WMU J Marit Affairs (2016) 15:317–336
DOI 10.1007/s13437-015-0094-0

IAMU SECTION ARTICLE

Authentic assessment in seafarer education: using literature review to investigate its validity and reliability through rubrics

Samrat Ghosh[1] · Marcus Bowles[1] · Dev Ranmuthugala[1] · Ben Brooks[1]

Received: 27 May 2015 / Accepted: 5 October 2015 / Published online: 14 October 2015

Abstract With the Standards of Training, Certification and Watchkeeping Convention 1995 (STCW'95) moving seafarer training towards outcome-based education (OBE), emphasis has shifted to assessment practices that will allow seafarer students to demonstrate their ability to perform workplace tasks at standards described in the STCW Code. This paper argues that authentic assessment comprising of performance-based tasks applied in real-world and meaningful contexts can provide a holistic approach to competence assessment for seafarers. But, authentic assessment can capture essential aspects of workplace tasks and result in consistency of student performance in different contexts only if they are valid and reliable. Rubrics as assessment tools are known to increase validity and reliability of assessments; however, it can do so only if different aspects of its own validity and reliability have been addressed. A literature review undertaken for this paper has uncovered an absence of academic investigation and empirical study on the different aspects of validity and reliability of authentic assessment through assessment rubrics. Moreover, there exists an even greater absence of global research on authentic assessment in the area of seafarer training. Through an investigation of authentic assessment, this research has uncovered the importance of using valid and reliable rubrics in order to improve not only the assessment process but also the tools and methods used to support the valid, reliable, and authentic assessment of outcomes achieved in the learning process. Future research aims to offer insights into improving the validity and reliability of rubrics and to empirically investigate how they can be used in authentic assessment within the confines of the STCW Code, in particular, to improve seafarer training practices, student engagement, resulting learning outcomes, and employer and regulator

✉ Samrat Ghosh
sghosh@utas.edu.au

[1] Australian Maritime College (UTAS), Launceston 7250, Australia

satisfaction with the attainment of the standards stipulated in the STCW Code to produce an evidence of competence.

Keywords Authentic assessment · Seafarer education and training · Rubrics · Validity · Reliability

1 Introduction

In education, assessment can be defined as 'a systematic collection, review, and use of information' (Walvoord 2004) to acquire feedback about a student's progress and achievements, the effectiveness of teaching and instruction, and the attainment of course outcomes (University of Tasmania (UTAS) 2011), while fulfilling the overall goal of improving student learning (Palomba and Banta 1999). In outcome-based education (OBE) such as vocational education and training (VET) or competency-based training (CBT), assessments also provide feedback about the attainment of minimum standards by students that are essentially required for the workplace (Brady 1997; p.10). Standards in such cases become the outcomes (Burke 2011) or more correctly 'learning outcomes' establishing what the students should be able to demonstrate at the end of the learning period (Driscoll and Wood 2007). Students direct their learning efforts towards 'outcome' attainment, and assessors are guided on what they are supposed to measure via assessments. The evidence produced from the assessments can be used by educators to not only improve teaching practices by identifying learning needs but also to meet accountability requirements by providing feedback to stakeholders on the learners' progress towards achievement of standards (Brindley 1998).

Standards for the occupational practice of seafaring are provided through the Standards of Training, Certification and Watchkeeping (STCW) Code of the STCW Convention that was introduced by the International Maritime Organization (IMO) in 1978 (then known as STCW'78). The STCW'78 was essentially knowledge-based comprising a syllabus for a quantifying examination instead of focusing on skills and abilities necessary to perform workplace tasks (Morrison 1997). The IMO revised the STCW Code through the 1995 amendments (since known as STCW'95) intending to fundamentally improve the training mandate by making it outcome-based. As a requirement of OBE and for the purposes of the certification and licencing, seafarers are required to demonstrate the achievement of the STCW standards through assessments.

Demonstration of attainment of competence that resembles workplace standards may require assessments that assess not only students' progress against outcome attainment but also their ability to perform workplace tasks. Evidence produced through traditional assessment tasks such as multiple choice questions or oral examinations can provide indicators for students' mastery of content knowledge but may not be able to adequately capture different aspects of a complex student performance resembling workplace tasks (Montgomery 2002). Such performance can be captured through assessment rubrics which comprise of individual and essential dimensions of performance known as criteria along with standards for levels of performance against those criteria (Jonsson and Svingby 2007). Rubrics involve creating a standard and a descriptive statement that illustrates how the standard is to be achieved (Cooper and

Gargan 2009). Rubrics may report on outcome attainment, but the validation of attainment is achieved through the assessment process (Davis et al. 2007).

To determine if the intended outcomes have been achieved and to collate evidence of the same, assessors need to decide whether the selected assessment methods will adequately allow for evaluation and demonstration of the students' learning outcomes (Moskal 2000). The quality of the information provided on outcomes attainment by the rubrics will only be as good as the assessments on which the reporting is based (Brindley 1998). The ability to perform workplace tasks should be assessed through assessment methods that resemble professional scenarios. Hence, fidelity of context to conditions in which the professional skill would be applied becomes an important element of assessment methods adopted. Such performance-based assessments applied in real-world contexts have often been described as authentic assessments (Herrington and Herrington 1998; Reeves and Okey 1996; Wiggins 1993; Meyer 1992).

However, fidelity of context cannot alone assure that essential aspects and constructs of professional competencies are being accurately assessed. Assessments should be valid and reliable to do so. Validity refers to the extent to which the evidence produced through assessments supports the inferences made about the student's competencies and whether such inferences are being interpreted in appropriate contexts (Moskal and Leydens 2000). On the other hand, reliability refers to the consistency of assessment scores obtained every time the same competencies are assessed irrespective of the scorer, time period between the assessments, and the contextual and individual learning variables under which the assessments occur (Moskal and Leydens 2000). Rubrics provide clear statements on learning and performance expectations for both educators and students. Such statements can then be used to assess if intended outcomes were achieved by students, educators, and assessors. Hence, rubrics are highly regarded as tools that increase validity and reliability in assessments (Rezaei and Lovorn 2010; Jonsson and Svingby 2007; Silvestri and Oescher 2006).

