Effectiveness of Peer Review in Teaching and Learning User Centered Conceptual Design Among Large Cohorts of Information Technology Students

Farshid Anvari
Department of Computing
Macquarie University
Macquarie Park, Australia
farshid.anvari@acm.org

Hien Minh Thi Tran
Minh Hien Pty Ltd
Sydney, Australia
hientran@minh-hien.com

Deborah Richards
Department of Computing
Macquarie University
Macquarie Park, Australia
deborah.richards@mq.edu.au

***Abstract*—Conceptual design often sets the direction of the development of a product. Educating Software Engineering and Information Technology (IT) students to consider the needs of users during the conceptual design stage is important. The assessment of conceptual design is a time consuming process. When budgets are limited, financial resources are inadequately allocated to allow staff to spend quality time assessing the conceptual design artifacts of large classes. We present a methodology to teach, capture and evaluate conceptual design artifacts for large classes. In two separate studies, two groups of over 185 undergraduate IT students reviewed their peers' design artifacts using a comprehensive rubric. Their reviews were later checked by two researchers using the same rubric. It was found that not only were the student-peers' reviews close to the researchers' reviews, it was possible to give students valuable and timely feedback and scaffold them to reflect, an essential characteristic for professional competence. In addition, we found that assessment by staff was not feasible due to inadequate resources. We conclude that for large classes, conceptual design artifacts can be evaluated and valuable feedback provided in a timely manner by peers with the guidance of a comprehensive rubric.**

Keywords—conceptual design assessment, peer review, reflective practice, persona, software engineering education, information technology education

I. INTRODUCTION

Scholars agree that there are multiple stages to design, which are often iterative and non-linear; however, the conceptual design is the first stage of the design where important decisions about the software design and the resultant application are made [20, 30, 51]. Creativity and innovation drive the economy and industry [59] and creativity happens during the conceptual design stage [5, 37]. Software engineers and Information Technology (IT) designers are often required to produce innovative solutions or improve on existing products before the problems are defined properly [1, 28]. Additionally, programmers and software engineers in industry often submit their code for review by their colleagues [27]. Hence, Software Engineering and IT students need to receive training in conceptual design and learn to review their peers' design artifacts objectively.

The conceptual design phase is an intense phase where ideas are generated at a rapid pace and the notes captured are often incoherent [5]. Schön [60] considers "reflection-in-action" as the shift of attention from analysis of the known problem to an occurrence of a new resultant situation which is an ideation or a solution to the problem [1, 60]. As early as 2000, researchers have been trying to devise tools to capture this phase of design [39]. Conducting conceptual design in a classroom or laboratory is similar to reflection-in-action that takes place in the midst of action [60]. When a design process is conducted after the ideation stage, the designer reflects on the ideations and often provides more explanations. The design artifacts tend to resemble detailed design. The activities are similar to "reflection-on-action" as the designers are reflecting on their initial ideations and revisiting their initial experiences of their design [60].

Incorporating the needs of users during the design phase, User-Centered Design (UCD), would result in products that have higher usability values [51]. In User-Centered Conceptual Design (UCCD), the design is conceived to serve the needs of users.

An important aspect of the education of IT students is proper evaluation of their conceptual design artifacts and provision of timely feedback on their work [19, 69]. Conceptual design artifacts that are produced in brief timeframes (e.g. during a class or laboratory session) require extra assessment effort by a marker, in comparison with a well-documented design. The marker-teacher is required to carefully read the artifacts and decipher any new and innovative design that may be embedded in the student's working notes as symbols, dot points or short phrases. Since conceptual design is innovative and not a copy of an existing design, it is unlikely that the markers can use a template to assess the artifacts quickly. Teaching staff at higher education institutions, due to limited time, often allocate the tasks of marking undergraduate students' assignments to casual staff who may not be paid for the time which they need to spend [53, 68]. The other compounding factor is markers often do not have adequate expertise to assess the conceptual design artifacts. An added issue is that students' results must be made available before the end of semester, which further limits the level of detailed assessment of the conceptual design that can be provided.

In this paper, we present peer review as an alternative method to staff assessment of conceptual design artifacts. In two separate studies of two large classes of over 200 students each, the students produced design artifacts, peers evaluated and provided comments on the design artifacts in a timely manner. During the second study, a university staff member was provided with a marking rubric and asked to assess the design artifacts from all of the workshop classes (the activity was worth 5% of the total assessment). The alternative was to ask tutors to mark their own classes but with 30 in each class, this was not viable based on the number of hours they were employed. The staff member not only had no time to provide any comments, there was wide variation between the staff review and the peers' or researchers' review. In addition

www.minh-hien.com/academic – Preprint: 43rd ACM/IEEE ICSE-SEET 2021

to providing timely and reasonable assessment, we found that peer reviewing had the added benefit that the students had multiple opportunities for learning and reflection. We expect that the learning experience would prepare them for their future in becoming reflective practitioners during their careers.

II. Background Information

The background information covers the following areas: (1) the difference between conceptual design and other phases of design, (2) reflective practices, personas and their roles during UCCD, (3) the assessment of conceptual design, rubrics and peer review.

Researchers (e.g. [5, 37, 38]) who studied the design process agree that, although the process is divided into a number of stages, it is neither a linear process nor are the stages clearly distinct. Conceptual design is the first stage of the design during which the cognitive processes are intense [30]. Horváth [37] proposed that the conceptual design is a multifaceted creative process that happens after a problem definition or a needs assessment such as "market-product-technology investigation or product idea generation" [37, p. 2].

Capturing conceptual design as it happens (enlightenment) consists of a number of dot points or unformed sentences or sketches [62] as "enlightenment is normally a short, critical period because the idea is easily lost" [5, p. 147]. This is because the cognitive process is very active and hence ideas are generated in a more rapid succession [63, 64].

Conceptual design has been researched in numerous domains and different models have been proposed for studying and capturing it [31, 61, 70]. Researchers also studied conceptual design ability as an individual trait [25]. Our research has indicated that a lot of work needs to be done in developing tools for facilitating, capturing and reproducing conceptual design; most researchers rely only on written text, speaking and sketching to capture conceptual design [2, 45].

Schön [60] introduced the concept of reflection in practice and defined reflection-in-action as the thinking and reflecting that happens whilst one is actively doing the task and, reflection-on-action as the thinking and reflecting that occurs after working on the task. Killion and Todnem [41] introduced reflection-for-action, which assists in improving future performance in a similar task. Beckwith [15] developed a framework, reflection-for-learning, which assists the reflective students to achieve higher order learning.

Hence, during the initial conception of a design in the classroom, a student needs to reflect-in-action. While documenting the conceptual design, the student is reflecting-on-action. While reviewing their peers' artifacts and when contemplating on their own artifacts, they are reflecting-for-action. Reflective practices are important in software engineering as it helps the engineers to draw inferences from their experiences for more complex situations [23]. Harlim, Belski, Lemckert, Jenkins and Lang-Lemckert [34] found that experienced engineers tend to reflect continuously to improve their performance whereas novice engineers reflect when they have made a mistake and hence recommended the introduction of reflective practices in education. Reflective practices can be learnt by scaffolding but the learner must be open to introspection and gain the capacity for abstract learning [22].

In reflective practice, the designer is conscious of the users of the product [35, 60]. The concept of UCD considers users as the center of the design [51]. However, for the design activities of Software Engineers and IT students when the users are unavailable, a persona, an archetypical user, is used to represent the users [21]. Personas have been used for teaching design in higher educational institutions [40, 48, 65-67]. The use of personas not only assists designers to design with the target end users in mind but also allows them to practice reflection in their design [11].

To assess students' output fairly and objectively, rubrics are used [18]. Further, students use rubrics to decide what they need to study and how to study to achieve them [4]. The Revised Bloom's Taxonomy groups knowledge into four categories and cognitive processes into six categories. The four categories of knowledge are: factual, conceptual, procedural and metacognitive and the six categories of cognitive processes are: remembering, understanding, applying, analyzing, evaluating and creating [3]. Students use meta-cognitive knowledge to select a learning methodology and approaches depending on the assessment techniques and purpose of learning [3].

Researchers have used the Revised Bloom's Taxonomy to evaluate participant's cognitive activities. Lu and Churchill [49] used the Revised Bloom's Taxonomy to rate students' cognitive activities analyzing, evaluating and synthesizing information for interpreting and meaning-making during social interaction for learning.

