SOTL IN THE SOUTH

A journal dedicated to the scholarship of teaching and learning in the 'global south'

Volume 3, Issue 1 April 2019

Pages: 6-25

Developing a rubric to assess critical thinking in a multidisciplinary context in higher education

Sadia Muzaffar Bhutta	Institute for Educational Development*	sadia.bhutta@aku.edu
Sahreen Chauhan	Network of Teaching and Learning, Office of the Provost*	sahreen.chauhan@aku.edu
Syeda Kauser Ali	Department for Educational Development, Faculty of Health Sciences*	syeda.ali@aku.edu
Raisa Gul	School of Nursing and Midwifery (Visiting Faculty)* Faculty of Nursing and Midwifery, Shifa Tameer-e-Millat University	dean.fnm@stmu.edu.pk
Shanaz Cassum	School of Nursing and Midwifery*	shanaz.cassum@aku.edu
Tashmin Khamis	Networks of Quality, Teaching and Learning, Office of the Provost*	tashmin.khamis@aku.edu

*Aga Khan University

ABSTRACT

Critical thinking (CT) is a generic attribute that is greatly valued across academic disciplines in higher education, and around the globe. It is also defined as one of the graduate attributes of higher education for the sample private university where this research was conducted, as it is perceived that CT helps the graduate to become 'engaged citizens' in the twenty-first century. Despite the well-documented importance of CT, its assessment remains a challenge. This study addresses this challenge through the systematic development and field-testing of a rubric for assessing critical thinking in a multidisciplinary context in higher education. A multidisciplinary group of faculty (i.e. education, nursing, medicine) from the sample university partnered with a policy research group in Canada to translate this plan into action. The development of the assessment tool followed a multi-step process including: (i) *identification of the main elements of CT*; (ii) *choice of a rubric format*; (iii) *adaptation of the currently available relevant rubrics*; and, (iv) *field testing and establishment of the reliability of the rubric*. The process resulted in the development of a holistic template, the Assessment of Critical Thinking (ACT) rubric. Two versions of the rubric have been field tested on a sample (n=59) of students drawn from different entities of the sample university. The data collected was subjected to psychometric analysis which yielded satisfactory results. This was a modest attempt

to develop an assessment tool to guide multidisciplinary faculty members in teaching and assessing CT by assisting them to make decisions about the level of their students' CT skills through a combination of numerical scores and qualitative description. It may also empower them to make self-initiated, conscious efforts to improve their classroom practice with reference to CT. The ACT rubric provides an anchoring point to start working on the daunting yet doable task of developing and fine-tuning both the assessment measures of CT and interventions to promote CT based on the assessment findings. Future research may not only provide robust evidence of the reliability and validity of the ACT rubric for a larger and varied sample but also help in making informed decisions to enhance teaching and learning of CT across entities of the sample University.

Introduction

Critical thinking (CT) is a generic attribute that is greatly valued across the academic disciplines. It is a high priority on both employability and citizenship agendas (Lowden, Hall, Elliot & Lewin 2011) and has been described as one of the most widely discussed concepts in education and educational reform (Atkinson 1997). It is also recognised as one of the most important skills in contemporary higher education that contributes to academic and career success (Liu, Frankel & Roohr 2014). Moreover, the concept of CT has acquired a prominent place in discussions about, and policies regarding, the concepts of generic education skills and university graduate attributes. These discussions and policies have led to the development and implementation of stand-alone courses on critical thinking in higher education institutions (Moore 2011). Furthermore, teaching of critical thinking has led to the emergence of a variety of assessments to test students' acquisition of critical thinking. Despite the well-documented importance of CT around the globe, its assessment in the sample university remains a challenge, which this study aims to address. This project proposes the development of an indigenous framework for the sample university to measure generic learning outcomes at the programme level with a particular focus on CT. In particular, the purpose of this study is to develop and field-test a rubric designed to assess the critical thinking skills of students across graduate programmes in the given sample university.

As part of this project, faculty members from different departments of the sample university (i.e. education, medicine and nursing) along with the university's Networks of Quality, Teaching and Learning collaborated with a leading higher education policy research group in Canada. This policy research group has extensive experience in working with higher education institutes in Canada to develop a range of assessment tools to measure generic learning outcomes, including CT (e.g. Nunley, Bers & Manning 2011; George Brown College 2015; Kapelus, Miyagi & Scovill 2017; Pichette & Watkins 2018; McKeown & Biss 2018). This symbiotic partnership benefitted the sample university which was able to learn from and build on the experiences of the research group, and contributed to their repertoire of assessment tools for higher cognitive abilities in the context of a developing country. This was a modest attempt to develop an assessment tool which would guide multidisciplinary faculty members in teaching and assessing CT by assisting them to make decisions about the level of their students' CT skills through a combination of numerical scores and qualitative descriptions. It may also

assist faculty members and others to make self-initiated conscious efforts to improve classroom practice with reference to CT. The tool could also be shared with other higher education institutes across the country and could be utilised broadly for teaching and assessment. Furthermore, such a CT assessment tool could be used by faculty members who would like to investigate their own practices related to teaching and learning of CT, which would lead to further improvement in the curriculum, pedagogical strategies as well as student learning outcomes.

Literature review

The 'what' and 'why' of critical thinking

Researchers have estimated the half-life of information to be between five and ten years (Crow 1989; Nelson 1989; Robinson 2011). This "refers to the time it takes for one half of the knowledge in a given field to become obsolete" (Crow 1989: 9). Arguably, in a situation where half the information acquired in pursuing a degree will be obsolete within a decade, 'what' a student knows becomes less significant than the life-long skills in evaluating sources of information that graduates bring to the workplace (Robinson 2011). Where students lack CT skills, they run the risk of "having all of the answers but still not knowing what the answers mean" (Halpern 1998: 450).

CT is one of the key goals of higher education because teaching students to think critically is an intrinsic good, which may provide students with a more analytical outlook (Moore 2011). A strong capability to think critically would not only help to engage with the knowledge in academia but is also imperative in becoming an 'engaged citizen' in the twenty-first century as the world is becoming increasingly technical and complex (Barnett 1997; Halpern 2010; Cottrell 2011). In order to survive in this era, CT is one of the essential metacognitive skills to perform on-the-job multifaceted tasks by manipulating abstract and multifarious ideas, obtaining new information competently, and remaining flexible enough to identify the need for continuing change for lifelong learning (Pithers & Soden 2000; Halpern 2010).

