An empirical evaluation of a Reflective Writing Framework (RWF) for Reflective Writing in Computer Science Education

Huda Alrashidi
Computer Science Department
University of Warwick
Coventry, UK
h.alrashidi@warwick.ac.uk
h.alrashidi01@gmail.com

Thomas Daniel Ullman
Institution Institute of Educational Technology
The Open University
Milton Keynes, UK
t.ullmann@open.ac.uk

Mike Joy
Computer Science Department
University of Warwick
Coventry, UK
h.alrashidi@warwick.ac.uk
m.s.joy@warwick.ac.uk

***Abstract*- Reflective writing is vital in learning professional skills and particularly in Computer Science (CS). However, there appears to be a lack of literature relating to a reliable of the current reflective writing frameworks used for CS education. This paper describes a novel Reflective Writing Framework (RWF) that has been applied to the manual content analysis of CS students' reflective writings. The paper aims to empirically examine this RWF in terms of reflection indicators and levels that can be manually assessed by raters. The results of this manual content analysis showed empirically that the coding scheme of the RWF is valid and reliable. The results were that the inter-rater reliability (IRR) of the overall coding scheme of the RWF increased from 0.5 to 0.8 in pilot tests and ranged between 0.40 and 0.75 in terms reflection indicators and levels; substantially, this means ranged from moderate to substantial agreement. This research contributes to CS education an RWF that can be reliably annotated and was validated by CS experts.**

Keywords*—Computer Science Education, Reflective writing, Framework, Coding Scheme, Annotation.*

I. INTRODUCTION

Higher education disciplines have been including reflective writing in their educational programs: for instance, medicine [1], teachers' pre-service training [2], management [3], and computer science (CS) [4, 5]. Reflective writing is an important part of the learning process as experienced in professional education due to the fact that it enables students to express and to record thoughts concerning their specific, subject-based experience; this will assist them in improving their skills and understandings.

However, despite the desirability of the use of reflective writing activities in CS education, the using of reflection is not universally accept by either students or instructors [4].

Therefore, it is not surprising that CS students lack knowledge of any scheme which would help them to write reflectively. A recent study by Demmans Epp, Akcayir and Phirangee [5] found that the quality of student reflective writing did not change between their first and last reflection assignments – as based on peer review feedback. This suggests the need for an investigation into providing guidelines relating to the various stages of students' reflective work which must be followed in CS education [4-6].

It is essential to have a valid and reliable coding scheme which represents the dimensions and depth of reflection of products of reflective writing "to have some means of identifying reflective thought and a measure of the depth of reflective thinking" [7].

Recent studies in CS education have shown that existing research regarding reflective writing is constrained to the use of existing reflection frameworks [5, 6] that are not even tailored for application in CS. These theoretical frameworks have not shown good results when applied in the CS context. The theoretical framework [8, 44] used in here on the other hand was constructed by investigating by CS instructors' views on the criteria which can be used to assess reflective writing on their subject area and then validating the framework based on such instructors' expertise [9, 45] and empirically validating

The RWF adopted here was evaluated via the manual content analysis method. This paper presents an overview of the content analysis of the RWF by measuring the reliability and validity of the coding scheme used in it to assess reflective writing. It also examines the relationships between indicators of reflection and depth of reflection.

This study focused on two research questions, as follows. (RQ1) What is the relationship between the seven reflection indicators (descriptive, understandings, feelings, reasoning, perspective, new learning, and future action) and the three-levels of reflection (non-reflective, reflective, and critically reflective) used in the RWF? (RQ2) Can we reliably differentiate the seven indicators and three-levels of reflection of the RWF?

The paper is organized as follows. Section 2 presents the background in terms of reflective writing frameworks and content analysis methods. Section 3 describes the method used to assess the reliability and validity of the framework. Section 4 shows the result from the pilot tests and the validity and reliability of the manual annotation. Section 5 presents the discussion and conclusions.

I. BACKGROUND

A. The Reflective Writing Frameworks (RWF)

Feedback from the instructors based on reflective writing frameworks can help students to develop their learning process. These RWF will generally categorize sections of a student's writing into either reflective or non-reflective. Reflection (in writing) is a complex process due to the introspective nature of the task and the personal aspects of writing based on one's experience. According to Boud, Keogh and Walker [10] reflection is a complex process which is affected by many dimensions of living and interacts with

the cognitive process. Recognizing the various processes which are involved can help to understand this kind of cognition. Examples of these processes are judgment, evaluation, reasoning, problem-solving and memorizing [11]. In CS education, there is evidence that reflection is used in the development of problem-solving and analytical skills; in particular, metacognition skills development can be important when learning to code programs [15, 16]. The frameworks in the literature recognize at least two values of depth of reflective writing, the low being descriptive, also known as non-reflective, and the higher being the reflective level [17]. However, in general, frameworks propose at least three levels of reflection: descriptive, reflective and finally critically reflective [18].

