International Journal of Learning, Teaching and Educational Research
Vol. 20, No. 2, pp. 81-103, February 2021
https://doi.org/10.26803/ijlter.20.2.5

A Competence Model to Assess and Develop Designing Competence Assessment Tool

Do Huong Tra
Hanoi National University of Education, Hanoi, Vietnam
https://orcid.org/0000-0001-5905-7667

Nguyen Thi Dieu Linh
Hanoi National University of Education, Hanoi, Vietnam
https://orcid.org/0000-0002-7342-423X

Abstract. Studies showed that competency-based assessment improvement had generated a much greater impact on students' achievements on standardized tests than other forms of educational activities. However, studies also indicated a number of challenges for teachers when designing competence assessment tools (CAT), especially in building assessment tasks that replicate real-life practice. There have been many different models for teachers' assessment competencies, but the competency of designing competence assessment tools (CDCAT) was not paid much attention. In hope to develop a competency model that would serve as a supporting role in developing the CDCAT for pre-service teachers and teachers, this study used a multi-step development process to construct a teachable model that reflectived the complexity of designing assessment tools. The model consisted of 12 behavioural indicators spreading across four dimensions informed by the existing literature and empirical findings in particular contexts. To guarantee the content value of the proposed model, the research twice used the expert method by two panels. The reliability of the model was tested by analyzing the data collected from the survey with students. Interesting findings were met, and the outlined CDCAT model assisted pre-service teachers in solving issues related to their assessment competence. The model was intended for educational researchers, educators, teachers, and policy makers to support teachers' assessment competence concerning the current accountability model across educational systems. Specific implications for developing pre-service teachers' CDCAT were discussed, followed by suggestions for future studies.

Keywords: assessment tool; CDCAT; competence model; designing

1. Introduction

To help students develop 21st-century skills, teachers should be equipped with knowledge and skills to select, adapt, and design classroom assessment tools

following curriculum-based competencies. Black and Wiliam (1998), in their meta-analysis of 250 empirical studies, indicated that the formative assessments brought about a positive impact on teachers' everyday teaching practice which could be seen in students' achievement on standardized tests, in which students performed better and more quickly in solving complex tasks. In other words, they proved to possess more effective strategies in dealing with problems. This study also showed firm evidence that the assessment encouraged self-monitoring, fostered self-reflection, and generated a greater level of students' commitment to their learning. Thus, the assessment competencies contributed a great deal to the students' academic success.

A main body of study indicated that teachers have not been provided with proper training and, therefore, were incapable of designing effective assessment tools (Bol et al., 1998). It has been shown in many empirical studies worldwide (DeLuca & Klinger, 2010; Volante & Fazio, 2007) that a large number of teachers demonstrated a weak competence assessment, including CDCAT. These studies at the same time determined that these teachers needed to improve their assessment competence via training on classroom assessment and testing.

In Vietnam, there is very little time to develop assessment competence for pedagogical students in the curricula of pedagogical universities. Therefore, despite being trained by the Vietnamese Ministry of Education and Training on the innovation of assessment and testing competencies, teachers still face many difficulties in designing competency assessment tools (Linh & Tra, 2016). This indicated the important role of initial training at pedagogical universities. Recognizing the importance of innovation in assessment and testing in general and CDCAT in particular, some pedagogical universities have included the module of "Assessment and Testing in Education" in their curricula.

According to Griffins (2015), the training and development of any competencies should be based on the model of that competence. Building a detailed model for CDCAT that included behavioural indicators and quality criteria can help to orient learning activities towards building each behavioural indicator. Such a model was also the basis for designing tools to evaluate the levels that learners achieved, thereby adjusting learning activities.

Many assessment literacy standards for teachers around the world and many studies on competence assessment models were available (Wiggins & McTighe, 2005; Brookhart, 2011; DeLuca & Klinger, 2010; Fulcher, 2012; Alonzo, 2016). In these publications, the CDCAT was only one component of such models. However, according to the present literature review, these studies were fairly and widely conducted and therefore provided little information on behavioural indicators of CDCAT. Besides, these said models have not been appropriately considered the complexity of designing assessment tasks in competency assessments.

Therefore, this study aimed to construct a teachable model that is reflective of the complexity of designing assessment tools, including behavioural indicators

and quality criteria that can help lecturers determine the objectives and design activities for developing pre-service teachers' CDCAT. The use of the multi-step process to build the model as well as the model itself will be covered in the next sections of this article.

2. Study Methods and their Procedure

Vital components of the competence assessment model proposed by Griffin (2015), which was adopted in this study, include elements, behavioural indicators, and quality criteria. There were three popular methods to build the competence model:

(1) Traditional survey with (a) defining the structure to be measured, (b) building the item pool, (c) obtaining experts' review to the items, (d) testing the items, (e) making factor analysis (DeVellis, 2003);

(2) Measuring competence with (a) defining the structure to be measured, (b) observing and interpreting performance, and (c) statistically modeling the reliability and validity of the scores produced (Shavelson 2013);

(3) Delphi study with brainstorming, narrowing down and ranking (Okoli, 2014; Naresh Giangrande et al, 2019).

On a basis of studying and synthesizing these methods, the study has used a multi-step development process to develop a CDCAT model, in which the researchers have used all the three methods above, but the Delphi method was the mainstream. Specifically, the following steps have been performed:

(i) Reviewing the existing literature on the model of assessment competencies, procedures, and literacy standards for teachers in the past and present to determine the components and behavioural indicators of the model was the first step in framing the CDCAT model. It is significant to be aware of the required national quality assurance system standards based on the perceptions of the requirements of teachers' assessment competency standards to ensure the quality of countries, review of existing studies on assessment competency and assessment tool design process, and accounting for content analysis to identify the most important criteria to form the basic structure of the CDCAT model.

(ii) Consulting experts (for the first time) on the components and behavioural indicators of the model was the first phase of the Delphi method (Okoli, 2014). Brainstorming could be an efficient way to gather and combine expert opinions without seeking convergence, but emphasizing the originality and diversity of the ideas was worthy. Therefore, the investigators have organized two sessions; brainstorming in the first session and applying the expert method to determine the value of the model in the second session. The researchers chose five experts, active lecturers, who taught the module of "Assessment and Testing in Education" or performed the study on assessments. Experts who worked independently were asked to list relevant factors in a random order. Then, the investigators consolidated the lists from all experts and the list gained from the literature review, removed exact duplicates, and guaranteed terminology inconsistency. This list was sent to five experts for validation, and they were asked to give opinions and suggestions for changes if any. When receiving the

written response from them, the researchers directly discussed with them to clarify their opinions as well as investigators' arguments.