The concept of CT has become more and more a part of higher educational discourses in the last two decades. These discourses have led to controversies about the concept itself. While there is a common consensus on the importance of teaching CT in higher education there is a good deal of disagreement and dispute on many aspects including (i) the definition of CT, and (ii) CT as a generic or subject-specific attribute (McPeck 1981; Ennis 1985; Jones 2004; Davies 2006; Moore 2011, 2013). A central issue is the question of what critical thinking is, exactly. Numerous definitions are suggested and debated ranging from a rather generic view-point of CT as "reflective and reasonable thinking that is focused on deciding what to believe or do" (Ennis 1985: 45) to specific definitions for disciplines such as history, physics, education, law and medicine (Jones 2004). This debate has not only affected curriculum planning but also assessment of CT. Despite a wide array of definitions and citations of a variety of aspects used to define CT, there are some constructs which are found to be common across a number of conceptions. Some of these include: inquiring, problem-solving, argument analysis and construction, uncovering and evaluating assumptions, justification, interpretation, and synthesis (Liu, Frankel & Roohr 2014; James, Hughes & Cappa 2010; Paul & Elder 2009; Bloom and Krathwohl 1956).

The other major controversy that appears in the literature about the teaching and assessment of CT is its nature: whether critical thinking is best taught generally or in specific contexts of a discipline such

as history, medicines, law, and education (Moore 2011; Liu, Frankel & Roohr 2014). The 'generalists' have viewed CT as a universal, general skill or ability that can be applied to any discipline (Ennis 1985, 1990, 1997; Davies 2006). On the other hand, the 'specifists' have conceptualised CT as a skill specific to a context and discipline (McPeck 1981, 1990, 1992; Moore 2004, 2011). The debate on this long-standing controversy is important to understanding the nature and patterns of thinking; however, it is not mandatory to take one position against the other. There is plenty of support for infusion of the two approaches (Swartz & Perkins 1989; Swartz & Parks 1994; Reed & Kromrey 2001, Davies 2006; Cassum, Profetto-McGrath, Gul, Ashraf & Syeda 2013). Since the study presented in this paper aims to develop a common CT assessment tool for diverse disciplines, the authors embrace the idea of educating students for 'multifaceted critical thinking' and the concept of CT that resonates with that of the proponents of 'infusion' (Davies 2006).

Measurement of critical thinking: existing tools

A plethora of assessment tools has been developed and validated around the globe to gauge development of CT (Watson & Glaser 1980; Ennis & Weir 1985; Ennis, Millman & Tomko 1985; Facione 1990; Facione & Facione 1992; Halpern 2010; Council for Aid to Education 2013). This includes a number of assessment methods ranging from standardised tests such as multiple choice questions (MCQs)(Spicer & Hanks 1995) to more performance-oriented, authentic and open-ended techniques such as interviews, case studies or portfolios (Horsburgh 1999). Some of the examples of widely used assessment tools include: the California Critical Thinking Disposition Inventory (CCTDI – Facione & Facione 1992); the California Critical Thinking Skills Test (CCTST – Facione 1990), the Watson-Glaser Critical Thinking Appraisal (WGCTA – Watson & Glaser 1980); the Ennis-Weir Critical Thinking Essay Test (Ennis & Weir 1985); the Collegiate Learning Assessment (CLA+ – Council For Aid to Education 2013); the Collegiate Assessment of Academic Proficiency (CAAP – Collegiate Assessment of Academic Proficiency 2012) and the Halpern Critical Thinking Assessment (HCTA - Halpern 2010). These widely used assessments tools reflect and overlap in a number of key themes related to CT such as logical reasoning, analysis, argumentation and evaluation. Moreover, the audience for these tests varies from junior and senior school students to university students and adults.

A number of rubrics have been developed and validated to assess CT skills and dispositions. These rubrics have been applied to classroom activities, oral presentations and written assignments. Arguably, the use of a scoring rubric is not based on a specific subject or level of education but is dependent upon the purpose of the assessment and the definition of CT. Rubrics may be analytical, detailing the specific tasks of an assignment and describing the performance criteria for each task, or may be holistic in nature, for assessment of a broader category of tasks (Moskal 2000). Hence, rubrics may be specific to a test used to assess relevant content or may be of a general nature to assess a competence that cuts across disciplines.

Rubrics have been criticised for being task-specific and having a lack of focus on essential skill(s) that determine mastery. Lengthy descriptors also tend to dissuade teachers from using them and a balance has to be maintained between them being too short or too lengthy, with a focus on usability for assessment and learning (Popham 1997). Despite this critique, rubrics are a frequently used assessment measure in higher education, both for cognitive and generic skills, to maximise objective scoring (Reddy & Andrade 2010; Jackson 2014; Velasco-Martinez & Tojar-Hurtado 2018). In other words, rubrics provide criteria to an individual assessor for consistent marking and enhance inter-rater

consistency in case of multiple assessors (Popham 1997). Rubrics consist of evaluative criteria, descriptions of levels of performance, and the numerical score assigned to the different levels of performance. Given the nature of rubrics, they allow assessors to make decisions about the level of cognitive or generic skills through a combination of numerical scores and qualitative descriptions (Bhutta 2006). This combination of quantitative-qualitative criteria not only facilitates scoring by specifying the criteria to be attained by the students for a particular score but also provides feedback to the students concerning how to improve their performance. To use rubrics effectively, it is important to share these tools with students at the start of the semester, so they can also use them for self- and peer-assessments and to enhance their learning (Dawson 2017: Hafner & Hafner 2003).

A literature search revealed some of the commonly used rubrics to assess CT across disciplines and levels of education. Some of these examples include: the Holistic Critical Thinking Scoring Rubric (Facione & Facione 1994); the Paul-Elder Critical Thinking Rubric (Paul & Elder 2009); the Valid Assessment of Learning in Undergraduate Education (VALUE – Association of American Colleges and Universities 2007); the Critical Legal Thinking rubric (CLT – James, Hughes & Cappa 2010); the Assessment Rubric for Critical Thinking (ARC – St. Petersburg College 2008) and the Critical Thinking Assessment Rubric (CTA – George Brown College 2015).