This paper establishes the importance of using rubrics as an authentic assessment instrument for assessing outcomes that represent workplace tasks. Authentic assessment is defined collating all the characteristics used by major authors in the field. Validity and reliability are then established as essential criteria for measuring the effectiveness of assessment methods by researchers. Based on an extensive literature review in the area of authentic assessment, this paper explores the practices adopted in the past to improve the validity and reliability of authentic assessment when rubrics are used as an assessment instrument. The review uncovers a lack of holistic approach in addressing both validity and reliability aspects of authentic assessment and an absence of global research on authentic assessment in the field of seafarer education and training.

2 Definitions

2.1 Authentic assessment

The idea of 'authenticity' in education was conceived and developed in response to increasing accountability to stakeholders. The movement started in the 1980s in the high schools of USA. The term 'authentic' was first linked to student achievement by

Archbald and Newmann (1988), requiring them to demonstrate outcomes beyond the school learning environment in an applied/work context. Wiggins (1989) related the term to student assessment while promoting authentic assessment as a process that required student performances (Wiggins 1990) at standards expected in the professional field. Unlike traditional tests that produced transcripts with unclear information of actual competence, evidence of student performance at workplace standards would improve accountability to stakeholders.

Authentic assessment is often used interchangeably with performance assessment as it imbibes some of the characteristics of the latter, but they are not synonymous (Marzano et al. 1993). For example, all authentic assessments require a performance of some kind, but not all performance-based assessments are conducted in authentic or real-world contexts (Meyer 1992). Palm (2008) provides a detailed classification of meanings describing the similarities and wide range of differences between the meanings of each concept. Authentic and performance assessments are known as types of 'alternative assessments' to traditional assessments (*Dikli* 2003). Traditional assessments include pen and paper testing, multiple choice questions (MCQs), and oral examinations. Cumming and Maxwell (1999) show that characteristics of authentic assessment can also be found in other assessments, such as problem-based and competency-based assessments, but provide clear distinction between them. For example, they explain that authentic assessment is based on theories of learning where performance of tasks occurs in genuine workplace or contextually similar situations. On the other hand, competency-based assessments are based on the theory of vocational education where assessment tasks should represent workplace tasks but can be performed in individual components and not necessarily integrated into one holistic task. Authentic assessments have also been called dynamic assessments (Chance 1997; Butler 1999) due to its dynamic nature of evolving to address student learning needs.

This paper defines authentic assessment by collating the characteristics provided by the most commonly cited authors in the area (Table 1). The exact number of citations for the individual papers has been obtained from the website of Google Scholar.

Based on the characteristics provided in Table 1, authentic assessment herein will encompass tasks *resulting in outcomes* in a *real-world context* that require an *integration of competence* to solve *forward looking questions* and *ill-structured problems*, processes that require *performance criteria to be provided beforehand* and *evidence of competence to be collected by the student*, and outcomes that result in *valid and reliable student performance, contextual and multiple evidence of competence, higher student engagement, and transfer of skills to different contexts.*

2.2 Rubrics

Rubrics (an example shown in Table 2) are assessment tools that comprise of individual and essential dimensions of performance known as criteria along with standards for levels of performance against those criteria (Jonsson and Svingby 2007). Although the terms 'criteria' and 'standard' is sometimes used interchangeably, they have distinct meanings (Sadler 2005). The definitions provided by Sadler (2005) and Spady (1994) provide a robust basis for distinguishing the terms. Standards are defined as levels of definite attainment and sets of qualities established by authority, custom, or consensus by which student performance is judged, whereas criteria are essential attributes or rules

Table 1 Characteristics of authentic assessment defined by most commonly cited authors

Major author/year	No. of citations	Context of study	Real-world context	Integration of competence	Known performance criteria	Valid and reliable performance	Forward looking questions
Wiggins (1989)	939	High schools	✓	✓	✓	✓	✓
Wiggins (1990)	383		✓	✓	✓	✓	✓
Wiggins (1993)	364		✓	✓	✓	✓	✓
Wiggins (1998)	956		✓	✓	✓	✓	✓
Archbald (1991)	27	Schools			✓		
Darling-Hammond and Snyder (2000)	397	Teacher education	✓	✓	✓		
Gulikers et al. (2004a)	212	Nursing education	✓	✓	✓		
Gulikers et al. (2004b)	17		✓	✓	✓		
Gulikers et al. (2006)	33		✓	✓	✓		
Gulikers et al. (2008)	23		✓	✓	✓		
Gulikers (2006)	10	Nursing education	✓	✓	✓		

Major author/year	III-structured problems	Evidence of competence collected by student	Tasks resulting in outcomes	Contextual evidence of competence	Multiple indicators of competence	Promote student engagement	Allow transfer to different contexts
Wiggins (1989)	✓	✓			✓		
Wiggins (1990)	✓	✓			✓		
Wiggins (1993)	✓	✓			✓		
Wiggins (1998)	✓	✓			✓		
Archbald (1991)			✓			✓	✓

Table 1 (continued)

Major author/year	III-structured problems	Evidence of competence collected by student	Tasks resulting in outcomes	Contextual evidence of competence	Multiple indicators of competence	Promote student engagement	Allow transfer to different contexts
Darling-Hammond and Snyder (2000)					✓		
Gulikers et al. (2004a)			✓	✓			
Gulikers et al. (2004b)							
Gulikers et al. (2006)			✓	✓			
Gulikers et al. (2008)			✓	✓			
Gulikers (2006)			✓	✓			

used for judging the completeness and quality of standards. Table 2 provides an example of how a rubric may be designed for the unit of competence of 'Prevent, control, and fight fires on board' at the operational level from the STCW'95 Code. The move of seafarer training to OBE has shifted the emphasis to demonstration of competence requiring the intended learning outcomes (ILOs) be established and communicated to students beforehand to make the learning process transparent (Biggs and Tang 2007). As assessment rubrics communicate standards and the feedback for its achievement, they are an essential tool to OBE (Reddy 2007).

Without rubrics, students have no guidelines towards achievement or to understand the teacher's feedback comments (Montgomery 2002) on outcomes achieved. For example, using a focus group discussion involving fourteen undergraduate students, Andrade and Du (2005) found the use of rubrics to be very effective in providing performance expectations and feedback about achievement of standards in teacher education. However, using rubrics to communicate standards achieved by students in professional education also requires assessment methods such as authentic assessment that can capture such standards.

