Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 11, September 2018 ISSN 1531-7714

An Effective Rubric Norming Process

Kevin Schoepp, *Independent Researcher*
Maurice Danaher, *Zayed University*
Ashley Ater Kranov, *Washington State University*

Within higher education, rubric use is expanding. Whereas some years ago the topic of rubrics may have been of interest only to faculty in colleges of education, in recent years the focus on teaching and learning and the emphasis from accrediting bodies has elevated the importance of rubrics across disciplines and different types of assessment. One of the key aspects to successful implementation of a shared rubric is the process known as norming, calibrating, or moderating rubrics, an oft-neglected area in rubric literature. Norming should be a collaborative process built around knowledge of the rubric and meaningful discussion leading to evidence-driven consensus, but actual examples of norming are rarely available to university faculty. This paper describes the steps involved in a successful consensus-driven norming process in higher education using one particular rubric, the Computing Professional Skills Assessment (CPSA). The steps are: 1) document preparation; 2) rubric review; 3) initial reading and scoring of one learning outcome; 4) initial sharing/recording of results; 5) initial consensus development and adjusting of results; 6) initial reading and scoring of remaining learning outcomes; 7) reading and scoring of remaining transcripts; 8) sharing/recording results; 9) development of consensus and adjusting of results. This norming process, though used for the CPSA, is transferable to other rubrics where faculty have come together to collaborate on grading a shared assignment. It is most appropriate for higher education where, more often than not, faculty independence requires consensus over directive.

The prevalence of rubric use in higher education is increasing. Not many years ago mentioning rubrics to faculty members in many fields may have brought forth looks of confusion, consternation, or disinterest. Today, however, the topic of rubrics can be found as part of regular faculty development programs, as standard expectations from accreditors, and as the focus of major cross-disciplinary higher education projects such as the Association of American Colleges and Universities (AAC&U) VALUE rubrics (Association of American Colleges and Universities, 2014). Rubrics are now seen as a way to bring to the surface and make transparent the criteria that faculty members value from assignments which can then serve as a pre-assignment guide, post-assignment assessment, and a feedback tool for students. Nonetheless, critics of rubric use exist, often arguing that rubrics may disrespect a faculty member's evaluative expertise or that the focus on specific criteria, to the exclusion of other criteria, limits or constrains creativity which makes the assignment and feedback inflexible. Bloxham, den-Outer, Hudson, and Price (2016) for example, have argued that with detailed assessment criteria, it "is likely to make marking an overly onerous process, limit independent thought and originality in students and encourage middling grades if individual criteria are scored" (p. 479). Though these voices of dissent continue to grow weaker, they remain a reality in higher education because of the degree of independence often granted to faculty as subject matter experts.

This paper describes the process known as norming, calibrating, or moderating rubrics. The act of rubric norming is defined as an iterative process in which raters assess samples of student work against criteria

presented in a rubric to establish an accepted level of consistency in marking. It is a collaborative process that requires discussion leading to evidence-driven consensus- a procedure where examples from student work are used to justify scores leading to a shared understanding amongst raters. Norming should be done any time faculty members are implementing a shared rubric across multiple sections or when faculty members are sharing the assessment of a single group of students. The norming process is important because it helps faculty members gain a shared understanding of the rubric criteria employed and of performance standards and thresholds. It is the crucial, oft neglected, post-development phase of rubric implementation which should occur prior to conducting any analyses into rubric reliability. While it is most certainly true that the quality of a rubric impacts its reliability and that interrater reliability statistics play an important role in the analysis of a rubric, good norming is an essential step in the successful deployment of shared rubrics. Norming is the rubric implementation phase which is often given only a passing reference in the literature. While there exists a plethora of rubric development literature (Burke, 2010; Stevens, & Levi, 2013), and a wide range of articles into rubric reliability issues (Bresciani et al., 2009; Jonsson & Svingby, 2007; Stemler, 2004), in only a few cases has the norming process been described (Crisp, 2017; Holmes & Oakleaf, 2013). It is almost as if good norming is a known, well-understood process when the reality is that unless a faculty member has been taught or has participated in a well-organized and structured norming session, they may be unsure of how to proceed and require guidance if they are to lead or participate effectively in a norming session.

In 2014 when we started working on our rubric, the Computing Professional Skills Assessment (CPSA), and investigating effective norming, we found that there were gaps in the literature. Since Holmes and Oakleaf (2013) had provided a useful set of rules for the norming process, and some of us had participated in other norming sessions, we instigated a process based on what we had learned and added to it as we continued to use and refine the CPSA. Over the following years we refined our rubric, method, and norming process. In 2017 Crisp published a norming paper that had a number of steps that we had already implemented such as raters justifying scores through specific language in the rubric and in student work, and raters continuing the discussions until consensus has been achieved. While we agree with most of the guidance put forth by both Crisp (2017) and Holmes and Oakleaf (2013), we have included more consensus building discussions into our process. The process includes an initial norming session, ratings, then an additional evidence-based discussion to ensure we achieve consensus. We now have an effective and successful norming process that works because it has proven to be both reliable (Danaher, Schoepp, & Ater Kranov, 2016) and valid (Danaher, Schoepp, & Ater Kranov, 2018). This paper presents a specific step-by-step example of our norming process that includes many transferable aspects.