Critical analysis of these rubrics highlights that different response categories have been defined to gauge the different levels of performance. Some of the examples include: 'inadequate to exemplary' in the CT assessment rubric (George Brown College 2015); 'benchmark to capstone' in the VALUE rubric (Association of American Colleges and Universities 2009); and 'not present to exemplary' in the ARC (St. Petersburg College 2008). Any of these response categories can be aligned with the need of a particular undertaking (e.g. research, teaching) if they are defined clearly. Additionally, the most frequently identified characteristics or criteria of learning related to CT highlighted in these rubrics are: communication, analysis, evaluation, synthesis and reflection. Evidently, critical thinking is a multidimensional construct and, accordingly, the assessment of CT is necessarily multidimensional (Cassum *et al.* 2013). These common constructs and scoring patterns highlighted in the literature guided the development of the rubric discussed in this paper.

Aim of the project

This project involved the development of an indigenous framework for the sample university in order to measure generic learning outcomes at the programme level with a particular focus on CT. In particular, the study aimed to develop a multidisciplinary rubric to assess the CT skills of students based on a written assignment, and to establish its psychometric properties. The sample higher education institute is a multi-site university with a major focus on three major disciplines (i.e. education, nursing and medicine). Therefore, a multidisciplinary team worked together on this project to represent all three departments. The researchers' own experiences of developing, implementing and evaluating stand-alone CT courses, as well as integrating aspects of CT in curriculum (e.g. argumentation as a strategy to teach science education and problem-based learning), contributed to the process. Some of the team members have also published in the area of teaching and integrating CT (Cassum *et al.* 2013; Bhutta & Anwar, forthcoming).

Critical Thinking project: an analytical account and discussion

The development of an assessment tool is a multi-step process where the researchers go back and forth between steps during the process to develop a comprehensive instrument. Development of the 'Assessment of Critical Thinking' (ACT) rubric consisted of four interlinked phases: (i) identifying main elements of CT; (ii) deciding on the CT rubric format; (iii) adapting CT assessment rubric; as well as, (iv) field-testing and establishing reliability. A summary of these iterative steps is presented as a conceptual framework in Figure 1 (Bhutta 2006).

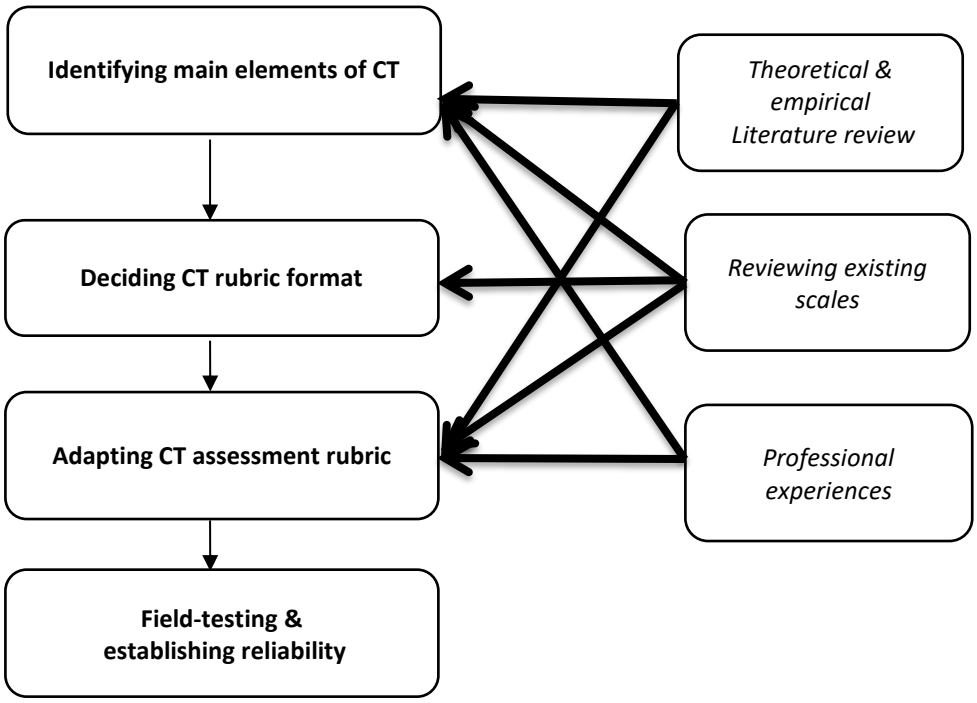

Figure 1: Steps in developing the ACT rubric (adapted from Bhutta 2006).

Identifying main elements of CT

The first step in this study was describing, mapping and specifying the indicators for CT. This was an essential step for ensuring accuracy of what the rubric is intended to measure (Robson & McCartan 2016; Fraenkel, Wallen & Hyun 2011). In order to identify the main elements of CT, literature was reviewed at two levels: theoretical literature (e.g. James, Hughes & Cappa 2010; Paul & Elder 2009; Facione & Facione 1994; Bloom and Krathwohl 1956), and existing measures of CT (e.g. George Brown College 2015; Liu, Frankel & Roohr 2014; Paul & Elder 2009; St. Petersburg College 2008; Association of American Colleges and University 2007). The former enabled an understanding of the philosophical underpinnings of CT while the latter helped to identify the patterns various researchers have used to translate these theoretical constructs into measurable indicators. As mentioned earlier, the researchers' own experiences of developing, implementing and evaluating CT courses and publishing in this area contributed to the process. A blend of information gained from these three resources (i.e. theoretical literature, existing scales and professional experience) helped to draw some main themes which encompass various aspects of CT. Some of the examples revealed in the literature can be classified into key skills (e.g. information seeking, analysis, logical reasoning, evaluation, synthesis, inference, explanation) and dispositional components (e.g. inquisitiveness, open-mindedness) of CT.

Based on the aim of the study and numerous rounds of informal discussion, within the team and in formal meetings with the higher education policy research group in Canada, the researchers decided to delimit these themes, as the main focus of the study was 'CT skills'. The overall outcome of this step was the development of a CT measurement framework, which comprised five broader indicators: (i) communication (i.e. comprehension of given situation/problem); (ii) analysis (i.e. compare and contrast the possible solutions); (iii) evaluation (i.e. select and defend the best solution); (iv) synthesis (i.e. suggest ways to improve the selected solution); and, (v) reflection (i.e. reflection on one's own thinking). After the five main areas were identified, the next step was to decide on the design of an assessment tool which could best represent these major indicators.

Deciding on the CT rubric format

A number of assessment tools have been developed and validated around the globe to gauge development of CT (Watson & Glaser 1980; Ennis & Weir 1985; Facione & Facione 1992; Halpern 2010). The format of these assessment tools ranges from standardised tests such as MCQs to performance-oriented and open-ended techniques such as interviews, case studies and portfolios. The marking of standardised tests is usually considered straightforward as it is relatively easy to make unbiased decisions about respondents' performance. On the other hand, higher-level learning skills which are the focus of the ACT rubric (e.g. analysis, synthesis, evaluation) do not readily lend themselves to objective examination (Rochford & Borchert 2011). Arguably, open-ended techniques such as case studies and portfolios may call for objective criteria for consistent assessment across markers and over time (Gearhart & Wolf 1997).

