Global Journal of Engineering Education

# Reliability and validity of the computing professional skills assessment

**Maurice Danaher†, Kevin Schoepp‡ & Anthony Rhodes†**

Zayed University, Abu Dhabi, Unite Arab Emirates†
Independent Researcher, Tulum, Mexico‡

ABSTRACT: The computing professional skills assessment (CPSA) is a way to assess the non-technical student learning outcomes for the Accreditation Board for Engineering and Technology (ABET) in the discipline of computing. These outcomes, also known as 21st Century, transferable or general education learning outcomes are recognised as essential for employment, but they have proven a challenge to assess in a direct and integrated manner. The CPSA overcomes this challenge with its scenario-based, small group, on-line discussion, where faculty raters assess the discussion transcripts according to the criteria presented in the six-part CPSA rubric. The method has been used with more than 600 computing students over a five-year period. Here, the authors present results on the reliability and validity of the instrument. Reliability was examined through evidence-based rater discussions and analysis of interrater reliability. Validity was examined through construct, content, criterion related and concurrent forms of validity. The results provide evidence that the instrument is reliable and valid.

Keywords: Accreditation, quality, learning outcomes, programme evaluation, 21st Century skills

## INTRODUCTION

The computing professional skills assessment (CPSA) is a method to assess the non-technical student learning outcomes specified by the Computing Accreditation Commission (CAC) of the Accreditation Board for Engineering and Technology (ABET). Often referred to as 21st Century, transferable or general education learning outcomes, the professional skills outcomes are outcomes, such as an ability to communicate, problem solve or work in teams that cross disciplinary boundaries. They have been cited by employers as essential skills [1], and in fact these outcomes have at times been prioritised ahead of technical skills [2].

First launched in 2013, the CPSA is a scenario-based, small group, on-line discussion, where students read a short computing-related article, and are then asked to discuss and develop solutions to the challenges raised in the scenario. The discussion transcripts are then rated at the group level by a team of faculty using the six-section analytic CPSA rubric that is aligned to the ABET professional skills outcomes. The CPSA is the only method currently available in the literature that can assess student attainment of all professional skills through a direct measure of assessment, rather than through an indirect measure, such as a perception survey or reflective essay. Further, unlike other approaches, it assesses all of the six skills with the one assessment.

The method is used to measure the attainment level of the students at each year in a programme. A measurement method needs to be both reliable and valid if its results are to be considered trustworthy [3]. In this article, the authors present an evaluation of the CPSA's reliability and validity as it currently stands after over five years of use with more than 600 computing students. The authors also demonstrate appropriate reliability and validity protocols that can and should be used when developing an assessment scored by a rubric.

## THE COMPUTING PROFESSIONAL SKILLS ASSESSMENT

The purpose of the CPSA is to determine the degree to which cohorts of students have attained the professional skills. Since it is assessment of students at the group level, it is ideal for programme assessment. The main components of the CPSA are:

1. a written scenario with a set of guiding questions;

2. a scenario development checklist;
3. the CPSA rubric including instructions; and
4. the method of implementation.

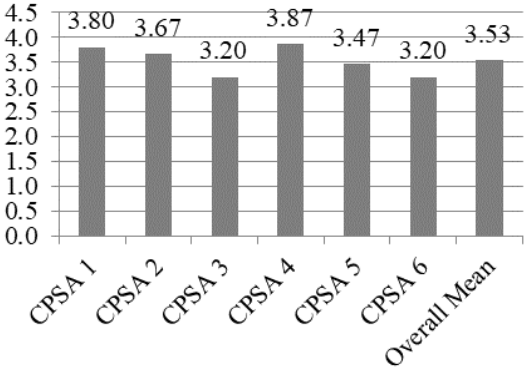
The six ABET CAC outcomes and the aligned CPSA professional skills that constitute the CPSA rubric are shown in Table 1. The CPSA professional skills have been slightly modified from ABET's versions to better fit the assessment task. Groups of students receive a score from the rubric on each of the six professional skills, and the rubric itself has criteria and sets of descriptors along a six-point scale: 0 - missing, 1 - emerging, 2 - developing, 3 - practicing, 4 - maturing, 5 - mastering. These levels of attainment are approximately aligned with year of study in the computing programme. For example, a score of 1 is the expected performance of first-year students, a score of 2 is the expected performance of second-year students, and so forth. As this is programme level assessment, scores from multiple raters across all of the groups are brought together and averaged to produce a set of scores for the entire cohort. This understanding is important later when concurrent validity of the CPSA is analysed. As Figure 1 shows from a recent cohort of fourth-year students, mean scores are calculated for each of the professional skills in order to determine areas of strengths and weaknesses - key elements in assessing programme effectiveness. For the complete CPSA instrument, including, but not limited to, the scenarios, rubric, and detailed implementation instructions, please refer to the research project Web site.