Reflective writing frameworks can be developed based on the depth (levels) of reflection, or on the indicators of reflection, or on both [19]. For example, Wong, Kember, Chung and Yan [20] adapted the Mezirow's [21] framework that uses three levels (indicating depth of reflection), these are: non-reflector, reflector, critical reflector. The former authors defined a set of breadth elements to as evidence for each level, and these were based on the model that was proposed by Boud, Keogh and Walker [10]. The breadth elements that were included were: attending to feelings, association, integration, validation, appropriation, and outcome of reflection. According to Wong, Kember, Chung and Yan [20], these elements, as compared to the levels as proposed in other papers, are "more problematic and considerably less reliable". Experiments were carried out to the analyze the journal writing of forty-five nursing students after they had been introduced to the concepts of reflective writing using these aforementioned categories [20]. The conclusion that was presented was that the results of journal writing can be used as evidence for the presence or absence of reflection. Ip, Lui, Chien, Lee, Lam and Lee [22] also presented a model which employed three reflective levels, as presented by Wong, Kember, Chung and Yan [20]; these were as follows: non-reflector, reflector, critical reflector. However, this model also defined five indicators that represented the critical considerations encountered in nursing practice: aesthetics, personal issues, ethics, empirics, and reflexivity. The experiments were conducted on the written work of thirty-eight undergraduate students in a nursing program who attended a workshop on reflection skills. The empirical evaluation of the students' writings was conducted by two raters in terms of the overall reflection level of the writing produced before, during and after the program, but did not include a detailed classification of sentences and paragraphs. The results did not focus on the applicability of the model, instead it attempted to compare the levels of the individual students in terms of their reflective writing, before and after participating in the program. This focus was aimed at verifying the effects of the structured education program (which was intended to improve the reflection abilities of the students).

Ullmann [23] developed a reflection framework that used the common indicators derived from 24 models of reflective writing; Ullmann's model, intended to assess reflection text, uses two reflection levels (reflective and non-reflective) and eight indicators: reflection, experience, feelings, personal belief, difficulties, perspective, lesson learned, and future intention. Ullmann's [23] framework was empirically evaluated by reporting on the performance, in terms of reliability, of the manual content analysis by calculating Cohen's kappa in relation to it; this ranged between 0.48 and 0.98.

Ullmann, Wild and Scott [24] had argued that there was a variety of contexts in which reflection research is embedded (e.g., medicine, psychology, vocational education), and that certain indicators of reflection may be more important in a given context than in others. Thus, different fields, or more generally, contexts, require different reflection models. This has meant that the frameworks/models relating to reflective writing have been developed to be tailored to one particular context, task, or purpose, although, clearly, many frameworks agree in terms of their levels-of-reflection hierarchies, some having four [21] or three levels [20], and some two as we have seen. Each framework can be differentiated by looking at the indicators which will be more specific to the context of the reflective writing.

B. Reflection in CS education

In CS education, Reflection can be seen as the means and problem-solving can be seen as the outcome in many situations pertaining to CS. Reflection can often benefit students by improving their understanding of software development in terms of how developers develop and use algorithms. Dewey [25] described reflection as a way for solving problems. According to Donald Schön [26] described a reflection as reflection action that entails reflection on the means of testing and retesting the interpretation of the experience when solving of problems. Problem-solving whereby practitioners check and re-test their understandings as they attempt to solve problems which are at first ill-defined.

Reflection, in this context, is not viewed as an end in itself; it is a platform for change, a transformational mechanism [21], that can trigger inner personal changes relating to how challenges are to be addressed, experienced and talked about in the future.

George [4] claimed that "reflection in scientific disciplines may be different in type to the type of reflections made in humanities because of the nature of the underlying knowledge". This study also indicated that it is necessary to teach problem-solving and reasoning skills in the course of CS education in order to improve students' awareness of how to learn from a situation they are presented with – such as how to deal with finding the right sequence of steps to reach a goal or how to identify the roots of a problem and not be led astray by their own initial feelings about the situation [6].

Moreover, most frameworks examine how the writer expresses their feelings about a specific experience [27]. In CS education, this indicator (feelings) is not necessarily important to the problem-solving and decision-making skills; these tend to be focused more on reasoning, pragmatism, and the development of computational thinking than on emotions, that are, anyway, not fundamental to the reflective writing process. These led to emphasize that there is a need for a robust and valid reflection framework specifically for assessing the reflective writing, associated specifically with CS education [4].

However, in CS education, there has been a lack of empirical effort directed at the assessing students' reflective writing, based on an appropriate framework. However, Chng [6] implemented a framework of reflective writing for CS education based on Kolb [28] by combining Kolb's

experiential learning components with concepts related to the problem-solving process. Four 'actions' were defined: active experimentation, gaining experience, observation/reflection, and abstraction/conceptualization. She found that students were often asked to reflect without being guided by educators on how to reflect. In contrast, the Chng's [6] framework focused on how the student should reflect by defining a series of processes. Kolb's model is considered to represent a cyclical process of learning that involves many stages [28]. Demmans Epp, Akcayir and Phirangee [5] investigated reflective writing in CS education by asking students to write reflectively based on a set of questions relating to reflection and referring the student to examples of reflective thinking, in line with [26] and [21]. They recommended investigating new methods, in terms of timing and coding schemes, in order to support the student to reflect usefully with respect to CS education [5]. However, in neither study were the reflection indicators and levels identified, and the reliability of the results could not be determined. This makes it difficult to apply Chng's [6] modified framework or use the set of questions defined by [5]. Thus, these theoretical frameworks are hard to apply, as compared to the theoretical framework proposed in this paper. Moreover, within these studies, there is missing information regarding the reliability of each indicator of the dimensions of the framework; indeed, even the validity of these indicators is not verified. This information is vital for the understanding of the quality of any framework.