A scan of the nature of assignments given to students at the sample university highlighted that the essay-type academic paper was a common thread across all three departments. Furthermore, discussion within the research group revealed that faculty members at the university are acquainted and comfortable with the use of rubrics to assess essay-type assignments. Interestingly, we also came to know that many self-developed rubrics are being used by faculties in their own disciplines. Thus, in order to enhance the probability of engagement on the part of potential users, it was imperative to align the new initiative (i.e. ACT rubric) with existing practice, and devise a common, contextually grounded rubric that measures actual CT constructs applicable for use in all three disciplines at the sample university.

Keeping in mind the current assessment setup, the research team decided to use scenarios as a trigger with guiding questions that encompass the five major CT constructs identified for this study. Respondents were expected to organise their discussion around the problem highlighted in the text using questions as guidelines. In order to assess their responses, a rubric was considered the appropriate option as this helps to articulate the standards expected of open-ended tasks, including case studies. It is important to note that the scenario may have focused on the content of a particular discipline, but the rubric provided generic criteria to score students' responses in order to gauge their level of CT.

As stated previously, the literature identifies multiple ways to define levels of performance for CT through rubrics. Some of these include: 'inadequate to exemplary' (George Brown College 2015), 'benchmark to capstone' (Association of American Colleges and Universities 2009), 'not present to exemplary' (St. Petersburg College 2008). These examples provided a guideline for devising a

contextually relevant yet expressive format for the rubric developed as part of this study. As a result of multiple rounds of discussions, the team was inclined to use more expressive qualitative categories which ranged from 'not present (0)' to 'exemplary (3)' with two categories in-between (emerging=1; and developing=2), with descriptors anchoring each category.

Adapting CT assessment rubric

In order to develop the rubric, the researchers were faced with two options: (i) to generate item pools based on various existing tools and theoretical literature, or (ii) to adapt a relevant existing measure (Robson & McCartan 2016; Fraenkel, Wallen & Hyun 2011; Punch 2000). The team opted for the latter as the literature search yielded a compendium of rubrics which had been used to measure CT across contexts, levels of education (undergraduate to graduate) and disciplines (e.g. law, medicine, education, nursing). Some of these measures were found to be relevant to the needs of the sample university (e.g. St. George Brown College 2015, Association of American Colleges and Universities 2009, James, Hughes & Cappa 2010; Paul & Elder 2009, St. Petersburg College 2008). After multiple discussions, the team reached a consensus in favour of using the Assessment Rubric for Critical Thinking, or ARC (St. Petersburg College 2008), as a template to modify and develop into a relevant tool for assessing CT of learners for all three disciplines represented at the sample university. The reason for selecting ARC was twofold: (i) scenarios were used as a prompt with guiding questions that provided a framework for thinking; (ii) a pre-designed rubric associated with these scenarios encompassed the major CT indicators defined for the study. This combination provided a pathway for initiating a CT project in which the data collection was independent of regular assignments yet followed a pattern with which respondents were acquainted (i.e. essay-type assignments with guiding questions). It was anticipated that once the CT rubric was developed and the psychometric properties established, it could be implemented to assess the critical thinking of learners across different disciplines.

Scenario review: First of all, a scenario applicable to students of all disciplines was identified, such as types of sampling in a research context. Since the scoring on the rubric depends on the scenario it was imperative to get the views of the potential respondents on the text of the scenario and related questions (Robson & McCartan 2016; Fraenkel, Wallen & Hyun 2011). Selected scenarios along with the guiding questions were sent to higher education students who were available and who agreed to participate in the study. The selected scenarios and guiding questions were sent to five graduates. They completed the task and engaged in discussions with the researchers on various aspects of the scenario such as clarity, content and time required for completion. There was general consensus among the respondents that the scenarios were interesting and that the guiding questions helped to trigger thinking. An example of the evaluation scenario and the guiding questions is included in Appendix A. Each question in the scenario represents a distinct construct of CT identified in the original rubric, namely, communication (Q1), analysis (Q2), problem solving (Q3 – but subsequently discarded, as described below), evaluation (Q4), synthesis (Q5) and reflection (Q6). The scenario may vary from situation to situation, but the guiding questions remain the same.

Rubric review: Following the 'scenario-review' stage, a two-tier approach was employed to adapt and refine the rubric. In order to be consistent, data were collected for both tiers at the end of the academic year. Since the research methods course is common across academic departments, a scenario related to sampling issues in research was considered suitable for adaptation.

For tier 1, data were collected from the students (n=37) enrolled in two courses in one of the departments. A fraction of cases (n=7) were randomly selected from the students' completed tasks to assess in light of the original rubric. Arguably, these small numbers could not identify all the potential problems of an assessment tool, but the number was deemed sufficient as it represented the sub-population of the intended target population (Dillman 2000). A smaller sample was selected for this exercise to enhance in-depth qualitative analysis by having multiple rounds of discussion for each case. In other words, depth was preferred over breadth as the latter was not the focus at this stage.

A thorough discussion after assessing each of these cases helped to identify areas of concern, which led to modification. For example, a gradual incline in quality was identified by the researchers as one of the issues, especially at the higher end of the scale. In the original rubric, there were five response categories including *not present, emerging, developing, proficient* and *exemplary.* Notes from the researchers' diary referred to this dilemma:

> We started with [a] five-point scale including '0' or 'not present' and marked seven papers accordingly. Initially, none of the participants scored beyond 'emerging' or '1'. However, a better example from a student made us think differently … a critical analysis revealed an issue in gradual incline in quality … especially for the last two response categories (i.e. proficient and exemplary) … in one of the constructs (analysis), the difference between descriptors anchored on *proficient* and *exemplary* was rather blurred … for example, 'logical reasoning to make inferences' was a requirement to qualify for *proficient* category while 'specific inductive or deductive reasoning to make inferences' was anchored on the higher end- *exemplary*. Even after a lot of discussion, marking papers and reviewing other available rubrics for references – the differentiation remained blurred … similar issues emerged in almost all the constructs ... merging the last two categories (*proficient* and *exemplary)* might be a viable solution (Researchers' notes, August 2016).