Table 2 Example of how a rubric may be constructed for the STCW unit of competence of 'Prevent, control, and fight fires on board' at the operational level

Criteria	Standard 1 (performance deemed insufficient to be competent at operational level)	Standard 2 (performance meeting minimum required to be deemed competent at operational level)	Standard 3 (performance beyond minimum required to be deemed competent at operational level)
Identify the class of fire and choose the correct extinguishing system	Unable to identify the class of fire and/or choose the correct extinguishing system	Able to identify the class of fire and choose the correct extinguishing system	Able to identify the class of fire and choose the correct extinguishing system; Able to identify alternate extinguishing systems for the class of fire
Operate the fire extinguisher as per the manufacturer's instructions	Unable to operate the extinguisher as per the manufacturer's instructions	Able to operate the fire extinguisher as per the manufacturer's instructions	Able to operate the fire extinguisher as per the manufacturer's instructions; Able to demonstrate adoption of measures to prevent the spread of fire and its reoccurrence
Wear the fireman's outfit as per the manufacturer's instructions and extinguish the fire	Unable to wear the fireman's outfit as per the manufacturer's instructions and extinguish the fire	Able to wear the fireman's outfit as per the manufacturer's instructions and extinguish the fire	Able to wear and use the fireman's outfit as per the manufacturer's instructions and extinguish the fire; Able to demonstrate adoption of measures for the care and maintenance of the fireman's outfit for reuse

Traditional assessments such as multiple choice questions and oral examinations assess the ability to recall facts and some of the applied skills (Archbald 1991) but fail to assess essential behaviour-based attributes (Wiggins 1992). An individual must develop along with technical skills and knowledge that together define professional competence (Sampson and Fytros 2008). Assessment of professional competence can be captured through authentic assessment tasks based on meaningful contexts and applied in real-world or contextually resembling real-world settings. However, professional competence is developed and assessed under specific contexts in educational settings. Transfer of performance or competence to perform individual components of a task to a holistic performance of the task where integration of competence is required cannot be assumed (Cumming and Maxwell 1999). According to Cumming and Maxwell (1999), learning and assessment need to be contextualised to make it relevant and meaningful for students. Meaningful context cannot only provide motivational benefits to student learning but also a clear understanding of learning that can or cannot be transferred to different contextual scenarios. If real-life contexts and complexities (task-centred approach) cannot be created in assessments, they should then focus on the selected constructs (construct-centred approach) of knowledge and skills (Messick 1994). For example, assessments designed in maritime education and training (MET) institutes may not be able to assess a student's competence to manage large crowds as is required on passenger ships, but they may be designed to assess a student's competence to do so through their ability to analyse risks associated with such management or developing crowd management plans. Although such assessments may take place in controlled situations, the authenticity will be reflected through ways in which the same skills would be applied in real-life contexts (Messick 1994). The standard of learning achieved in the real-world contexts may be communicated via rubrics, making it an important authentic assessment instrument for assessing outcomes that represent workplace tasks.

3 Authentic assessment

3.1 Aligning authentic assessment with rubrics

One of the key characteristics requires authentic assessment to provide performance criteria to students beforehand, which can be done through the use of rubrics. Provision of clear expectations of standards of performance via rubrics allows students to learn and educators to adopt appropriate instructional strategies to guide students towards the achievement of the desired outcomes (Archbald 1991). The use of summative examinations at the end of the learning period represents the final judgement of the students' performance and is often too late to make any changes to the learning strategies. Authentic assessment methods that are based on ongoing use of formative assessments may be more suitable to provide diagnostic feedback and make adjustments to improve the learning process (Burke 2011).

Hence, the alignment of the learning, teaching, and instruction process towards the achievement of outcomes creates constructive alignment (Biggs and Tang 2007). Constructive alignment comes from the constructivist theory (Biggs and Tang 2011), where the student is not a mere receiver of knowledge but is also actively involved in the construction of it while progressing in learning. Newmann et al. (1996, 1995) and Cooperstein and Kocevar-Weidinger (2004) connected authentic assessment to the constructivist way of learning. Although principles of constructivism can allow everyone to construct meaningful learning, Newmann et al. (1996) recommended that high intellectual standards provided through rubrics in authentic assessment can promote highly intellectual construction of knowledge and meaning leading to superior learning and performance that would require students to use higher-order cognitive skills. In the current educational environment of the twenty-first century, assessments should not only capture the content knowledge or the professional skills but also higher-order skills (Burke 2011) of problem-solving, critical thinking, leadership, and team-working. According to Wiggins (1989), assessments should not only monitor standards but also set them to reveal achievement of higher-order skills which may not be quantified but is a necessity in a work context. Traditional assessments are not always performance-based nor can they be always creatively designed to encourage demonstration of higher-order skills. For example, a study by Brawley (2009) that involved authentic assessment of twenty-four students in early childhood showed that authentic assessments, when designed properly, are a better way to determine the higher-order thinking skills (as defined by Bloom's taxonomy) required to complete a task. Creating authentic experience for students correctly becomes central to designing authentic assessment.

3.2 Validity and reliability of authentic assessment

Advances in technology such as simulators, web-learning, multimedia, etc. have allowed many researchers (Neely and Tucker 2012; Neo et al. 2012; Osborne et al. 2013; Scholtz 2007) to use such technology in the area of authentic assessment to create authentic experiences that can replicate real-world tasks for the students. However, Messick (1996) was not convinced that authentic assessments can ever fully represent real-world tasks in educational settings. Messick believed assessments are prone to threats of validity which emphasises the appropriateness of assessment tasks as effective measures of intended learning outcomes (Rhodes and Finley 2013). Because authentic assessments have a high fidelity to real-world contexts, it does not necessarily lead to the conclusion that they are more valid than traditional examinations. Assessment methods should be judged by established criteria for judging the technical adequacy of measures. Key among these criteria are the concepts of validity and reliability (Linn et al. 1991).

Validity and reliability are crucial to the acceptance of authentic assessment (or rubrics as an assessment tool) as an accurate measure of knowledge, skills, and behaviours (Stevens 2013). There are numerous extraneous variables that affect the validity and reliability of the rubrics when used an assessment instrument (Taylor 2011). If these variables are not addressed, then the validity and reliability of the assessment and the resulting outcomes becomes questionable (Olfos and Zulantay 2007).