Table 1: Outcome alignment.

| ABET CAC outcomes* | CPSA professional skills learning outcomes |
|---|---|
| b. An ability to analyse a problem, and identify and define the computing requirements appropriate to its solution. | CPSA 1. Students problem-solve from a computing perspective. |
| d. An ability to function effectively on teams to accomplish a common goal. | CPSA 2. Students work together to accomplish shared goals. |
| e. An understanding of professional, ethical, legal, security and social issues and responsibilities. | CPSA 3. Students consider ethical, legal and security aspects. |
| f. An ability to communicate effectively. | CPSA 4. Students communicate professionally in writing. |
| g. An ability to analyse the local and global impact of computing on individuals, organisations and society. | CPSA 5. Students analyse the impacts of computing solutions at local and global levels. |
| h. Recognition of the need for, and an ability to, engage in continuing professional development. | CPSA 6. Students interpret, represent and seek information. |

* Labelled in accordance with ABET's alphabetical labelling

Figure 1: 4th year mean scores.




## RELIABILITY AND VALIDITY

From its inception in 2013 until today, the CPSA has been continually refined, so as to be a reliable and valid method of assessment. Instrument reliability has been measured previously through a check on interrater reliability [4], and in this article has been repeated, but in a far more comprehensive manner. In addition, evidence-based rater discussions leading to consensus have also been implemented as a way to ensure reliability of ratings. Instrument validity is essential because …*without valid assessment, students lack the opportunity to demonstrate their learning, making it difficult to assess whether the standard is being met or not met* [5]. For CPSA instrument validity a number of forms of validity have been implemented and examined. In this article, the authors describe the ways in which construct, content, criterion related and concurrent forms of validity have been applied. Together, the reliability and validity protocols that have been employed offer evidence that the CPSA is both a reliable and valid method of assessment.

## RELIABILITY

Before the validity of any instrument can be determined, the reliability of the instrument must be established. A measure can be reliable but not valid, but it cannot be valid if unreliable. Moskal and Leydens describe a reliable instrument as one that produces the same scores repeatedly - it is a consistent instrument [6]. For example, if a student took the same examination more than once without any interventions or interference from previous attempts, one would expect the student to produce a very similar score. If they did, the instrument could be considered reliable. If the student produced wildly different scores, the instrument would be inconsistent, and therefore, lacking reliability. The two methods used to determine CPSA reliability are evidence-based rater discussions and analysis of interrater reliability.

Evidence-based rater discussions are utilised after the initial round of transcript scoring has been tabulated. Where there is a disagreement of more than 1 point on the rubric, raters are required to justify their score by referencing examples from the student transcripts. Dialog between raters occurs until consensus is approached. Though perfect alignment is not always possible, consensus to within 1 point is required. Given the nature of peer review that exists within higher education, the evidence-based rater discussions have proved to be a useful way to build a shared understanding of the instrument and to obtain consensus [7].

The second way in which CPSA reliability has been determined is through calculation of interrater reliability. Because multiple raters are used to score the same CPSA transcripts from student groups, interrater reliability or the percentage of agreement amongst raters is the method used to determine this measure of reliability. Because of the relatively small sample sizes, that is the limited number of group transcripts scored at one time, interrater reliability was calculated through the simplest of methods. This was done by counting the number of transcripts assigned identical scores, and then dividing this by the total number of transcripts scored. This was done for each CPSA learning outcome, and then the overall mean score was calculated. The minimum target for acceptable interrater reliability was the 70% threshold put forth by Stemler [8]. A previous investigation into CPSA interrater reliability in 2016 was determined to be 75% [4].