The most important aspects of reflection need to focus on quality of the framework for CS according to Hazzan and Tomayko [41] are: (1) the complexity inherent in the development of software systems, which requires the developer to improve their understanding of their own mental processes, and such can be achieved by applying a reflection approach, in order to teach developers how to think effectively; and (2) the role of communication among teams and with customers which requires developers to improve their communication skills, and such can be achieved by enhancing one's own mental processes. In this study the focus is on measuring the quality of the reflection framework in CS education.

C. Content Analysis

Content analysis is a research technique that involves using a specific coding scheme for annotating and summarising and so reporting on the content of text documents [29]. Manual content analysis is a common method which has been used to create, via analysis and evaluation, formative assessments of students' reflective writing [30-32]. These studies demonstrated the use of a coding scheme for assessing or evaluating student written reflection through manual content analysis. In order to apply manual content analysis, researchers must define the unit of analysis, e.g., sentence, paragraph, message, or document. This is then taken to represent the level of object to be annotated. A code is then manually assigned to each of these in the text to be analysed. It recommended that a smaller unit than the whole document should be used for manual coding schemes in order to assist in measuring the dimensions of a piece of reflective writing in a more detailed fashion [33-35]. For this reason, the use of the sentence as a unit of analysis was chosen here.

The most widespread measure of the quality of a manual content analysis is IRR; this measures consistency between raters in order to verify the performance of the coding process for assessing the reflective text. However, using a sophisticated coding scheme has been shown to exacerbate the pressure on instructors [36]. This issue has led to an emphasis, here, on the wider future aim of using this quality framework to automate reflective writing assessment.

A coding scheme is used to help raters annotate the dataset easily. The coding scheme has been adapted from Alrashidi, Joy, Ullmann, and Almujally [44] to assess the reflective texts of students based on CS instructors to determine the reflection depth and indicators. Three levels of reflection depth and seven indicators were used in the adapted coding scheme; definitions were constructed for each coding category (see Appendix). The manual content analysis aimed to classify sentences appropriately that related specifically to the reflection indicators and so to the levels defined in the RWF.

II. METHODOLOGY

A. Dataset

The dataset used in this study – of 174 different reflective writing documents – was employed for use by raters for the initial annotation as well as for the annotation coding scheme refinement. The data were collected from 174 third- and fourth-year CS student projects undertaken during the academic years 2013 through 2016 at the first author's university and anonymized before analysis. 1200 sentences were selected randomly for the annotation task. The students were asked to write reflectively about themselves in terms of their contributions, technical achievements, time management skills, limitations, lessons learned, and future work.

B. Annotation participants

IRR measurement was undertaken in order to revise the manual annotation process over the four pilot annotations. For all four pilots we recruited four raters based on their experience of assessing formative reflective writing, and their knowledge of reflective writing. This was so that they would be of the greatest assistance in producing reliable and clear coding scheme (known also guideline) – which would be based on their (the raters') comments and suggestions. The aim was to ensure the production of a high quality annotated reflective writing dataset for CS education based on the RWF. In order to ensure the availability of competent raters, each was given adequate training. The annotation coding scheme components are shown in the Appendix.

Many challenges exist in relation to manual annotation. First, the researchers had to perform several rounds of manual coding in order to ensure consistency between the raters. This involved using several pilot tests through which to define and refine the coding scheme so that the annotations could be made reliably and would precisely discriminate between different categories of text in the dataset is discussed below.

C. Annotation of the content analysis workflow

Manual annotation is a critical issue due to the fact that it dictates the quality and consistency of the annotated dataset. Moreover, in order to examine how easily the coding scheme could be applied by other raters, the coding scheme consists of rules that the raters must follow when annotating a dataset. The final coding scheme is the outcome of and iterative process. The coding scheme was revised after each pilot, based on the four raters' comments. These raters were hired

to manually annotate 1200 sentences according to the coding scheme. In Fig 1, the annotation workflow intended to ensure IRR.