Experts were invited to consider whether the model had only behavioral indicators related to the basic structure of CDCAT, also whether it could provide teachers the foundation to help the student along the developmental progressions and enable the identification of indicative behaviors that could be used for interpreting students' performance. At the end of this step, refining the final version of the consolidated lists took place.

(iii) A survey on self-assessment of pre-service teachers at Hanoi National University of Education (HNUE) was conducted to determine the content validity of the initial model. The HNUE's students come from many different regions of the country and the students participating in the survey were selected at random to avoid regional bias (20% male and 80% female). The survey was conducted with 60 pre-service teachers in the second semester of the third academic year when they completed the module "Assessment and Testing in Education". Each student was required to complete a questionnaire, in which they had to assess the level of confidence for each behavioral indicator of the CDCAT. Before the survey, the questionnaire was piloted for with pre-service teachers to check whether the proposals of the scale created interpretation problems for people who were not familiar with the CDCAT model. Based on the data obtained from semi-structured interviews and first stage data, the researchers analyzed and modified, consolidated, or removed original sentence structures, terminology or wording to keep pace with sample observations. Students involved in developing models increased their usability and helped students understand, develop, and use their understanding of task requirements. It was expected to help them "think critically about their work." (Huba & Freed, 2000). Therefore, before filling in the questionnaire, students were asked to carefully read the behavioural indicator of the CDCAT and suggested changes in the irrational points. This step could be considered to be a single phase in the traditional survey approach. However, for the small sample, it was sufficient to calculate the Cronbach Alpha coefficient. Instead of the second step in the Delphi study, this step was intended to reduce the number of factors.

(iv) The following step concerned with reviewing the literature to propose quality criteria. At this phase, each evaluation task was analized in the assessment tools to identify quality criteria that were disclosed through each assessment task. The quality criteria of each behavior indicator were aggregated and reviewed to determine if they adequately reflected the achievable levels by each behavior indicator. Typical quality criteria were chosen for reference when proposing quality criteria.

(v) In this phase, the investigators analyzed the collected practical data during training and retraining of the CDCAT and pre-service teachers, in combination with interviews with pre-service teachers' cognition to propose quality criteria. This step could be considered to be the observation phase of performance and interpretation of performance in line with the second approach. To limit shortcomings of many large-scale surveys, Shavelson (2013) indicated that even when using statistical models, there is a need for evidence of cognitive validity;

the tasks evoked the types of thinking and reasoning that became part of the inference to make judgments about competence. In this sense, such evidence was collected through the "think-aloud" method, whereby students had told their thoughts as they performed a task (Leighton, 2004). The results of think-aloud with pre-service teachers about their perceptions were used to collect information on how students participated in the design of assessment tools to identify the key factors for their success It also helped to determine the level of internal development of learners concerning their obtained product. Students' product samples represented different levels of performance and showed up observable criteria in behaviours (Tierney & Simon, 2004). Therefore, the assessment tools designed by the learner were classified according to the levels of each behavioural indicator. To determine learners' levels, the product samples selected for analysis included students from the lower, middle, and upper groups of the class in terms of overall learning progress. The results of this analysis were combined with the results of the above stated step to propose quality criteria.

(vi) With regard to the above steps, consulting experts (for the second time) on the complete CDCAT was concsidered. This step was similar to phase 3 of Delphi method. In order to strike a balance between "generalized wording" for increased applicability and "detailed description" for reliability, there should be cooperation between interdisciplinary instructors when describing the criteria (Suskie, 2004). 23 experts and active lecturers, who taught "Assessment and Testing in Education" for students from nine faculties at pedagogical universities (Math, Physics, Chemistry, Technology and Education, Biology, Literary, History, Geography, Information Technology) and experts in the Center for Educational Assessment and Quality assurance, were selected. The oldest was 60 years old and the youngest expert was 33 (M =42.69; SD = 8.99). The minimum period of teaching experience among the expert samples was eight years and the maximum was 35. In total, 11 women and 12 men from HNUE, VNU University of Education, and also Vinh University participated in the study. Experts were encouraged to examine the suitability of the model's quality criteria with the students' product and reviewed if the model had enough difficulty levels to distinguish the students' CDCAT levels. The following questions were asked for experts for reflect:

1. Do the criteria of the model address all aspects intended to be measured in the assignments given by you?;
2. Are all the important criteria relevant to the assessment method being evaluated through the model?; and
3. Do the criteria reflect competencies that will suggest success on future or related performances?" (Moskal & Leydens, 2000).

If the experts disagreed with any points in terms of content and expression, asking them for advice on how best to adjust it was considered. A direct discussion with the experts also took place upon receiving their written responses. Besides, the investigators shared feedback with each panelist and asked them to re-rank each list. This was repeated until the members reached an agreement or had a high consensus.

3. Content and Research Results

3.1. The concept of competence of designing assessment tools

Assessment competence is related to the understanding and appropriate use of assessment practices, as well as the theoretical and philosophical background in measuring student learning (Stiggins 2002; Volante & Fazio 2007). Another simpler definition was provided by the North Central Regional Educational Laboratory (2016), in which "Assessment competence is the willingness of an educator to design, implement and discuss assessment strategies". The above definition indicated that the CDCAT is a component of assessment competence. In recent studies on teacher competencies, assessment competence has been mentioned (Caena, 2011; OECD, 2018). However, since assessment competence is only one dimension of teacher competence, CDCAT is also only one component of the assessment competence, this is why CDCAT has not been covered in detail in these studies.

Therefore, to define and clarify the model of CDCAT, an analysis of the Stiggins assessment competence model and 12 assessment literacy standards for teachers (from 1990 to present) from six geographic regions was conducted (USA, UK, Canada, Australia, Europe, and New Zealand). Through this analysis, four themes representing contemporary aspects of CDCAT were identified:(1) Defining the purposes and objectives of the competency assessment activities; (2) Planning the development of the assessment tool; (3) Developing assessment tool; and (4) Testing and editing the assessment tools. What had been reviewed suggested that competence of designing assessment tools was to connect assessments for clear purposes, apply proper assessment methods, and develop quality assessment exercises and scoring criteria appropriately.