Biggs and Tang [17] who considered knowledge to be holistic, designed their rubric to assign categories which are more suited to use across a semester-long subject. Some students are interested only in learning material superficially in order to score marks [17]. Thus, design of the rubric and allocation of marks must encourage and assess engagement with the learning material at the level of depth required for meeting the learning outcomes of the unit. A rubric is a good tool to achieve this as it aligns marks to the requirements of the course [4].

A review of the literature by Richards [57] concluded that a well-constructed rubric can be a tool for successful peer and self-review. Hafner and Hafner [33] used a well-structured rubric (five grades for five elements of the task with explanations for each) for peer and staff assessment of oral presentations by biology students, and found that it produces consistent results. The rubric for assessing Software Engineering students' conceptual designs must distinguish between the elements of innovation as well as improvements made to an existing design [50] and be balanced in its level of detail so that it can be used easily [14]. The assessable material has to be divided into distinguishable sections so that the knowledge and the cognitive processes contained in each section can be easily classified and assessed. Bailey and Szabo [14] devised a rubric that is linked to different levels of Bloom's Taxonomy [3]: remembering, understanding, applying, analyzing, evaluating, and creating.

To prepare students for realistic work situations, Software Engineering and IT students must learn how to design whilst participating in activities such as studios or professional meetings [26, 28]. Professionals are often

required to review their peers' designs [27]. It is essential for undergraduate Software Engineering and IT students to learn the art of UCCD and peer reviewing, because by learning to use personas, the students become familiar with the concept. Hence when working in industry, their willingness to work with personas increases [58]. Furthermore, as peer review requires the reviewers to evaluate their peers' work at a higher rung of Bloom's Taxonomy, hence, it promotes a deep learning strategy.

Peer review has been studied by many researchers in technical and engineering fields: Richards [57] in assessing a project-based subject for senior students, used peer review assessment for internal project groups and found these to be consistent with staff assessments. Garousi [29] applied peer review for a project-based design subject during the final year Software Engineering course. Kwan and Leung [44] engaged a group of Hospitality Industry students to do peer review and concluded that the learning benefit outweighs risk of erroneous assessment, particularly when it is a small percentage of overall assessment.

In massive online systems, self and peer assessment is often used for materials submitted. To train the students in assessing their peers' work, Kulkarni et al. [43] used a system in which the students initially assessed teachers' assessed materials and hence, their competence was assessed before they could assess other students.

The challenge of assessing conceptual design is to detect creative ideas that are relevant and feasible [5]. Due to the time and cost associated with marking students' assessments, research to automate the assessment of student outputs that are not quizzes is continuing e.g. [2, 16, 46] yet these are not used for conceptual design assessments.

III. Research Objective

Our research objective in this paper is to discover the assessment strategy that results in a fair assessment of the conceptual design artifacts and provides feedback to students in a timely manner. Towards this objective, we formulated the following research question:

How can large classes of Software Engineering and Information Technology students be taught User-Centered Conceptual Design and assessed efficiently?

IV. Methodology

To answer the above research question, we conducted two studies involving the use of personas for conceptual design and peer review of the design artifacts. Fig. 1 shows our research model and study design to achieve the objective in this paper for the two studies with the exception that in the first study, the marker's assessment was carried out by the researchers.

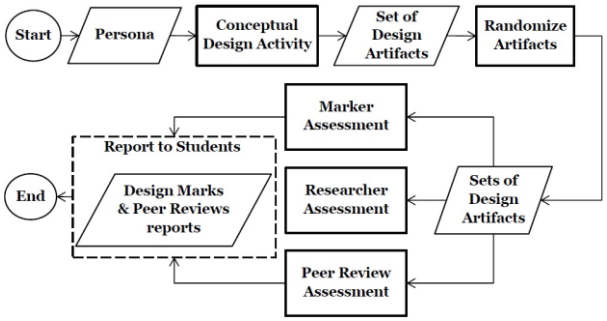

Fig. 1. Research Model (relevant parts)

In this paper, we focus on the conceptual design activity, peer review activity, marker's assessment, researchers' assessment and peer reviews' assessment from parts of the studies that were conducted in 2017 (Study A) and in 2018 (Study B). This paper also reports on the parts of a bigger project that was designed to answer a wider objective and was conducted over a number of years [12].

In Study A, we used four personas (Henry, Henrik, Hank and Harry) and a specification (as a control group). In Study B, we used six personas (Paul, Peter, William, Minh, Thuy, Chi). All personas were authored to be varying in knowledge and cognitive processes. The personas had a common problem. In Study A, the personas had problems with managing their finances and, in Study B, the personas wished to improve their English language skills. Our objective was not to study the design of a financial or a linguistic application but to study the learning and teaching of conceptual design, UCD with the help of personas and the assessment of the conceptual design artifacts. In this paper, we concentrate on the processes for the assessment of the design artifacts.

In both studies, we gave the students one persona (selected randomly from the set of personas) for which to design an application that met the needs of that persona. For Study A, the set also included one specification for the control group. Students in the control group were given a specification instead of a persona, and thus had to design to the needs of an imaginary user.

We designed our data collection for the design activities in two phases. The students were required to conceive a design that helped a given persona who had a problem (or design according to the specification). In the first phase, the ideation phase, the students produced their conceptual design and a scenario in which the persona could use the application that they were designing. The purpose of writing a scenario was for students to demonstrate (to themselves as well as to the reader) that their designs were feasible. The ideation was done under induced time pressure – a clock counted the time down from 15 minutes. Once the students submitted their initial design and the scenario, they could not modify them. The majority of the students completed the first phase within time.

In the second phase, the documentation of the ideation phase, the students were asked to provide further explanations about their designed applications. The students produced four separate artifacts: (1) for the architecture of the designed application (connectivity and platform of operation); (2) features of the designed application that met the objective of the design and hence solved the problem that the persona was having (e.g. the features that were related to a finance or linguistic application); (3) features of the application that met the needs of the persona according to traits of the persona; (4) a more detailed scenario in which the persona used the application. In the case of students who received the specification, instead of persona, they were asked to provide solutions to the problems that a user of their designed application would have. Thus, we have captured ideation and documentation.

As the study was part of the course, the students were expected to participate in the design and the peer review

activities. The data presented in this paper are from students who gave their consent to participate in our research.

The conceptual design activity for both studies was worth 5% of the total unit assessment. In Study A, the maximum amount of credit allocated by peer review was scaled to 2% of the unit (or 40% of the allocated assessment) and the students reviewed four artifacts from the second phase, the documentation phase, only. The amount of credit allocated by the marker (the main researcher) was scaled to 3% of the unit (or 60% of the allocated assessment). In Study B, the maximum amount of credit allocated by peer review and the marker (the official marker appointed by the Unit Conveners) were each scaled to 2.5% of the unit (or 50% of the allocated assessment) and the students peer reviewed all the artifacts.

For peer review, we anonymized the design artifacts and combined them and created sets of design artifacts. The set contained four artifacts in Study A and five artifacts in Study B. Each student was asked to peer review five sets (Study A) or six sets (Study B) of design artifacts, one set of artifacts for each persona. (For Study A, each student-reviewer received four personas and one specification). The conceptual design artifacts were provided at random. No time limit was set for peer review and students peer reviewed in their own time. The peer review was done according to a rubric. The rubric was designed considering the Revised Bloom's Taxonomy and the UCD methodologies. Hence, both innovation and the needs of personas were considered important aspects for evaluating the design artifacts. Innovative ideas that considered the traits of the given persona attracted the highest marks. The rubric and the personas used in teaching User Centered Conceptual Design are available from the first author [7]. To ensure that the assessment was insightful, the reviewers were required to provide the reasons for their assessments as they allocated marks to each of the artifacts. The students' artifacts were also assessed by an independent marker to allocate the balance of the marks. The students received a report of the assessments and the reasons.

Thus, the design of our studies, which was about the teaching of conceptual design, incorporated a reflective practitioners' perspective [1, 11, 60]. To scaffold the students for reflection, in the first phase, which was reflection-in-action, the students faced a problem for the first time and produced a design in a short time. In the second phase, which was reflection-on-action, the students elaborated how their design might address the needs of the persona (or specification) not only by designing a solution for the problems they faced but also how their needs were met according to the persona's traits. The peer review was designed as reflection-for-action [41], where the students, during review, contemplated on their peers' design and learnt concepts for future design. We have developed these processes and procedures for more than five years [8, 9].