Rubrics

Though resistance remains, it could be argued that the use of rubrics to assess student learning is becoming mainstream. A recent google scholar search for the term *rubric* brings up approximately 347,000 results, and a keyword search within *Practical Assessment, Research & Evaluation* brings up 54 articles (November 21, 2017). Within American higher education, there is a strong movement towards the utilization of rubrics as a way to assess student attainment of learning outcomes at the program or institution level. Through the VALUE project, a set of 16 rubrics have been drafted, edited, and implemented across the country (Association of American Colleges and Universities, 2014). In addition, a review of literature from both regional and disciplinary US-based accreditors, for example, the Accreditation Board for Engineering and Technology (ABET), the Western Association of Schools and Colleges, the Middle States Commission on Higher Education, and the Council for the Accreditation of Educator Preparation, found that rubrics are heavily promoted as a trustworthy method of assessment.

There are two types of rubrics for evaluating students' work: analytic rubrics and holistic rubrics. We are focused on analytic rubrics in this study. Analytic rubrics offer unique hierarchical descriptors of student work along specific assessment criteria. They let the rater select the appropriate descriptor for each of the criteria, making the results focused and meaningful. They provide a complete picture of student performance across an entire spectrum of criteria, so besides providing an overall score, they offer guidance into areas of both strength and weakness (see Figure 1).

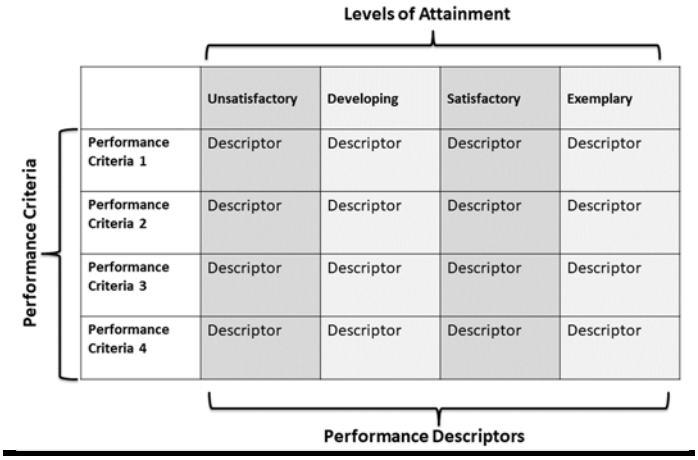

Figure 1. Analytic Rubric

This differs from holistic rubrics which, while including performance descriptors, only provide a meaningful overall score because there can be a misalignment between some performance descriptors and the overall score assigned (see Figure 2). For example, a holistic rubric would have a set of different performance descriptors within a single level of attainment. The problem is that a student may align with 3 performance descriptors at one level and 2 at another, so the problem becomes determining their actual level of attainment.

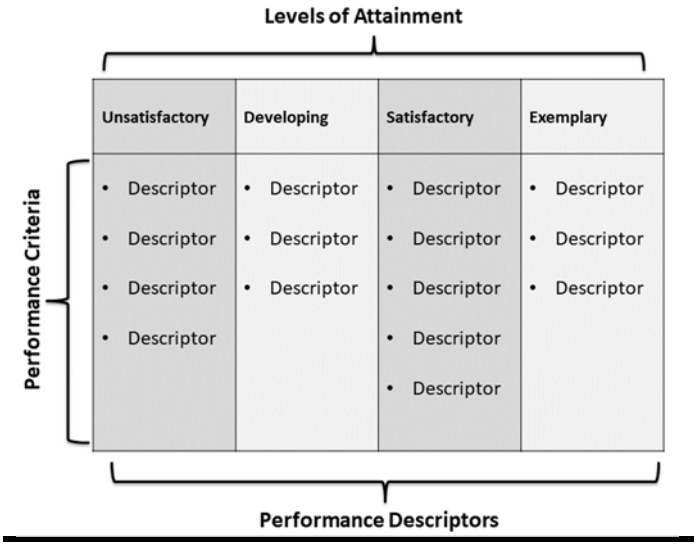

Figure 2. Holistic Rubric

Many educational assessments require the use of a judge or rater to evaluate a response to items that are subjective or qualitative in nature such as an open-ended response, a behavioral observation, or an essay which purports to demonstrate the attainment of a particular learning outcome. Any time a judge or rater is involved in such assessments, there is a degree of subjectivity and in turn variability which cannot be ignored. To help control or mitigate against the inherent subjectivity of open-ended responses, rubrics should be utilized to improve the consistency and reliability of raters, thereby increasing the objectivity of results (Tierney & Simon, 2004). O'Connell et al. (2016) recently found a great deal of variability in scoring amongst accounting faculty but also discovered that if raters undergo a norming program, this variability is cut considerably.

Literature Review

Though rubrics have been strongly promoted across disciplines in higher education, in recent years, there is a paucity of literature on aspects of their use, in particular on norming. Research has been published on reliability studies and a little has been published on components of rubrics and common problems. Recently a few studies have discussed the benefits of the norming process and a few others briefly outlined the norming process.