3.2. Identification of Competence Components and Behavioural Indicators

As stated in part 2, the study of the CDCAT model was conducted in 6 steps. The results obtained through each step were as follows:

Step 1: The literature review to determine the components and behavioral indicators of the model

To determine the components and behavioural indicators of the model, all criteria related to CDCAT from previous publications were synthesized. All publications built assessment competency models to guide the activities fostering teachers' assessment competencies in general, but each publication focused on specific points, depending on the evaluation trends at that time. Therefore, in this section, to take a complete look at the CDCAT, it was crucial to synthesize all the criteria related to the CDCAT from the publications.

i. The CDCAT in the models of assessment competence

The criteria relating to CDCAT in the publications were picked up and classified into topics representing contemporary aspects of CDCAT. For overlapped standards, more general criteria were chosen to include in the synthesized table, so as not to miss out on criteria related to CDCAT. None of the standards mentioned comprehensively covered all the important teacher assessment competencies. Except for the Standards for Teacher Competence in the American Educational Assessment of Students published in 1990 (ASTCEAS), most of these standards were only outlined generally, with incomplete requirements for

teachers' assessment competency, and components of the CDCAT were not clear. This study considered models of teachers' assessment competency proposed by Stiggins (1999), DeLuca and Klinger (2010), Brookhart (2011), Fulcher (2012), Alonzo (2016), and ASTCEAS (1990) to determine the components of the CDCAT.

Table 1 shows the results of the analysis of those models. For each model of teachers' assessment competency, its components were classified into four groups corresponding to the four components of CDCAT (Row 1).

Table 1: Comparison of components from the existing model of Teacher Assessment competency

	I. Determine the purposes and objectives of the competency assessment activities	**II. Plan the development of the assessment tool**	**III. Develop anassessment tool**	**IV. Trial and finalize the assessment tool**	**Conduct assessment and use assessment results**
ASTCEAS (1990)		Choosing assessment methods.	Buildingassessment methodsCreating valid pupil gradingprocedures.	Finding out unethical, illegal, and inappropriate assessment methods	Administering, scoring, and interpreting. Using assessment results when making decisions. Communicating assessment results.
Stiggins (1999)	Connecting assessment to clear purposes.	Applying proper assessment methods. Clarifying Achievementexpectations.	Preparing quality assessment exercises,scoring criteria and sampling in an appropriate manner.	Avoiding bias in the assessment.	Using an assessment asthe instructional intervention. Communicating effectively about student achievement.
DeLuca & Klinger (2010)	Assessing under philosophies ofclassroom assessment and philosophies of large-scale assessment.	Using and being aware of learning practices and theory. Using types of assessment.	Designing and marking the test. Taking theoretical principles in assessment of learning and assessment for learning. Applying and having technical knowledge of summative assessment practice and assessment item formats.	Using statistical techniques for assessment. Defining reliability and validity issues.	Implementing provincially mandated assessment practices. Giving the rationale for assessment decisions and practices.

Brookhart (2011)	- Defining clearlythe teachingknowledge and method to achieve; - Identifying the learning outcomes aboutcurriculum goals andstandards	Planning strategies for discussion about the learning outcomes that are going to be assessed with the students	- Obtaining a good understanding of the efficiency of how the available assessment alternatives work; - Building a tool that can assess the students' performance through a scoring system with helpful data.	- Learning about whether or not the assessment tools the teachers are using meet the intended learning outcomes including the required knowledge and thinking skills.	Providing effective, useful feedback on student work. - Offeringconstructive feedback on students' performance; Helping students use assessment information to make sound educational decisions. - Supportingstudents to make decision on their assessment - Giving ethical considerations on the administration of the assessment being used.
Fulcher (2012)		Learning about and applyinglarge-scalestandardized testing,classroom testing, and washback.	Designing and developing the test.	Being valid and reliable	
Alonzo 2016			Designing assessment tasks. Using rubrics to assess students' learning.	Considering factors that affect students' performance. Securing task completion in any form. Holding up dialogue/ conversation with colleagues to ensure consistent, fair, and comparable judgment of students'learning.	

ii. The behavioural indicators of the CDCAT in the processes of designing assessments

The models in the literature listed in table 1 did not demonstrate the behavioural indicators of the CDCAT. Therefore, a process-based approach based on analysis of the thinking process was chosen when designing an assessment tool to infer the behavioural indicators of the CDCAT. In this regard, this study analysed the processes of designing assessments that have been popularly used in research by Stiggins (1987), Grant Wiggins (2005), Brookhart (2011), Tillema (2011), and National Register for Vocational Education and Training in Australia (2012). To

draw out behavioural indicators for the CDCAT model (table 2), the researchers compared the actions for each step of one process with another.

Table 2: Comparison of steps from the existing process of assessment design

	I. Determine the purposes and objectives of the competency assessment activities	**II. Plan the development of the assessment tool**	**III. Develop anassessment tool**	**IV. Trial and finalize the assessment tool**	**Conduct assessment and use assessment results**
Stiggins (1987)	Clarifyingreason (s) for assessment • Specifying decision(s) to be made • Nominatingde cisionmaker(s) • Giving used to be made of results • Describing students to be assessed	Defining performance to be evaluated • Providing the detailed content or skill focus of the assessment • Selecting the type of performance to be evaluated • Listingperforma nce criteria	Designingexercise s • Selectingtypes of exercises • Determining obtrusiveness of assessment • Defining the amount of evidence needed Making performance rating plan • Choosing the type of score needed • Specifying who is to rate performance • Clarifying the score recording method		
Wiggins & McTighe (2005)	Identifying desired results • the big ideas • desired specifi understandings • predictable misunderstanding s	Making out acceptable evidence • that students will demonstrate achievement of the desired results • the way students will reflect upon and self-assess their learning	• Determining performance tasks students will completedemonstr ating the desired understanding • Setting the criteria by which performances will be judged		
Brookhart (2010)	• Specifying the kind of thinking and the content you wish to see evidence for.	• Making decision on what you will take as evidence that the student has exhibited this kind of thinking about the	• Designing assessment task. Planninga balance of content and thinking by an assessment blueprint.		