The first author had researched evaluation of the conceptual design artifacts since 2011 [6, 8, 9]. Initially he assessed all design artifacts. He found that the process was time consuming and suitable only for research purposes. In order to have the design artifacts assessed and comments given to students in a timely manner, we investigated the possibility of having students review their peers' design artifacts. Study A was our first attempt. The success of Study A led to Study B.

V. Threats to the Validity of the Experiment and Measures to Overcome These

The internal and external threats to the validity of the study were identified as shown below. Measures were taken to mitigate these threats. We will discuss further some of the threats in the Results section.

A. Internal Threats

The first threat was that the students could show bias or favor their friends' conceptual design [54, 55]. The effect of this threat was minimized by removing the identity of the students, who designed the artifacts, and the reviewers. Thus, all comments the reviewers made or assessments they provided were anonymous. All students were informed that the process was double blind and all identifications such as names and students' IDs were removed and replaced with Participant IDs. Participant IDs were never disclosed to students. The students could not attempt to find out the details of any reviewer including their own details, as they did not know their own Participant ID. We, thus, mitigated this threat. The assessment was 5% of the unit mark, the class size was greater than 200 students, the duration of peer review was about one week and the students did not have access to their own artifacts. Hence, we did not expect that possible speculation by students about the identity of participants would be practicable. Three students accidentally received their own design to review but only one student identified it; we will discuss this in the results and discussion sections. We believe that we mitigated this threat by implementing a double blind process.

The second threat was that the students who are not attentive do not evaluate the design artifacts properly. One of the requirements of the peer review was that every student had to provide a comment for each mark they have assigned. This would lessen the effect of this threat. We found that by asking students to provide reasoning for the marks they gave, most students were vigilant. We noticed that the majority of students demonstrated engagement in the review process and multiple reviews were completed for almost all artifacts. It was a simple process to identify and exclude data supplied by a few students who were not attentive as their peer review times were very short and they allocated marks unreasonably. As expected, this threat was minimized.

The third threat was that the students' English language proficiency might not have been adequate for the review. The design of the experiment required that each artifact be reviewed by a number of other students. It was expected that this process would minimize this threat. As part of their admission to the course in a university in an English speaking country, the students were expected to have proficiency in English language. We asked students for the length of time they have spoken and written in English so that we could pay close attention to reviewers who had written in English for less than one year. We found that their knowledge of English language was adequate for the task.

The fourth threat was that the students were unfamiliar with doing conceptual design, as they were not taught this content in lectures. We mitigated this threat by having a tutorial early in the survey, which provided the basic knowledge of conceptual design. During the second phase of the design activity, we made provisions for the students to document their design in sections (architecture and connectivity, etc.) and elaborate each section separately.

Hence, these measures provided reasonable knowledge and the ability for students to mitigate this threat. The design artifacts indicated that the majority did not have issues with understanding the requirement of the studies.

The fifth threat was that the students were unfamiliar with assessing their peers' design artifacts. During the peer review, we provided a comprehensive rubric to each student, which covered all sections of the design artifacts. However, we took a cautious approach, with fall back strategies, as we will highlight in the discussion section. From the comments the students provided, we found that the students used the rubric to review their peers and increased their own knowledge at the same time. Some students commented positively about the rubric and some students suggested ways to improve the rubric. None of the students commented that they did not have confidence in reviewing their peers.

The sixth threat was that the students would provide solutions for each part that only met the rubric requirements. As the artifacts were separate items, they may be unrelated to each other (i.e. the design is not holistic). One of our pedagogical objectives was to make the student familiar with aspects of UCCD. We would credit the students when they demonstrated that they understood and applied the UCCD principles. By reviewing their peers' design, they have the chance to learn further the UCCD principles and holistic design. However, from reading the design artifacts, we observed that the majority of students provided holistic and thoughtful solutions.

The seventh threat was that the students could not design according to the principles of usability because many students did not know about them. As the marks allocated for the study was only 5% of the unit and the allotted time was two one-hour sessions, hence the learning and teaching was limited. The students were asked to produce a conceptual design and hence it was not at the level of detail for them to consider interface design. However, the students learnt the need for understanding the human element in software engineering. Our questionnaires were aimed at whetting their appetite.

The eighth threat was that both the researchers and the student reviewers could be wrong in assigning marks. To overcome this threat, we relied on the explanations that were provided while the marks were assigned. The peer review marks were the average of the marks given by a number of reviewers for each set of artifacts. We provided a rubric that we developed over a number of years to all who assigned marks. If there was a need, we would arrange for learned colleagues to advise us. We will discuss this threat in detail in the discussion section.

The ninth threat was that when the students were exposed to one persona or specification (control students), they might be at a disadvantage because of the particular traits of the persona or because the student is part of the control group. From the design artifacts and comments left by the students, it appears that the students accepted that in real life they have to design for people whose traits they do not like. From a pedagogical perspective, during peer review, all students were exposed to all personas (and the specification for Study A) and one set of their peers' design artifacts, hence they had equal opportunities to learn.

B. External Threats

A threat that may affect generalizability was that students may think that their reviews were being monitored, hence, that knowledge might influence their behavior. To overcome this threat, the students were required to take part in the peer review activity but the peer reviews were not counted towards the marks they obtained. We will discuss this threat further in the discussion section.

The other external threats relate to the inability to generalize the conclusions of this study due to limited sample size. We have data from two large groups of students. We plan to run this study with other groups from different units of studies at different universities and in different countries.

VI. Results

We present the results of statistical analysis of the evaluations made by peer review, the researchers' review and an official marker; we considered each set of data as a sample data and made the following comparisons: for Study A – peer review versus the researchers review – and for Study B – peer review versus the researchers review and peer review versus the official marker. During the discussion, we will expand on our results. In choosing methods for analysis of our data, we relied on literature (e.g. [13, 52]) which, through studies that included simulation, have shown that for large datasets, parametric statistics can be used for analysis of Likert scales and non-normally distributed data. (According to central limit theorem, when the sample sizes are greater than 25, the means of large number of samples are normally distributed regardless of the distribution of the population.) For comparing different raters, we used Intra-Class Correlation (ICC) as "the only measure that works well when ratings are on a continuous scale" [32, p. 8]. In order to use the correct ICC formula, we used the guiding principles provided by Koo and Li [42] to select model, type and definition. We selected a two-way random effect model (generalizable to number of raters who have similar characteristics), the mean value of multiple raters as the type (mean of numerous raters) and a consistency definition (as the raters' assessment can be close but not exact). For each of the two sample comparisons, we also provided a t-test of the variation in means of the two samples, an F test of the variance of the two samples and Cronbach's alpha for comparison.

We used the Microsoft Excel application (Excel) for the linear analysis and R program for the t-test. For linear analysis, we graphically displayed our results as scatter plot and used the "line of best fit" or trendline. We calculated the slope of the trendline, its intercept with the variable plotted along y-axis and R^2 value (R^2 is the reliability of the trendline – when this number is closer to one, the trendline is a better representation of the relationship between the two variables). We conducted t-test to compare two samples of data and we checked if they are from the same population. The null hypothesis is that the difference between the means of the two samples of measured values under study is not significantly different or the two samples are from the same population. The p-value indicates the confidence level at which the null hypothesis can be accepted or rejected [24]. For ICC we used, "icc" function (R program, "irr v0.84.1" package).

A. Demographics

The participants in both studies were students attending a university in a Western country and doing a second year IT subject which taught them the engineering, design and development of software applications. Table 1 presents the demographics of the students. From Table 1, the majority of the participants were from English speaking backgrounds or were fluent in the English language. Relevant data supplied by the participants who took part in the design and the peer review activities, which were used for peer review assessment, are reported in this paper.

B. Assessment by Peer Review and Student Report

Each student who participated in the design was expected to assess a number of their peers' design artifacts using the rubric (five sets of design artifacts in Study A and six sets of design artifacts in Study B). In checking the time to complete the activity, we noticed that the students were keen to participate in the peer review task and some took longer than the expected time. Table 2 provides statistics for the peer reviews. For Study B, the average time to complete the peer review task (six sets of designs artifacts for six personas) was about 55 minutes (each student spent 549 seconds on average to review one set of design artifacts), each set of design artifacts was peer reviewed 4.2 times with a standard deviation of 1.7. Similar results are reported for Study A. In addition to peer reviews, students spent additional time to read the rubric, the personas, and answer questions about the personas (and the specification) and their design experiences.