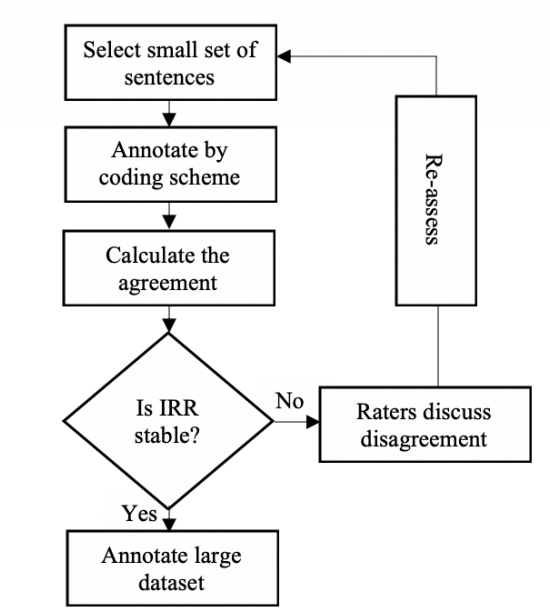

Fig 1: Annotation Workflow for IRR

The annotation workflow steps are as follows;

(1) Select a test set randomly of a specific size from the dataset.

(2) Annotate this test set using two or more independent raters based on the coding scheme which has already been defined into the following steps; (a) read the sentence; (b) fragment to focus on the meaning to assess the presence of indicators to determine which level is presented; (c) to rate each sentence according to categorize the sentences that seem to have one or more reflection indicators; and (d) to rate the sentence into one reflective level.

(3) Calculate the IRR between their separate coding using the chosen IRR metric. The IRR metric and set the stable value which represents acceptable agreement, to infer the stable degree of reliability for the consistency between raters on annotating. This step was done in two rounds: the first round had raters independently annotate each sentence (see step 2). In the second round, the raters came together to discuss disagreement sentences to come up with a consensus on the definition of the coding scheme categories. Then they independently re-annotated disagreements sentences (from the first round). Disagreement sentences in the second round were resolved by another rater (Judger). If all raters classified a sentence differently, the sentence was removed from the annotated dataset. The IRR was calculated kappa statistics (κ) such as Cohen's κ for two raters and/or Fleiss's κ for more than two raters to measure the consistency of agreement. The R package is used for the computation of κ and Spearman's correlation.

(4) Compare with the minimum value of the IRR measure yielded in Step 3 as follows: (a) if the IRR calculated is unstable value then steps 1, 2, and 3 are re-assessed; (b) if the IRR is stable value, the coding is considered to be reliable. This means that a single rater can annotate the rest of the data set independently.

III. Result

A. Manual Annotation of pilot tests

Four pilot studies were carried out before the actual annotation dataset to achieve high quality of annotated dataset. Before the pilot tests, the raters were introduced to the following; Firstly, the raters explored the interpretation of the coding scheme independently annotating a short document based on the RWF coding schemes. Secondly, we began to explore the raters' views on whether they could be reliable apply the RWF to the dataset. This step was proposed by Abou Baker El-Dib [30] who argued that there is a need to consult the expert raters who are creating a manual content analysis; such a consultation can result in a reviewing and reporting by raters which will assist in improvements to the coding scheme. The focus was to train the raters to be familiar with and proficient in the use, of the coding scheme employed for this study. A first pilot test, twenty sentences were annotated using the initial coding scheme annotation design by four independent raters, after the ratings were explained. It was found that the annotation coding scheme which had been given to the raters had some points of vagueness. For example, the raters suggested that attention should be given to providing examples, from the CS dataset, of each level and indicator which they are expected to use. They also identified that attention should be given to clarify the feelings indicators, by adding further explanations. Via this process, the coding scheme annotation design was altered. The altered coding scheme annotation design was then used in the second annotation pilot test by three raters over 40 randomly selected sentences. After modifying and retesting the RWF coding scheme with respect to the first pilot test, we then further reevaluated and redesigned the coding scheme by adding examples from the dataset with respect to the second. After these modifications which were made to the RWF coding schemes, a consensus was obtained regarding the three levels and seven indicators. Following this, only minor improvements to the framework were made. An example of such a minor amendment was the re-drafting of the definitions of the levels of reflection, to make these more accurate; another example was the addition of an example of each specific indicator. The final version of annotation designs the coding scheme was used in the third and fourth annotation pilot tests. Three independent raters used the RWF and further tested it by applying it to 100 and then 400 randomly selected sentences. Table 1 shows the values of IRR indicator, κ. With this process, we achieved the κ values of 0.87, 0.78 and 0.80 these values are, substantial to almost perfect agreement [38]. The observation of the pilot studies resulted in the final annotation design of coding scheme that used as a guideline to annotate the dataset by raters to achieve high quality of annotation dataset.

TABLE 1. THE IRR COMPUTED FOR EACH ITERATION OF THE RWF

Date	#Pilot Test	#Sen	#Raters	κ
October 2018	1	20	4	0.52
January 2018	2	40	3	0.73
March 2019	3	100	2	0.87
May 2019	4	200	3	0.78
July 2019	4	400	3	0.80

B. Validity

Krippendorff [29] discusses face validity, which "is "obvious" or "common truth." We appeal to face validity when we accept research findings because they "make sense" - that is, because they are plausible and believable "prima-facia" – usually without having to give or expecting to hear detailed reasons (p.313)." Spearman's ρ rank was chosen to measure the strength of association between the reflection indicators and the levels of depth.

