic cyber environments.
[19]	Security Metrics	Business Alignment	Aligning metrics with business goals	Supporting orgaizational objectives
[20]	Software Quality	Quality Characteristics	Mapping ISO standard to measures	Bridging theory with practical use
[21]	Education	Learning Analysis	System developing a with GQM	Guiding data collection
[22]	Regression Testing	Research vs. Practice	Comparing approaches with GQM	Lack of structured measurement in practical applications

III. THE ADAPTED GQM MODEL FOR THE LEARNING PROCESS

In the educational context, aligning instructional strategies with precise learning objectives is crucial for promoting effective student outcomes. To facilitate this alignment, structured methodologies like the GQM approach are invaluable. This model acts as a systematic framework that bridges high-level educational goals with specific research questions and measurable metrics, enabling educators to make informed, data-driven decisions during both the design and implementation phases.

The primary aim of adopting the GQM model in this study is to ensure that the content of each assessment method is directly linked to clearly defined metrics for each course, guiding the selection of appropriate and impactful learning experiences. This approach not only tailors teaching strategies to emphasize desired learning outcomes but also supports a flexible curriculum development process that effectively meets specific needs for:

- Clarity: Clearly defines expected competencies.
- Support: Outlines learning activities that aid learners in achieving their goals.
- Evaluation: Measures learners' performance against established assessment criteria.

The GQM model consists of three main components as shown in Fig. 1:

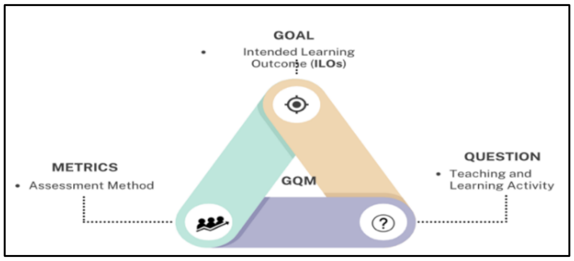

GOAL
• Intended Learning Outcome (ILOs)
METRICS
• Assessment Method
GQM
QUESTION
• Teaching and Learning Activity

Fig. 1. GQM model.

- Goals: Define what learners should know, understand, or be able to do by the end of their courses.
- Questions: Pose specific evaluation questions that assess whether the set goals are being met.
- Metrics: Establish measurable criteria quantifying the extent to which learners achieve the desired outcomes.

Fig. 2 presents a structure clarifying the distinct roles of learning activities and assessment methods in the educational process:

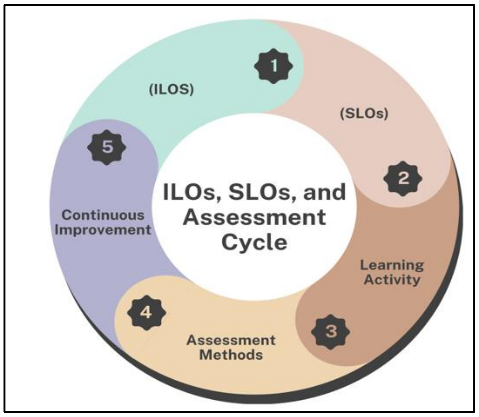


Fig. 2. ILOs, SLOs, and assessment cycle.

- Intended Learning Outcomes (ILOs): Goals set by educators regarding what students should achieve.
- Student Learning Outcomes (SLOs): Specific outcomes expected from students, derived from ILOs.
- Learning Activities: Engagement strategies designed to help students achieve SLOs (e.g., lectures, group projects, discussions).

- Assessment Methods: Tools used to measure whether students have achieved SLOs (e.g., quizzes, exams, portfolios).
- Continuous Improvement: Feedback from assessments informs revisions to ILOs, SLOs.

A. Mapping GQM Components to Educational Process

Table II illustrates the mapping of the GQM components to educational steps, alongside descriptions:

Table III is an example illustrating how a single goal can be effectively broken down into both overarching and specific learning objectives, ensuring coherence in the educational process.

Following this structure, the GQM framework is adapted to the learning process to achieve both validity and reliability in assessing student learning outcomes in SE. This adaptation involves focusing on broad aims, specifying the aspects of learning to be assessed, listing questions that gauge student mastery of specific skills, and providing measurable data points to evaluate student performance and understanding.