		appropriate content			
Tillema, (2011)	• Indicating purpose or goal of the assessment.		• Designing or selecting an assessment task. • Setting criteria for the assessment task.		• Administering the assessment. • Scoring the assessment. • Making appraisal or ''grading of the assessment''. • Giving feedback and further promotion of learning
Australia(2012)*	1.1 Identifying the target group of candidates, purpose of assessment tools, and contexts 1.2 Accessing relevant benchmarks for assessment and interpret them 1.3 Indicating, obtaining, and interpreting organizational, legal, and ethical requirements 1.4 Specifying other related documentation	2.1 Selecting assessment methods that support the collection of definedevidenc e. 2.2Encouraging candidates to show or support their claim through selected assessment methods 2.3 Considering different assessment instruments for the selected assessment methods 2.4 Reviewing how the assessment instruments will be administered	3.1 Developing specific assessment instruments 3.2 Defining and preparing clear and specific procedures instructing assessor 3.3 Considering requirements of assessment system policies and procedures, addressing storage and retrieval needs, and reviewing, evaluating control procedures as part of this process	4.1 Checking draft assessment tools against evaluationcriteri a and amend as required 4.2 Testing assessment tools to validate content and applicability 4.3 Collecting and making written feedback 4.4 Amending the final toolbased on an analysis offeedback 4.5 Appropriately formatting and filing finalized assessment toolaccording to assessment system policies and procedures as well as organizational, legal, and ethical requirements	

*Note. *Adapted from TAEASS502B Design and develop assessment tools, Australia Goverment Department of Education, Employment and workplace Relations (2012). All rights reserved.*

iii. Synthesize criteria related to the CDCAT and propose the model

The model will be used to design tasks in teaching and assessing the CDCAT. Hence, the model needs to support students and teachers in tracking their competence development and analyzing the key components of their work. The model proposed based on the CAT design process can meet this requirement.

Table 2 showed the results of content analysis of the assessment tool design process. After analyzing the common and reasonable points of these processes, the following points were added as follows:

- Each assessment activity should have contextualized and diversified purposes. To ensure this, it is necessary to specify the purposes into specific goals. In this regard, teachers can only define goals that are suitable to the situation when they clearly define its characteristics. Such a conception was confirmed by Herppich et al. (2017) who claimed that an assessment-competent teacher should be able to master a wide range of assessment-related situations relevant to the teaching profession.Therefore, at this stage, after determining the purpose of the assessment, these two steps were included: (1) Identifying the characteristics of the situations; and (2) Determining the objectives of the assessment tasks.

- "Competence involves putting into action conceptual knowledge, procedural knowledge, and attitudes to be able to resolve a particular situation" (Ananiadou & Claro, 2009). Nevertheless, it is difficult to select authentic situations to design competency assessment tasks that must meet the key criteria on Accuracy, Generality, and Extrapolation (Gulikers et al., 2005). The above processes have not considered this complexity. When analyzing the thinking process in designing assessment tasks, the following two steps were put forward before the step of designing the assessment task: (1) Specifying the type of information to be used; and (2) Searching for the type of information to be used.

Corresponding to the steps in the assessment designing process, behavioural indicators in the competence model for designing assessment tools were proposed as shown in Figure 1.

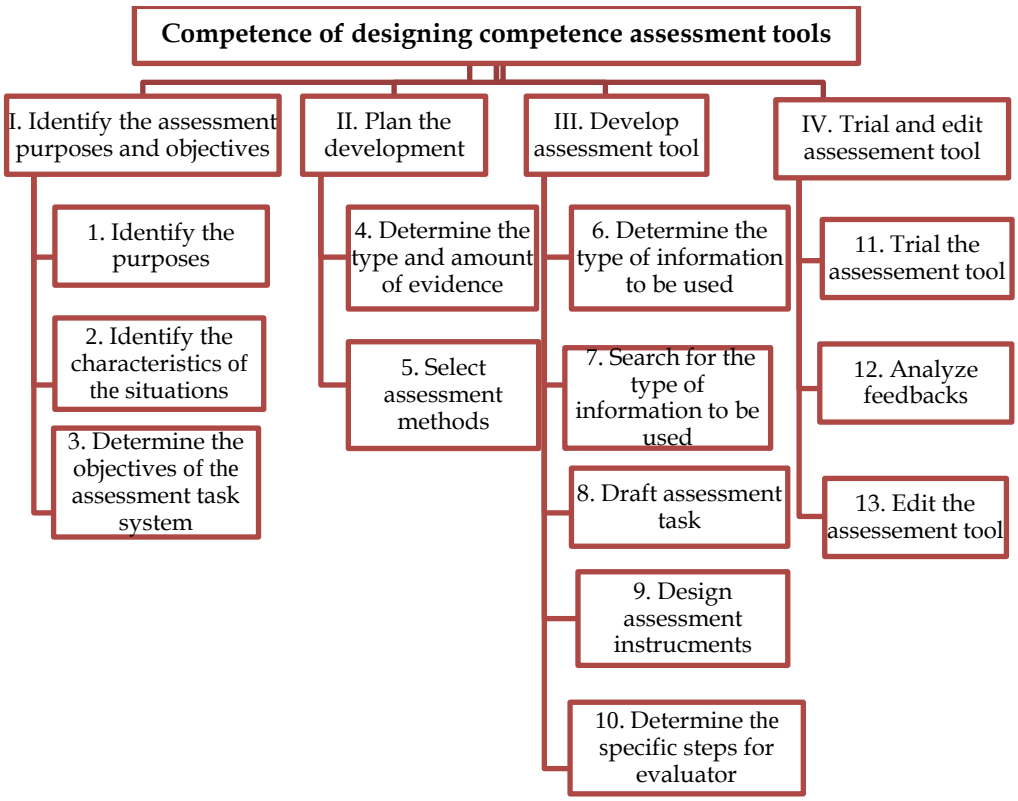

Figure 1: An initial tentative model of the CDCAT

Because the selected literature wasn't a sufficient basis to offer quality criteria, a tentative model was proposed, just by including components and behavioural indicators.

Step 2: Consult experts on the components and behavioural indicators of the model
The brainstorming results were analyzed and were as follows: all the ideas presented by three out of five experts matched the behavioural indicators of 1, 5, 8, 9, 10, 11, 12, 13 of the model proposed in step 1. The remaining two experts proposed adding two behavioural indicators namely: weighting and coding test results. However, after discussing, the below agreement were reached out:

- For weighting, the weight of each behavioural indicator was determined in the competence model to be assessed, and the designer of the assessment tools implemented it as a simple procedure.
- As for coding, it is possible that coding is defined in the 10th behavioural indicator. Before implementing behavioural indicator 12, coding is just a set of predefined steps that did not demonstrate the use of assessment knowledge.
- As far as analyzing is concerned, when analyzing feedback from people involved in the experimental process, the designer should use knowledge of reliability, validity, difficulty, discrimination, and some data processing softwares.

From the above analysis, the model was retained as shown in Figure 1. Then, the model was sent to the five experts followed by a direct discussion. The results revealed that all the experts agreed with the model.