As it was required to send the reports to students before the end of the semester, the final reports were emailed to each student showing all peer reviewers' evaluations (the average of the peer reviewers' evaluation was the mark assigned by the peer review scaled appropriately). In Study A, the students received the marker's assessments and their total final mark by email. In Study B, the students received the peer review assessments and comments by email. They were directed to check their final marks when the official marks became available.

C. Assessment by Researchers

The main researcher (the first author) independently evaluated each design artifact thoroughly using the same rubric as the students used in peer reviews. All the assessments of the designs were done independently. The second author randomly selected 20% of the design artifacts and independently marked the artifacts. After marking the artifacts, she checked the assigned marks given by the main researcher as well as the comments entered. She disagreed with 2.5% of the assessments. These were discussed and a consensus was reached. For Study B, the assessment process took 67 hours to complete for 265 students' conceptual design artifacts without including the additional time spent for rechecking the marks and other administrative work. The same process was followed and a similar amount of time was spent to assess the design artifacts in Study A.

D. Comparison of Peer Review and Researchers Evaluation

In both studies, after we finalized all the markings, we statistically evaluated the relationship between the peer reviews and the researchers' evaluation. Fig. 2 and Fig. 3 present a graph of the marks assigned by the peer reviewers and the researchers' evaluation (review) for Study A and Study B, respectively.

From Fig. 3, the fitted linear trend line has a slope of 0.92 and the intercept of 6.14 %. The R^2 value is 0.85. Table 3 shows the statistical evaluation of the marks. Our null hypothesis is that the two samples, the evaluation by peer review and the evaluation by researchers review are from the same population.

TABLE I. DEMOGRAPHICS FOR STUDIES A AND B

Item	Category	Study A		Study B	
		Design Activity	*Peer Review Activity*	*Design Activity*	*Peer Review Activity*
Total		245	221*	265	216
Gender	Male	172	147	170	115
	Female	45	38	58	52
	Unknown	28	36	37	49
Completed both#		186	186	187	187
Did only peer review		-	35	-	29
Fluency in English	Native Speaker	145	128	148	104
	3 years or more	62	47	76	61
	1-3 years	8	8	3	1
	less than 1 year	2	2	1	1
	Unknown	28	36	37	49

Note: * 4 Students did not provide complete dataset but their data for all statistics in this paper were valid.
Number of students who completed both design activity and peer review activity.

TABLE II. PEER REVIEW STATISTICS

Item	Statistic	Study A	Study B
Sets of design artifacts #	Total	245	265
Number of peer reviews for each set of artifacts	Mean	3.8	4.2
	SD	1.9	1.7
Mean Time of peer reviews for each set of artifacts *	Mean (Sec)	655	549
	SD (Sec)	644	552

Notes: SD – Standard Deviation
Sec – Seconds
set of artifacts refers to the number of artifacts that each student produced during their design and were available for peer review (4 artifacts during Study A, 5 artifacts during Study B)
* When the mean time was large, the time was corrected to Mean + 2SD of original calculations. The figures do not include time for reading instructions, Personas and the rubric, and answering questions.

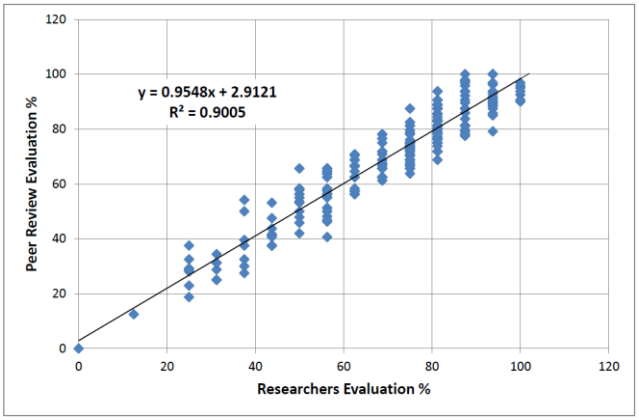

Fig. 2. Peer Review and Researchers Evaluation – Study A

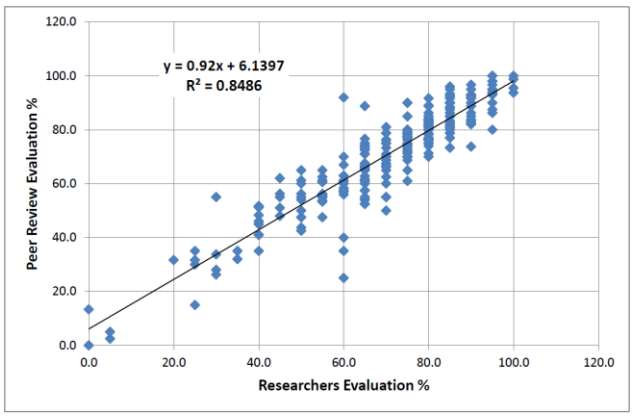

Fig. 3. Peer Review and Researchers Evaluation – Study B

TABLE III. PEER REVIEW VERSESE RESEARCHERS REVIEW AND MARKER – STUDY A AND STUDY B

Item	Statistic	Study A	Study B	
Evaluator		***Researchers***	***Researchers***	***Official Marker^***
Sets of Design artifacts	Total	245	265	259
Evaluator's Evaluation	Mean (%)	69.92	69.98	94.66
	SE (%)	1.23	1.14	1.71
Peer Review Evaluation	Mean (%)	69.67	70.53	70.60
	SE (%)	1.23	1.14	1.16
t-test * (paired t-test between peer review and evaluator's evaluation)	p-value	0.53~	0.23~	0.00@
	t	0.63~	-1.20~	9.85@
	df	244~	264~	258@
Intra-Class Correlation#	ICC value	0.974~	0.959~	-1.32@
	Confidence interval	0.97, 0.98	0.95, 0.97	-1.96, -0.81
F-Test$	F	0.99~	1.00~	2.15@
	df	244, 244	264,264	258, 258
	p-value	0.92~	0.98~	<0.001@
Cronbach's alpha	raw_alpha	0.97~	0.96~	-1.32@
	Confidence interval	0.97, 0.98	0.95, 0.97	-1.83, -0.80

Notes:
^ data drived from offical markes supplied by the Unit Convenor by the formula: official marks = total offical final marks - peer review marks.
* Null Hypothesis: the mean of the two samples are same
$ Null Hypothesis: the variance of the two samples are same
~ the two samples are: peer review and researchers reviews
@ the two samples are: peer review and offical marks
# two-way random effect model, mean value of multiple raters type and consistency definition.
df – degrees of freedom
p-value – probability value

E. Assessment by Staff as a Marker

In Study B, an official marker, one of the Unit Conveners, assessed the conceptual design artifacts. The Unit Convener provided the final total marks (official results). The marks given included the peer review marks. No further details were made available. Hence, we deduced that the difference of the final mark given and the average of the peer review marks was the official marker's assessment for each student's design artifact. Table 3 shows the statistical evaluation of the marks given by the official marker and by the peer reviews for the students that are reported in this paper. The marks for six students were not available in the official marker's results. However, the six students produced valid design artifacts and these were evaluated by their peers. Their data are included in the researchers' statistics but not in the official marker's statistics. The null hypothesis in t-test is that the difference between the means of the two samples, the marks given by the official marker and by the peer review for the students that are reported in this paper is not significantly different.

F. Students reviewed their own design

The sets of design artifacts were collated into sets of design artifacts for review (five sets in Study A and six sets in Study B) and were loaded onto Qualtrics [56]. The students accessed the website at their own convenience and downloaded their sets of artifacts for review. However, there was a small chance that some of the students would receive their own design artifacts to peer review. Table 4 shows that in Study B three students reviewed their own design artifacts.

VII. DISCUSSION OF RESULTS

In this section, we first discuss the design of the studies and then discuss the results, particularly focusing on mitigation of some of the threats to the validity of the study and answering our research question.