TABLE 2. THE CORRELATION BETWEEN THE REFLECTION' LEVELS AND THE INDICATORS

Indicators	Non Reflective	Reflective	Critically Reflective
Descriptive	0.91	-0.63	-0.27
Understanding	-0.14	0.63	0.43
Feelings	-0.45	0.53	0.24
Reasoning	-0.56	0.46	0.07
Perspectives	-0.3	0.02	0.33
New Learning	-0.25	-0.39	0.82
Future Action	-0.16	-0.23	0.50

In the table 2, all rank correlations are statistically significant between the reflection levels competency and the seven indicators are statically significant. This suggests that some indicators positively correlate with each reflection level. The strength of the correlation among the seven indicators of dimension and the three levels of depth was assessed using Spearman's ρ rank, ranging between -1 and +1. The results show that values approaching -1, indicating of course a negative correlation, were encountered. For instance, there were negative values representing the correlation between the descriptive indicator and the reflective and the critically reflective levels. Values approaching +1 indicate a positive correlation (which validates the strength of the corresponding relationships between indicators and the reflection levels). For example, the non-reflective level had a very high association with the descriptive indicator but a negative association with the rest of the indicators. In contrast, the reflective level had a low to moderate association with all the indicators. For example, reasoning, understanding, and feelings correlated moderately with the reflective level. while descriptive, new learning, and future action correlated negatively with the reflective level. The critically reflective level correlated strongly with new learning and moderately with future action, reasoning, understanding, perspective, and feelings. Reflecting critically involves looking back on previous experiences or on what has already been learned. This (Spearman's) result showed that text at the critically reflective level can include evidence of all the indicators in the reflective level plus evidence of new learning and/or the future action indicator. This demonstrates a direct mapping between the reflection indicators and the reflection level (depth) [20, 30]. It is important to note that these correlations statistically indicated a variety of types of correlation, and this (valid) situation was due to the large sample size and the unit of analysis in use.

C. Reliability

Ensuring IRR is a major problem in the study of content. Indeed, this is considered to be the primary measure of objectivity in content studies and is defined as stated by Rourke, Anderson, Garrison and Archer [39] "the extent to which different coders, each coding the same content, come to the same coding decisions." Many studies do not report the IRR of the coding scheme that they employ. According to Lombard, Snyder-Duch and Bracken [40] indicated this "can be seen as the consequence of a lack of detailed and practical guidelines and tools available to researchers regarding reliability." This means that it is vital to report the reliability of the coding scheme applied to a manual content analysis; this enables the researchers to consider how the established human raters can distinguish well between the indicators specified in the coding scheme; this improves the reliability of the research results [17].

Table 3 shows the values of IRR of indicators and levels of reflection. For the seven indicators achieved the κ values ranged between 0.46 to 0.75 these means moderate to substantial agreements. And for the three reflection' levels achieved the κ values ranged between 0.40 to 0.72 these values are fair to substantial agreement [38]. In the statistical test a p-value is generated to determine the significance of the results weather correlate with the selected variables. For example, to determine whether the indicators and levels of reflection are correlated with the rater's agreement or not. The indicators and levels of reflection are highly significant with the raters' agreement with amounts to p-value = 0.00.

TABLE 3. INTER-RATER RELIABILITY AND P-VALUE OF THE INDICATORS AND LEVELS OF REFLECTION IN THE RWF OF THE MANUAL ANNOTATION

Indicators and levels	#Raters	#Sentence	κ	p-value
Descriptive	3	1114	0.55	0.00
Feelings	3	1128	0.64	0.00
Understanding	3	1128	0.46	0.00
Reasoning	3	1128	0.64	0.00
Perspective	3	1128	0.58	0.00
New Learning	3	1128	0.69	0.00
Future Action	3	1128	0.75	0.00
Non Reflective	3	509	0.72	0.00
Reflective	3	509	0.42	0.00
Critically Reflective	3	269	0.40	0.00

IV. THE RWF

As regards CS students, Stone and Madigan [42] investigated the idea that the depth of reflection a student is capable of follows a hierarchy, and that a student can usually write more insightfully and to a greater depth at each reflection. This means that students, generally, start by writing descriptive texts then proceed to be more insightful each time they practice reflective writing until they reach the critically reflective level.

According to the framework of Alrashidi, Joy, Ullmann, and Almujally [44] there are three levels of reflection depth (non-reflective, reflective, and critically Reflective) and these are linked to a specific indicator of reflection. These levels are described in the following.

A. Non Reflective Level

Text at the non-reflective level will show evidence of the descriptive indicator, involving a reporting of something, like an event or a theory, without an elaboration in terms of how, why and what the impact of the event or theory might be. The students usually start writing at this level then gradually progress to the reflective levels [8]. This fragment is from the CS dataset and includes evidence of the **descriptive** indicator: "*This project involved developing a digital music recommender system with the specific goal of solving the network effect issues evident in current recommender systems.*"

B. Reflective Writing Level

The *reflective writing level* is demonstrated when the writer mulls over reasons, discusses alternatives, presents conjectures and exhibits other products of deep cognition [14].