In light of the rigorous discussion based on data collected, the last two categories were merged to be identified as *exemplary.* Another major change was made at the construct level. The last construct *reflection* was excluded in light of the scoring pattern. A floor-effect was observed in assigning a score to the students' responses where the questions related to this construct were either not responded to or irrelevant responses were presented. Tier 1 resulted in modifications in *level of performance* (merging of two response categories) and in *construct* (exclusion of one of the constructs, i.e. reflection). The revised ACT rubric was used for tier 2.

For tier 2, the rest of the cases of the same data set were analysed using the revised rubric. Rigorous discussion was a continued feature of this process of analysis. At this stage, the discussion on basic features of the rubric, such as content, clarity and gradual incline in quality was extended to wider utility issues of the rubric in terms of its efficacy for teaching, assessment and research. This psychometric-utility nexus led to further changes or modifications in the rubric at three levels including construct, explanation of the construct and level of performance. A comparative overview of the original and modified versions of the rubric is included as Appendix B.

The original rubric has five distinct response categories arranged on a continuum (*not present, emerging, developing, proficient, and exemplary*). The modified version for tier 1 comprised a 4-point

scale (*not present, emerging, developing, and exemplary*), which was retained for tier 2. As mentioned earlier, this adjustment was an outcome of data-based discussions.

The number of constructs in the modified version of the rubric was reduced to five because the two constructs (*problem solving* and *evaluation*) were merged into one category (*evaluation*). The issue of retaining or discarding 'problem solving' as a construct was debated repeatedly during tier 2. Problem solving may be considered an expected outcome of critical thinking; therefore, it may be appropriate not to use this nomenclature. That said, the description of this construct ('select and defend your chosen solution') was retained under 'evaluation'. Reflection was included in the rubric again for pragmatic reasons. There was a consensus that, "since the rubric will be used for multiple purposes including teaching, assessment and research … and reflection is at the heart of the teaching-learning processes … the construct [reflection] should not be excluded ... as it would help to sensitise practices to improve CT" (Researchers' notes of an informal discussion, November 2016). For the rest of the constructs, the main gist of the descriptors anchored on each point was retained with some additional explanation for further clarification and to make it user-friendly.

Some changes were also made in the descriptors anchoring different response categories in order to make a clear distinction between adjacent categories. For example, the modified version which was a product of tier 2 was field tested on students (n=22) of another discipline of the sample university. In order to maintain consistency within the first round of data collection, the same scenario was used as a prompt. Furthermore, the data were collected at the end of the academic year. Data collected at two points were put to statistical analyses.

Field-Testing and establishing reliability

The data collected from two groups as part of tier 1 (n=37) and tier 2 (n=22) were analysed to gather statistical evidence for the ACT rubric. These data sets were subjected to various statistical tests in order to assess inter-rater reliability (i.e. Weighted Kappa) and internal consistency (i.e. Cronbach's alpha, item-total correlation). The purpose of subjecting data to the various statistical tests was to strengthen the evidence for the reliability of the ACT rubric. In addition, these measures were used to maximise the accuracy of measurement by minimising the sources of error as much as possible and obtaining an estimate of how much error remained (Black 1999; Robson & McCartan 2016). In order to compute the reliability coefficient, SPSS 19.0 was used. However, SPSS does not provide an option for computing "Weighted Kappa"; therefore, these calculations were done manually.

Inter-rater reliability: A scan of inter-rater agreement provides evidence of reliability and is presented in Table 1. Evidently, the ACT rubric demonstrates adequate (Weighted Kappa=0.60) to good inter-rater reliability (Weighted Kappa= 0.94) for Tier 1 (Tabachnick & Fidell 2001). The modified version of the ACT further improved the reliability score which fell in the range of satisfactory (Weighted Kappa = 0.76) to excellent (Weighted Kappa = 1.00) (Roberts & McNamee 1998; Fleiss 1971; Cohen 1960). A visible improvement was observed in the construct of *communication*. In contrast, a decline can be noticed in two constructs (*evaluation* and *synthesis*) between tier 1 and tier 2. However, the values remain in the satisfactory range. Reviewing these results, the product of tier 1 and tier 2 almost equally qualify for the 'satisfactory' range of inter-rater reliability. Nevertheless, as discussed earlier, the product of tier 2 has more clarity in terms of language (e.g. defining the descriptors, guiding questions) hence may have more acceptance for teaching, research and assessment.

Table 1. Weighted Kappa for Tier 1 and Tier 2

Constructs	Weighted Kappa (Tier 1)	Weighted Kappa (Tier 2)
Communication	0.60	0.80
Analysis	0.79	0.80
Problem solving	0.91	N.A.
Evaluation	0.94	0.76
Synthesis	0.93	0.80
Reflection	N.A	1.00

Internal consistency: Internal consistency was computed for both Tier 1 and Tier 2 through Cronbach's alpha and item-total correlation. Results of Cronbach's alpha show an improvement from adequate (α= 0.76) to good (α= 0.82). A scan of item-total correlation coefficients further strengthened the evidence of reliability as presented in Table 2. All the values of item-total correlation are above the cut-off point (i.e. 0.3) except communication in tier 1 which barely meets the standard (Field 2005, 2009, 2013). The results specify that each construct in the rubric contributes to the overall score. In other words, if participants perform well on a particular construct, they will have higher overall CT scores.

Table 2. Item-total correlation for Tier 1 and Tier 2

Constructs	Item-total correlation (Tier 1)	Item-total correlation (Tier 2)
Communication	0.33	0.41
Analysis	0.62	0.62
Problem solving	0.47	N.A
Evaluation	0.69	0.60
Synthesis	0.59	0.74
Reflection	N.A	0.73

Table 3 presents evidence related to internal consistency which was accumulated by developing a correlation matrix for both tier 1 and tier 2. The patterns of association among selected CT constructs across tiers revealed that, in general, the constructs associate with each other well except *communication*. Communication focuses on defining the problem in simple words with examples. Arguably, this is one of the fundamental skills which need to be developed to narrate a problem comprehensively before deconstructing the given issue (*analysis*), discussing the best alternatives with justification (*evaluation*), critiquing the proposed solution along with suggestions for further improvement (*synthesis*) and reflecting on the process (*reflection*). Of the three common constructs across tier 1 and tier 2, the association of *communication* seems to have improved with two other processes (i.e. *Analysis and Synthesis*). This trend is encouraging in terms of improved association among constructs; however, further evidence needs to be collected to strengthen correlation of *communication* with the other constructs.