In Study A, which was our first attempt at peer review, we did not wish any student to be disadvantaged, and hence we allocated only a portion of the artifacts for peer review; we allocated 40% of the marks by peer review and 60% by the main researcher review which was randomly checked by another researcher. If the peer review proved unsatisfactory, we would moderate the marks. In an unlikely situation where most reviews were erroneous, we would scale the researcher review to 100% and disregard the peer reviews. However, the quality of reviews and dedication of the majority of students as shown in Figure 2 and Table 3 demonstrated that peer review was feasible. Hence, in Study B, we increased the assessment by peer review to 50% and had students review all design artifacts that the students produced. In Study B, we sought assistance from the Unit Conveners to mark the remaining 50% of the assessments to remove possible researcher bias.

We have the null hypothesis that the two samples, the evaluation by peer review and evaluation by the researchers are from the same population or the difference between the two samples' means is not significantly different. According

TABLE IV. PEER REVIEWED OWN DESIGN

PID	Identified*	Self Review	Peer Review #	Researchers Review
1810026	Yes	100.0%	86.3%	80.0%
1810085	No	70.0%	72.5%	65.0%
1810220	No	90.0%	87.5%	80.0%

Notes: PID – Participant ID
* Student indicated that this is her/his own design
# The average marks for Peer Review includes Self Review

to Table 3, our results show that for Study A and Study B, we cannot reject the null hypothesis because the difference between the means of the samples of the review by peers and review by the researchers is not significant. However, Table 3 shows that in Study B, we can reject the null hypothesis that the difference between the means of the two samples, the marks given by the official marker and by the peer review, is significantly different. Hence, the peer review is likely to have produced valid evaluations of the design artifacts.

Table 3 also shows the ICC for the Study A and Study B. For both studies, comparing the researchers' reviews and the peer review, the ICC is within the confidence interval. In study A, the ICC is 0.974 and the confidence interval is a narrow range 0.97 and 0.98. In study B, the ICC is 0.959 and the confidence interval is a narrow range 0.95 and 0.97.

However, the calculations for the official marker produced negative values, indicative of negative correlation between data. Similarly we evaluated Cronbach's alpha which shows that the values for Study A and Study B are within the confidence interval for the researchers and the peer review but the calculations produced negative results comparing the official marker with the peer reviews. From the F-test, the homogeneity of variance for both the peer review and the researchers' reviews, for Study A, $F(244,244) = 0.99$, p-value=0.92 and for Study B, $F(264,264) = 1.00$, p-value=0.98 cannot be rejected. But homogeneity of variance for the official marker and peer review for Study B, $F(258,258)=2.15$, p-value=0.0 can be rejected.

Fig. 2 and Fig. 3 show the slope and the intercept of the trend line and R^2 value (reliability of the trend line). When slope = 1.0, the intercept = 0.0 and $R^2 = 1.0$, the peer review evaluation follows the researchers' evaluations. As shown in Fig. 3, the data for fitted linear trend line: slope = 0.92, intercept = 6.14 % and $R^2 = 0.85$. Table 3 presents the statistics for the two samples of Study B: the peer review evaluation and the researchers' evaluations. The results show that the sample means are close (the means are 70.53 and 69.98 for the peer review and the researchers' evaluations respectively) with small Standard Error (1.14 for both samples). The p-value of 0.23 in the t-test studies shows that our null hypothesis for t-test (the difference between the means of the peer reviews and the researchers' reviews is not significantly different) cannot be rejected. We have similar results for Study A. Due to the size of the samples, we have reasonable confidence in our statistics. Thus, from Fig. 2, Fig 3 and Table 3, we gain confidence that statistically the peer review and the researchers' evaluations for our samples are similar [24]. However, Table 3 shows that the peer reviews and the marker's assessment do not agree. Hence, in the following paragraphs we will discuss the validity of the assessments and conclude accordingly.

We found that it was time consuming to evaluate the conceptual design artifacts and provide meaningful comments. Paying someone, or a team of people, to spend 67 hours (the time spent by the first author, as the main researcher) for marking an assessment task worth 5%, is not viable. The marking budget for the second year undergraduate IT units at the university where this research was conducted was 30 minutes per student in total, giving a total of 132.50 hours for all assessments (minus the final exam) for 265 students. In reading the design artifacts, there was clear evidence that the responses were examples of "reaction on the job" [36, 37, 60]. Hence, proper assessment of conceptual design required time to read the artifacts carefully and decipher innovative ideas within the artifacts. Due to budget constraints, teaching staff at higher education institutions often have limited time to mark and provide feedback for conceptual design artifacts of students from large classes. The National Tertiary Education Union (NTEU) in their submission to the Australian Parliament quotes from a casual academic "We are allocated 22.5 minutes to mark a 1500 word assignment. 22.5 minutes is entirely unrealistic: I have spent up to 15 hours of unpaid time per subject to complete marking" [53, p. 15]. Hence, in large classes, the educational institutions are more likely to rely on methods that lend themselves to automated marking such as quizzes [46].

Even if funding is available, another compounding factor is finding markers with adequate expertise to assess the conceptual design artifacts of hundreds of students. Furthermore, as all students' results must be made available within a fixed deadline, at least before the end of semester, staff have limited time to provide detailed assessment of conceptual designs. As researchers, we were ready to spend the required time and effort and independently check the design artifacts. University staff are assumed to have the most in-depth and accurate knowledge and often their marking is set as the standard in ground truth assessment e.g. [43]. However, teaching staff may not necessarily be better at assessing design creativity and value compared with knowledgeable peers – as conceptual design is like a trait – that can be characterized as an open/broadminded person who widely reads and explores ideas and who is typically more creative than the person who is good at doing tasks procedurally [25].

From Table 1, in Study A 35 students and in Study B 29 students participated in peer review without having participated in the design activity and they produced valid reviews. They were aware that no credit would be given for participating in peer review. Their participation demonstrates that the students were keen and wished to learn. From Table 2, the marks were 5% of the unit marks and yet the time that a number of the students spent for the peer review activity was more than an hour and they did not receive any mark for peer review. They would not have spent the time if they did not enjoy it or did not perceive its value. This is further evidence that the students desired to learn and hence would not have been biased.

In the eighth threat we highlighted that both the researchers and the student reviewers could be wrong in assigning marks. We used statistical measures to evaluate the reliability of the evaluation of the artifacts. For ascertaining the validity of the evaluation of artifacts, we relied on the comments that were provided for each evaluation. As we had no comments from the official marker, so we cannot judge how the marker assessed the conceptual design. Our experiences during our studies were in line with previous researcher's comment: "Most computer science academics lead double lives, leading their research lives and their teaching lives according to different mindsets. In their research lives they read literature, attend conferences, and publish, in a repeating cycle, with the individuals of a research community building upon each other's work. In contrast, the teaching lives of most computer scientists are relatively self-contained, even among active members of the SIGCSE community" [47, p. 146] .

Our rubric was instrumental in producing consistent results. The peer reviewer's comments and their assessments show that most students took the task seriously, reviewed their peers' design artifacts objectively and provided meaningful comments. Selections of the reviewers' comments are included in Table 5 (Study A) and Table 6 (Study B). In both tables, 'Post Thought' refers to the last text box we provided at the completion of the peer reviews with the heading 'any thoughts or comments' for students to enter whatever they wished. Table 5 and Table 6 indicate that through the peer review process, many students learnt from each other and they learnt whether the conceptual design artifacts met the human aspects of software engineering. From the level of maturity demonstrated in the comments made by the students who participated in the peer review, we expect that they will be able to carry the learning benefits into their professional life [58]. Further, even if our results of the peer review assessment are not valid, as pointed out by Kwan and Leung [44], the educational benefits of our methodology, as reflected above, outweighs any small threat caused due to validity in the assessments. However, our results show that peer review using our rubric provided for consistent (reliable) evaluation of the design artifacts. Hence, the second-year undergraduate IT students could review their peer's elaborated design artifacts with the help of a rubric, as the designers divided their design into sections and therefore had clarified their thinking. Our findings are in line with Hafner and Hafner [33].

Table 4 shows the students who accidentally reviewed their own design artifacts and Table 6 provides selected comments the reviewers made. Notably in the list are Participant IDs 1810085 and 1810220 who not only were fair in reviewing themselves but also provided reasoning which demonstrated their learning. Thus, majority of the students would assess the design artifacts fairly even if they identified their friends through their design signatures. Participant ID 1810026 did not see that the peer review could also be a self-review. She/he was the only student who was negative about the study and also the only student who manifested bias in her/his comments.