At the most rudimentary level of understanding rubrics, Popham (1997) posited that they have "three essential features: evaluative criteria, quality definitions, and a scoring strategy" (p. 72). The evaluative criteria are what matters in any given assignment. The quality definitions are the descriptions of performance of each identified criterion, while the scoring strategy refers to the aforementioned format of either holistic or analytic. The way in which rubrics were discussed did not progress much beyond this for a number of years. Part of the progression included a paper (Tierney & Simon, 2004) that focused on consistency in performance descriptors, specifically basic consistency- keeping the attributes examined the same in each descriptor, and negative/positive consistency- keeping the language used positive throughout the continuum of learning. Though the authors recognized the importance of exemplars or benchmarks to operationalize meaning, they stressed that accurate and consistent wording of performance descriptors is still critical. A later paper (Goldberg, 2014) expanded on this work and provided a comprehensive description of the rubric revision process. It addressed a number of problems that commonly plague rubrics, issues like lack of consistency and parallelism, redundancy in descriptors, and unevenness in incremental levels. Goldberg attempted to make faculty aware of common problems, so as to increase the probability that these issues could be

identified as part of a rubric revision process, or that they could be avoided altogether as a rubric is being developed. Progressing far beyond this, Dawson (2017) developed a 14-part framework to describe rubrics in far greater detail. The framework included elements such as whether or not a rubric is task-specific or generic, whether or not it is shared with students or kept secret, types of feedback given to the learner, approaches taken to ensure reliability and validity, considerations about the complexity of judgment, and the number and type of quality levels. As with Goldberg, the framework was developed to help eliminate confusion or misunderstanding that can often take place when rubric discussions occur.

Norming is now recognized as an important aspect of rubric use and is crucial to the reliability and validity of a rubric. Without such a process, deployment of a rubric may be a waste of time, or severely limit its effectiveness (Holmes & Oakleaf, 2013). As Jonsson and Svingby (2007) concluded through an examination of over 75 rubric studies, rater training (i.e. norming) and the use of exemplars will increase, though never fully eliminate rubric reliability issues. For example, Boulet, Rebbecchi, Denton, Mckinley, and Whelan (2004) emphasized that rater training was the most important step in assuring reliable assessments. In a pilot study assessing the ability of medical students to accurately summarize patient data, there was a 76% variance between raters. After two rounds of training, the rating variance was reduced to 12%. In a similar manner, a more recent rubric norming paper that described an experimental design with a control and treatment group using accounting faculty raters found that there was a great deal of variability in ratings, but that with a consensus norming workshop, variability amongst raters halved (O'Connell, et al., 2016). An important element to the norming process is range finding, the process where range finders, benchmarks, or exemplars of student work are used to "operationalize the concepts described in the language of the scoring rubric… [and] define the standards of performance for a given assessment and serve as the rubric's surrogate reference points, against which all samples are judged" (Osborn Popp, Ryan, Thompson, Behrens, 2003, p. 3). Studies (Geisinger & Foley, 2010; Osborn Popp, Ryan, Thompson, Behrens, 2003; Wang, Engelhard, Raczynski, Song, & Wolfe, 2017) have found that exemplar papers play a key role in scoring outcomes and that poor benchmarks can limit the effectiveness of the entire scoring process. Hence, the range finding process, where experts identify exemplars and benchmarks of student work, is critical to rubric implementation. One study stands out in contradiction to the accepted belief that norming is required to ensure interrater reliability. The study had a rubric administered to more than 200 student presentations, and researchers (Bresciani et al., 2009) found a remarkable level of rater agreement-Cronbach's alpha correlations up to 0.77. This high level of agreement was achieved with raters only given the rubric and an instruction sheet. Because of this, the authors speculated if "informal norming may have occurred and therefore influenced the statistical level of agreement" (p. 4).

In a review of rubric use in higher education, Reddy and Andrade (2010) found a myriad of rubric papers, but they were unable to identify any studies which described the process of norming in any meaningful manner. Hence, it has often been necessary to turn to look outside of the traditional peer reviewed publications for such literature. In particular, two handbooks or guides have been especially relevant and useful. The first, is a handbook about interrater reliability in the evaluation of teachers. In it, Graham, Milanowski and Miller (2012) stated that well-designed rater training improves agreement and that it must be built around developing a common understanding amongst raters. Though they include a guide sheet for leading training to foster interrater reliability, it is rather limited, and does not clearly describe the process. It does, however, include topics such as clarifying the rubrics and answering questions about wording and explaining common errors which do have roles in norming. In one of the better online guides about the norming process, the University of Hawaii Manoa's Assessment Office (2013) described how consensus can be developed through discussion where raters who gave different scores should explain their judgments by referring explicitly to the rubric. While this is certainly part of the process, explicit reference to aspects of the text or artifact under review is also imperative and not included in the guide.