TABLE II. MAPPING GQM COMPNENTS TO EDUCATIONAL STEPS

GQM Component	Educational Step	Description
Goals	1) ILOs	High-level educational objectives set by educators.
	2) SLOs	Specific outcomes derived from ILOs that students are expected to achieve.
Questions	3) Learning Activities	Evaluation questions that guide the design of learning activities.
Metrics	4) Assessment Methods	Measurable criteria to evaluate achievement of SLOs.
	5) Continuous Improvement	Data from assessments used to inform revisions to ILOs, SLOs, and teaching methods.

TABLE III. EXAMPLE OF MAPPING GQM TO EDUCATIONAL STEPS

GQM Component	Educational Step	Description
Goal: Students will understand the foundational principles of SE.	**1)ILO**: What is SE?	This ILO reflects the goal by focusing on a broad understanding of the discipline. It sets the expectation for students to grasp the fundamental concepts of SE, aligning with the overall educational aim of instilling foundational knowledge.
	2)SLO: Explain the concept of a software lifecycle with an example, including the deliverables produced from each phase.	This SLO operationalizes the goal by specifying a measurable outcome that students must demonstrate. It details how students will apply their understanding of SE principles by explaining a specific concept (the software lifecycle) and providing an example, thus ensuring that the goal translates into tangible student performance.
Question: How effectively are students grasping the foundational principles of SE?	**3) Learning Activity**: Collaborative Group Project. Students work in teams to analyze a SE case study, identify the foundational principles at play, and present their findings.	By asking the question, educators can focus on specific, measurable outcomes that reflect student learning. • The **Collaborative Group Project** directly supports the ILO (What is SE?) by requiring students to explore and articulate foundational concepts in a practical context. • It also aligns with the SLO (Explain the concept of a software...) by prompting students to apply their knowledge to a real-world scenario, demonstrating their understanding of key principles, including the software lifecycle.
Metrics: What percentage of students can accurately explain foundational principles of SE and apply them in practical scenarios?	4) **Assessment Methods:** Rubric-Based Assessment of Group Project Presentations. Bu using a detailed rubric to evaluate students' presentations on their case studies, focusing on their ability to explain foundational principles and apply them to real-world scenarios.	The **rubric-based assessment** aligns with the metrics by providing a structured way to evaluate how well students articulate and apply foundational principles. The rubric can include criteria such as clarity of explanation, relevance to SE concepts, and ability to connect theory to practice.
	5) **Continuous Improvement**: Peer and Instructor Feedback. After presentations, both peers and instructors provide feedback based on the rubric criteria, highlighting strengths and areas for improvement.	**Continuous feedback** from both instructors and peers enhances learning by providing students with insights into their performance. This feedback can inform students about specific areas to improve, fostering a growth mindset and encouraging them to engage more deeply with the material.

IV. PROPOSED FRAMEWORK FOR COURSE ASSESSMENT DESIGN AND EVALUATION BASED ON GQM

This section introduces the application of the GQM approach to streamline the assessment process in a SE course. It integrates Intended Learning Outcomes (ILOs), Student Learning Outcomes (SLOs), learning activities, and assessment methods, as depicted in Fig. 3. This integration enhances data-driven decision-making aimed at improving student performance in Computer Science.

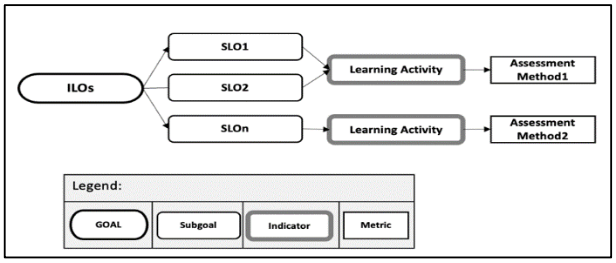


Fig. 3. Hierarchical of GQM in learning process.

A. *Course Overview*

The CPCS 351 SE course prepares students to understand fundamental concepts in SE, particularly in system analysis, and equips them with the skills to design medium-scale software systems and apply engineering principles in practical scenarios. The case study aims to align each assessment method with the course's articulation metrics, facilitating the selection of suitable learning experiences.

The ILOs of the proposed design span three learning domains: knowledge, cognitive abilities, and interpersonal skills. The assessment methods include a final exam comprising multiple-choice questions (MCQs) and essay questions. MCQs effectively evaluate knowledge areas, while essay questions provide a broader assessment of skills but are more time-consuming.

B. *GQM Template Design*

The proposed GQM template is structured into basic and advanced levels. The basic level addresses overarching goals, associated questions, and the assessment methods employed. The advanced level provides a detailed matrix that aids in evaluating how well the assessment methods fulfill the SLOs.