Step 3: Determine the structural validity of the CDCAT tentative model
The results of a self-assessment of 60 pre-service teachers were analyzed to determine the structural validity. The Cronbach's Alpha coefficient was initially calculated for all 13 behavioural indicators (Figure 1). The variable VAR0007 corresponding to the seventh behavioural indicator had a low corrected item-total correlation (0,268), yet this behavioural indicator was found necessary for the process of designing a competency assessment tool. However, a person who can search for information but did not have the necessary knowledge of assessments cannot design competency assessment tools, while a person with an average ability to search for information can do this if they had assessment knowledge. Thus, the experimental results were appropriate. A process-based approach was then chosen to determine the behavioural indicators of the competency model for designing competency assessments. This analysis indicated that keeping track of cognitive processes and trying to introduce the full stages of the design process without appropriately considering the necessary knowledge was a mistake. Right after that, with each behavioural indicator, the investigators proceeded to identify all the necessary knowledge to carry out those behaviours to avoid the above mistake. As a result, the remaining behavioural indicators did not have the same problems as indicator 7.

After removing the VAR0007 variable, the results showed that the Cronbach's Alpha coefficient was 0.876, in the range from 0.8 to 1, indicating that the scale was very good (See figure 2). The corrected item-total correlation coefficient of all behavioural indicators was greater than 0.3 which denoted that all

behavioural indicators were satisfactory and no behavioural indicators measured another competence. The statistical analysis showed that the above 12 behavioural indicators were well correlated with each other and measured the same variable, the CDCAT of pre-service teachers.

Reliability Statistics

Cronbach's Alpha	N of Items
,876	12

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
VAR00001	37,8305	35,660	,553	,867
VAR00002	38,1525	36,994	,502	,870
VAR00003	38,3729	36,169	,562	,867
VAR00004	38,4576	36,425	,492	,871
VAR00005	37,6949	35,009	,670	,861
VAR00006	37,7966	34,337	,616	,863
VAR00008	37,9831	35,707	,606	,864
VAR00009	37,6610	34,676	,572	,866
VAR00010	37,9492	35,118	,517	,870
VAR00011	38,3559	35,578	,603	,864
VAR00012	38,1186	34,934	,599	,864
VAR00013	38,4068	34,625	,553	,868

Figure 2: Cronbach's Alpha coefficient for 12 behavioural indicators

3.3. Identification of the Quality Criteria

Step 4: Review of the literature to propose quality criteria

To propose the quality criteria, the following studies on assessment tools were analyzed:

• Three tools designed based on the content of the ASTCEAS were: questionnaires of Likert-type items (Zhang & Burry-stock, 1997), questionnaires of multiple-choice questions (Mertler & Campbell, 2005), and True/False questions (Daniel & King, 1998).

• An instrument (DeLuca et al., 2016) consisted of both Likert-type items and multiple-choice questions focusing on the contents of the US 2014 Standards for Educational and Psychological Testing.

• A rubric assessing the teacher's assessment competency (Alonzo, 2016) and a questionnaire (Jarr, 2012) consisted of 15 sentences designed on Likert-type items.

• Other tools to assess teacher assessment in specific subjects (McGee, 2012; Perry, 2013; Bremner, 2014; Gutierrez, 2014; Nary, 2014; Vidacovich, 2015; Xu & Brown, 2016; Hammami, 2016; González, 2018).

The results of the synthesis of the quality criteria of each behavioural indicator evaluated by the above assessment tools were as follows:

- The tools focus a lot on evaluating some content in ASTCEAS such as item analysis, select assessment methods, and tools, interpretation ofthescore, select the test sample.These are, in turn, some essential aspects of the quality criteria of the behavioral indicators of 12, 5, 10, 9, 11 in Figure 1. However, these tools had not yet fully assessed the aspects and levels of these behavioural indicators. The same thing happened with behavioural indicators in 1, 2, 3, 4, 7, and 13.

- Some quality criteria were at the high level of the behavioural indicator 8, for example: "My methods and types of assessment allowed students to demonstrate their learning in diverse ways" or "I had thought deeply about my approach to assessment", which were evaluated in the Approaches to Classroom Assessment Instrument (DeLuca et al., 2016).

These quality criteria were useful references in proposing quality criteria for the model. The remaining quality criteria were determined through the analysis of actual data in step 5.

Step 5: Analyze Practical Data
With this model, it was expected by the researchers to have 4 levels of quality criteria. Therefore, the practical data outlined below were classified into four levels. In the summer of 2014, the Ministry of Education of Vietnam organized training courses on teachers' assessment competencies in the whole country. A survey of 382 teachers was conducted to identify the level of difficulty that teachers encountered in designing competency assessment tools (Linh & Tra, 2016). According to the assessment tools designed by those teachers, the researchers' observations during the teacher training process, and the level of difficulty were categorized in 4 levels (beginning, developing, accomplished, and exemplary) as a basis for proposing quality criteria.

During the academic year 2015-2016, in the module of "Assessment and Testing in Education", 56 pre-service teachers from the Physics Faculty at HNUE were required to make their portfolios. Before each class, pre-service teachers had to read the lesson documents and determine the objectives of that lesson. After class, they recorded in their portfolios what they had learned, the difficulties they encountered, and what else they could learn. As mentioned above, the assessment tools designed by pre-service teachers, observations during the teaching process, and the data in theportfolios were also classified into 4 levels.

For three consecutive years (2015-2017), data were collected during the process of supervising nine pre-service teachers to do graduation on these topics of designing competency assessment tools. A direct discussion took place with these pre-service teachers to identify their difficulties, suggest directions, and monitor changes to their problem-solving processes. These discussions happened every time they designed new tools. First, for each behavioural indicator, pre-service teachers were grouped according to the four levels. When comparing their assessment tools with the results of cognitive interviews, learners' internal development level were determined relative to their external performance and key factors for their success whenin designing assessment tools.

The analysis of student artifacts and in-depth discussions among lecturers ensured the consistency between the descriptions in the model built based on the researcher's inference and experience with practice (Reddy, 2011). Therefore, after independent analysis, quality criteria and the SOLO taxonomy were suggested and discussed respectively to propose sound quality criteria (Biggs &

Collis, 1982). Basically, the SOLO taxonomy can be used to build scores or codes to determine the subtle perception of the subjects (in this case, the designers of competency assessment tools). This rating scale was also a simple, reliable, and easy-to-use model that matched the quality rating of those behavioural indicators. At the end of this step, combined with the result of step 4, the researchers outlined 48 quality criteria of 12 behavioural indicators mentioned in step 3. To retest the claritied behavioural indicators, all the quality criteria were practiced the scoring samples of student work.