Within peer reviewed literature, Finley (2011) described a cross-disciplinary norming process with participants located throughout the US implementing one of the VALUE rubrics. In the process, faculty raters first familiarized themselves with the rubric and then scored samples as part of an initial norming round. Their scores were compared to the scores set by a team of experts and if there was alignment, they scored two

additional samples. If there was a lack of alignment, raters discussed the discrepancy with an expert and then participated in the same process again before progressing to two additional samples. Though not overly robust, the method seemed to work. However, the discussions with experts, where the real norming process occurred, was not described in detail. Holmes and Oakleaf (2013) wrote a more detailed and very practical article with a generic set of rules for successful rubric norming which included: a facilitator must be in charge; the facilitator should explain their scores; let the raters try a few ratings independently; discuss, explain, and reconcile; let the raters try a few more; repeat the process. The paper provided an honest introspective analysis of the challenges that can be faced when norming rubrics with colleagues. Another more recent publication offered a 7-step rubric norming process (Crisp, 2017). The steps in the norming process were: 1. review of the task; 2. examination of prompt and student work; 3. clarifying questions; 4. rubric clarification; 5. read and score; 6. discussion; 7. debrief. If followed, the steps to effective norming provided by Crisp would enable a faculty member to facilitate a norming session.

Computing Professional Skills Assessment

In this paper the rubric utilized to demonstrate an effective norming process is known as the CPSA rubric. The CPSA is both a method and a rubric and while the rubric is used to assess discussion transcripts, the norming process itself is fairly transferable to more traditional assessments such as essays, presentations or projects. A short history of the CPSA, the method, and the rubric will be provided in order to provide the necessary context to make the norming process clear.

The CPSA has its roots as far back as 2008 when researchers (Ater Kranov, Hauser, Olsen, & Girardeau, 2008) created the Engineering Professional Skills Assessment (EPSA). The EPSA was designed to simultaneously measure the 21st century, transferable, or professional skills learning outcomes that were expected from ABET- accredited engineering programs. ABET had been on the forefront of integrating these types of learning outcomes into programs since the release of the *Engineering Criteria 2000* document because they recognized that the teaching and learning of these learning outcomes in technical programs was often a weakness, and because these outcomes were vitally important to employment (ABET, n.d.). The EPSA was a scenario-based face-to-face small group discussion where students read a short engineering-related article and were then asked to discuss and come up with solutions to the problems posed in the article. Discussions were recorded, transcribed, and then assessed using the learning outcomes expressed in the EPSA rubric. Scoring was done at the group level, rather than at the individual level because it was program effectiveness, not individual student performance that was under scrutiny.

The CPSA was developed because the EPSA was engineering specific, yet CPSA developers were keen to attempt a similar, but more rigorous, method with students in a computing program. The CPSA has been iteratively developed over a number of years and rounds of implementation. Up until now, more than 400 students and about 10 faculty members have used the CPSA with the first implementation occurring in 2014 (Ater Kranov, Danaher & Schoepp, 2014). Fundamentally, the CPSA method and rubric are similar to what was done with the EPSA, but there are a number of significant differences specifically with the scenarios, the medium of the assessment, and the rubric itself. In terms of similarities, the CPSA is a small group scenario-based discussion that simultaneously measures all of ABET's professional skills. Regarding the differences, scenarios are all computing focused and since the students are all second language learners, scenarios have been written to grade 12 on the Flesch-Kincaid scale (Kincaid, Fishburne Jr, Rogers, & Chissom, 1975). Where the EPSA utilized face-to-face discussion, the CPSA is conducted through an asynchronous online discussion board. This was done because a discussion board gives a ready-made transcript, and the asynchronous medium was thought to be more suitable for second language learners because they have more time to analyse the problem and develop and craft their responses. In the discussion, lasting 12 days, each student makes a minimum of five posts of around 200 words. As there are five students per group the total transcript usually consists of around 25 posts. The CPSA rubric assesses the professional skills learning outcomes particular to ABET's Computing Accreditation Commission (CAC). It does this through alignment to a set of six CPSA outcomes that have been slightly modified from those of ABET so as to better fit the context of the CPSA. The CPSA outcomes, the definition of the outcomes, and the rubric itself represent an attempt to granularize broad ABET CAC outcomes. The six outcomes are:

CPSA 1 Students will be able to problem-solve from a computing perspective.
CPSA 2 Students will be able to work together to perform a specific task.
CPSA 3 Students will be able to evaluate professional, ethical, legal and security considerations when solving a problem.
CPSA 4 Students will be able to communicate professionally in writing.
CPSA 5 Students will be able to analyse the local and global impacts of computing.
CPSA 6 Students will be able to recognize when they need to seek further information to extend their knowledge.

Each section of the rubric assesses one of the learning outcomes and includes the name of the learning outcome, space for the rater's score, a detailed definition of the outcome, the six-level rubric, the performance indicators, the performance descriptors, and a comments area. See figure 3 for an example of one of the sections. The easy-to-read single page version of the rubric has been added as an appendix.

The Norming Process

In this section we present the details of our norming process which was refined over a period of about three years. We started from the literature available at the time, in particular the set of rules by Holmes and Oakleaf (2013), and additionally our own previous experience. We were particularly concerned with developing a norming process which would emphasize rater consensus. We believe our emphasis on consensus can mitigate two of the major concerns brought forth by critics Bloxham, den-Outer, Hudson, and Price (2016) in which they claim raters do not have a shared understanding of either the criteria or of standards. In fact, Watty et al. (2014) posit that this sort of intensive consensus development can help academics achieve a shared understanding of standards and criteria. Crisp's (2017) outline of norming is in line with our process, though ours involves more consensus building discussions. We believe our additional evidence-based discussions to facilitate consensus is paramount to meet the expectations of university faculty regarding peer review and collaboration.