C. *Basic Level Template*

The basic level template features a two-dimensional matrix that aligning the curriculum content with the total number of evaluation items, as shown in Fig. 4. Rows represent the GQM framework stages of the learning process, and columns cover various attributes, including course information, assessment methods, question types, topic organization, CLOs, ILOs, and learning domains.

Steps for designing course assessments include:

1) General Information:
- Course Name and Number.
- Assessment Methods: Options like Exam 1, Exam 2, Final Exam, Projects.
- Activity Method: MCQs or Essays.
- Total Time Allocated for MCQ and Essay Exams.
- Total Marks of the assessment method.

2) Type of Questions:
- Define the question types used, such as MCQs, essays, etc.

3) Topics Organization:
- Sequentially number topics over a 14-week term.
- Divide each topic into ILOs addressing specific topic aspects.

4) Course Learning Outcomes (CLOs):
- Map each CLO to a set of ILOs relevant to topics, guiding topic coverage depth.

5) Learning Domains:
- Incorporate domains relevant to the Computer Science department, such as Knowledge (K), Cognitive Skills (C), and Interpersonal Skills (IP).

6) Exam Item Importance Scale (IMS):
- Score each ILO tested by MCQ or essay, categorizing importance from 'Most Important' to 'Not Included in the Exam'.

7) Total Item Importance Scale (TIMS):
- Calculate total importance for MCQs and essays by summing ILO values across topics.

8) Duration Time Needed (DT):
- Estimate the required duration to complete each type of question, based on expert input.

Process of weight assignment, question type determination, and quality assurance:

1) Weight Assignment:
- Assign weights to each ILO reflecting their importance, ensuring total ILO weight sums to 100%.

2) Question Type Determination:
- Categorize assessment questions by type and align them with learning domains.

3) Quality Assurance:
- Review and validate the alignment between assessment methods, ILOs, and SLOs. Conduct pilot testing for refinement.

D. *Advanced Level Template*

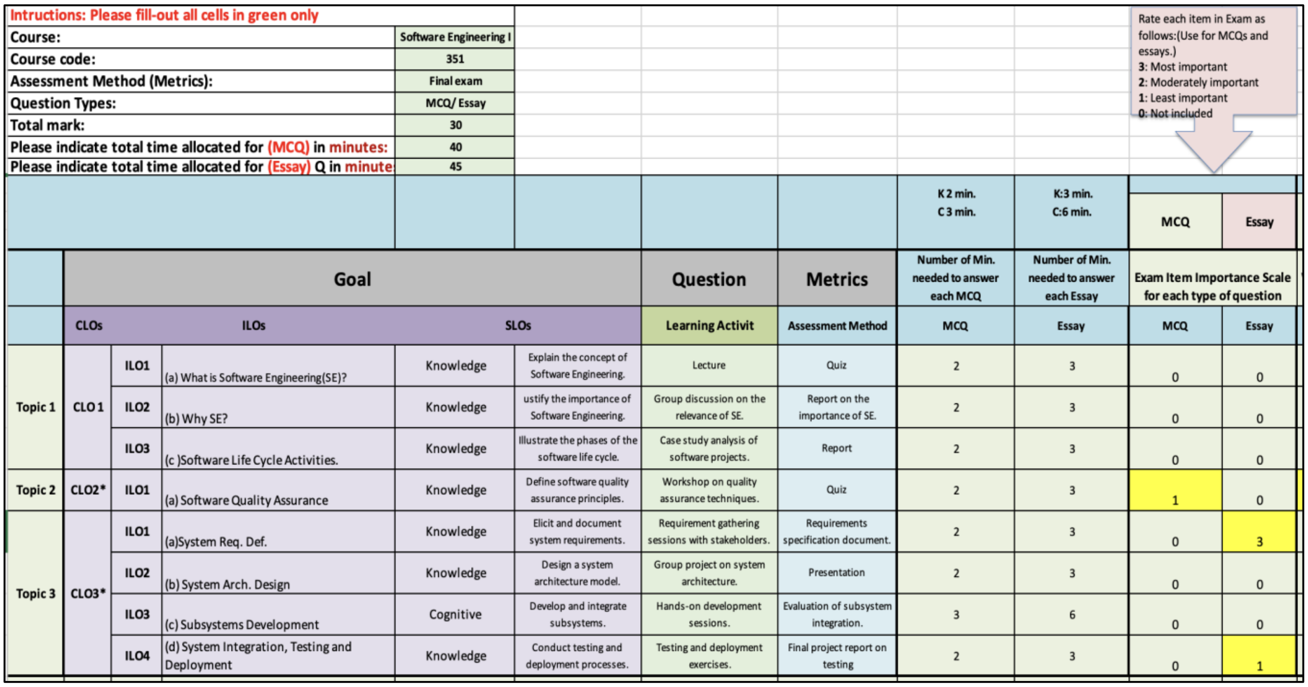
The advanced-level template, shown in Fig. 5, is generated based on:

1) Calculate the Weight Level (WL):
- Determine the importance score of each ILO, ensuring the total weight level for each exam type equals one.