Due to space limitations we listed a limited number of comments in Table 5 and Table 6. We have also published 37 quotes from students for Study B [10, Table 6 and Table 7]. (In these tables, "Solution Thoughts" refer to comments students made as the last activity during the design and "Post Thought" during peer review.) None of the students expressed any comment that would make us conclude that the students lacked confidence in marking or were not clear about the objective of the study.

Table 5 and Table 6 show that the majority of the students referred to the persona by name and considered the needs of a persona important in their evaluation of the design. This is in line with reflective thinking [60] and UCD methodology [21, 51], the knowledge and skills we sought our students to gain. In answering our research question, our results show that the students can evaluate their peers' conceptual design using our rubric and provide valid and insightful comments for the students' benefit in a timely manner. Our results are consistent with previous research [4, 33, 57] though none of that work involved the use of peer review for learning conceptual design – other than the research that the authors of this paper did [10].

In summary, the students in both Study A and Study B have benefited from exposure to professional tools and techniques. They learned user-centered conceptual design using personas and gained exposure to a reflective concept in their learning activities: While they took part in conceptual design activities, they reflected-in-action as well as reflected-on-action and while they reviewed their peers' conceptual design artifacts they reflected-on-action and reflected-for-action [41, 60]. These activities were to encourage students to act like professional software engineers who "rethink their professional creations during and after the accomplishment of the creation process" [35, p. 161]. Thus, reflective practices assist students to learn at the high levels of the Revised Bloom's Taxonomy [15]. Hence, our methodology would particularly be beneficial not only in a large cohort of undergraduate IT and software engineering students where financial resources are limited but also in teaching subjects where reflective learning is desired [1].

VIII. Conclusion and Future Research

Within prevailing budget constraints, it is not practical for staff to evaluate conceptual design artifacts of large classes of undergraduate students and provide them with detailed feedback of their work [53, 68]. We can conjecture that the teaching and assessment of conceptual design is not commonly undertaken by higher education institutions for undergraduate IT students due to lack of resources. Our research shows that Software Engineering and Information Technology undergraduate students can review their peers, provide timely evaluation of conceptual design artifacts and produce valid assessment results using a comprehensive rubric.

Our data show that students learn while designing, and further enhance their learning during peer reviewing. The use of personas is an effective tool for keeping students focused on a target user while designing. Our methodology teaches UCCD to undergraduate students at the highest rung of cognitive processes as we provide them with an opportunity to reflect-in-action, reflect-on-action and reflect-for-action about their conceptual design artifacts using personas [3, 41, 60].

The novelty of our research is that we introduce to undergraduate Information Technology and Software Engineering students tools and methodologies that are used by professional software engineers and user-centered designers, and expose them to reflective practices that are essential characteristics for professional competence [1, 11, 34, 36, 51] with using little extra resources.

We plan to further this research with different cohorts of students from other higher education institutions in other countries and cultures. We plan to extend our research to include postgraduate students and professional people from various industries [7].

IX. Data Availablity

The rubric and the personas can be found in [7]. Please address your inquiries to the first author.

Acknowledgment

We thank Paul Vincent for proofreading this paper.

TABLE V. PEER REVIEWER'S COMMENTS (STUDY A).

PID	Artifact%	Persona#	Marks *	Peer Reviewer's Comment (Their comments are not edited)
1710139	Unique Feature	Henry	0.75	The unique of the mobile application has Henry put into the accounts accurately and keep his mind set on his spendings on savings and expenses such as rent and bills. Also, he can apply it to a draft as quick budget just in case for safety spending. …
1710081	Scenario	Henry	1	good answer. it provides the reason and the structure on how to using the software.
1710071	Architecture	Henrik	0.50	only mentions one platform for the app to run in but does not elaborate on structure of app
1710205	Architecture	Hank	0.75	This is a very detailed response for the reasoning however, the architecture and connectivity are not explained enough. Is this stand alone? or does this access other apps such as a calendar?
1710116	Unique Feature	Harry	0.75	Mentioned a unique feature and explained how it works in detail.
1710038	Specfc to Persona	Harry	1	Meeting of holistic persona's requirements are adequately identified, with strong reasoning provided exhibiting applications function in relation to Harry's flaws.
1710086	Scenario Writing	Specifica-tion	0.50	Only a preliminary scenario is given with brief details in the most superfluous interactions between user and application. No true detail in how the features of the application work together in a detailed scenario.
1710105	Architecture	Specifica-tion	1	Cloud besed service and different interface on different devices. Easy to access anywhere and anytime
1710049	Post Thought	-	-	Interesting to see the range of responses and solutions. Especially when it comes to understandings of what is reasonable with interactions to other systems.
1710081	Post Thought	-	-	i think those suggestion are really good, compared with my assignment. in this review, i truly understand how to do a design and compare other design. / wish i can do again of my assigement.
1710091	Post Thought	-	-	It was interesting reading some of the responses of fellow students and a worthwhile experience looking at things from a a non-self review context.

Notes: PID - Participant ID
* Maximum 1 mark
# Persona or specification for whom the design was prepared
% Post Thought is the student's thinking at the end of peer reviews.

TABLE VI. PEER REVIEWER'S COMMENTS. (STUDY B)

PID	Artifact%	Persona#	Marks*	Peer Reviewer's Comment (Their comments are not edited)
1810090	Ideation	Chi	1.0	This conceptual design is good at help Chi pronounce English words. Chi also can easily understand words by translating words to her native language.
1810233	Ideation	Thuy	0.5	There should be more elaboration on specifics and examples rather than just stating what the person is looking for, the 0.15 [sic] marks would be awarded if it was elaborated and a better example was given.
1810062	Ideation	Thuy	0.5	What is written is very hard to understand. The translator is a good idea. And it shows that that Thuy's description has been read, mentioning she uses her mobile phone. However, does not delve deeper than that.
1810283	Ideation	Minh	0.75	I still find it quite vague what the videos would actually include. Great idea but not elaborated on for full marks.
1810051	Ideation	Minh	1.0	The design is suitable for someone who is keen on learning new things everyday and wants to have proper time management like Minh.
1810072	Ideation	Minh	1.0	My reason for giving the mark is that Minh can search on the relevant topics on "Informinhtion" Software application after coming from the library. She can use this application n her mobile phone as she don not have computer. She gets all the relevant information from most relevant to least relevant. She can also download the articles from the application as well as the videos so that when her dormitory wi-fi is not working then also she can access the information.
1810019	Ideation	Paul	0.25	The concept design doesn't help Paul achieve professional English skills by reading, writing and speaking.
1810131	Ideation	Paul	1.0	conceptual design is elaborated and a reasonable scenario is provided. From entries, it can be deduced that the application meets the Holistic Persona's requirements, considering her/his traits and is feasible
1810123	Ideation	Peter	0.75	Good ideas on the conceptual design. It will help Peter advance in his studies because the design is built for Peter.
1810198	Ideation	Will	1.0	The designer is EXTREMELY engaged in understanding the traits of Will as the user of the application as well as his needs.
1810083	Ideation	Will	1.0	The conceptual design is extensively elaborated upon and it is clear the requirements are understood as it is reflected in the design. It takes into account the traits of Will, such as him wanting to have clues on how to use the application (the participant included a Help tab).
1810085	Unique Features	Chi	0.5	the design names the list of unique features without more explanations and their performance and processes. there is no detailed information for these features.
1810220	Architecture	Peter	0.75	not much about platform - structure and connectivity good though
1810026	Unique Features	William	1.0	Not much of a peer review if I am reviewing myself
1810233	Post Thought	-	-	This was an interesting assessment, viewing peoples work and ideas which were somewhat very similar was very interesting to view.
1810175	Post Thought	-	-	I would prefer designing an app for Paul and Minh because I feel like I can relate to them more than the others so I feel like I would be able to help them out.

Notes: PID - Participant ID
* Maximum 1 mark. The marks were assigned by the reviewers.
# Persona for whom the design was prepared
% Post Thought is the student's thinking at the end of peer reviews.

REFERENCES

[1] R. S. Adams, J. Turns, and C. J. Atman, "Educating effective engineering designers: the role of reflective practice," Design Studies vol. 24, 3 (2003/05/01/), pp. 275-294, 2003. DOI= http://dx.doi.org/https://doi.org/10.1016/S0142-694X(02)00056-X.