For the CPSA norming process, at least three raters are always used in order to increase the reliability and validity of the ratings. The steps taken in the process are:

1) document preparation;
2) rubric review;
3) initial reading and scoring of one learning outcome;
4) initial sharing/recording of results;
5) initial consensus development and adjusting of results;
6) initial reading and scoring of remaining learning outcomes;
7) reading and scoring of remaining transcripts;
8) sharing/recording results;
9) development of consensus and adjusting of results.

After the set up and initial round to establish a baseline consensus amongst raters, steps 6-9 are where the majority of the ratings occur. Because of the initial

CPSA 2. Students will be able to work together to perform a specific task.
Rater Score for Skill_____

Definition: Students understand the task, as outlined by the prompts, and work to complete it. Their discussion is guided by the prompts. Students work together to address the problems raised in the scenario by acknowledging, building on, clarifying and/or critiquing each other's ideas. Students encourage participation of all team members.

	0 - Missing	1 - Emerging	2 - Developing	3 - Practicing	4 - Maturing	5 - Mastering
Task Orientation	Student discussion is not guided by the prompts.	Students use only a portion of the prompts to guide their discussion. Students get off task. They may be unaware that they have gotten off task or may work to get back on task but unsuccessfully.		Students use the entire set of prompts to guide their discussion. Students recognize when they get off task and work to get back on task.		Student discussion is closely aligned to the entire set of prompts. Students plan their discussion according to the prompts in order to ensure completion and thorough consideration.
Discussion	Students do not acknowledge or encourage participation of others.	Students may pose individual opinions without linking to what others say. Students acknowledge the ideas of others, but may too hastily defer to an opinion.		Students acknowledge, build on, clarify and/or critique and others ideas with some success. Students encourage participation of others to come to consensus.		Students clearly encourage participation from all group members, generate ideas together, actively help each other, and clarify and/or critique each other's ideas.
Comments:						

Figure 3. CPSA Rubric Example

baseline there is usually a high degree of interrater reliability from step 6 onwards. In order for a good norming process to occur, raters need to participate in the process with an open mind and a willingness to re-examine some of their ratings based on evidence and peer consensus. During consensus-building discussions, lead raters may need to remind themselves that when it comes to norming a rubric, consensus is indeed the goal. For rubric assessments to be valid, they must first be reliable, and that means multiple raters must be able to provide consistent scores. Achieving consistent ratings may be a challenge, but it is through the evidence-based discussions that an acceptable level of agreement can occur.

1) Document preparation

Though this may seem obvious, without properly organized documentation for each of the raters, confusion can easily ensue. For the CPSA, the transcripts must be downloaded from the discussion board in chronological order, so that each of the posts appears after the one that came directly before it. Next any identifying headings from the posts must be removed, then posts are numbered and page numbers added. The importance of post numbers and page numbers is that they will make it easier when referring to specific parts of the text during the rating and consensus building process. A labeled copy of each transcript must be provided to each rater. A set of rubrics must also be distributed to the raters, so that they have at least one rubric for each transcript.

2) Rubric review

The entire rubric needs to be reviewed by each of the raters, especially if raters have not assisted in the creation of the rubric. Any definitions, performance indicators, or descriptors that bring forth any concerns or questions from the raters need to be discussed and clarified. Without a shared understanding of the rubric, there is little chance to ensure a trustworthy assessment process. A short time limit of approximately 10 minutes should be set to allow the raters to review the rubric, and time should be watched if clarification discussions ensue.

3) Initial reading and scoring of one learning outcome

This is basically the trial run before assessing all of the learning outcomes and the entire set of transcripts. With the constructs contained in the rubric in mind, raters should read the first transcript and score the first learning outcome. To accurately assess the transcripts, raters should underline, highlight, and make notations throughout the transcript, any time they identify text that pertains to CPSA learning outcomes. This is also the time to add comments and refer to page numbers in the comments section of the rubric and possibly highlight or underline pertinent descriptors within the rubric. Crisp (2017) emphasizes these aspects in order to justify scores- all of this is done to facilitate evidence-based assessment. Next, a score should be written for each performance indicator and an overall score for the learning outcome added to the correct page of the rubric. A time limit of no more than 20 minutes should be given for this phase, otherwise the scoring process can become burdensome.

4) Initial sharing/recording of results

Each of the raters reads their score for the initial learning outcome, and the results should be recorded onto a spreadsheet.

5) Initial consensus development and adjusting of results

Once the scores from the first learning outcome have been added to the spreadsheet, results must be reviewed for any scores of more than 1 point difference on the rubric. A difference of one is considered acceptable since the rubric is a six-point scale. Where a discrepancy in scores exists, an evidence-based discussion must occur. This means, "resolving issues centered on either the meaning of the rubric or the merit and validity of the evidence in the student work until consensus is reached" (Crisp, 2017, p. 12). This is where the previous work in highlighting, underlining, commenting, and so forth becomes crucially important. As explained by Holmes and Oakleaf (2013), it is in this discussion where the lead rater can serve as a role model and use phrases like *I have given a score of _ because posts _ and _ were good examples of the descriptor at this level.* The use of evidence throughout the discussion should lead to having the scores on a single learning outcome to within 1 point of each other. Where changes in scores have been made, the spreadsheet must be updated. This phase, while having the potential to be contentious, has never become so in our experience because of the focus on evidence.