2) Calculate the Duration of Each Item (DD):
- Compute as WL multiplied by the allocated exam time.

3) Determine the Number of Items in Each Exam Set (NI):
- Calculate by dividing DD by the time required for each domain level.

4) Adjust the Number of Items in Each Exam Set:
- Round decimal values to the nearest integer and map ILOs accordingly.



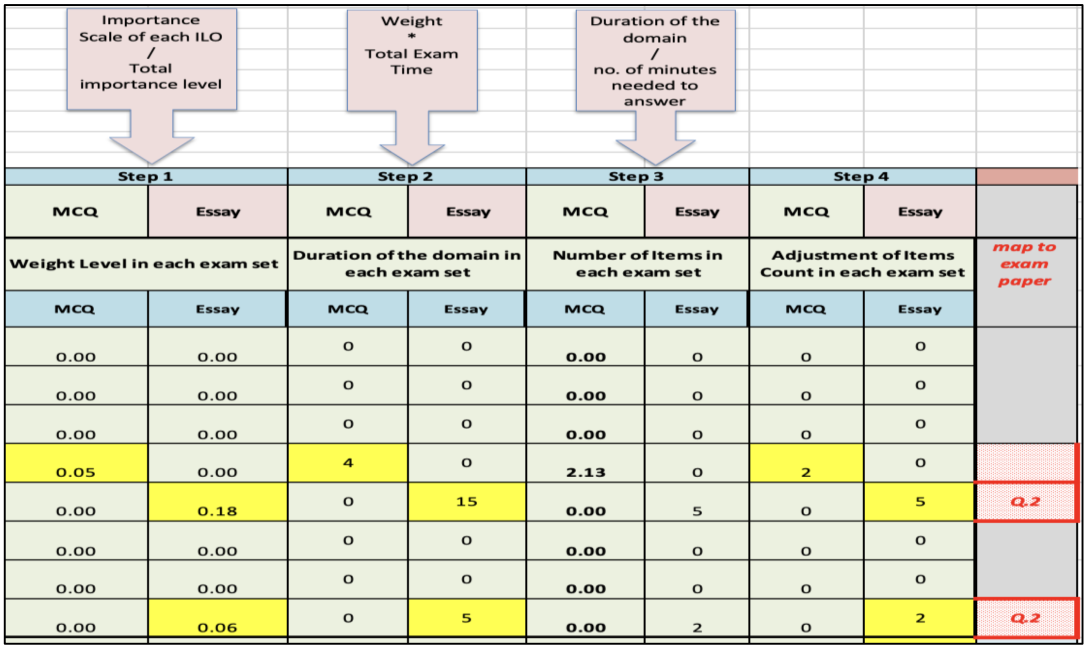
Intructions: Please fill-out all cells in green only

Course:	Software Engineering I
Course code:	351
Assessment Method (Metrics):	Final exam
Question Types:	MCQ/ Essay
Total mark:	30
Please indicate total time allocated for (MCQ) in minutes:	40
Please indicate total time allocated for (Essay) Q in minute:	45

Rate each item in Exam as follows:(Use for MCQs and essays.)
3: Most important
2: Moderately important
1: Least important
0: Not included