3.3 Validity and reliability of rubrics

In the area of education, validity is not seen as a property of the assessment but how the results have been interpreted (Jonsson and Svingby 2007). Validity refers to the degree to which evidence produced from assessments support the interpretations made about student's competencies. Table 3 describes the three types of evidence that are commonly examined to support the validity of an assessment instrument: content, criterion, and construct (Moskal and Leydens 2000).

It is extremely difficult to construct an assessment which is truly valid in measuring what it is supposed to measure (Finch 2002). For example, an assessment designed to assess a student's ability to fight fires may not be able to effectively measure personal or professional behaviours (such as creativity and critical thinking) associated with the task performance. According to Messick (1996), it is hard for assessments to achieve complete validity, but he believed that the threats to validity can be minimised by ensuring that assessments do not contain anything that is irrelevant to the measurement of the desired outcomes. For example, assessments designed to assess a student's ability to fight fires should not include pen and paper testing in classrooms which are irrelevant to the measurement of either the task performance or behaviours associated with it.

Does this mean that relevant and authentic scenarios can insure validity?

Capturing a more authentic performance does not insure validity (Stevens 2013). For example, HoepfL (2000) pointed out that creating standards for authentic assessments is a challenging task which may suffer from 'Construct underrepresentation' if the standards fail to assess essential dimensions of knowledge and skills or 'Construct-irrelevant variance' if the standards require tasks that are not relevant to measuring the desired competencies (Messick 1995). Assessments are valid if they effectively measure the intended learning outcome it was designed to assess. Whether assessments effectively measure the intended learning, outcomes cannot be based on the subjective judgement of whether questions appear to do so, known as face validity (Drost 2011). Drost (2011) explains that although face validity is important for credibility to stakeholders, it is the weakest and least scientific form of establishing validity for assessments.

For effective measurement, outcomes should be accompanied by the essential criteria and the levels of performance by which the performance would be judged (Mueller 2005). The criteria and the levels are usually combined into a rubric, which forms a scoring guide for the assessment making it easier for educators to define what is being measured through assessments and how the score is to be interpreted (Emery 2001). Scoring without specific guidelines may lead to subjective judgements. Rubrics can be used to improve the objectivity of scoring by specifying the same criteria and standards to be applied to all students' work for scoring by either individual or multiple

Table 3 Three types of evidence commonly examined to support the validity of an assessment

Validity		
Content validity: extent to which the assessment instrument provides a representative sample of the content domain in the area of interest (Lynch 2003).	Criterion validity: extent to which a student's performance on a test accurately predicts the student's performance on an external criterion (Lynch 2003).	Constuct validity: extent to which the assessment measures the theoretical construct on processes that are internal to an individual (Moskal and Leydens 2000).

assessors (Dennison et al. 2015). For example, according to Jonsson and Svingby (2007), one widely cited effect of rubrics in the areas of authentic and performance-based assessments is the consistency of judgement and scoring across students, tasks, and different raters (scorers). Consistency of assessment scores obtained every time the same competencies are assessed irrespective of the scorer, time period between the assessments, and the context under which the assessments occurred is referred to as reliability (Moskal and Leydens 2000). Table 4 provides the different types of reliability testing conducted in the area of education.

Ideally, an assessment should produce similar results independent of the scorer and the context of assessment. But, is this obtainable?

The more consistent the scores are over different scorers and contexts, the more reliable the assessment is thought to be. Methodologically, sound assessment instruments should have acceptable levels of both validity and reliability (Rhodes and Finley 2013). For example, the study by Vendlinski et al. (2002) used rubrics to authentically assess 134 first-year high school chemistry students to achieve valid inferences of a student's content understanding, while not allowing the score to be affected by gender, ethnic, or socioeconomic bias.

The validity of the results and the strength of the rubric as an assessment instrument are evidenced by positive results on a variety of reliability tests (Diller and Phelps 2008). Performance-based assessments like authentic assessment face the problem of obtaining reliability (Lynch 2003). Issues such as lack of reliability, inconsistency in assessment design and grading, and potential for grading bias remain important challenges with authentic assessment (Rhodes and Finley 2013). Authentic assessments represent real-world tasks as valid indicators of workplace competence which should be consistent irrespective of the context or scorer. Such consistency can only be proved through reliability. Hence, authentic assessments should achieve both validity and reliability.

Because it can be difficult to establish whether an assessment instrument truly captures the outcome for which it is intended or whether the outcome can be consistently measured, it is preferable for instruments to demonstrate more than one type of validity (Rhodes and Finley 2013) and reliability. There are numerous aspects of validity and reliability investigated and reported in the literature on assessment. They may be discussed selectively, but none should be ignored (Jonsson and Svingby 2007). Although rubrics do not make assessment valid, addressing different aspects

Table 4 Different types of reliability testing used in student assessments

Reliability			
Inter/intra-rater: variations in rate's judgments across raters, known as inter-rater reliability, or in the consistency of one single rater, called intra-rater reliability (Jonsson and Svingby 2007).	Test-retest: consistency of results when the same test is administered after a specific period (Drost 2011).	Spite-half: two tests and two measures assessing the same construct (Drost 2011).	Internal consistency:how well the different components of the assessment measure a particular construct (Drost 2011).

empirically could make assessments more valid and reliable for its intended purpose, eliciting the required performance (Jonsson 2008). There is sparse research focussing on the quality of rubrics as a valid and reliable assessment tool (Stellmack et al. 2009). Hence, a literature review in the area of authentic assessment was carried out to reveal if a holistic approach to improving its validity and reliability through rubrics has been used by past researchers in the area.

4 Classification of literature

The classification is based on a review of 124 articles which included books, chapters in books, conference papers and proceedings, government documents, journals, reports, thesis, and other articles classified as generic. The articles were chosen after a web-based search on popular websites such as Google, Google Chrome, and Google Scholar as well as the library database of the University of Tasmania. The University of Tasmania uses popular search systems such as ProQuest and Web of Science which enabled to widen the search of articles. Articles were also found by the snowballing technique based on a search through citations in articles discovered through online search. The online search used the phrases 'authentic assessment', 'authenticity in assessment', and 'authentic+assessment'. Hence, all reviewed articles contain both the words 'authentic' and 'assessment' or 'authenticity' and 'assessment', the exception being the articles by Wiggins (1998) and BoarerPitchford (2010). While the former was chosen based on the fact that Wiggins is the most cited author in the area of authentic assessment, the latter was selected due to the discussion of authentic assessment in the research. The articles span from 1989 (when authentic assessment was first introduced) to 2015 (when this paper was being written). An effort was made to obtain as many articles as possible through the above methods.