As outlined above, the CPSA ratings process involves initial ratings, an evidence-based rater discussion, and then a second set of consensus scores. Table 2 presents the interrater reliability percentages for both of these rounds in the most recent implementation of the CPSA, which was a fourth-year course. In each case, the interrater reliability increased in the second set of consensus ratings with the overall cohort mean score increasing from 66% to 83%. Though the initial ratings, prior to evidence-based discussions, were below the 70% target, Cherry and Meyer suggest that a lower threshold for interrater reliability is acceptable if assessments are conducted at the group, as in the CPSA, rather than individual level [9]. Even though Krippendorff argues that *…the only publishable reliability is the one measured before the reconciliation of disagreements* [10], the evidence-based rater discussions that are built into the CPSA assessment method make this a moot point. Post-consensus, ratings ranged from 73 to 100%, and with an overall cohort mean score of 83%, there is an increase in interrater reliability from the value of 75% in 2016.

Table 2: 4th year pre- and post-interrater agreement percentages.

| | CPSA 1 | CPSA 2 | CPSA 3 | CPSA 4 | CPSA 5 | CPSA 6 | Overall cohort mean |
|---|---|---|---|---|---|---|---|
| 2018 initial ratings | 53 | 67 | 53 | 87 | 60 | 73 | 66 |
| 2018 final ratings | 80 | 73 | 73 | 93 | 100 | 80 | 83 |
| Increase | 27 | 6 | 20 | 6 | 40 | 7 | 17 |

Table 3: Multiple cohort interrater agreement percentages.

| | CPSA 1 | CPSA 2 | CPSA 3 | CPSA 4 | CPSA 5 | CPSA 6 | Overall cohort mean |
|---|---|---|---|---|---|---|---|
| 2016 - 3rd year | 72 | 61 | 67 | 83 | 83 | 83 | 75 |
| 2016 - Masters | 67 | 84 | 92 | 92 | 75 | 59 | 78 |
| 2018 - 2nd year | 100 | 87 | 87 | 87 | 100 | 100 | 93 |
| 2018 - 4th year | 80 | 73 | 73 | 93 | 100 | 80 | 83 |
| Outcome mean | 84 | 81 | 76 | 91 | 88 | 84 | 82 |

Because *interrater reliability refers to the level of agreement between a particular set of judges using a particular instrument at a particular time* [8], it is important to examine interrater reliability at regular intervals. Table 3 shows the interrater reliability rates from 2016 to 2018, which cover second, third, fourth and Masters levels. As is clear, the overall cohort mean percentages of agreement have always been above the 70% threshold with a low of 75% recorded in 2016 with a cohort of third-year students, and a high of 93% recorded with a cohort of second-year students in 2018. Examining the interrater reliability for each of the separate CPSA learning outcomes, in only four cases has the 70% threshold not been met out of a total of 24 data points, and on four occasions 100% interrater reliability was achieved. Over time the implementations of the CPSA from 2016 to 2018 show that not only has the overall cohort mean percentage of agreement exceeded expected thresholds, but also each individual CPSA learning outcome mean (see last row in Table 3) exceeded the expected thresholds. This demonstrates strong evidence for the reliability of the instrument.