Table 3. Correlation matrix for Tier 1 and Tier 2

CONSTRUCTS	Communication	Analysis	Problem Solving	Evaluation	Synthesis	Reflection
Communication	1					
Analysis (T1)	0.20	1				
Analysis (T2)	0.35					
Problem solving	0.02	0.48*	1			
Evaluation (T1)	0.48*	0.57**	0.39*	1		
Evaluation (T2)	0.23	0.59**	N.A			
Synthesis (T1)	0.25	0.52*	0.43*	0.48*	1	
Synthesis (T2)	0.48*	0.47*	N.A	0.50*		
Reflection	0.30	0.55**	N.A	0.59**	0.80**	1

Note:** 'problem solving' for tier 1 (T1) only and 'reflection' for tier 2 (T2) only p<0.05*; p<0.01

In general, these results reveal sufficient reliability estimates for the ACT rubric, indicating a satisfactory beginning (Field 2009, 2013). The final version of the ACT rubric has five constructs (five items) which are arranged on a 4-point scale to assess a scenario-based written assignment with guiding questions. One may argue that the number of items in the ACT rubric is not enough. However, it could be counter-argued that items are never sufficient to measure the breadth and depth of any construct (Bhutta 2002) and more so with CT, as there is no common consensus on the definition of CT (Jones 2004; Davies 2006; Cassum *et al.* 2013; Moore 2013). That said, a search of the literature revealed some common constructs on which the ACT rubric drew (George Brown College 2015; Liu, Frankel, & Roohr 2014; James, Hughes & Cappa 2010; Paul & Elder 2009; Facione & Facione 1994).

Furthermore, adding too many constructs would have implications for the length of the written assignment as well as its assessment. The field-test revealed that, on average, respondents took 60 minutes to read the scenario and guiding questions, understand the text and write their responses. If the tool is to be used for teaching and learning processes, this might be an optimum amount of time which can be allocated in the course to gauge students' levels of CT at different points during the course. Also, if using it as a research tool, one would avoid asking for too much of the respondents' time, as this would put an unnecessary burden on them (Far 2018; Gul & Ali 2010). That said, the structure of the ACT rubric can be modified to suit the needs of course assignments by assigning different weightings to various constructs or adding other aspects (e.g. language). Also, the rubric can be used across disciplines for assessing CT and can be extended to align with various assessment and teaching and learning needs. Needless to say, the validity and reliability has to be ascertained for new disciplines and groups.

Conclusion

Despite the well-documented importance of CT, its assessment remains a challenge for faculty in higher education. This study attempts to address this challenge by developing and assessing the

efficacy of an assessment tool (the ACT rubric) which would assist multidisciplinary faculty members in enhancing their teaching and assessment practices regarding CT.

The research team has learnt a significant amount during the process of developing the ACT rubric, but acknowledges that much more needs to be done to strengthen this form of assessment. Psychometric properties of the rubric developed as part of the study need to be rigorously evaluated through repeated rounds of assessments on a larger and more varied probability sample across campuses and sites of the sample university (Moy & Murphy 2016; Ha, Hu, Fang, Henize, Park, Stana & Zhang 2015). Arguably, assessment without corrective action would be rightfully considered as empty gesture. Faculty members need to meet regularly to discuss assessment processes and results, devise appropriate interventions for the academic entities, and engage in improving understanding, teaching and assessment of CT. This tool provides an anchoring point to start working on the daunting yet doable task of developing and fine-tuning the assessment measures and interventions based on the assessment findings.

With a particular focus on the ACT rubric, it would be advisable to work in content-specific as well as interdisciplinary groups to develop and maintain a bank of scenarios which can be used as prompts for the teaching and assessment of CT (Davies 2006; Moore 2004, 2011). The users of the ACT rubric need to think of possibilities of sharing the criteria with students – for them to assess their own progress in critical thinking (Wang 2017) and allow for self-regulated learning (Pandero & Alonso-Tapia 2013). Needless to say, change is a continuous process (Fullan 1999) and needs to be monitored efficiently, in this case, by analysing trend data to establish the efficacy of ACT rubric-based interventions. Future undertakings – in both teaching and research – may be benefitted not only by collecting reliability and validity data but also answering some of the teaching-learning questions, including: 'Do the CT-focused interventions significantly improve learning in general and CT in particular over time?' and 'Is one CT teaching approach more effective than the other?'. The investigations guided by these general types of questions would help in generating much-needed contextually relevant information about the teaching, learning and assessment of CT.

In conclusion, the ACT rubric *provides* faculty members with a template to teach and assess CT; *guides* them to make decisions about the level of their students' CT skills through a combination of numerical scores and qualitative description; *assists* them in taking practical steps for further improvement of CT skills of the students; *extends* their thinking about CT; and *empowers* them to make self-initiated conscious efforts to improve their classroom practice with reference to CT.

Note: readers may contact the primary author for a copy of the ACT rubric.

References

Association of American Colleges and Universities (AAC&U). 2009. *Critical Thinking VALUE rubric.* [O]. Available: https://www.aacu.org/value/rubrics/critical-thinking. Accessed 28 March 2019.

Atkinson, D. 1997. A critical approach to critical thinking in TESOL. *TESOL Quarterly* 31(1):71-94.

Barnett, R. 1997. *Higher education: a critical business*. Buckingham: Society for Research into Higher Education and Open University Press.

Bhutta, SM. 2002. Developing a health education Child-to-Child classroom profile. Unpublished MSc Thesis, the University of Oxford, Oxford, UK

Bhutta, SM. 2006. Health Education Classroom Practice in Primary Classroom in Pakistan. Unpublished DPhil Thesis, the University of Oxford, Oxford, UK.

Bhutta, SM & Anwar, NP. Forthcoming. Promoting argumentation in science education classrooms through socio-scientific issues: a case of teacher education classroom. In Bashiruddin, A & Rizvi, NF. *Signature Pedagogy*. Karachi: Oxford University Press.

Black, TR. 1999. *Doing quantitative research in Social Science: an integrated approach to research design, measurement and statistics*. London: Sage Publications.

Bloom, B & Krathwohl, DR. 1956. *Taxonomy of educational objectives: the classification of educational goals, Handbook I: cognitive domain*. New York: Longmans.

Cassum, SH, Profetto-McGrath, J, Gul, RB, Ashraf, D & Syeda, K. 2013. Multidimensionality of critical thinking: a holistic perspective from multidisciplinary educators in Karachi, Pakistan. *Journal of Nursing Education and Practice* 3(7):9-23.