The purpose of the classification was to highlight the different types of validity and reliability demonstrated in past research, when authentic assessment was implemented with the use of rubrics. As a result, articles where authentic assessment was implemented without the use of rubrics were excluded from the classification. Table 5 provides a snapshot of the criteria used for the inclusion and exclusion of articles from the classification.

The articles included in the classification were reviewed (Table 6) to investigate the extent of validity and reliability testing of rubrics in the past, when used as an authentic assessment instrument by researchers for student assessments in various areas of education and training.

Table 5 The criteria used to select articles for classification

Total number of articles selected for the review	124
Articles excluded based on the non-implementation of authentic assessment (includes theory discussion, theoretical models/frameworks, data collected via interviews; focus groups; and surveys only)	83
Articles excluded based on implementation of authentic assessment but without the use of rubrics	24
Articles included based on implementation of authentic assessment with the use of rubrics	17

5 Gaps found from the literature classification

The intention of the literature classification was to find out the extent of investigation that has been carried out in the area of testing validity and reliability of rubrics as authentic assessment tools. Reliability and validity problems are found to be very typical of authentic assessment (Olfos and Zulantay 2007). It is often assumed that reliability is achieved concurrently with validity, due to which it may be ignored or accepted with low levels in traditional assessments (Olfos and Zulantay 2007). This was evident in the study by Olfos and Zulantay (2007) which showed a lack of reliability but showed evidence of validity. So, reliability is often accepted as a necessary condition of validity (Olfos and Zulantay 2007). However, in cases of authentic assessment, reliability cannot be ignored or accepted with low levels as a trade-off between validity and reliability (Jonsson 2008). Reliability mainly indicates consistency of performance which is essential for workplace-based tasks.

The most obvious gap found in this respect reflects an absence of both validity and reliability testing in some studies such as Todorov and Brousseau (1998), Emery (2001), Vendlinski et al. (2002), and Brawley (2009). Reliability and validity are crucial to the acceptance of authentic assessment as an accurate measure of knowledge, skills, and behaviours (Stevens 2013). There are numerous extraneous variables that affect the validity and reliability of the rubrics when used an assessment instrument (Taylor 2011). If these variables are not addressed, then the validity and reliability of the assessment and the resulting outcomes becomes questionable (Olfos and Zulantay 2007). Fook and Sidhu (2010) believe that there is a general lack of research in exploring practices that can improve validity and reliability of assessments through criteria and standards provided in rubrics. The classification reveals that past research in the area of authentic assessment has addressed typically only one or two aspects of validity and reliability while others have not been investigated. The validity was mostly achieved through a review by field experts as evident in the studies by Moon et al. (2005), Fatonah et al. (2013), Olfos and Zulantay (2007), Johnson (2007), Taylor (2011), and Lang II (2012). Barring one study by Jonsson (2008), none of the studies in the classification demonstrated construct validity. A lack of construct validity may indicate that that underlying psychological variables such as problem-solving, social interaction, and communication which are required universally in most professions were not adequately assessed in these cases.

Some studies revealed other types of validity, such as face and convergent validity, which were not categorised under the three common types of evidence required to support the validity of an assessment instrument. While face validity is the weakest and least scientific form of establishing validity, convergent validity was explained by Cassidy (2009) as a subcategory of construct validity that seeks 'agreement between a theoretical concept and a specific measuring instrument'. The review revealed that some researchers like Cassidy (2009) use a pre-tested instrument expecting the same validity and reliability as obtained in previous studies. However, if using a pre-existing instrument, it is essential for researchers to establish the instrument's validity and reliability in the context of their own research (Burton and Mazerollw 2011).

A common method for establishing reliability for rubrics is revealed to be through inter-rater scoring or internal consistency reliability. Reliability in authentic assessments has often demonstrated by a variety of statistical measures and coefficients as evidenced by the studies ofJohnson (2007), Lang II (2012), Olfos and Zulantay (2007), and Diller and Phelps (2008). According to Lovorn and Rezaei (2011), simply using rubrics do not improve the reliability of the assessment. Reliability can only be improved if rubric users are well trained on its development and use. Raters/Scorers need to be involved in the development of rubrics or else it takes time for them to understand its purpose and implementation (Diller and Phelps 2008). For example, the study by Lovorn and Rezaei (2011) involved the training of 55 teachers in rubric use to find a resulting increase of reliability in writing assignments. However, many of the studies such as Moon et al. (2005), Olfos and Zulantay (2007), and Diller and Phelps (2008) do not mention any training for rubric users before they were administered. In the study by Taylor (2011), teacher development workshops were carried out to minimise threats to internal validity only. However, according to Taylor (2011), training conducted for rubrics development or use should be consistent for all involved. Differing approaches in terms of context, standards, or application can impact the results of research data and create problems with validity.

The classification also reveals an absence of research of authentic assessment in the field of seafarer education and training. Past research (Bell and Bell 2003; Cassidy 2009; Wellington et al. 2002) showed that authentic assessment has been implemented to investigate its impact on achievement of educational or professional standards, constructive alignment of instruction processes with assessment, and achievement of professional competence (including demonstration of essential behaviours). Similar research is needed but has been largely ignored in the area of seafarer education.

6 Conclusion

The move of the STCW'95 code towards OBE highlights the need of assessment practices that allow demonstration of learning outcomes by seafarer students through performances in real-world or contextually similar settings provided by authentic assessment.

To validate if intended outcomes are being measured consistently through assessments, authentic assessments need to achieve validity and reliability through clear statements of learning expectations provided by assessment rubrics. The validity and reliability of the rubric is not only essential for the validation of outcomes attainment but also for the rubric to be accepted as an instrument of authentic assessment that can effectively measure outcomes. An extensive literature review in the area of authentic assessment revealed a lack of research in a holistic approach to addressing different aspects of validity and reliability of rubrics when used as an authentic assessment instrument. The absence of a robust framework challenges and undermines the resulting outcomes from the learning and teaching experience attained by past researchers who based their findings using rubrics that addressed only selected aspects of validity and reliability. While addressing different aspects of validity will identify and assess the content and essential underlying constructs of professional competence in different contextual scenarios, different aspects of reliability will assure consistency in performance. Overall, this will ensure a holistic approach to competence assessment at a standard expected in employment.