## CONSTRUCT VALIDITY

Construct validity refers to whether the scores of a test or instrument measure the distinct dimension (construct) they are intended to measure [11]. Specific to rubrics, Moskal and Leydens pointed out that it is essential that a rubric's evaluation criteria only measure elements that concern the construct, rather than those superfluous to the construct [6]. For example, if a construct being assessed is to measure the ability of students to effectively communicate through written text, there should be no criteria or descriptors that focus on the impact of computing locally and globally. The construct under examination, effective written communication, needs to be the focus of the criteria and descriptors. With the CPSA, an iterative process has been undertaken to ensure the construct validity of the rubric. This has led to a significant number of modifications to the rubric over time. Every time a rating session has been conducted, if any issues surrounding the rubric definitions, criteria or descriptors occur, these are noted and changes are proposed as deemed necessary. For example, one of the criteria from *CPSA 2. Students work together to accomplish shared goals* was labelled *Task Orientation* which at times caused confusion. Because of this, it has been relabelled to the more accurate phrase - *Prompts* since the focus of the criterion is to determine the degree to which students respond to the discussion prompts. Over time from 2014 to 2018 the rubric, while staying true to its underlying constructs, has undergone numerous modifications and improvements to increase its construct validity.

## CONTENT VALIDITY

Closely related to construct validity, content validity refers to the degree to which an assessment is aligned with the content domain it seeks to measure [11]. An initial step in the process to ensure content validity is to accurately define the constructs under examination, and then to determine the purpose of the assessment. This is important, because one could imagine an assessment that accurately measures the ability to problem solve, but if that assessment is then used to measure someone's ability communicate effectively, there is an obvious misalignment. Experts are often utilised to evaluate the content validity of an instrument.

Specific to the CPSA, the constructs under examination are adaptations of ABET's CAC student outcomes, and these adaptations also include an expanded definition to further define the content. For example, the alignment between the outcomes for CAC e and CPSA 3 is as follows:

| | |
|---|---|
| *ABET CAC e* | An understanding of professional, ethical, legal, security and social issues and responsibilities. |
| *CPSA 3* | Students consider ethical, legal and security aspects. |
| *CPSA 3 Definition* | Students identify relevant ethical, legal and security aspects in their discussion of problems and potential solutions. |

The purpose of ABET's outcomes are to show what students are able to do at the completion of a programme; hence, the purpose of the CPSA is to assess student attainment of the learning outcomes at the programme level, not at the level of the individual student. Experts also played a major role in the development of the CPSA. All three CPSA authors drafted and re-worked the CPSA learning outcomes and the definitions, because of their subject matter expertise. Another way in which subject matter expertise contributed to instrument content validity was that a language expert was engaged to review the language of the scenarios to ensure they were appropriate to the reading levels of the students. Because the students were second language learners, readability was set to grade 12 on the Flesch-Kincaid scale. The authors wanted to ensure that the CPSA assessed the six learning outcomes and not the reading ability of the students.

## CRITERION RELATED VALIDITY

Moskal and Leydens describe the evidence for criterion related validity as evidence that shows how the results of an assessment relate to future events [6]. Within ABET's realm, this is often the validity of an assessment as it relates to skills and knowledge within the pertinent professions. Moskal and Leydens state that in a rubric-based assessment *…the scoring criteria should address the components of the assessment activity that are directly related to practices in the field. In other words, high scores on the assessment activity should suggest high performance outside the classroom or at the future work place* [6].

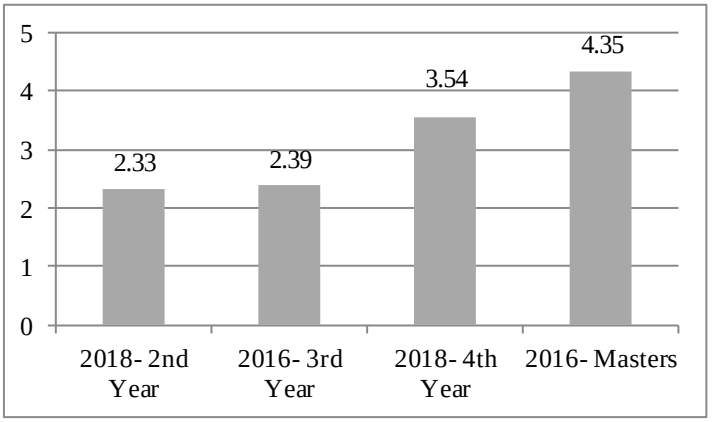
The tight alignment between the CPSA learning outcomes and the ABET CAC student outcomes makes criterion related validity a strength of the CPSA. The ABET outcomes have been developed as a benchmark of what graduates are expected to do at the completion of an academic programme; more specifically, that accredited academic programmes *produce graduates prepared to enter a global workforce* [12]. The CPSA assesses the professional skills aspect for computing students through the tight ABET CAC alignment.