6) *Initial reading and scoring of remaining learning outcomes*

After having completed scoring and reaching consensus for the first learning outcome, the process is repeated for the remaining five learning outcomes on the selected transcript. Since the raters are already familiar with the transcript and the rubric, a 5-10 minute time limit for scoring each learning outcome is adequate. By the end of this step, each rater will have assessed the 6 learning outcomes for the selected transcript, evidenced-based discussions will have occurred, and the spreadsheet will have a completed set of scores for an entire transcript.

7) *Reading and scoring of remaining transcripts*

Having established a good understanding of the rubric and having reached consensus on the first transcript, the reading and scoring process should be repeated for the remaining transcripts. A time limit for scoring should be agreed upon to increase efficiency.

8) *Sharing/recording results*

Like in step 4, each of the raters need to read their scores for each of the transcript's learning outcomes and the results should be recorded onto a spreadsheet once all of the ratings have been completed.

9) *Consensus development and adjusting of results*

After all of the scores have been added to the spreadsheet, results must be reviewed for any scores of more than 1-point difference on the rubric. When such differences exist, another evidence-based discussion must occur and all of the scores on a learning outcome should be brought to within 1 point of each other. Again, raters need to justify their scores by pointing to the transcripts and to specific language in the rubric. Where changes in scores have been made, the spreadsheet must be updated. Consensus to within 1 point of one another is very achievable through the evidence-based discussion. A score is only accurate and defendable if evidence can be shown to support it. Progressing beyond guidance from either Holmes and Oakleaf (2013) or Crisp (2017), it is this repeated evidence-based discussion that makes this process so meaningful for university faculty because it develops a shared understanding.

One of the challenges we have found during our rating sessions is that when a rater is not aligned with the other raters, they are often consistently low or consistently high. To address this issue, we start by referring to evidence in the transcripts. This is where the comments, highlighting and underlining of transcripts to share as evidence with other raters is crucial. By pointing to examples in the transcripts, we are usually able to come to a consensus and adjust the outlier scores throughout the ratings process. Reviewing the language of the rubric is also useful because it can often clarify meaning, but if disagreement remains, the rubric is highlighted for future revision. This issue has shown us the value of range finders, benchmarks, or exemplars of student work in operationalizing the concepts described in the rubric. Because of this, we are in the process of drafting a CPSA administration manual that includes examples from student transcripts that represent ratings on particular outcomes. Though exemplars are not always possible for university faculty to include in a norming session, they can help facilitate understanding and expectations of student work as it pertains to the language of the rubric.

Interrater Reliability

Though not the focus of this paper, it is important to discuss interrater reliability because of the key role it plays in evaluating the effectiveness of the implementation of a rubric. Whereas rubric norming attempts to align rubric raters prior to the formal rating process, interrater reliability statistics are a check on the ratings after the fact. For the purposes of this paper, it demonstrates the success of both the development of the rubric and, more crucially, the norming process.

When working with the CPSA rubric, interrater reliability has been determined through the most basic method- a count of ratings receiving the same scores divided by the total number of ratings completed. This measure of interrater reliability has been shown to be the most commonly applied when calculated to exact or adjacent agreement (Jonsson & Svingby, 2007). The target for agreement is 100%, but Stemler's (2004) guidance that agreement between raters should reach at least 70% has been adopted. Over the past few years with the CPSA rubric, interrater reliability has been calculated twice as a check on the norming process. In the earlier study, Danaher, Schoepp, and Ater Kranov (2016) found that the cumulative level of agreement was 75%; however, while 3 of the learning outcomes had an 83% agreement (communication, local and global impact, and professional development), both the

teamwork (61%) and ethics (67%) had a level of agreement of less than the desired 70%. Because of the difference and questions about the clarity of the rubric, further refinement of the rubric was conducted. In a later, as of yet unpublished, study on interrater reliability that used the work from approximately 25 students distributed amongst 5 groups across 3 separate classes, interrater reliability ranged from 87-100% with a cumulative agreement of 90%. The 90% mirrors the findings of Jonsson and Svingby (2007) when adjacent scoring was the utilized method. Overall, the interrater reliability calculations have shown the efficacy of the rubric and the norming process.

Conclusion

The importance of the norming process cannot be overstated any time faculty are going to utilize a shared rubric. It is a crucial process that should be implemented because it "does reduce variability across graders and also builds grader confidence" (O'Connell, et al., 2016, p. 331). Without clear descriptions and authentic examples of the norming process, faculty cannot be expected to do norming and do it well, and as Crisp (2017) noted in describing a norming session at her institution, even a two-hour session will probably lead to more reliable scoring. Through detailing the CPSA norming process practiced and refined over a number of years, it is hoped that this paper has contributed in a practical manner to helping faculty understand both the need for norming and the norming process itself. A clear strength of the CPSA process is the emphasis on evidence-based discussions which are used to promote consensus amongst faculty raters. Being consensus-driven means that a shared understanding of criteria and standards is developed to the betterment of the assessment.