Past research provides theoretical justification and empirical evidence of the value of authentic assessment when educators are seeking to:

(1) Obtain evidence of the development and achievement of professional competence,
(2) Raise the standards of student performance and achievement,
(3) Measure the effectiveness of the teaching and learning,
(4) Develop higher-order and critical thinking skills in students, and
(5) Successfully align learning, teaching, and instruction with assessment.

The above outcomes together with a holistic approach to competence assessment will also benefit seafarer education and training. While knowledge-based components may continue to be assessed via traditional examinations, application of skills in real-world contexts will engage seafarer students through meaningful and relevant learning. Authentic assessments will go beyond meaningful contexts and also require seafarer students to integrate competence acquired for different STCW tasks for a holistic workplace-based performance. For example, assessment for the STCW task of ‘planning and conducting a passage and determine position’ may be designed to integrate components from other STCW tasks such as ‘maintain a safe navigational watch’, ‘use of ECDIS to maintain the safety of navigation’, and ‘manoeuvre the ship’. Assimilating, analysing, and integrating information from different units of competence will make the seafarers active participants in the process of learning and enhance student engagement. Demonstrating competence in authentic contexts will provide seafarer students with an understanding of how skills acquired in classrooms may be transferred at the workplace. Using pre-established performance criteria, students will frequently reflect on their current level of learning and compare it with the level required at the workplace, allowing them to develop strategies for raising their standards of performance.

The review reveals that there is a lack of global research on authentic assessment in the field of seafarer education and training. Further research needs to establish how to use authentic assessment within the confines of the STCW Code to improve:

(1) Student engagement,
(2) Transfer of competence, and
(3) Standards of performance.

Inherent to such future research, investigations shall also reveal ways to:

(1) Increase the validity and reliability of rubrics as an authentic assessment instrument and
(2) Use rubrics as an authentic assessment instrument to satisfy employer and regulator expectations with the attainment of the standards stipulated in the STCW Code.

Appendix

Table 6 Classification of literature based on the extent of validity and reliability testing of rubrics when used as an authentic assessment instrument

Author/ year	Context of study	Type of validity demonstrated	Type of reliability demonstrated	Techniques/ coefficients for validity/reliability of rubrics	Reason for using scoring rubrics
Todorov and Brousseau (1998)	School students	None	None	None	Evidence of achievement of content standards
Emery (2001)	School students	None	None	None	Improving student performance through scoring rubrics
Wellington et al. (2002)	School students	None	None	None	To provide a correlation between different measures of student understanding
Moon et al. (2005)	School students	Content validity	Inter-rater reliability	Reliability through Kappa formula; reliability through Kappa formula	To provide quantifiable information about student learning and instruction process
Johnson (2007)	School students	(1) Face validity (2) Content validity (3) Content relevance	Internal consistency reliability	Validity through field experts; reliability through Kuder-Richardson #20 (KR20)	To compare student achievement scores on authentic assessment with that on traditional assessments
Olfos and Zulantay (2007)	School students	Concurrent validity	Internal consistency reliability	Validity through criteria of judges, parallel instruments, and non-obstructive data; reliability through Rho of spearman, index *r* of Pearson, Cronbach's alpha	To improve the validity and reliability of the web-based authentic assessment system
Anders Jonsson (2008)	University students (teacher education)	(1) Face validity (2) Construct validity	(1) Internal consistency reliability (2) Inter-rater reliability (3) Rank Correlation	Face validity through student interviews; content validity through experts' validation; internal consistency reliability through Cronbach's Spearman's rho; Rank Correlation through Pearson's *r*	To assess student performance and self-assessment skills of students in authentic assessment
Diller and Phelps (2008)	University programme (information literacy)	Validity demonstrated through reliability tests	Internal consistency reliability	Multivariate, item correlation, factor analysis; Cronbach's alpha	To assess the effectiveness of the course programme through authentic assessment
Brawley (2009)	School students	None	None	None	To assess if authentic assessment requires higher-order thinking skills than traditional assessments

Table 6 (continued)

Author/ year	Context of study	Type of validity demonstrated	Type of reliability demonstrated	Techniques/ coefficients for validity/reliability of rubrics	Reason for using scoring rubrics
Cassidy (2009)	Elementary school teachers	Convergent validity	Internal consistency reliability	Validity established based on previous use; reliability provided through multiple assessment tasks	To measure relationship between teacher effectiveness (in terms of level of instructional quality) and student achievement through authentic assessment scores
Taylor (2011)	School students	Internal validity	Inter-rater reliability	Threats to internal validity minimised through teacher development workshops, feedback from parents and students; reliability obtained through multiple raters	To measure achievement of learning objectives through interdisciplinary authentic assessment
Azim and Khan (2012)	School students	None	None	None	To assess students' knowledge, higher-order skills, and performance through authentic assessment
Lang II (2012)	University students (teacher education)	(1) Content validity	(1) Internal consistency reliability (2) Item bias	Validity through field experts; reliability obtained using Kuder-Richardson index (KR20); item bias through Mantel-Haenzel chi-square and an unnamed statistical method	To compare validity between authentic assessment and feedback tool by articulating lecturer's expectations from students
Mccarthy (2013)	University students (business graduates)	None	None	None	To use as a self assessment and feedback tool by articulating lecturer's expectations from students
Blackburn and Kelsey (2013)	School students	None	None	None	To assess student performance in authentic assessment
Fatonah et al. (2013)	School students	Content validity	(1) inter-rater reliability (2) instrument reliability	Validity through field experts using Aikends validity; inter-rater reliability using Kappa formula; instrument reliability using Alpha formula, and factor analysis using SPSS and Lisrel	To access student performance in a proposed authentic assessment model
Hensel and Stanley (2014)	University students (nursing education)	None	Inter-rater reliability	Achievement of reliability implied text; empirical measures and data not available	To score student performance in a stimulated authentic assessment task

References

Andrade H, Du Y (2005) Student perspectives on rubric-referenced assessment. Practical Assessment, Res & Eval 10:1–11

Archbald DA (1991) Authentic assessment: principles, practices, and issues. Sch Psychol Q 6:279–293

Archbald DA, Newmann FM (1988) Beyond standardized testing: assessing authentic academic achievement in the secondary school. Reston, National Association of Secondary Principals

Azim, S. & Khan, M. 2012. Authentic assessment: an instructional tool to enhance students learning. Academic Research International, 2

Bell, A. & Bell, M. 2003. Developing authentic assessment methods from a multiple intelligences perspective. Master of Arts Action Research Project, Saint Xavier University and SkyLight Professional Development Field-based Master's Program.