Step 6: Consult experts on the complete competence model

To ensure the validity of the model, the latter was reviewed by experts for the second time. Behavioural indicators 6 and 8 (Figure 1) required creativity. Based on the practical data above and the rubric suggested by Alonzo (2016), corresponding to each behaviour, the highest level "Instructing the behavior in a professional way for colleagues" was designed. To reach this level for each behaviour, pre-service teachers needed to draw up logical methods and rules when performing the behavior. This also corresponded to the highest level in the SOLO taxonomy. However, 52% of experts said that this was not reasonable because, in achieving those quality criteria, pre-service teachers also would be able to present information. After discussing behavioral indicators 8 (Figure 1), the highest level of the SOLO taxonomy (Extended Abstract) was expressed by "Draft assessment tasks and reflect on the implementation process to withdraw appropriate rules for the next time".

Other feedback from the experts focused on terms and expressions, particularly for indicator which called for clarification (30% experts). Analyzing these responses led to some modifications to the quality criteria for the CDCAT as shown in Table 3.

Table 3: The full model of the CDCAT

Components	Behavioural indicators	Quality criteria
I. Determine the purposes and objectives of the competency assessment activities	1. Identifythe purposes of using the assessment tool	1.1. State the purpose of using the assessment tool in a general way.
		1.2. State clearly the familiar purposes of using the assessment tool.
		1.3. State clearly and fully the purposes of using the assessment tool.
		1.4. State, classify and rank the purposes of using the assessment tool.
	2. Determine the characteristics of the situation using the tool	2.1. Identify some common factors (time, acquired knowledge).
		2.2. Identify some basic factors, consistent with the assessment purposes (student level, reading comprehension, number of students, facilities).
		2.3. Fully determine the factors to be considered, appropriate for assessment purposes (health and psychology, language, and student readiness).

		2.4. Determine, classify and rank the factors to be considered.
	3. Determine the objectives of the assessment task system	3.1. Determine the objective, only focusing on some common factors (acquired knowledge and time).
		3.2. Determine the objective, focusing on the competence model to be assessed.
		3.3. Determine the objective, ensuring the purposes of assessment.
		3.4. Determine the objective that is relevant to the situation in which the tool is being used.
II. Plan the development of the assessment tool	4. Determine the type of evidence and the amount of evidence to be collected to assess learners' competencies	4.1. Determine the type and amount of evidence needed to assess some behavioral indicators.
		4.2. Determine the type and amount of evidence needed to separately assess each behavioral indicator.
		4.3. Determine the type and amount of evidence with attention to the relationships between behavioral indicators.
		4.4. Determine the type and amount of evidence that fully meets the assessment objectives.
	5. Select assessment methods supporting the collection of such evidence	5.1. Select some assessment methods in accordance with the collection of some types of evidence.
		5.2. Select assessment methods supporting the separate collection of each type of evidence.
		5.3. Select assessment methods with attention to the combination of collecting different types of evidence.
		5.4. Select assessment methods that achieve full assessment objectives.
III. Develop anassessment tool	6. Determine the type of information used to draft assessment tasks	6.1. Only find assessment tasks that are available for use directly.
		6.2. Determine the characteristics of the information that can be immediately utilized to draft assessment tasks.
		6.3. Determine the characteristics of information that can be utilized to find ideas to generate other assessment tasks.
		6.4. Determine the type of information and reflect on the implementation process to withdraw appropriate rules for the future.
	7. Draft assessment tasks.	7.1. Draft assessment tasks only to obtain some simple assessment objectives.
		7.2. Draft assessment tasks to achieve the full assessment objectives.
		7.3. Draft multidimensional assessment tasks that allow students to come up with different

		ways to express ideas.
		7.4. Draft assessment tasks and reflect on the implementation process to withdraw appropriate rules for the next time.
	8. Design assessment instruments (scale, checklist, rubric, etc.) to assess the evidence obtained.	8.1. Select suitable tools.
		8.2. Select appropriate assessment instruments to assess the evidence obtained.
		8.3. Develop assessment instruments that meet some basic criteria according to the assessment theory.
		8.4. Develop appropriate assessment instruments to assess the evidence and create learning opportunities for students.
	9. Determine the specific steps that evaluators should take to manage and use the tool.	9.1. The content of some steps to manage and use the tool is reasonable.
		9.2. The content of all the steps to manage and use the tool is reasonable.
		9.3. The order of steps to manage and use the tool is basically reasonable but not optimal.
		9.4. The order and content of the steps are optimal (saving time, facilitating implementation, and reducing errors).
IV. Trial and finalize the assessment tool	10. Trial of the assessment tool	10.1. Select the appropriate test sample.
		10.2. Choose the right method for collecting experimental information.
		10.3. Determine the factors affecting the readiness of the subjects participating in the experiment.
		10.4. Eliminate or minimize all factors that influence the readiness of the subjects participating in the experiment.
	11. Analyze feedback from people involved in the experimental process.	11.1. Consider feedback and provide a general comment about the tool
		11.2. Select appropriate feedback analysis methods.
		11.3. Separately analyze each type of feedback.
		11.4. Analyze types of feedback with attention to the relationships between them.
	12. Finalize the assessment tool	12.1. Consider a number of factors influencing the accuracy of the assessment tool
		12.2. Determine factors that may affect the accuracy and optimality of the assessment, (assessment duration, task difficulty level, language, design, and readiness of subjects in the experiment).
		12.3. Modify the assessment tool in order to fix some of those factors.
		12.4. Appropriately finalize the assessment tool.

4. Discussion of the Obtained Results

According to the review in step 1, behavioural indicators 2 and 3 were not explicitly stated in studies and assessment standards. Among the publications analyzed, only the publication of National Register for Vocational Education and Training in Australia (2012) addressed identifying the group of students to be assessed when designing the assessment tool in general, but not to mention the specific factors that need attention to define goals when designing assessment tools such as: student level, reading comprehension, health and psychology, and student readiness, etc.

However, Assessment Literacy Inventory (Mertler & Campbell, 2005) designed based on ASTCEAS had assessed some aspects of behavioural indicators 2 and 3. This proved that there was the presence of these behavioural indicators (2 and 3) in the tacit knowledge of the assessors (Mertler & Campbell, 2005). In particular, the behavioural indicators 8 proposed by the research had not been mentioned in previous studies, but it was often grouped with behavioural indicators 9. It was the grouping of many behavioural indicators together that caused failure for many lecturers to accurately identify which behavioural indicators were responsible when designing assessment tools. As a result, it made it difficult to enhance teachers' competencies.