References

ABET. (n.d.). *History*. Retrieved from http://www.abet.org/about-abet/history/

Association of American Colleges and Universities (2014). VALUE Rubric. Retrieved from https://www.aacu.org/value-rubrics

Ater Kranov, A., Danaher, M., & Schoepp, K. (2014). A Direct Method for Teaching and Measuring Engineering Professional Skills for Global Workplace Competency: Adaptations to Computing at a University in the United Arab Emirates. Proceedings of 2014 IEEE International Conference on Interactive Collaborative Learning, Dubai, UAE, pp. 29-36. Retrieved from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7017937

Ater Kranov, A., Hauser, C., Olsen, R. G. & Girardeau, L. (2008). A direct method for teaching and assessing professional skills in engineering programs. Proceedings from the American Society for Engineering Education Annual Conference and Exposition, Pittsburgh, PA. Retrieved from www.asee.org/documents/conferences/annual/2008/ashley.pdf

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria, *Assessment & Evaluation in Higher Education*, 41:3, 466-481, DOI:10.1080/02602938.2015.1024607

Boulet, J.R., Rebbecchi, T.A., Denton, E.C., Mckinley, D., & Whelan, G.P. (2004). Assessing the written communication skills of medical school graduates. *Advances in Health Sciences Education*, 9, 47–6.

Bresciani, M.J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., & Hickmot, J. (2009). *Practical Assessment, Research & Evaluation*, 14(12). Retrieved from http://pareonline.net/getvn.asp?v=14&n=12

Burke, K. (2010). From standards to rubrics in six steps: Tools for assessing student learning. London, UK: Corwin Press.

Crisp, E. A. (2017). Calibration: Are you seeing what I's seeing? *Intersection, Winter* 1(3), 7-13.

Danaher, M., Schoepp, K., & Ater Kranov, A. (2018). The Computing Professional Skills Assessment method. Manuscript submitted for publication.

Danaher, M., Schoepp, K., & Ater Kranov, A. (2016). A new approach for assessing ABET's professional skills. *World Transactions on Engineering and Technology Education*, 14(3), 355-360. Retrieved from http://www.wiete.com.au/journals/WTE&TE/Pages/Vol.14,%20No.3%20(2016)/04-Danaher-M.pdf

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice, *Assessment & Evaluation in Higher Education*, 42(3), 347-360, DOI: 10.1080/02602938.2015.1111294

Finley, A. P. (2011). How reliable are the VALUE Rubrics? Peer Review, Fall/Winter, 31-33. Retrieved from http://209.29.151.145/peerreview/2011-2012/fall-winter/finley

Geisinger, K. F., & Foley, B. P. (2010). Considerations on the Validation of the Scoring of the 2010 FCAT

Writing Test, Buros Center for Testing The University of Nebraska-Lincoln, Retrieved from http://www.fldoe.org/core/fileparse.php/7490/urlt/fcatwritingreport.doc

Goldberg, G. L. (2014). Revising an Engineering Design Rubric: A Case Study Illustrating Principles and Practices to Ensure Technical Quality of Rubrics. *Practical Assessment, Research & Evaluation*, 19(8). Retrieved from http://pareonline.net/getvn.asp?v=19&n=8

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and Promoting Interrater Agreement of Teacher and Principal Performance Ratings. Center for Education Compensation Reform. Retrieved from http://cecr.ed.gov/pdfs/Inter_Rater.pdf

Holmes, C., & Oakleaf, M. (2013). The official (and unofficial) rules for norming rubrics successfully. The *Journal of Academic Librarianship*, 39, 599-602.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2, 130–144.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (No. RBR-8-75). Naval Technical Training Command Millington TN Research Branch. Retrieved from http://www.dtic.mil/get-tr-doc/pdf?AD=ADA006655

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., & Watty, K. (2016). Does calibration reduce variability in the assessment of accounting learning outcomes?, *Assessment & Evaluation in Higher Education*, 41(3), 331-349, DOI: 10.1080/02602938.2015.1008398

Osborn Popp, S. E., Ryan, J. M., Thompson, M. S., & Behrens, J. T. (2003). Operationalizing the Rubric: The Effect of Benchmark Selection on the Assessed Quality of Writing. Paper presented at the Annual Meeting of the American Educational Research Organization, Chicago, IL. Retrieved from https://files.eric.ed.gov/fulltext/ED481661.pdf

Popham, W. J. 1997. What's Wrong – and What's Right – with Rubrics. *Educational Leadership*, 55(2), 72–75.

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

Rhodes, T. (2009). Assessing outcomes and improving achievement: Tips and tools for using the rubrics. Washington, DC: Association of American Colleges and Universities.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4

Stevens, D. D., & Levi, A. J. (2013). Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning. Sterling, Virginia: Stylus Publishing.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, 9(2), Retrieved from http://PAREonline.net/getvn.asp?v=9&n=2 .

University of Hawaii Manoa. (2013). Creating and Using Rubrics. Retrieved from http://manoa.hawaii.edu/assessment/howto/rubrics.htm

Watty, K., Freeman, M., Howieson, B., Hancock, P., O'Connell, B., De Lange, P., & Abraham, A. (2014). Social moderation, assessment and assuring standards for accounting graduates. *Assessment & Evaluation in Higher Education*, 39(4), 461-478.

Wang, J., Engelhard Jr, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, 33, 36-47.