## CONCURRENT VALIDITY

A type of criterion related validity, concurrent validity is any *…operationalization's ability to distinguish between groups that it should theoretically be able to distinguish between* [13]. Hence, the CPSA rubric should be able to differentiate between student groups in the first, second, third or fourth years of undergraduate study and at the Masters level, because the target levels of attainment within the rubric are roughly aligned to the year of study. For example,

first-year students should achieve: 1 - emerging; second-year students, 2 - developing; third-year students, 3 - practicing; fourth-year students, 4 - maturing and Masters students, 5 - mastering. Of course, perfect alignment between year of study and rubric level cannot be expected, because of variability in student performance. It is actually expected that different cohorts of students perform better or worse than projected. Nonetheless, some level of alignment is expected if an instrument, such as the CPSA exhibits concurrent validity. Using the overall cohort means, Figure 2 shows how the anticipated alignment has actually occurred for two reasons. First, there is an increase in student scores as cohorts progress through the academic programme. Second, scores are near the expected numerical target. Only between the second and third-year cohorts, are the scores extremely similar to one another. It could be argued that the second-year cohort performed above expectations since they broke the 2.0 threshold, while the third-year cohort performed slightly below what was anticipated in that they were 0.61 away from a 3.0. Nonetheless, a rate of increase is present from year to year and scores fall within anticipated ranges. A major argument against concurrent validity would exist if second-year students received higher scores than Masters students, for example.

Figure 2: Concurrent validity overall cohort means.




Further examination of concurrent validity is demonstrated through Table 4 which shows the mean scores and rank (1 for lowest mean and 4 for highest mean) for every cohort across each of the CPSA learning outcomes. For CPSA 1 and 2, the ranks were as expected, only the scores for the second-year students were higher than the target of 2.0 with scores of 2.20 and 2.40, respectively. Both the ranks and range of scores were exactly as anticipated for CPSA 3. An obvious anomaly emerged from CPSA 4, because the second-year students, not only outrank the third-year students, their mean score is also above the 3.0 threshold which is the target for third-year students. CPSA 5 again followed the anticipated pattern of mean scores and ranks with no apparent inconsistencies. With CPSA 6, there were a number of unexpected results. First, Masters students did not exceed 4.0 and were more aligned with expectations of fourth-year undergraduates with a mean score of 3.67. Second, second-year students outranked the third-year students. Finally, third-year students failed to meet the target of even second-year students with a 1.67, and second-year students with a 2.80 were more in line with anticipated scores for third-year students. Overall, given that some cohorts could be expected to either underperform or over perform, results from Table 4 are quite aligned with performance expectations and provide a strong case for the CPSA demonstrating concurrent validity.

Table 4: Concurrent validity mean scores and rank.

| | CPSA 1 | CPSA 2 | CPSA 3 | CPSA 4 | CPSA 5 | CPSA 6 |
|---|---|---|---|---|---|---|
| 2018 - 2nd year | 2.20- 1 | 2.40- 1 | 2.00-1 | 3.20-2 | 1.40- 1 | 2.80-2 |
| 2016 - 3rd year | 2.33-2 | 2.67-2 | 2.67-2 | 2.83- 1 | 2.17-2 | 1.67-1 |
| 2018 - 4th year | 3.80-3 | 3.67-3 | 3.20-3 | 3.87-3 | 3.47-3 | 3.20-3 |
| 2016 - Masters | 4.67-4 | 4.25-4 | 4.67-4 | 4.58-4 | 4.25-4 | 3.67- 4 |

## CONCLUSIONS

Computing education needs quality assessment methods that provide reliable and valid evidence of student learning of the professional skills. The development of the CPSA and the work done to ensure that it is a reliable and valid instrument has been undertaken to help meet this need. Here, the authors have outlined the CPSA method. They have demonstrated CPSA reliability through the implementation of evidence-based rater discussions and a check on interrater reliability, and established CPSA validity through construct, content, criterion related and concurrent forms of validity. Readers are invited to visit the Web site - www.cpsa.ae - where complete details of the method may be found, as well as all resources needed for its use.