Collegiate Assessment of Academic Proficiency (CAAP). 2012. ACT CAAP technical handbook 2011–2012. [0]. Available: http://www.aum.edu/docs/default-source/OIE/student-achievement/caap-technical-handbook.pdf Accessed 24 January 2017.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37-46.

Cottrell, S. (ed.) 2011. *Critical thinking skills: developing generic analysis and argument*. Second Edition. Basingstoke: Macmillan.

Council for Aid to Education (CAE). 2013. CLA+overview. [O]. Available: http://cae.org/ Accessed 29 March 2016

Crow, LW. 1989. The why, what, how and who of critical thinking and other higher order thinking skills. In Crow, LW. (eds.) *Enhancing critical thinking in the sciences.* Second Edition. Houston: Baylor College of Medicine, 9-16.

Davies, WM. 2006. An 'infusion' approach to critical thinking: Moore on the critical thinking debate. *Higher Education, Research and Development* 25(2):179-193.

Dawson, P. 2017. Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education* 42(3): 347-360.

Dillman, D. 2000. *Mail and internet surveys: the tailored design method*. New York: Wiley.

Ennis, RH. 1985. Critical thinking and the curriculum. *Phi Kappa Phi Journal* 65(1): 28-31.

Ennis, RH. 1990. The extent to which critical thinking is subject-specific: further clarification. *Educational Researcher* 19(4): 13-16.

Ennis, RH. 1997. Incorporating critical thinking in the curriculum: an introduction to some basic issues. *Inquiry 1*6(3): 1-9.

Ennis, RH, Millman, J & Tomko, TN. 1985. *Cornell critical thinking tests*. Pacific Grove, CA: Midwest Publications.

Ennis, RH & Weir, E. 1985. *The Ennis–Weir critical thinking essay test*. Pacific Grove, CA: Midwest Publications.

Facione, P. 1990. *Critical thinking: a statement of expert consensus for purposes of educational assessment and instructions. Research findings and recommendations*. Millbrae, CA: California Academic Press.

Facione, P & Facione, N. 1992. *The California Critical Thinking Dispositions Inventory (CCTDI) and the CCTDI Test Manual*. Millbrae, CA: California Academic Press.

Facione, P & Facione, N. 1994. *How to use the holistic critical thinking scoring rubric*. [0]. Available: https://goo.gl/ivRwK4
Accessed 29 March 2016.

Far, PK. 2018. Challenges of recruitment and retention of university students as research participants: lessons learned from a pilot study. *Journal of the Australian Library and Information Association* 67(3): 278-292.

Field, A. 2005. *Discovering Statistics using SPSS.* Second Edition. London: Sage.

Field, A. 2009. *Discovering Statistics using SPSS.* Third Edition. London: Sage.

Field, A. 2013. *Discovering Statistics using IBM stats SPSS.* Fourth Edition. London: Sage.

Fleiss, JL. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(1): 378-382.

Fraenkel, JR, Wallen, NE & Hyun, HH. 2011. *How to design and evaluate research in education*. New York: McGraw-Hill.

Fullan, MG. 1999. *Change forces: the sequel.* Philadelphia, PA: Falmer Press.

Gearhart, M & Wolf, SA. 1997. Issues in portfolio assessment: assessing writing processes from their products. *Educational Assessment* 4(4): 265-296.

George Brown College. 2015. Critical thinking: learning, teaching, and assessing. A teachers' handbook. [O]. Available: https://trello-attachments.s3.amazonaws.com/5893848b33718d1aff2646a9/594be57df1ca2e7d99f043bc/9dc9a6ebe6bf9e12c626caea832b59eb/Critical-Thinking-Learning-Teaching-and-Assessment.pdf
Accessed 10 October 2015

Gul, RB & Ali, PA. 2010. Clinical trials: the challenge of recruitment and retention of participants. *Journal of Clinical Nursing* 19(1-2): 227-233.

Ha, L, Hu, X, Fang, L, Henize, S, Park, S, Stana, A & Zhang, X. 2015. Use of survey research in top mass communication journals 2001–2010 and the total survey error paradigm. *Review of Communication* 15(1): 39-59.

Hafner, JC & Hafner, PM. 2003. Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating. *International Journal of Science Education* 25(12): 1509-1528.

Halpern, DF. 1998. Teaching critical thinking for transfer across domains: dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist* 53(1): 449-455.

Halpern, DF. 2010. Halpern critical thinking assessment manual. Vienna, Austria: Schuhfried GmbH.

Horsburgh, M. 1999. Quality monitoring in higher education: the impact on student learning. *Quality in Higher Education* 5(1): 9-25.

Jackson, D. 2014. The use of rubrics in benchmarking and assessing employability skills. *Journal of Management Education*. DOI: 10.1177/105256291351143.

James, N, Hughes, C & Cappa, C. 2010. Conceptualising, developing and assessing critical thinking in law. *Teaching in Higher Education* 15(3): 285-297.

Jones, A. 2004. Teaching critical thinking: an investigation of a task in introductory macroeconomics. *Higher Education, Research and Development* 23(2): 167-181.

Kapelus, G, Miyagi, N & Scovill, V. 2017. *Building Capacity to Measure Essential Employability Skills: A Focus on Critical Thinking*. Toronto: Higher Education Quality Council of Ontario.

Liu, OL, Frankel, L & Roohr, KC. 2014. *Assessing critical thinking in higher education: current state and directions for next-generation assessment*. ETS Research Report: Wiley Online Library.

Lowden, K, Hall, S, Elliot, D & Lewin, J. 2011. *Employers' Perceptions of the Employability Skills of New Graduates Research Commissioned*. London: Edge Foundation.

McKeown, J & Biss, LD. 2018. *HEQCO's Guide to Developing Valid and Reliable Rubrics.* Toronto: Higher Education Quality Council of Ontario.

McPeck, J. 1981. *Critical Thinking and Education.* New York: St Martin's Press.

McPeck, J. 1990. *Teaching critical thinking: dialogue and dialectic.* New York: Routledge.

McPeck, J. 1992. Thoughts on subject specificity. In Norris, S. (eds.) *The generalizability of critical thinking: multiple perspectives on an educational ideal.* New York: Teachers College Press, 198-205.

Moore, T. 2004. The critical thinking debate: how general are general thinking skills? *Higher Education, Research and Development* 23(1): 3-18.

Moore, TJ. 2011. Critical thinking and disciplinary thinking: a continuing debate. *Higher Education, Research and Development* 30(3): 261-274.

Moore, T. 2013. Critical thinking: seven definitions in search of a concept. *Studies in Higher Education* 38(4): 506-522.