A student or developer may, when writing at this level, identify relationships between feelings and/or understandings relating to prior knowledge with such feelings, and/or understandings relating to new knowledge, by engaging in explanations that add value to their acquired knowledge. Here is a fragment/sentence from the CS dataset which includes evidence of the **understanding** and **reasoning** indicators: "*Both approaches have completed their work but I now feel that it is much healthier and easier emotional to continually work at a steady pace, meaning that you are not away from the project for an extended amount of time and therefore do not need to reacquaint yourself with its intricacies.*" These findings are in line with Hazzan and Tomayko [41] as regards the complexity inherent in the development of software systems, which requires the developer to improve their understanding of their own mental processes by raising questions about how to think. George [5] added that, in CS, a knowledge of the relationship between problem–solving and reasoning can be applied to practical problems.

C. Critically Reflective Level

In this context, text at the critically reflective level will describe the outcomes that the writer expects from the application of new ideas or learning, rather than just reporting learning experiences in a descriptive manner. This level can be achieved by stating what has been learned and, importantly, how to deal with related experiences in the future. The following fragment of critically reflective writing is from the CS dataset in which the writers provide evidence indicating their **new learning**, **perspective,** and **reasoning**: *"Whilst I think we worked well as a team, I found MemberA's and MemberB reluctance to manage and criticise my work to be relatively frustrating, and I would have preferred to be free from the urge to pursue some of my more abstruse ideas (such as pursuing CRF-based annotation), and my tendency to build overwrought, critically-engineered, overly adaptable software systems.*"

V. Discussion and Conclusion

Up to our knowledge, this is the first research that has investigated on assessing CS students' reflective writing by using the RWF coding scheme which developed for this context.

In this study, the aim was to analyze CS students' reflective writing by using manual content analysis as a procedure for applying RWF as a coding scheme to assess students' reflection written.

The findings relating to the manual annotation of the CS dataset are in line with the aspects of reflection outlined by Hazzan and Tomayko [41] associated with the context of software developers, that they must become aware of their communication skills by increasing their own mental processing skills and their ability to think deeply about situations and about how to deal with difficulties.

A. The relationship between reflection levels and indicators

The reflective writing literature showed that reflection is a multi-dimension that described the levels and indicators. The RWF captured two dimensions with three levels and seven indicators that are aligned with the common indicators in reflection literature.

The RQ1 was answered by using manual content analysis approach that the dataset annotated to measure the correlation (Spearman's test) between two dimensional these are the reflection levels and reflection indicators – in accordance with the RWF. The results of correlation showed a positive correlation that the reflection indicators are, indeed, related to the reflection levels, which is as expected by the framework and a sign of empirical validity.

Some of the indicators were closely linked to specific reflection levels. For example, the reflective level was related closely to understanding, the feelings and reasoning indicators were closely linked with the critically reflective level, which was most strongly correlated with the new learning, and future action indicators.

However, the perspective indicator showed only a weak correlation with the reflective and critically reflective levels. Our findings are consistent with the results of the correlations provided by Ullmann [17, 43]. Ullmann [43] found that associated the weak correlation between reflection and perspective because in this indicator the writer in the dataset collected "shifts focus away from the thought sphere of oneself to other perspectives." The coding scheme was successfully used in the manual content analysis via annotation used to assess of reflective writing in CS education.

B. The reliability of manual annotation of the RWF

Many studies did not report the validity of their manual annotation in relation to reflection frameworks, being content with reporting on reliability only [33]. The recent research in this area did not report empirical evidence of the reliability or validity of coding scheme that used to assess reflection in CS education such as [6, 5].

The RQ2 was answered by employing independent raters to manually annotate sentences according to the presence or absence of seven indicators and the three levels, as defined by the RWF, to assess the reflective writing. The IRR estimates showed that the manual raters can reliably annotate sentences according to reflection the three levels and seven indicators of the RWF.

The evaluation of the quality of the RWF showed that the theoretical framework for reflective writing is reliable and valid framework for analyzing student's reflective writing. The evaluation also showed that our RWF is not only

theoretical sound, but also showed evidence of empirical validity of manual annotation.

For practical implication, the RWF can be used as a rubric for teachers to assess the quality of students' reflective writing. For example, the educator can use the RWF as guideline to classify student's reflection' depth in their own reflective writing.

Moreover, the annotated dataset will be used to implement automated reflective writing analysis to help to understand the automated analysis by using machine learning algorithm.

ACKNOWLEDGEMENT

The research was partially funded by Kuwait Foundation for the Advancement of Sciences (KFAS) under project code "CB19-68SM-01."

REFERENCES

[1] H. S. Wald, J. M. Borkan, J. S. Taylor, D. Anthony, and S. P. Reis, "Fostering and evaluating reflective capacity in medical education: developing the REFLECT rubric for assessing reflective writing," *Acad Med,* vol. 87, no. 1, pp. 41-50, Jan, 2012.

[2] E. Cohen-Sayag, and D. Fischl, "Reflective Writing in Pre-Service Teachers' Teaching: What Does It Promote?," *Australian Journal of Teacher Education,* vol. 37, no. 10, pp. 2, 2012.