								K 2 min. C 3 min.	K:3 min. C:6 min.	MCQ	Essay
	Goal					Question	Metrics	Number of Min. needed to answer each MCQ	Number of Min. needed to answer each Essay	Exam Item Importance Scale for each type of question	
	CLOs	ILOs			SLOs	Learning Activit	Assessment Method	MCQ	Essay	MCQ	Essay
Topic 1	CLO 1	ILO1	(a) What is Software Engineering(SE)?	Knowledge	Explain the concept of Software Engineering.	Lecture	Quiz	2	3	0	0
		ILO2	(b) Why SE?	Knowledge	ustify the importance of Software Engineering.	Group discussion on the relevance of SE.	Report on the importance of SE.	2	3	0	0
		ILO3	(c)Software Life Cycle Activities.	Knowledge	Illustrate the phases of the software life cycle.	Case study analysis of software projects.	Report	2	3	0	0
Topic 2	CLO2*	ILO1	(a) Software Quality Assurance	Knowledge	Define software quality assurance principles.	Workshop on quality assurance techniques.	Quiz	2	3	1	0
Topic 3	CLO3*	ILO1	(a)System Req. Def.	Knowledge	Elicit and document system requirements.	Requirement gathering sessions with stakeholders.	Requirements specification document.	2	3	0	3
		ILO2	(b) System Arch. Design	Knowledge	Design a system architecture model.	Group project on system architecture.	Presentation	2	3	0	0
		ILO3	(c) Subsystems Development	Cognitive	Develop and integrate subsystems.	Hands-on development sessions.	Evaluation of subsystem integration.	3	6	0	0
		ILO4	(d) System Integration, Testing and Deployment	Knowledge	Conduct testing and deployment processes.	Testing and deployment exercises.	Final project report on testing	2	3	0	1

Fig. 4. Sample of the proposed basic level template.



Importance Scale of each ILO / Total importance level

Weight * Total Exam Time

Duration of the domain / no. of minutes needed to answer

Step 1		Step 2		Step 3		Step 4		
MCQ	Essay	MCQ	Essay	MCQ	Essay	MCQ	Essay	
Weight Level in each exam set		Duration of the domain in each exam set		Number of Items in each exam set		Adjustment of Items Count in each exam set		map to exam paper
MCQ	Essay	MCQ	Essay	MCQ	Essay	MCQ	Essay	
0.00	0.00	0	0	0.00	0	0	0	
0.00	0.00	0	0	0.00	0	0	0	
0.00	0.00	0	0	0.00	0	0	0	
0.05	0.00	4	0	2.13	0	2	0	
0.00	0.18	0	15	0.00	5	0	5	Q.2
0.00	0.00	0	0	0.00	0	0	0	
0.00	0.00	0	0	0.00	0	0	0	
0.00	0.06	0	5	0.00	2	0	2	Q.2

Fig. 5. Sample of the proposed advanced level template.

E. Mapping ILOs to Assessment Methods

This structured template aligns ILOs, SLOs, learning activities, and assessment methods, facilitating a data-driven approach to evaluate student performance and identify improvement areas. By adhering to these steps, the assessment process remains clear, organized, and aligned with educational goals, ensuring effective evaluation and continual enhancement of the learning experience.

V. Validation of the GQM Approach

After finalizing the GQM template, it was essential to conduct a pilot test to assess its effectiveness and implement any necessary adjustments. The pilot test involved an expert user—a professor from King Abdulaziz University (KAU) with over ten years of experience in teaching and designing courses. This evaluation took place in the professor's office at KAU during standard office hours, following a pre-arranged appointment.

The feedback received was highly positive, affirming that the creation of a new template tailored to course-specific ILOs and faculty requirements represents a significant advancement. The professor endorsed the application of the template (referenced in Fig. 4) and provided insights on optimizing the allocation of time per domain based on their expertise. A formal evaluation was deemed necessary to further validate the template’s impact on enhancing exam design processes.

A. Experimental Assessment

This assessment focused on a comparative study involving female student groups from two consecutive semesters enrolled in the CPCS 351 course at the Department of Computer Science, FCIT, KAU.

In the first semester, a cohort of 59 female students completed their final exam using traditional design questions that were not intended to be aligned with the IILOs. The exam was constructed without the support of the newly proposed template, which could have improved this alignment.

In contrast, the second semester featured a larger cohort of 98 female students whose exam was constructed using the new template. This template was specifically designed to ensure that all questions were meticulously aligned with the defined ILOs.

Both exams were crafted by the same course coordinator, an experienced educator who had taught the course for several years. The study's independent variables were the number of question items aligned with the selected ILOs, as depicted in Fig. 4. The dependent variable was the students' performance, providing a direct measure of the template’s educational impact. The exam questions for both groups were designed to maintain consistency in content and format, ensuring that each set matched in difficulty. This careful alignment is essential for validating the effectiveness of the GQM framework, as it enables a reliable assessment of whether any improvements in exam outcomes stem from the framework's implementation rather than variations in exam difficulty.

B. Testing Environment

The exams were conducted in a controlled environment, specifically a closed room on the second floor of the FCIT building at KAU. Each student from both semesters received a standardized final exam, tailored to their respective study conditions. This setting ensured that the testing conditions were consistent across both groups, facilitating a reliable comparison of the template’s effectiveness.