Moskal, BM. 2000. Scoring rubrics: what, when and how? *Practical Assessment, Research & Evaluation* 7(3).

Moy, P & Murphy, J. 2016. Problems and prospects in survey research. *Journalism & Mass Communication Quarterly* 93(1): 16-37.

Nelson, RR. 1989. What is private and what is public about technology? *Science, Technology and Human Values* 14: 229-241.

Nunley, C, Bers, T & Manning, T. 2011. *Learning Outcomes Assessment in Community Colleges*. University of Illinois, IL: National Institute for Learning Outcomes. [O]. Available: http://www.learningoutcomesassessment.org/documents/CommunityCollege.pdf Accessed 17 January 2019

Panadero, E & Alonso-Tapia, J. 2013. Self-assessment: theoretical and practical connotations: when it happens, how is it acquired and what to do to develop it in our students. *Electronic Journal of Research in Educational Psychology* 11 (2): 551–576.

Paul, R & Elder, L. 2009. *Critical Thinking competency standards: Standards, principles, performance, indicators, and Outcomes with a Critical Thinking Master Rubric.* Dillon Beach, CA: Foundation for Critical Thinking Press.

Pichette, J & Watkins, EK. 2018. *Learning from the Queen's University Assessment Experience: Considerations for selecting an appropriate skills measurement tool*. Toronto: Higher Education Quality Council of Ontario.

Pithers, RT & Soden, R. 2000. Critical thinking in education: a review. *Educational Research* 42(3): 237-249.

Popham, WJ. 1997. What's wrong – and what's right – with rubrics. *Educational Leadership* 55(2): 72-75.

Punch, K. 2000. *Introduction to Social Research*: *quantitative and qualitative approaches*. London: Sage

Reddy, YM & Andrade, H. 2010. A review of rubric use in higher education. *Assessment & Evaluation in Higher Education* 35(4): 435-448.

Reed, JH & Kromrey, JD. 2001. Teaching critical thinking as a community college history course: empirical evidence from infusing Paul's model. *College Student Journal.* 35(2): 201-215.

Roberts, C & McNamee, R. 1998. A matrix of Kappa-type coefficients to assess the reliability of nominal scales. *Statistics in Medicine* 17(4): 471-488.

Robinson, SR. 2011. Teaching logic and teaching critical thinking: revisiting McPeck. *Higher Education Research & Development* 30(3): 275-287.

Robson, C & McCartan, K. 2016. *Real world research*. Oxford: John Wiley & Sons.

Rochford, L & Borchert, PS. 2011. Assessing higher level learning: developing rubrics for case analysis. *Journal of Education for Business* 86(5): 258-265.

Spicer, KL & Hanks, WE. 1995. Multiple measures of Critical Thinking Skills and Pre-Disposition in assessment of critical thinking. Paper presented at 81st annual meeting of the Speech Communication Association, 18-21 November San Antonio, TX.

St. Petersburg College. 2008. Assessment Rubric for Critical Thinking. [0]. Available: https://go.spcollege.edu/CriticalThinking/students/rubrics.htm Accessed 29 March 2016

Swartz, RJ & Perkins, DN. 1989. *Teaching thinking: issue and approaches*. Cheltenham: Hawker Brownlow Education.

Swartz, RJ & Parks, S. 1994. *Infusing the teaching of critical thinking into content instructions*. Pacific Grove, CA: Critical Thinking Books & Software.

Tabachnick, BG & Fidell, LS. 2001. *Using Multivariate Statistics*. Boston, MA: Allyn & Bacon.

Velasco-Martinez, L-C & Tojar-Hurtado, J-C. 2018. Competency-based evaluation in higher education-design and use of competence rubrics by university educators. *International Education Studies* 11(2): 118-132.

Wang, W. 2017. Using rubrics in student self-assessment: student perceptions in the English as a foreign language writing context. A*ssessment & Evaluation in Higher Education* 42(8): 1280-1292.

Watson, G & Glaser, EM. 1980. *Watson-Glaser Critical Thinking Appraisal, forms A and B Manual*. San Antonio TX: the Psychological Corporation.

Appendix A: a sample scenario

Read the following scenario carefully and respond to the questions given after it.
Scenario:
You are attending a conference with other professionals in the early childhood care and education field. A slide is used in one of the presentations which show Infant Mortality rates for countries around the world. In the discussion that followed, several members of your group had opinions on the best way to change the standing for Pakistan.

- One member suggested increased parent education since she felt future parents were unaware of what they should do before their baby is born.
- Another felt that education was not the primary solution since many families know what to do, but do not have access to prenatal (before birth) care.
- Still another felt that the focus should be on a specific solution such as substance abuse prevention since that is the cause of many premature births.
- One member then suggested improvements in neonatal (a period of 40 days after birth) intensive care programs in all community hospitals.

Guiding Questions:

Q1. Define the problem in your own words.
Q2. Compare and contrast the available solutions in the scenario**.**
Q3. Select and defend what you think is the 'best solution' among those presented in the scenario.
Q4. Identify any weaknesses this 'best solution' may have.
Q5. Suggest ways to strengthen or improve this 'best solution'.
Q6. Reflect on your own thought process after completing the assignment.
a. *What did you learn from this process?*
b. *What would you do differently next time to improve?*

Adapted from https://go.spcollege.edu/CriticalThinking/students/rubrics.htm

Appendix B: a comparative overview of original and modified versions of the rubric

(ARC) Original			**Modified (ACT)**		
5-point scale (not present. emerging, developing, proficient, exemplary)	**Construct**	**Explanation**	**4-point scale (not present. emerging, developing, exemplary)**	**Construct**	**Explanation**
	Communication	Define problem in your own words		Communication	Define problem clearly in your own words with supporting examples, if applicable.
	Analysis	Compare & contrast the available solutions.		Analysis	Discuss the pros and cons of the given situations (solutions/alternatives) with reasoning.
	Problem Solving	Select & defend your chosen solution.		Evaluation	Select the best solution for the given situation with justification.
	Evaluation	Identify weaknesses in your chosen solution.		Synthesis	Suggest ways to improve/strengthen your chosen solution.
	Synthesis	Suggest ways to improve/strengthen your chosen solution.		Reflection	Reflect on the challenges faced in answering the questions related to the given scenario/situations (e.g. What did you learn from this process? What would you do differently next time to improve?)
	Reflection	Reflect on your own thought process (e.g. What did you learn from this process? What would you do differently next time to improve?)			