Thus, the proposal of the current new behavioural indicators explicitly showed the tacit knowledge of teachers. This will support teachers and learners in detecting their own problems to improve their teaching and learning practices. These behavioural indicators were consistent with the results of the students' product and thinking analysis. They have also been confirmed through the expert method and Cronbach alpha analysis results. Moreover, methods of collecting and analyzing students' thinking and data processing to find new content for the model in this study could suggest a new way for teacher educators to build a model of other teacher competencies and facilitates verifying or improving the competency model.

5. Conclusion

To construct a reflective model on the complexity of designing assessment tools, the literature on the process of designing the assessment model and data on students' thinking were synthesized and analyzed, respectively. The validity and reliability of the model were proved by analyzing the data from the survey with students and twice using the expert method by two panels. Overall, this study provided initial validity and reliability evidence to support the usefulness of the CDCAT model. The construction of this new model was approached in a comprehensive, mixed process (including both qualitative and quantitative) taking into account all the factors related to CDCAT. Research added to the literature the model that clarify the cognitive aspects of CAT by designing and highlighting tacit knowledge used for designing CAT. The results of this research will be the basis for the development of CDCAT for pre-service teachers. Based on behavioural indicators of the model, a lecturer can plan

training and design tasks for pre-service teachers. The level achieved by each pre-service teacher can be determined through the application of the quality criteria of the CDCAT. The next step was to design learning tasks for each behavioural group to train pre-service teachers so that they could achieve higher quality criteria. The CDCAT model had a significant value in providing additional insights into the difficulties in the cognition of teachers and the nature of teachers' development in their designing CAT and making these visible for teacher-educators and pre-service teachers.

6. Implications for Teacher Education

One of the most obvious uses of the model was to provide a basis for directing teacher competency development. There is evidence that depending on teachers' identified approaches to assessment, areas of confidence, and professional development priorities and preferences, the teacher had approaches to specialized learning or maintained other learning goals in the assessment learning (Linh & Tra, 2016). The tool provided diagnostic information about teachers' CDCAT as the foundation for developing differentiated and targeted professional learning. By viewing the development of the CDCAT as a mixture of dimensions rather than aggregating scores to obtain an overall score, the model can assist educators-teachers and teachers in responding and guiding adjustments. The quality criteria that clearly defined teachers' CDCAT expectations can be used as tools for monitoring changes and, more importantly, as learning aids. Besides, models can be used to enhance instruction by providing educators, mentors, and students with descriptions of common concepts and language to foster discussion and feedback. Furthermore, the quality criteria described in the CDCAT model can elicit teacher dialogue with questions about incorporating theory and practice into classroom assessment.

In addition, by using models and looking at patterns collected during teacher development, educators-teachers can determine when a teacher's CDCAT is growing faster and changing slowly. Also, researchers can use the model and examine samples collected from different teachers, in different contexts, to discover factors that affected teachers' CDCAT development.

The strength of this model was that it complemented the model in some behavioural indicators to clarify the cognitive aspects of CAT design (indicator 8) and highlight tacit knowledge used for designing assessment tools (indicator 2,3). With the outlined 48 quality criteria, this model also provided more detailed information about the nature of teacher development as they became more competent in CAT designing. However, because it required a lot of actions when assessing teachers' CDCAT to report in detail the level of teacher achievement in each behavior indicator, using the model in teacher training was more appropriate than in summary assessment.

7. Research Limitations

The limitation of the study was that the model of CDCAT was built through the analysis of small data.Although research design with this small sample can deeply analyze the inner learner's activities through a think-aloud and analyzing

the learning portfolio, further research should continue by measuring CDCAT on a wider sample, varied pre-service teachers, and using the Rasch model for analysis.In comparison with factor analysis, the use of the Rasch model will give more meaningful results to quantify the construct being studied (William, 2016). The strength of the Rasch model lies in its ability to determine if the items are not related and item fit (Randall & Engelhard, 2010). A factor analytic approach is more appropriate if the focus of the study is to account for and establish the multiple dimensions of the construct (Sick, 2011). The CDCAT is one dimension of assessment competence, so it was consistent with the Rasch model. If studied with larger samples, the model of CDCAT could be standardized and the developmental model could be built to support the CDCAT training for teachers and pre-service teachers.

8. References

Alonzo, D. A. (2016). *Development and application of a teacher assessment for learning literacy tool* [Doctoral dissertation, University of New South Wales].

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington: National Council on Measurement in Education.

Ananiadou, K., & Claro, M. (2009). 21st Century skills and competences for new millennium learners in OECD countries. https://doi.org/10.1787/218525261154

Assessment Reform Group. (2008). *Changing assessment practices: Process, principles and standards*. http://www.aaia.org.uk/content/uploads/2010/06/ARIA-Changing-Assessment-Practice-Pamphlet-Final.pdf

Association for Educational Assessment-Europe. (2012). *European framework of standards for educational assessment 1.0*. Rome: Edizioni Nuova Cultura.

Australia Goverment Department of Education, Employment and workplace Relations. (2012). *TAEASS502B Design and develop assessment tools*. https://training.gov.au/TrainingComponentFiles/TAE10/TAEASS502B_R1.pdf

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, 5, 7–74.

Bol, L., Stephenson, P. L., O'connell, A. A., & Nunnery, J. A. (1998). Influence of experience, grade level, and subject area on teachers' assessment practices. *The Journal of Educational Research*, *91*(6), 323–330. http://doi.org/10.1080/00220679809597562.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how?. *CBE Life Sci Education*, *15*(4). http://doi.org/10.1187/cbe.16-04-0148

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practices, 30*(1), 3-12.

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD Alexandria, Virginia USA.

Caena, F. (2011). *Literature review teachers' core competences: requirements and development*. https://www.researchgate.net/publication/344906332_Literature_review_Teachers'_core_competences_requirements_and_development

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice,*17, 419–438. http://doi.org/10.1080/0969594X.2010.516643

DeLuca, C., LaPointe, D. M., & Luhanga, U. (2016). Approaches to classroom assessment inventory: A new instrument to support teacher assessment literacy. *Educational Assessment, 21(*4), 248-266. http://doi.org/10.1080/10627197.2016.1236677.