[3] J. Betts*, "Theology, therapy or picket line? What's the 'good'of reflective practice in management education?," *Reflective Practice,* vol. 5, no. 2, pp. 239-251, 2004.

[4] S. E. George, "Learning and the reflective journal in computer science," *Australian Computer Science Communications,* vol. 24, no. 1, pp. 77-86, 2002.

[5] C. Demmans Epp, G. Akcayir, and K. Phirangee, "Think twice: exploring the effect of reflective practices with peer review on reflective writing and writing quality in computer-science education," *Reflective Practice,* vol. 20, no. 4, pp. 533-547, 2019.

[6] S. I. Chng, "Incorporating reflection into computing classes: models and challenges," *Reflective Practice,* vol. 19, no. 3, pp. 358-375, 2018.

[7] D. Kember, "Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow," *International Journal of Lifelong Education,* vol. 18, no. 1, pp. 18-30, 1999.

[8] H. Alrashidi, T. D. Ullmann, M. Joy, and S. Ghonuaim, "A Framework for Assessing Reflective Writing Produced Within the Context of Computer Science Education," March, 2020.

[9] H. Alrashidi, T. D. Ullmann, S. Ghounaim, and M. Joy, "Understanding Reflective Writing Criteria in Computer Science Education from CS Educators in Higher Education."

[10] D. Boud, R. Keogh, and D. Walker, *Reflection:Turning experience into learning*: Routledge, 1985.

[11] G. Schraw, and R. S. Dennison, "Assessing metacognitive awareness," *Contemporary educational psychology,* vol. 19, no. 4, pp. 460-475, 1994.

[12] J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry," *American psychologist,* vol. 34, no. 10, pp. 906, 1979.

[13] M. Main, "Metacognitive knowledge, metacognitive monitoring, and singular (coherent) vs. multiple (incoherent) models of attachment," *Attachment across the life cycle,* vol. 127, pp. 159, 1991.

[14] L. Corno, "The metacognitive control components of self-regulated learning," *Contemporary educational psychology,* vol. 11, no. 4, pp. 333-346, 1986.

[15] A. Fekete, J. Kay, J. Kingston, and K. Wimalaratne, "Supporting reflection in introductory computer science." pp. 144-148.

[16] J. Stone, and E. Madigan, "Integrating reflective writing in CS/IS," *SIGCSE Bulletin,* vol. 39, pp. 42-45, 06/01, 2007.

[17] T. D. Ullmann, "Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches," *International Journal of Artificial Intelligence in Education*, pp. 1-41, 2019.

[18] J. Fook, S. White, and F. Gardner, "Critical reflection: a review of contemporary literature and understandings," *Critical reflection in health and social care,* vol. 3, pp. 20, 2006.

[19] J. Moon, "Using reflective learning to improve the impact of short courses and workshops," *Journal of Continuing Education in the Health Professions,* vol. 24, no. 1, pp. 4-11, 2004.

[20] F. K. Wong, D. Kember, L. Y. Chung, and L. Yan, "Assessing the level of student reflection from reflective journals," *Journal of advanced nursing,* vol. 22, no. 1, pp. 48-57, 1995.

[21] J. Mezirow, "How critical reflection triggers transformative learning," *Fostering critical reflection in adulthood,* vol. 1, pp. 20, 1990.

[22] W. Y. Ip, M. H. Lui, W. T. Chien, I. F. Lee, L. W. Lam, and D. Lee, "Promoting self-reflection in clinical practice among Chinese nursing undergraduates in Hong Kong," *Contemporary nurse,* vol. 41, no. 2, pp. 253-262, 2012.

[23] T. D. Ullmann, "Keywords of written reflection-a comparison between reflective and descriptive datasets." pp. 83-96.

[24] T. D. Ullmann, F. Wild, and P. Scott, "Comparing automatically detected reflective texts with human judgements," 2012.

[25] J. Dewey, *A restatement of the relation of reflective thinking to the educative process*: DC Heath, 1933.

[26] D. A. Schön, *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*: Jossey-Bass, 1987.

[27] D. Boud, R. Keogh, and D. Walker, *Reflection: Turning experience into learning*, New York: Nichols Publishing Company, 1985.

[28] D. A. Kolb, *Experiential learning: Experience as the source of learning and development*, p.^pp. 86–87 New Jersey: Prentice Hall, 1984.

[29] K. Krippendorff, "Reliability in content analysis," *Human communication research,* vol. 30, no. 3, pp. 411-433, 2004.

[30] M. Abou Baker El-Dib, "Levels of reflection in action research. An overview and an assessment tool," *Teaching and Teacher Education,* vol. 23, no. 1, pp. 24-35, 2007.

[31] D. Kember, J. McKay, K. Sinclair, and F. K. Y. Wong, "A four-category scheme for coding and assessing the level of reflection in written work," *Assessment & Evaluation in Higher Education,* vol. 33, no. 4, pp. 369-379, 2008.

[32] F. K. Wong, D. Kember, L. Y. Chung, and L. Y. CertEd, "Assessing the level of student reflection from reflective journals," *Journal of advanced nursing,* vol. 22, no. 1, pp. 48-57, 1995.