Appendix

CPSA 1. Students will be able to problem-solve from a computing perspective.					
0 - Missing	**1 - Emerging**	**2 - Developing**	**3 - Practicing**	**4 - Maturing**	**5 - Mastering**
Students do not identify the problem(s) in the scenario.	Students begin to define the problem(s). Potential solutions may be general and/or naive.		Students are generally successful in defining primary and secondary problems with reasonable accuracy and with justification. There is evidence that they have begun to formulate potential solutions from a computing perspective.		Students convincingly and accurately define the primary and secondary problems, providing justification. They suggest detailed and viable potential solutions from a computing perspective.
Students do not identify stakeholders.	Students identify the most obvious stakeholders. Students may state stakeholder perspectives in an inaccurate or limited way.		Students explain the perspectives of major relevant stakeholders and convey these with reasonable accuracy.		Students thoughtfully consider perspectives of diverse relevant stakeholders and articulate these with clarity and accuracy.
CPSA 2. Students will be able to work together to perform a specific task.					
0 - Missing	**1 - Emerging**	**2 - Developing**	**3 - Practicing**	**4 - Maturing**	**5 - Mastering**
Student discussion is not guided by the prompts.	Students use only a portion of the prompts to guide their discussion. Students get off task. They may be unaware that they have gotten off task or may work to get back on task but unsuccessfully.		Students use the entire set of prompts to guide their discussion. Students recognize when they get off task and work to get back on task.		Student discussion is closely aligned to the entire set of prompts. Students plan their discussion according to the prompts in order to ensure completion and thorough consideration.
Students do not acknowledge or encourage participation of others.	Students may pose individual opinions without linking to what others say. Students acknowledge the ideas of others but may too hastily defer to an opinion.		Students acknowledge, build on, clarify and/or critique and others ideas with some success. Students encourage participation of others to come to consensus.		Students clearly encourage participation from all group members, generate ideas together, actively help each other, and clarify and/or critique each other's ideas.
CPSA 3. Students will be able to evaluate, ethical, legal, and security considerations when solving a problem.					
0 - Missing	**1 - Emerging**	**2 - Developing**	**3 - Practicing**	**4 - Maturing**	**5 - Mastering**
Students do not identify ethical, legal, and security considerations.	Students give passing attention to related ethical considerations and/or may describe only the most obvious ethical considerations.		Students identify relevant ethical, legal, and security considerations in context of the problem(s).		Students clearly articulate relevant ethical, legal, and security considerations and evaluate them in the context of the problem(s).
CPSA 4. Students will be able to communicate professionally in writing.					
0 - Missing	**1 - Emerging**	**2 - Developing**	**3 - Practicing**	**4 - Maturing**	**5 - Mastering**
Students are unable to write in an accurate manner.	Student errors in grammar, punctuation, and spelling at times impedes the effectiveness of communication.		Students have few errors in grammar, punctuation, and spelling, so effective communication is seldomly impeded.		Students write clearly and have no discernable grammar, punctuation, or spelling errors.
Students do not demonstrate a professional vocabulary.	Students inconsistently demonstrate a professional vocabulary.		At times students demonstrate the vocabulary expected of a computing professional.		Students consistently demonstrate the vocabulary expected of a computing professional.
CPSA 5. Students will be able to analyze the local and global impacts of computing.					
0 - Missing	**1 - Emerging**	**2 - Developing**	**3 - Practicing**	**4 - Maturing**	**5 - Mastering**
Students do not consider either the local or global impacts of computing on	Students analyse local and/or global impacts of computing on individuals, organizations and society. Student analysis may be superficial.		Students analyse local and global impacts of computing on individuals, organizations and society. Students begin to recognize the associated		Students judiciously analyze local and global impacts of computing on individuals, organizations and society. Students recognize the associated complexities and interdependencies.

individuals, organizations and society.			complexities and interdependencies.		
CPSA 6. Students will be able to recognize when they need to seek further information to extend their knowledge.					
0 - Missing	**1 - Emerging**	**2 - Developing**	**3 - Practicing**	**4 - Maturing**	**5 - Mastering**
Students do not refer to or evaluate information presented.	Students refer to the information presented in the scenario. Students refer to the sources of information presented during the discussion.		Students evaluate the information presented in the scenario. Students evaluate the sources of information presented during the discussion.		Students critically evaluate information presented in the scenario and presented during the discussion. Examples include, but are not limited to: discussing potential and probable biases of the information sources, distinguishing fact from opinion in order to determine levels of information validity, analyzing implied information.
Students do not differentiate between what they do and do not know. Students do not demonstrate an awareness of the need to seek additional information.	Students begin to identify what they do and do not know. Students may acknowledge the need to seek additional information.		Students identify what they do and do not know. Students provide additional sources to support the discussion and extend their knowledge.		Students accurately identify the specific limits of their knowledge and how those limitations affect their analysis. Students actively seek relevant additional information and bring forth a variety of reliable sources to support the discussion and extend their knowledge.

Citation:

Schoepp, Kevin, Danaher, Maurice, & Ater Kranov, Ashley. (2018). An Effective Rubric Norming Process. *Practical Assessment, Research & Evaluation*, 23(11). Available online: http://pareonline.net/getvn.asp?v=23&n=11

Corresponding Author

Maurice Danaher
College of Technological Innovation
Zayed University
PO Box 144534
Abu Dhabi
United Arab Emirates

email: Maurice.Danaher (at) zu.ac.ae