[2] F. Ahmed, M. Fuge, S. Hunter, and S. Miller, "Unpacking Subjective Creativity Ratings: Using Embeddings to Explain and Measure Idea Novelty," In Proceedings of the ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Quebec City, Quebec, Canada, August 26–29 2018), Volume 7: 30th International Conference on Design Theory and Methodology. DOI= http://dx.doi.org/10.1115/DETC2018-85470.

[3] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock, A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives Complete Edition, Longman, Inc. 2001.

[4] H. L. Andrade and Y. Du, "Student perspectives on rubric-referenced assessment," Practical assessment, research & evaluation vol. 10, 3, pp. 1-11, 2005.

[5] M. M. Andreasen, C. T. Hansen, and P. Cash, "Conceptual design," Cham, Switzerland: Springer, 2015.

[6] F. Anvari, "Effectiveness of Persona with Personality on Conceptual Design and Requirements," In Computing Macquarie University, Sydney. 2016.

[7] F. Anvari, "Accademic materials," http://minh-hien.com/academic/ (Accessed: 2021/01/30).

[8] F. Anvari, D. Richards, M. Hitchens, and M. A. Babar, "Effectiveness of Persona with Personality Traits on Conceptual Design," 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (ICSE) Florence, 16-24 May 2015, pp. 263-272. DOI= http://dx.doi.org/10.1109/ICSE.2015.155.

[9] F. Anvari, D. Richards, M. Hitchens, M. A. Babar, H. M. T. Tran, and P. Busch, "An empirical investigation of the influence of persona with personality traits on conceptual design," Journal of Systems and Software vol. 134, Supplement C (2017/12/01/), pp. 324-339, 2017. DOI= http://dx.doi.org/https://doi.org/10.1016/j.jss.2017.09.020.

[10] F. Anvari, D. Richards, M. Hitchens, and H. M. T. Tran, "Teaching User Centered Conceptual Design Using Cross-Cultural Personas and Peer Reviews for a Large Cohort of Students," 2019 Proceedings of the 41st International Conference on Software Engineering: Software Engineering Education and Training Montreal, Quebec, Canada, 25-31 May 2019, IEEE Press, pp. 62-73. DOI= http://dx.doi.org/10.1109/ICSE-SEET.2019.00015.

[11] F. Anvari and H. M. T. Tran, "Holistic Personas and Reflective Concepts for Software Engineers," 2014 Proceedings of the 8th European Conference on IS Management and Evaluation: ECIME2014 Ghent, Belgium, pp. 20-28.

[12] F. Anvari, H. M. T. Tran, D. Richards, and M. Hitchens, "Towards a method for creating personas with knowledge and cognitive process for user centered design of a learning application," 2019 IEEE/ACM 12th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE) Montreal, QC, Canada, IEEE Press, pp. 123-130. DOI= http://dx.doi.org/10.1109/chase.2019.00037.

[13] P. Bacchetti, "Peer review of statistics in medical research: the other problem," British Medical Journal vol. 324, 7348, pp. 1271, 2002.

[14] R. Bailey and Z. Szabo, "Assessing engineering design process knowledge," International Journal of Engineering Education vol. 22, 3, pp. 508, 2007.

[15] P. Beckwith, "Developing higher order thinking in medical education through reflective learning and research," Journal of pedagogic development, 2016.

[16] R. E. Bennett, M. Steffen, M. K. Singley, M. Morley, and D. Jacquemin, "Evaluating an Automatically Scorable, Open-Ended Response Type for Measuring Mathematical Reasoning in Computer-Adaptive Tests," Journal of Educational Measurement vol. 34, 2 (1997/06/01), pp. 162-176, 1997. DOI= http://dx.doi.org/10.1111/j.1745-3984.1997.tb00512.x.

[17] J. Biggs and C. Tang, Teaching for quality learning at university, McGraw-Hill International. 2011.

[18] S. M. Brookhart, The Art and Science of Classroom Assessment. The Missing Part of Pedagogy. ASHE-ERIC Higher Education Report, Volume 27, Number 1, ERIC. 1999.

[19] R. A. Calvo and R. A. Ellis, "Students' Conceptions of Tutor and Automated Feedback in Professional Writing," Journal of Engineering Education vol. 99, 4 (2010/10/01), pp. 427-438, 2010. DOI= http://dx.doi.org/10.1002/j.2168-9830.2010.tb01072.x.

[20] H. Christiaans and R. A. Almendra, "Accessing decision-making in software design," Design Studies vol. 31, 6, pp. 641-662, 2010.

[21] A. Cooper, R. Reimann, D. Cronin, and C. Noessel, About Face: The Essentials of Interaction Design, Wiley Publishing. 2014.

[22] D. Coulson and M. Harvey, "Scaffolding student reflection for experience-based learning: a framework," Teaching in Higher Education vol. 18, 4 (2013/05/01), pp. 401-413, 2013. DOI= http://dx.doi.org/10.1080/13562517.2012.752726.

[23] T. Dybå, N. Maiden, and R. Glass, "The Reflective Software Engineer: Reflective Practice," IEEE Software vol. 31, 4, pp. 32-36, 2014. DOI= http://dx.doi.org/10.1109/MS.2014.97.

[24] A. Field, J. Miles, and Z. Field, Discovering Statistics Using R, SAGE Publications Ltd., London, UK. 2012.

[25] B. W. Field, "Visualization, intuition, and mathematics metrics as predictors of undergraduate engineering design performance," Journal of Mechanical Design vol. 129, 7, pp. 735-743, 2007.

[26] S. Finger, D. Gelman, A. Fay, and M. Szczerban, "Supporting collaborative learning in engineering design," 2005 Proceedings of the Ninth International Conference on Computer Supported Cooperative Work in Design, 2005., 24-26 May 2005, 2, pp. 990-995 Vol. 992. DOI= http://dx.doi.org/10.1109/CSCWD.2005.194322.

[27] B. Fitzgerald, K.-J. Stol, R. O'sullivan, and D. O'brien, "Scaling agile methods to regulated environments: an industry case study," In Proceedings of the 2013 35th International Conference on Software Engineering (San Francisco, CA, USA2013), IEEE Press, pp. 863-872. DOI= http://dx.doi.org/10.1109/ICSE.2013.6606635.

[28] D. Garlan, D. P. Gluch, and J. E. Tomayko, "Agents of change: educating software engineering leaders," Computer vol. 30, 11, pp. 59-65, 1997. DOI= http://dx.doi.org/10.1109/2.634865.

[29] V. Garousi, "Applying Peer Reviews in Software Engineering Education: An Experiment and Lessons Learned," IEEE Transactions on Education vol. 53, 2, pp. 182-193, 2010. DOI= http://dx.doi.org/10.1109/TE.2008.2010994.

[30] J. S. Gero and U. Kannengiesser, "The situated function–behaviour–structure framework," Design Studies vol. 25, 4, pp. 373-391, 2004.

[31] J. S. Gero and U. Kannengiesser, "A function–behavior–structure ontology of processes," Artificial Intelligence for Engineering Design, Analysis and Manufacturing vol. 21, 4, pp. 379-391, 2007. DOI= http://dx.doi.org/10.1017/S0890060407000340.

[32] M. Graham, A. Milanowski, and J. Miller, "Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings," Online Submission, 2012.

[33] J. Hafner and P. Hafner, "Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer - group rating," International Journal of Science Education vol. 25, 12 (2003/12/01), pp. 1509-1528, 2003. DOI= http://dx.doi.org/10.1080/0950069022000038268.

[34] J. Harlim, I. Belski, C. Lemckert, G. Jenkins, and S. Lang-Lemckert, "Educating a reflective engineer: learning from engineering experts," AAEE2013: Work Integrated Learning-Applying Theory to Practice in Engineering Education, pp. 1-9, 2013.

[35] O. Hazzan, "The reflective practitioner perspective in software engineering education," Journal of Systems and Software vol. 63, 3, pp. 161-171, 2002. DOI= http://dx.doi.org/https://doi.org/10.1016/S0164-1212(02)00012-2.

[36] O. Hazzan and J. Tomayko, "The reflective practitioner perspective in eXtreme Programming," In Extreme Programming and Agile Methods-XP/Agile Universe 2003 Springer, 2003. pp. 51-61.

[37] I. Horváth, "Conceptual design: inside and outside," 2000 Proceedings of the 2nd International Seminar and Workshop on Engineering Design in Integrated Product, Citeseer, pp. 63-72.