C. Hypothesis

To test the effectiveness of the GQM template in enhancing student understanding and performance by providing a structured and aligned assessment method, the following hypothesis is defined as follow:

- Null Hypothesis (H0): There is no significant difference in overall performance between students who took the final exam based on the proposed GQM template (Group 2) and students who took the final exam without the proposed template (Group 1).
- Alternative Hypothesis (H1): Students who took the final exam based on the proposed GQM template (Group 2) will demonstrate significantly higher overall performance compared to students who took the final exam without the proposed template (Group 1).

VI. Results

A. Data Description

The study involved two distinct groups. Group 1 comprised 59 participants, with data points ranging from 17.5 to 28. The mean and median values were calculated to provide a comprehensive analysis. Group 2 included 98 participants, exhibiting a broader data range from 9.25 to 28.75.

To assess the normality of the datasets, the Shapiro-Wilk test was conducted for both groups. For Group 1, the test yielded a p-value of 0.2348, indicating that the data follows a normal distribution, as this p-value exceeds the conventional alpha level of 0.05. Conversely, Group 2's p-value was approximately 0.00000104, suggesting a significant deviation from normality. This notable difference in distribution characteristics between the two groups implies that non-parametric methods should be utilized for comparative analyses, rather than parametric tests that assume normality.

B. Mann-Whitney U Test Execution

To evaluate the differences between the exam performances of the two cohorts, we conducted the Mann-Whitney U test using Python. This non-parametric test is appropriate for comparing two independent groups, especially when the data do not follow a normal distribution. The analysis involved the following steps:

1) Data preparation: We collected the exam scores from both cohorts.

2) Statistical analysis: We utilized the mannwhitneyu function from the scipy.stats module to calculate the U-statistic and the corresponding p-value. The Mann-Whitney U test yielded a U-statistic of 1902.5 and a p-value of 0.0183.

The U-statistic of 1902.5 indicates the rank-based test statistic calculated from the scores of both groups; a higher U-statistic value generally reflects a greater difference in scores between the two groups. Additionally, the p-value of 0.0183 is less than the conventional alpha level of 0.05, providing statistical evidence to reject the null hypothesis (H0) and support the alternative hypothesis (H1). This suggests that students who took the final exam based on the proposed GQM template (Group 2) demonstrated significantly higher overall performance compared to those who took the exam without the proposed template (Group 1).

C. Analysis of Variance (ANOVA)

Table IV presents a comparative analysis of two groups using ANOVA test. Group 1, consisting of 59 observations, has an average value of 23.18 with a variance of 9.85. In contrast, Group 2, with 98 observations, exhibits a higher average of 24.95 and a lower variance of 8.90. The F-statistics of 5.73 and the associated p-value of 0.018 indicate a statistically significant difference between the groups. These results indicate that the observed differences in average values are unlikely to be due to random variation.

TABLE IV. ANOVA TEST RESULT

Description	Count	Average	Variance	F-statistics	P-value
Group 1	59	23.18	9.85	5.73	0.018
Group 2	98	24.95	8.90		

D. 95% Confidence Intervals

Following the analysis of variance (ANOVA), the calculation of 95% confidence intervals for the means of the two groups provides valuable additional insights into the data. The confidence intervals for Group 1 (22.39, 23.68) and Group 2 (23.07, 24.33) allow us to quantify the uncertainty around the estimated means. Specifically, for Group 1, we can be 95% confident that the true population mean lies within the interval (22.39, 23.68), while for Group 2, the true mean is expected to fall between (23.07, 24.33).

These intervals not only highlight the range of plausible values for the population means but also facilitate a clearer understanding of the potential differences between the two groups. By providing a visual representation of the means and their associated uncertainty, the confidence intervals enhance our interpretation of the results, confirming the statistical significance observed in the ANOVA test and allowing for a more comprehensive discussion of the implications of these findings.

E. Cohen's d Value

The Cohen's d serves as a measure of the effect size, providing insight into the magnitude of the difference between the two groups. In this analysis, the calculated Cohen's d value of approximately -0.24 indicates a small effect size, suggesting that while there is a statistically significant difference between the groups, the practical significance of this difference is modest. The negative sign further confirms that, on average, Group 2 scored higher than Group 1.