[33] E. Poldner, P. Simons, G. Wijngaards, and M. Van der Schaaf, "Quantitative content analysis procedures to analyse students' reflective essays: A methodological review of psychometric and edumetric aspects," *Educational Research Review,* vol. 7, no. 1, pp. 19-37, 2012.

[34] A. Bell, J. Kelton, N. McDonagh, R. Mladenovic, and K. Morrison, "A critical evaluation of the usefulness of a coding scheme to categorise levels of reflective thinking," *Assessment & Evaluation in Higher Education,* vol. 36, no. 7, pp. 797-815, 2011.

[35] M. M. Plack, and L. Greenberg, "The reflective practitioner: reaching for excellence in practice," *Pediatrics,* vol. 116, no. 6, pp. 1546-52, Dec, 2005.

[36] S. Koole, T. Dornan, L. Aper, A. Scherpbier, M. Valcke, J. Cohen-Schotanus, and A. Derese, "Factors confounding the assessment of reflection: a critical review," *BMC Medical Education,* vol. 11, no. 1, pp. 104, 2011/12/28, 2011.

[37] B. R. Eagan, B. Rogers, R. Serlin, A. R. Ruis, G. Arastoopour Irgens, and D. W. Shaffer, "Can we rely on IRR? Testing the assumptions of inter-rater reliability."

[38] M. R. Lynn, "Determination and quantification of content validity," *Nursing research*, 1986.

[39] L. Rourke, T. Anderson, D. R. Garrison, and W. Archer, "Methodological issues in the content analysis of computer conference transcripts," 2001.

[40] M. Lombard, J. Snyder-Duch, and C. C. Bracken, "Content analysis in mass communication: Assessment and reporting of intercoder reliability," *Human communication research,* vol. 28, no. 4, pp. 587-604, 2002.

[41] O. Hazzan, and J. E. Tomayko, "Reflection and abstraction in learning software engineering's human aspects," *Computer,* vol. 38, no. 6, pp. 39-45, 2005.

[42] J. A. Stone, and E. M. Madigan, "Integrating reflective writing in CS/IS," *ACM SIGCSE Bulletin,* vol. 39, no. 2, pp. 42-45, 2007.

[43] T. D. Ullmann, "Automated detection of reflection in texts. A machine learning based approach," The Open University, 2015.

[44] H. Alrashidi, M. Joy, T. D. Ullmann, and N. Almujally, "Validating the Reflective Writing Framework (RWF) for Assessing Reflective Writing in Computer Science Education Through Manual Annotation," *Intelligent Tutoring Systems.* pp. 323-326.

[45] H. Alrashidi, M. Joy, T. D. Ullmann, and N. Almujally, "Educators' Validation on a Reflective Writing Framework (RWF) for Assessing Reflective Writing in Computer Science Education," *Intelligent Tutoring Systems.* pp. 316-322.

Appendix

The coding scheme of reflective writing levels in the RWF

Reflection' levels	Definition	Example	Explanation
Non-Reflective/ Descriptive	The non-reflective writing level is characterized by the mere description of things, like events or theories; such description will not be elaborated in terms of how, why or impact.	"The group leader was a different person in the first term from the one in the second."	This is the non reflective level because of its superficial description of fact, not backed with evidence.
Reflective	The reflective writing level is demonstrated when the writer mulls over reasons, discusses alternatives, presents conjectures and exhibits other products of deep cognition	"I think that I was in the best role for me, I led the team in ways of project management when required and helped a lot with teaching the others to use git and then later in the project completed some of the more complex tasks required on git."	This is the reflective level because there are elements of analysis and self-reflection on emotions.
Critically Reflective	Writing at the critically reflective writing level exhibits new ideas and decision making. This level is involved with providing the type of transformations of perspective that are unlikely to occur frequently, and often relate to modifications to a fundamental theory.	"If I had had a better idea of the scope of the components earlier on in the project, I perhaps could have offered my services to other group members working on complex components which unfortunately did not function."	This is the critically reflective level because the writer has come to the awareness of his/her action, indicating his/her new learning gain, either its positive or negative outcome thus suggesting alternative ideas.

The coding scheme of reflective writing indicators in the RWF

Reflection' indicators	Indicators	Examples
Descriptive	The writer reports a fact from experience and/or material.	"The second category of personal development afforded by the project is the technical skills that it has taught."
Understanding	The writer understands and/or analyse the experience.	"The process of making error corrections helped me understand better these complex structures in NLP"
Feeling	The writer identifies their own thoughts and feelings.	"It was the most fun and rewarding project I have ever undertaken, as well as being the most challenging and frustrating at times."
Reasoning	The writer explains the experience by giving reasons.	"I am a better Java programmer as a result of using the language for significant development."
Perspective	The writer shows awareness of alternatives.	"I am also now competent in Python, a language which I had never used before the project. In addition to this, more familiarity was gained with other software development tools; for example, Maven, IntelliJ, and Git."
New learning	The writer describes concrete learning.	"The biggest lesson I have learnt from this project is that it is highly important to not cast a continuous assessment aside."