## REFERENCES

1. National Association of Colleges and Employers, Employers Identify Four *Must Have* Career Readiness Competencies for College Graduates (2016), 31 March 2019, http://www.naceweb.org/career-readiness/competencies/employers-identify-four-must-have-career-readiness-competencies-for-college-graduates/

2. Association for American Colleges and Universities. It takes More than a Major: Employer Priorities for College Learning and Student Success (2013), 12 April 2019, http://www.aacu.org/leap/presidentstrust/compact/2013 SurveySummary.cfm
3. Rhodes, A., Danaher, M.M., Ater Kranov, A. and Isaacson, L., Measuring attainment of foundation skills in general education at a public university in the United Arab Emirates. *World Trans. on Engng. and Technol. Educ.*, 14, **4**, 506-512 (2016).
4. Danaher, M., Schoepp, K. and Ater Kranov, A., A new approach for assessing ABET's professional skills in computing. *World Trans. on Engng. and Technol. Educ.*, 14, **3,** 355-360 (2016).
5. Watty, K., Freeman, M., Howieson, B., Hancock, P., O'Connell, B., De Lange, P. and Abraham, A., Social moderation, assessment and assuring standards for accounting graduates. *Assessment & Evaluation in Higher Educ.*, 39, **4**, 461-478 (2014).
6. Moskal, B.M. and Leydens, J.A., Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, 7, **10** (2000).
7. Schoepp, K., Danaher, M. and Ater Kranov, A., An effective rubric norming process. *Practical Assessment, Research and Evaluation*, 23, **11** (2018).
8. Stemler, S.E., A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9, **4** (2004).
9. Cherry, R.D. and Meyer, P.R., *Reliability Issues in Holistic Assessment.* In: Williamson, M.W. and Huot, B.A. (Eds), Validating Holistic Scoring for Writing Assessment. Cresskill: Hampton Press (1993).
10. Krippendorff, K., *Content Analysis: an Introduction to its Methodology*. Thousand Oaks: Sage Publishers (2013).
11. Markus, K.A. and Lin, C., *Construct Validity.* In: Salkind, N.J. (Ed), The Encyclopedia of Research Design. Thousand Oaks: Sage Publishers (2012).
12. ABET, About ABET (n.d.), 23 June 2018, http://www.abet.org/about-abet/
13. Trochim, M.K., Measurement Validity Types (2006), 1 May 2019, http://www.socialresearchmethods.net/kb/ measval.php

BIOGRAPHIES

Maurice Danaher is an Associate Professor in the College of Technological Innovation at Zayed University, United Arab Emirates. He received his PhD in computing and information systems from Swinburne University of Technology, Melbourne, Australia. His research has been in the areas of information technology and education. In IT, he has published in security, artificial intelligence and computer graphics. In recent years his focus in education research has moved towards issues related to quality in education. He focusses on quality assessment, and teaching and assessing the 21st Century skills.

Kevin Schoepp is currently an independent researcher, but was the Director of Educational Effectiveness at Zayed University. His role in educational effectiveness included learning outcomes assessment, accreditation and programme review. He has a doctorate in higher education leadership and a Masters degree in educational technology from the University of Calgary, and he has an undergraduate and Masters degree from the University of Alberta. His current research interests are in the development of effective and sustainable assessment and accreditation processes, creating a culture of assessment to foster continuous improvement, and using on-line discussions to assess ABET's professional skills' learning outcomes.

Tony Rhodes holds a PhD in information security from Queensland University of Technology (QUT), Australia. He is currently employed as Chair, Department of Computing and Applied Technology, College of Technological Innovation, Zayed University, United Arab Emirates. In addition to his research interests in information security (security management, access control and database security), he is an active researcher in the domain of teaching and learning as it relates to the attainment of professional/employment skills required by undergraduates as they make the transition from higher education to the workplace.