This assessment aligns with the statistical significance indicated by the Mann-Whitney U test, which yielded a p-value of 0.018, and the ANOVA results, which produced an F-statistic of 5.73. Together, these findings lead us to reject the null hypothesis in favor of the alternative hypothesis, affirming that the differences between the groups are meaningful, albeit modest in magnitude. By incorporating Cohen's d, we not only confirm the presence of a significant difference but also contextualize the practical implications of that difference in the educational setting.

VII. DISCUSSION

In this study, we conducted a comprehensive statistical analysis to compare two distinct groups. Initial assessments of normality confirmed that the data were well-suited for parametric tests. Subsequent t-tests and ANOVA revealed significant variations in mean differences between the groups, findings that were further supported by the Mann-Whitney U test, a non-parametric alternative. These results align with previous research, such as that by Idahmash and Gravell [15], which demonstrated how structured methodologies like the Goal-Question-Metric (GQM) can enhance performance metrics in agile projects.

The 95% confidence intervals provided additional insight into the range within which the true means of the groups are likely to fall, enhancing our understanding of the data spread and variability. Moreover, the calculation of Cohen's d, resulting in a value of -0.24, indicated a small but notable effect size, where Group 2 consistently exhibited higher values than Group 1. This effect size, while statistically significant, suggests a modest practical significance, consistent with findings by Tahir et al. [16], who noted similar effect sizes in their assessments of software measurement practices.

In assessing the validity of our study's findings, several potential threats must be considered. First, internal validity could be compromised by selection biases or non-random assignment of participants to groups, which might influence outcomes if the groups are not equivalent at baseline. This concern echoes the challenges identified by Calvo and Beltrán [18], who highlighted the importance of adaptability in risk assessment frameworks. Additionally, while our measures are standardized, they might not fully capture the constructs in question, leading to potential measurement errors, a limitation also noted in the work of Philippou et al. [19].

Externally, the generalizability of our results may be limited by the specific sample used, which may not accurately represent the broader population. This limitation is particularly relevant in light of the findings from Falco and Robiolo [20], who emphasized the importance of contextualizing metrics for practical implementation. Furthermore, the reliance on statistical assumptions, such as those inherent in t-tests and ANOVA, may not hold true across all datasets, potentially affecting the robustness of our conclusions.

Lastly, the effect size, though statistically significant, was small, suggesting that while differences between groups are present, their practical significance may be modest. This aligns with the observations made by Hsueh et al. [21], who discussed the need for careful interpretation of metrics in educational settings. Together, these analyses provide a comprehensive understanding of our findings in relation to existing literature, highlighting both the contributions and limitations of our study.

VIII. Conclusion

This paper has critically examined the impact of the GQM framework on SE exam outcomes. The analysis reveals that implementing the GQM framework significantly enhances the clarity and focus of the learning objectives, which in turn positively affects student performance in exams.

The lessons learned from this study suggest that the structured approach of GQM not only aids educators in designing more effective assessments but also helps students in aligning their study strategies to meet specific learning goals. The findings underscore the potential of the GQM framework as a powerful tool in educational settings, particularly in disciplines that require high levels of analytical and problem-solving skills like SE.

Furthermore, the statistical evidence supports the hypothesis that systematic goal setting within educational frameworks can lead to improved educational outcomes. This insight is crucial for educators seeking methods to enhance instructional quality and for educational institutions aiming to boost academic performance.

In light of these findings, future research should d address the following:

- Investigating the GQM framework's applicability across various disciplines beyond SE could provide insights into its versatility and effectiveness in different educational contexts.
- Longitudinal studies are needed to assess the sustained impact of the GQM framework on student learning outcomes and curriculum development over time.
- Examining the potential for integrating the GQM framework with educational technologies, such as learning management systems, may improve data collection and analysis for continuous improvement.

References

[1] F. N. Colakoglu, A. Yazici, and A. Mishra, "Software product quality metrics: A systematic mapping study," IEEE Access, vol. 9, pp. 44647-44670, 2021.

[2] T. Galli, F. Chiclana, and F. Siewe, "Software product quality models, developments, trends, and evaluation," SN Computer Science, vol. 1, pp. 1-24, 2020.

[3] A. Yasin, R. Fatima, L. Wen, W. Afzal, M. Azhar, and R. Torkar, "On using grey literature and Google Scholar in systematic literature reviews in software engineering," IEEE Access, vol. 8, pp. 36226-36243, 2020.

[4] T. Tahir, T. Ghulam Rasool, and C. Gencel, "A systematic literature review on software measurement programs," Information and Software Technology, vol. 73, pp. 101-121, 2016.