Biggs, J. & Tang, C. 2007. Teaching for quality learning at university, Maidenhead Open University, McGraw Hill

Biggs J, Tang C (2011) Teaching for quality learning at university: what the student does. McGraw-Hill, England

Blackburn, J. J. & Kelsey, K. D. 2013. Understanding authentic assessment in a secondary agricultural mechanics laboratory: an instrumental case study. Journal of Human Sciences and Extension, 1

Boarerpitchford, J. K. 2010. An examination of the assessment practices of community college instructors. PhD, Caella University

Brady D (1997) Assessment and the curriculum. In: Cullingford C (ed) Assessment versus evaluation. Cassell, Great Britain

Brawley, N. 2009. Authentic assessment vs. traditional assessment: a comparative study. Bachelor of Science Honors, Coastal Carolina University

Brindley G (1998) Outcomes-based assessment and reporting in language learning programmes: a review of the issues. Lang Test 15:45–85

Butler KG (1999) Dynamic and authentic assessment of spoken and written language disorders. In: Pinto MGLC, Veloso J, Maia B (eds) Plenary lecture 5th international society of applied psycholinguistics, Portugal, 1999, pp 47–57

Burke K (2011) From standards to rubrics in six steps: tools for assessing student learning. California, Corwin

Burton LJ, Mazerollw SM (2011) Survey instrument validity part I: principles of survey instrument development and validation in athletic training education research. Athl Train Educ J 6:27–35

Cassidy, K. E. 2009. Using authentic intellectual assessment to determine level of instructional quality of teacher practice of new elementary school teachers based on teacher preparation route. Doctor of Education PhD, The George Washington University

Chance BL (1997) Experiences with authentic assessment techniques in an introductory statistics course. J of Stat Educ 5

Cooper BS, Gargan A (2009) Rubrics in education: old term, new meanings. Phi Delta Kappan 91:54–55

Cooperstein SE, Kocevar-Weidinger E (2004) Beyond active learning: a constructivist approach to learning. Reference Services Review 32:141–148

Cumming JJ, Maxwell GS (1999) Contextualising authentic assessment. Assessment in Education: Principles, Policies and Practices 6:177–194

Darling-Hammond L, Snyder J (2000) Authentic assessment of teaching in context. Teach Teach Educ 16: 523–545

Davis MH, Amin Z, Grande JP, O'Neill AE, Pawlina W, Viggiano TR, Zuberi R (2007) Case studies in outcome-based education. Med Teach 29:717–722

Dennison RD, Rosselli J, Dempsey A (2015) Evaluation beyond exams in nursing education: designing assignments and evaluating with rubrics. Springer Publishing Company, New York

Dikli S (2003) Assessment at a distance: traditional vs alternative assessments. The Turkish Online J of Educ Technol 2:13–19

Diller KR, Phelps SF (2008) Learning outcomes, portfolios, and rubrics, oh my! Authentic Assessment of an Information Literacy Program portal: Libraries and the Academy 8:75–89

Driscoll A, Wood S (2007) Developing outcomes-based assessment for learner centered education: a faculty introduction. Stylus Publishing, Virginia

Drost EA (2011) Validity and reliability in social science research. Educ Res and Perspect 38:105–123

Emery, D. E. 2001. Authentic Assessment in High School Science A Classroom Perspective. *In:* Shepardson, D. P. (ed.) Assessment in science: a guide to professional development and classroom practice. Indiana: Springer-Science+Business Media Dordrecht

Fatonah S, Suyata P, Prasetyo ZK (2013) Developing an authentic assessment model in elementary school science teaching. J of Educ and Practice 4:50–61

Finch AE (2002) Authentic assessment: implications for EFL performance testing in Korea. Secondary Education research 49:89–122

Fook CY, Sidhu GK (2010) Authentic assessment and pedagogical strategies in higher education. J of Socl Sciences 6:153–161

Gulikers, J. T. M. 2006. Authenticity is in the eye of the beholder: beliefs and perceptions of authentic assessment and the influence on student learning. PhD, OpenUniversiteitNederland

Gulikers JTM, Bastiaens TJ, Kirschner PA (2004a) A five-dimensional framework for authentic assessment. ETR&D-Educational Technology Research and Development 52:67–86

Gulikers JTM, Bastiaens TJ, Kirschner PA (2004b) Perceptions of authentic assessment: five dimensions of authenticity. Second Biannual Joint Northumbria/EARLI SIG assessment conference, Bergen

Gulikers JTM, Bastiaens TJ, Kirschner PA (2006) Authentic assessment, student and teacher perceptions: the practical value of the five dimensional framework. J of Vocational Educ and Training 58:337–357

Gulikers JTM, Bastiaens TJ, Kirschner PA, Kester L (2008) Authenticity is in the eye of the beholder: student and teacher perceptions of assessment authenticity. J of Vocational Education and Training 60:401–412

Hensel D, Stanley L (2014) Group simulation for "authentic" assessment in a maternal-child lecture course. Journal of the Scholarship of Teaching and Learning 14:61–70

Herrington J, Herrington A (1998) Authentic assessment and multimedia: how university students respond to a model of authentic assessment. Higher Education Research and Development Society of Australasia 17: 305–322

HoepfL, M. 2000. Large-Scale Authentic Assessment. *In:* CUSTER, R. L. (ed.) Using authentic assessment in vocational education. Columbus: Center on Education and Training for Employment

Johnson, Y. L. 2007. The efficacy of authentic assessment versus pencil and paper testing in evaluating student achievement in a basic technology course. PhD, Walden University

Jonsson, A. 2008. Educative assessment for/of teacher competency. A study of assessment and learning in the "interactive examination" for student teachers. PhD Doctoral Thesis, Malmo University

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

Lang II, T. R. 2012. An examination of the relationship between elementary education teacher candidates' authentic assessments and performance on the professional education subtests on the Florida Teacher Certification Exam (FTCE). Educational Specialist Graduate Theses and Dissertations, University of South Florida