[38] T. J. Howard, S. J. Culley, and E. Dekoninck, "Describing the creative design process by the integration of engineering design and cognitive psychology literature," Design Studies vol. 29, 2, pp. 160-180, 2008. DOI= http://dx.doi.org/10.1016/j.destud.2008.01.001.

[39] W. Hsu and B. Liu, "Conceptual design: issues and challenges," Computer-Aided Design vol. 32, 14 (2000/12/01/), pp. 849-850, 2000. DOI= http://dx.doi.org/https://doi.org/10.1016/S0010-4485(00)00074-9.

[40] M. C. Jones, I. R. Floyd, and M. B. Twidale, "Teaching design with personas," 2008 Proceedings of the Human Computer Interaction in Education Rome, Italy.

[41] J. P. Killion and G. R. Todnem, "A Process For Personal Theory Building. (Cover story)," Educational Leadership vol. 48, 6, pp. 14, 1991.

[42] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," Journal of Chiropractic Medicine vol. 15, 2 (2016/06/01/), pp. 155-163, 2016. DOI= http://dx.doi.org/https://doi.org/10.1016/j.jcm.2016.02.012.

[43] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer, "Peer and self assessment in massive online classes," ACM Trans. Comput.-Hum. Interact. vol. 20, 6, pp. 1-31, 2013. DOI= http://dx.doi.org/10.1145/2505057.

[44] K. P. Kwan and R. W. Leung, "Tutor Versus Peer Group Assessment of Student Performance in a Simulation Training Exercise," Assessment & Evaluation in Higher Education vol. 21, 3 (1996/09/01), pp. 205-214, 1996. DOI= http://dx.doi.org/10.1080/0260293960210301.

[45] H. Lipson and M. Shpitalni, "Conceptual design and analysis by sketching," Artificial Intelligence for Engineering Design, Analysis and Manufacturing vol. 14, 5, pp. 391-401, 2000. DOI= http://dx.doi.org/10.1017/S0890060400145044.

[46] R. Lister, "On blooming first year programming, and its blooming assessment," 2000 Australasian Conference on Computing Education Melbourne, Australia, ACM, pp. 158-162. DOI= http://dx.doi.org/10.1145/359369.359393.

[47] R. Lister, A. Berglund, T. Clear, J. Bergin, K. Garvin-Doxas, B. Hanks, L. Hitchner, A. Luxton-Reilly, K. Sanders, and C. Schulte, "Research perspectives on the objects-early debate," ACM SIGCSE Bulletin vol. 38, 4, pp. 146-165, 2006.

[48] F. Long, "Real or imaginary: The effectiveness of using personas in product design," 2009 Proceedings of the Irish Ergonomics Society Annual Conference, pp. 1-10.

[49] J. Lu and D. Churchill, "The effect of social interaction on learning engagement in a social networking environment," Interactive Learning Environments vol. 22, 4 (2014/07/04), pp. 401-417, 2014. DOI= http://dx.doi.org/10.1080/10494820.2012.680966.

[50] A. F. Mckenna, J. E. Colgate, S. H. Carr, and G. B. Olson, "IDEA: formalizing the foundation for an engineering design education," International Journal of Engineering Education vol. 22, 3, pp. 671, 2007.

[51] D. A. Norman, "Cognitive engineering," User centered system design, pp. 31-61, 1986.

[52] G. Norman, "Likert scales, levels of measurement and the “laws” of statistics," Advances in Health Sciences Education vol. 15, 5, pp. 625-632, 2010. DOI= http://dx.doi.org/10.1007/s10459-010-9222-y.

[53] NTEU, "Unlawful underpayment of employees’ remuneration (Wage Theft) Submission to the Senate Economics References Committee," https://www.nteu.org.au/library/download/id/10218 (Accessed: 31/01/2021).

[54] J. Pearce, R. Mulder, and C. Baik, Involving students in peer review: Case studies and practical strategies for university teaching, Centre for the Study of Higher Education, University of Melbourne. 2009.

[55] C. H. Peterson and N. A. Peterson, "Impact of Peer Evaluation Confidentiality on Student Marks," International Journal for the Scholarship of Teaching and Learning vol. 5 n2 Article 13, 2011.

[56] Qualtrics, "qualtrics.com," http://qualtrics.com (Accessed: 30/06/2020).

[57] D. Richards, "Designing Project-Based Courses with a Focus on Group Formation and Assessment," Trans. Comput. Educ. vol. 9, 1, pp. 1-40, 2009. DOI= http://dx.doi.org/10.1145/1513593.1513595.

[58] J. Salminen, S.-G. Jung, J. M. Santos, S. Chowdhury, and B. J. Jansen, "The Effect of Experience on Persona Perceptions," 2020 Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Honolulu, HI, USA, Association for Computing Machinery, pp. 1–9. DOI= http://dx.doi.org/10.1145/3334480.3382786.

[59] P. Schlesinger, "Creativity: from discourse to doctrine?," Screen vol. 48, 3, pp. 377-387, 2007.

[60] D. A. Schön, The reflective practitioner: How professionals think in action, Basic books. 1983.

[61] J. J. Shah, S. V. Kulkarni, and N. Vargas-Hernandez, "Evaluation of Idea Generation Methods for Conceptual Design: Effectiveness Metrics and Design of Experiments," Journal of Mechanical Design vol. 122, 4, pp. 377-384, 2000. DOI= http://dx.doi.org/10.1115/1.1315592.

[62] J. J. Shah, S. M. Smith, N. Vargas-Hernandez, D. R. Gerkens, and M. Wulan, "Empirical studies of design ideation: Alignment of design experiments with lab experiments," 2003 ASME 15th International Conference on Design Theory and Methodology, pp. 1-10. DOI= http://dx.doi.org/10.1115/DETC2003/DTM-48679.

[63] G. Sun, S. Yao, and J. A. Carretero, "Evaluating cognitive efficiency by measuring information contained in designers’ cognitive processes," 2013 ASME 25th International Conference on Design Theory and Methodology Portland, Oregon, USA, 2013, ASME. DOI= http://dx.doi.org/10.1115/DETC2013-13628.

[64] G. Sun, S. Yao, and J. A. Carretero, "Comparing Cognitive Efficiency of Experienced and Inexperienced Designers in Conceptual Design Processes," Journal of Cognitive Engineering and Decision Making vol. 8, 4 (2014/12/01), pp. 330-351, 2014. DOI= http://dx.doi.org/10.1177/1555343414540172.

[65] H. M. T. Tran, F. Anvari, and D. Richards, "Holistic Personas for Designers of a Context-Aware Accounting Information Systems e-Learning Application," EAI Endorsed Trans. Context-aware Syst. & Appl. vol. 4, pp. e4, 2018. DOI= http://dx.doi.org/10.4108/eai.18-6-2018.154822.

[66] N. M. C. Valentim, W. Silva, and T. Conte, "The students' perspectives on applying design thinking for the design of mobile applications," 2017 39th International Conference on Software Engineering: Software Engineering and Education Track Buenos Aires, Argentina, IEEE Press, pp. 77-86. DOI= http://dx.doi.org/10.1109/icse-seet.2017.10.

[67] B. Warin, C. Kolski, and C. Toffolon, "Living persona technique applied to HCI education," 2018 2018 IEEE Global Engineering Education Conference (EDUCON), 17-20 April 2018, pp. 51-59. DOI= http://dx.doi.org/10.1109/EDUCON.2018.8363208.

[68] J. Watts and N. Robertson, "Burnout in university teaching staff: a systematic literature review," Educational Research vol. 53, 1, pp. 33-50, 2011.

[69] P.-H. Wu, G.-J. Hwang, M. Milrad, H.-R. Ke, and Y.-M. Huang, "An innovative concept map approach for improving students' learning performance with an instant feedback mechanism," British Journal of Educational Technology vol. 43, 2 (2012/03/01), pp. 217-232, 2012. DOI= http://dx.doi.org/10.1111/j.1467-8535.2010.01167.x.

[70] L. Zhu, Y. He, and Y. Tang, "Using Wikipedia to Construct Product Conceptual Space," 2018 11th International Symposium on Computational Intelligence and Design (ISCID), 8-9 Dec. 2018, 02, pp. 103-106. DOI= http://dx.doi.org/10.1109/ISCID.2018.10124.