[5] V. R. Caldiera, V. R. Basili, and H. D. Rombach, "The Goal Question Metric Approach," in Encyclopedia of Software Engineering, 1994, pp. 528-532.

[6] A. Janes and G. Succi, Lean Software Development in Action, Springer Berlin Heidelberg, 2014. doi: 10.1007/978-3-642-00503-9. [Online]. Available: https://doi.org/10.1007/978-3-642-00503-9.

[7] R. Van Solingen, V. Basili, G. Caldiera, and H. D. Rombach, "Goal question metric (GQM) approach," Encyclopedia of Software Engineering, 2002.

[8] A. I. Gonzalez-Tablas Ferreres, K. Wouters, B. Ramos Alvarez, and A. Ribagorda Garnacho, “EVAWEB: A web-based assessment system to learn X.509/PKIX-based digital signatures,” IEEE Trans. Educ., vol. 50, no. 2, pp. 112–117, May 2007.

[9] E. Guzman and R. Conejo, “Self-assessment in a feasible, adaptive web-based testing system,” IEEE Trans. Educ., vol. 48, no. 2, pp. 688–695, Nov. 2005.

[10] National Commission for Academic Accreditation and Assessment, “Handbook for Quality Assurance and Accreditation in Saudi Arabia: Part 2,” Version 3, Oct. 2015. [Online]. Available: http://www.kfupm.edu.sa/deanships/dad/Documents/AAC/NCAAA%20Documents/H2.%20Handbook%20Part%202.pdf

[11] National Qualifications Framework for Higher Education in the Kingdom of Saudi Arabia, 2009. [Online]. Available: https://www.mu.edu.sa/sites/default/files/National%20Qualifications%20Framework%20for%20HE%20in%20KSA.pdf

[12] K. McLaughlin, S. Coderre, W. Woloschuk, and H. Mandin, “Does blueprint publication affect students’ perception of validity of the evaluation process?” Adv. Health Sci. Educ., vol. 10, pp. 15–22, 2005.

[13] P. Bridge, J. Musial, R. Frank, T. Roe, and S. Sawilowsky, “Measurement practices: Methods for developing content-valid student examinations,” Med. Teach., vol. 25, no. 4, pp. 414–421, 2003.

[14] Z. Meng, C. Zhang, B. Shen, and Y. Wei, “A GQM-based approach for software process patterns recommendation,” in Proc. SEKE, 2017, pp. 370–375.

[15] A. M. Aldahmash and A. Gravell, “Measuring success in agile software development projects: a GQM approach,” 2018.

[16] T. Tahir, G. Rasool, W. Mehmood, and C. Gencel, “An evaluation of software measurement processes in Pakistani software industry,” IEEE Access, vol. 6, pp. 57868–57896, 2018.

[17] M. Shojaeshafiei, "Analytic hierarchy process-based fuzzy measurement to quantify vulnerabilities of web applications," International Journal of Computer Networks & Communications (IJCNC), vol. 12, 2020.

[18] M. Calvo and M. Beltrán, "Applying the Goal, Question, Metric method to derive tailored dynamic cyber risk metrics," Information & Computer Security, vol. 32, no. 2, pp. 133-158, 2024.

[19] E. Philippou, S. Frey, and A. Rashid, “Contextualising and aligning security metrics and business objectives: A GQM-based methodology,” Computers & Security, vol. 88, p. 101634, 2020.

[20] M. Falco and G. Robiolo, “Product Quality Evaluation Method (PQEM): To understand the evolution of quality through the iterations of a software product,” Int. J. Software Eng. Appl. (IJSEA), vol. 12, 2021.

[21] N.-L. Hsueh, J.-J. Wang, and D. Bilegjargal, “Building learning analysis system with GQM methodology and ELK stack,” J. Internet Technol., vol. 24, no. 2, pp. 379–387, 2023.

[22] N. M. Minhas, T. R. Koppula, K. Petersen, and J. Börstler, “Using goal–question–metric to compare research and practice perspectives on regression testing,” J. Softw. Evol. Process, vol. 35, no. 2, p. e2506, 2023.

Author’s Profile

Reem Alnanih is an associate professor of Computer Science in Faculty of Computing and Information Technology (FCIT), at KAU, Jeddah, Saudi Arabia. She received her B.Sc. and M.Sc. degrees from Computer Science Department. She got her Ph.D. in Computer Science from Concordia University, Montreal, Canada. She currently serves as the Vice Dean of Scientific Research for Female at KAU. Her research interests include software engineering, HCI, quality measurement and related evaluation techniques.