Linn RL, Baker EL, Dunbar SB (1991) Complex, performance-based assessment: expectations and validation criteria. Educational Researcher 20:15–21

Lovorn MG, Rezaei AR (2011) Assessing the assessment: rubrics training for pre-service and new in-service teachers. Practical Assessment, Research & Evaluation 16:1–18

Lynch, R. 2003. Authentic, performance-based assessment in ESL/EFL reading instruction. Asian EFL Journal, 1–28

Marzano, R. J., Pickering, D. & Mctighe, J. 1993. Assessing student outcomes: performance VA, Association for Supervision and Curriculum Development

Mccarthy G (2013) Authentic assessment—key to learning. In: Doyle E, Buckley P, Carroll C (eds) Innovative business school teaching: engaging the millennial generation. Routledge, United Kingdom

Messick S (1994) The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher 23:13–23

Messick S (1995) Validity of psychological assessment: validation of references from person's responses and performances as scientific inquiry into score meaning. American Psychologist 50:741–749

Messick S (1996) Validity and washback in language testing. Language Testing 13:241–256

Meyer CA (1992) What's the difference between "authentic" and "performance" assessment? Educational leadership 49:39–40

Montgomery K (2002) Authentic tasks and rubrics: going beyond tradtional assessments in college teaching. College teaching 50:34–40

Moon TR, Brighton CM, Callahan CM, Robinson A (2005) Development of authentic assessments for the middle school classroom. The Journal of Secondary Gifted Education 16:119–133

Morrison WGS (1997) Competent crews = safer ships: an aid to understanding STCW 95. Sweden, WMU Publications, Malmo

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

Moskal BM, Leydens JA (2000) Scoring rubric development: validity and reliability. Practical Assessment, Research & Evaluation 7

Mueller, J. 2005. The authentic assessment toolbox: enhancing student learning through online faculty development. Journal of Online Learning and Teaching, 1

Neely P, Tucker J (2012) Using business simulations as authentic assessment tools. American Journal of Business Education 5:449–456

Neo M, Neo KT-K, Tan HY-J (2012) Applying authentic learning strategies in a multimedia and web learning environment (MWLE) Malaysian students' perspective. The Turkish Online Journal of Educational Technology 11:50–60

Newmann FM, Marks HM, Gamoran A (1996) Authentic pedagogy and student performance. American Journal of Education 104:280–312

Olfos R, Zulantay H (2007) Reliability and validity of authentic assessment in a web based course. Educational Technology & Society 10:156–173

Osborne R, Dunne E, Farrand P (2013) Integrating technologies into "authentic" assessment design: an affordances approach. Research in Learning Technology 21

Palm, T. 2008. Performance assessment and authentic assessment: a conceptual analysis of the literature. Practical Assessment, Research and Evaluation, 13.

Palomba CA, Banta TW (1999) Assessment essentials: planning, implementing, and improving assessment in higher education. Jossey-Bass, San Francisco

Reddy MY (2007) Rubrics and the enhancement of student learning. Educ Comm Tech J 7:3–17

Reeves, T. C. & Okey, J. R. 1996. Alternative assessment for constructivist learning environments, Englewood Cliffs, NJ, Educational Technology Publications

Rezaei AR, Lovorn M (2010) Reliability and validity of rubrics for assessment through writing. Assessing Writing 15:18–39

Rhodes TL, Finley A (2013) Using the VALUE rubrics for improvement of learning and authentic assessment. Association of American Colleges and Universities, United States of America

Sadler R (2005) Interpretations of criteria-based assessment and grading in higher education. Assessment & Evaluation in Higher Education 30:175–194

Sampson, D. & Fytros, D. 2008. Competence models in technology-enhanced competence-based learning. In: Adelsberger, H. H., Kinshuk, Pawlowski, Jan Martin (ed.) Handbook on information technologies for education and training. Springer Berlin Heidelberg.

Scholtz A (2007) An analysis of the impact of authentic assessment strategy on student performance in a technology-mediated constructivist classroom. A study revisited International Journal of Education and Development using Information and Communication Technology (IJEDICT) 3:42–53

Silvestri L, Oescher J (2006) Using rubrics to increase the reliability of assessment in health classes. International Electronic Journal of Health Education 9:25–30

Spady, W. G. 1994. Outcome-based education: critical issues and answers, Arlington, VA, American Association of School Administrators.

Stellmack MA, Konheim-Kalkstein YL, Manor JE, Masset AR, Schmitz JAP (2009) An assessment of reliability and validity of a rubric for grading APA-style introductions. Teaching of Psychology 36:102–107

Stevens, P. 2013. An examination of a teacher's use of authentic assessment in an urban middle school setting Doctor of Education PhD, Ohio University

Taylor, J. M. 2011. Interdisciplinary authentic assessment: cognitive expectations and student performance. Doctor of Education PhD, Pepperdine University

Todorov, K. R. & Brousseau, B. 1998. Authentic assessment of social studies. *In:* Education, M. D. O. (ed.). Lansing, Michigan

University OF Tasmania (UTAS) 2011. Guidelines for good assessment practice, Tasmania, Australia, UTAS.

Vendlinski, T., Underdahl, J., Simpson, E. & Stevens, R. Authentic assessment of student understanding in near-real time. NECC 2002: 23rd National Education Computing Conference Proceedings, 17–19 June 2002 San Antonio, Texas

Walvoord, B. E. 2004. Assessment clear and simple: a practical guide for institutions, departments and general education, San Francisco, Jossey-Bass

Wellington P, Thomas I, Powell I, Clarke B (2002) Authentic assessment applied to engineering and business undergraduate consulting teams. International Journal of Engineering education 18:168–179

Wiggins G (1989) A true test: toward more authentic and equitable assessment. The Phi Delta Kappan 70:703–713

Wiggins, G. 1990. The case for authentic assessment. Practical Assessment, Research and Evaluation, 2.

Wiggins G (1992) Creating tests worth taking. Educational Leadership 49:26–34

Wiggins G (1993) Assessment: authenticity, context and validity. Phi Delta Kappa International 75(200–208): 210–214

Wiggins G (1998) Educative assessment: designing assessments to inform and improve student performance. Jossey-Bass Publishers, San Francisco