Department for Education-United Kingdom. (2012). *Teachers' standards*. http://www.education.gov.uk/schools/teachingandlearning/reviewofstandards/a00205581/teachers-standards1-sep-2012.

Department of Education-Australia. (2012). *Australian professional standards for teachers*. http://www.teacherstandards.aitsl.edu.au/OrganisationStandards/Organisation

DeVellis, R. F. (2003). *Scale development: Theory and applications (*2nd ed.*)*. Thousand Oaks, CA: Sage Publications.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132. http://doi.org/10.1080/15434303.2011.642041

Gardner, J. (2006). Assessment for learning: a compelling conceptualization. In J. Gardner (Ed.), *Assessment and learning* (pp197–204). Sage.

González, E. F. (2018). *The impact of assessment training on English as a foreign language* [Doctoral dissertation, University of Southampton].

Griffins, P. (2015). *Assessment for teaching*. Cambridge University Press.

Gulikers, J. T. M., Bastiaens, T. J., & Martens, R. L. (2005). The surplus value of an authentic learning environment.*Computers in Human Behavior, 21*(3), 509–521. http://doi.org/10.1016/j.chb.2004.10.028

Gutierrez, S. L. (2014). *From national standards to classrooms: A case study of middle level teachers' assessment knowledge and practice* [Doctoral disseration, Western Michigan University].

Hammami, A. (2016). *ESL teacher profiles of ICT integration in their classroom practices and assessment activities: A portrait viewed through the lens of some Quebec teachers' social representations* [Doctoral dissertation, Université de Sherbrooke].

Herppich, S., Praetorius, A. K., Foster, N., Glogger-frey, I., Karst, K., Leutner, D., Behrmann, L., Bohmer, M., Ufer, S., Klug, G., Hetmanek, A., Ohle, A., Bohmer, I., Karing, C., Kaiser, J., & Sudkamp, A. (2017). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*, *76,* 181-193. http://doi.org/10.1016/j.tate.2017.12.001

Huba, M. E., & Freed, J. E. (2000). *Learner-centered cssessment on college campuses: Shifting the focus from teaching to learning*. Pearson.

Jarr, K. A. (2012). *Education practitioners' interpretation and use of assessment results* [Doctoral dissertation, University of Iowa]. http://ir.uiowa.edu/etd/3317

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15. http://doi.org/10.1111/j.1745-3992.2004.tb00164.x

Linh, N., & Tra, D. (2016). From surveying reality to proposing some solutions to develop competence of designing assessment tool in training and improving physics teacher. *HNUE journal of science, 8*, 213-225.

McGee, J. R. (2012). *Developing and validating a new instrument to measure the self-efficacy of elementary mathematics teachers* [*Doctoral* dissertation, University of North Carolina]**.**

Mertler, C. A., & Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the Assessment Literacy Inventory* [Paper presentation at meeting]. American Educational Research Association, Montreal, Canada.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research, and Evaluation*, *10*(*7*), 71-81. http://doi.org/10.7275/q7rm-gg74

Nary, T. (2014). *Development and validation of classroom assessment literacy scales: English as a foreign language (EFL) instructors in a Cambodian higher education setting* [Doctoral dissertation, College of Education Victoria University Melbourne, Australia].

National Register on Vocational Education and Training in Australia. (2012). *Design and develop assessment tools.* TAE10 Training and Education Training Package version 2.0. Commonwealth of Australia.

North Central Regional Educational Laboratory. (2016). *Indicator: Assessment*. http://www.ncrel.org/engauge/framewk/pro/literacy/prolitin.htm

Organisation for Economic Co-operation and Development (OECD). (2018). *Teaching and learning international survey*. http://www.oecd.org/education/school/TALIS-2018-MS-Teacher-Questionnaire-ENG.pdf

Okoli, C., & Pawlowski, S.D. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management*, *42*, 15–29. http://doi.org/10.1016/j.im.2003.11.002

Perry, M. L. (2013). *Teacher and principal assessment literacy* [Doctoral Dissertation, University of Montana].

Popham, W. J. (2013). *Classroom assessment: What teachers need to know* (7th ed.). Boston: Pearson.

Randall, J., & Engelhard, Jr. G. (2010). Using confirmatory factor analysis and the Rasch model to assess measurement invariance in a high stakes reading assessmen., *Applied Measurement in Education*, *23*(3), 286-306. http://doi.org/10.1080/08957347.2010.486289

Reddy, Y. M. (2011). Design and development of rubrics to improve assessment outcomes. *Quality Assurance in Education*, *19*(1), 84-104. http://doi.org/10.1108/09684881111107771

Shavelson, R. R. (2013). An approach to testing and modeling competences. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and measuring competencies in higher education: Tasks and challenges* (pp. 29-43).

Sick, J. (2011). Rasch measurement and factor analysis. *SHIKEN: ALT Testing and Evaluation SIG Newsletter, 15*(1), 15-17.

Stiggins, R. (1987). Design and development of performance asessment. *Educational Measurement: Issues and Practice*, *6*(3), 33-42. http://doi.org/10.1111/j.1745-3992.1987.tb00507.x

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758–765.

Stiggins, R. J. (1999b). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*(1), 23-27. http://doi.org/10.1111/j.1745-3992.1999.tb00004.x

Stiggins, R. J., Conklin,N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany: State University of New York Press.

Suskie, L. (2004), *Assessing student learning: A common sense guide*. Anker Publishing Company, Bolton, MA.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency ofperformance criteria across scale levels. *Practical Assessment, Research and Evaluation*, *9*(2). http://doi.org/10.7275/jtvt-wg68

Tillema, H., Leenknecht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning – A review of research studies. *Studies in Educational Evaluation, 37,* 25–34. http://doi.org/10.1016/j.stueduc.2011.03.004

Vidacovich, C. (2015). *Measuring teachers' knowledge and use of data and assessments: Creating a measure as a first step toward effective professional development* [Doctoral dissertation, University of Denver].

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education, 30*, 751–772. http://doi.org/10.2307/20466661

Wiggins, G., & McTighe, J. (2005). *Understanding by design*. Association for Supervision and Curriculum Development.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149-162. http://doi.org/10.1016/j.tate.2016.05.010

Zhang, Z., & Burry-stock, J. A. (1997). Assessment practices inventory: A multivariate analysis of teachers' perceived assessment competency. https://www.semanticscholar.org/paper/Assessment-Practices-Inventory%3A-A-Multivariate-of-Zhang-Burry Stock/2ade782d7f56a4143f74985e87ef786d90c9eeb3