Development and Validation of a Measurement Scale for Teacher Assessment Literacy in Higher Education in China

A Thesis Submitted to the College of Graduate and Postdoctoral Studies in Partial Fulfillment of the Requirements for Doctor of Philosophy in the Department of Educational Psychology and Special Education

University of Saskatchewan

Saskatoon

By

Qin Wang

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a graduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Educational Psychology and Special Education
28 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan, S7N 0X1
Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
105 Administration Place
Saskatoon, Saskatchewan, S7N 5A2
Canada

Abstract

The development and prevalence of online education have brought about new challenges to teachers' assessment practices. Especially, due to the outbreak of the coronavirus disease (COVID-19) in late 2019, online education was brought to the forefront of higher education. For instance, teachers need to offer online courses for students and implement online assessment tasks. This makes it necessary for teacher assessment literacy (AL) to cover the assessment competence in the digital environment. Furthermore, teacher AL was also conceptualized from a unidimensional to a multidimensional perspective. However, most AL measurement scales still adopted the 1990 Standards as the guiding framework. A measurement scale that can reflect current assessment needs is lacking. Against this background, the current study adopted Onwuegbuzie et al.'s (2010) instrument development and construct validation (IDCV) process as a mixed methods framework for developing and validating a measurement scale that can be used to measure higher education teachers' AL in both classroom and digital contexts. It contains 10 phases from identifying and conceptualizing the construct of the interest to validating and evaluating the process and product. In Phases 1 to 5, the measurement scale was developed based on the framework generated from a thorough literature review and focus group interviews. Phases 6 to 9 involve quantitative and qualitative data analyses and crossover analyses based on a sample of $N = 265$ higher education teachers from over 15 disciplines and five levels of universities in China. The pilot test validated the content-related validity, including item, face, and sampling validity of the measurement scale. The analyses of data from the field test indicate that the measurement scale has adequate reliability, and construct-related validity, including structural, convergent, and discriminant validity. In Phase 10, both the product and the process were comprehensively evaluated to help the researcher reflect on the IDCV process and discover

areas for further improvement of the instrument. The measurement scale developed and validated in this research can be used by researchers and practitioners to support assessment practices in higher education, for instance, to help promote interventions or diagnose problems in training pre- or in-service teachers on assessment.

Keywords: teacher assessment literacy, digital assessment literacy, higher education, instrument development, instrument validation

Acknowledgments

I would like to express my deepest appreciation to my supervisor, Dr. Amin Mousavi, who is a person with a strong sense of responsibility and academic literacy. During my PhD study, he gave me valuable feedback and support on my courses and research work. In learning and teaching, he not only helped me learn many courses but also supported me in preparing courses and completing teaching. In doing research, Dr. Mousavi provided valuable comments and suggestions for my papers and guided me from selecting journals to completing publication.

I'm also extremely grateful for the feedback and suggestions on my thesis from Dr. Laureen McIntyre, Dr. Pei-Yin Lin, Dr. Zhi Li, and Dr. Marguerite Koole. They generously provided knowledge and expertise for me to complete my thesis.

I am also grateful to my friends and my research partners, Dr. Chang Lu, Dr. Yizhu Gao, and Dr. Fu Chen, for their help and support in my PhD study and research work.

I would thank my family for giving me financial and emotional support. I would also like to thank my cat for all the entertainment and emotional support.

Lastly, I would thank all the experts who participated in my study and all the higher education teachers who completed the survey of my thesis.

Table of Contents

List of Figures

FIGURE PAGE

List of Tables

TABLE PAGE

List of Terms

ACRONYM	FULL NAME
AL	*Assessment Literacy*
DAL	*Digital Assessment Literacy*
IDCV	*Instrument Development and Construct Validation*
TALQ	*Teacher Assessment Literacy Questionnaire*
CALI	*Classroom Assessment Literacy Inventory*
ALI	*Assessment Literacy Inventory*
API	*Assessment Practices Inventory*
TCoA	*Teacher Conceptions of Assessment*
TCoA-IIIA	*Teacher Conceptions of Assessment Version III abridged inventory*
ACAI	*Approaches to Classroom Assessment Inventory*
MTMM	*Multitrait-Multimethod Matrix*

CHAPTER ONE: INTRODUCTION

Over the past few decades, the increasing global trend of school accountability has put great pressure on teachers, who have been repeatedly called to use assessment information in decision-making (Darling-Hammond, 2010; Schildkamp & Lai, 2013; Stiggins, 2017). Specifically, teachers are expected to be capable of making judgments about student learning including, but not limited to, administering, interpreting, and reporting assessment results and working ethically (Williams, 2015). Given this responsibility, it is commonly acknowledged that teachers should be knowledgeable of assessment and related skills (Brookhart, 2011; Popham, 2009). Such knowledge about assessment and the skills of integrating and using assessment to effectively measure and promote student learning has been defined as teachers' "assessment competency," or "assessment literacy (AL)" (Stiggins, 1991a; 1991b). Insufficient AL may result in teachers implementing unreliable assessments, thus making ill-informed instructional decisions which, in turn, lowers the quality of teaching and learning (Xu & Brown, 2017). The development of technologies also brought about new opportunities and challenges to teachers' assessment work, such as online learning platforms and various learning tools that have been used for collecting evidence of formative and summative assessment. Therefore, to help students achieve higher levels of learning achievement, teachers need to develop appropriate AL types and improve their AL levels (Stiggins, 1995).

Research Background

Over the past few decades, extensive research has been conducted to examine the definition, conceptualization, measurement, and development of teacher AL. Early research attempted to define and investigate AL focusing on measurement theory and practice (e.g., the Standards of AFT, NCME, NEA, 1990). This unidimensional perspective can be reflected in

early professional standards for assessment, for instance, the 1990 *Standards*, which includes seven assessment competencies relating to assessment knowledge and skills. The 1990 *Standards* serves as a starting point for subsequent related studies. The majority of research during this phase framed AL around the core theme of knowledge and skills in assessment. For instance, Stiggins (1991a; 1991b) listed the key actions the assessment literate teachers should do to ensure a sound assessment, including explaining and sharing with students the assessment content and purpose, designing an accurate assessment process, and reporting results in an effective way to all stakeholders of assessment. Popham (2011) described AL as teachers' understanding of the basic notions and procedures of assessment.

However, as exploration and investigation into what AL entails increased, AL has been examined from a unidimensional perspective (i.e., assessment knowledge and skills) to a multidimensional one, such as sociocultural and socio-emotional aspects (e.g., Brown, 2004; Brown et al., 2011; Fulcher, 2012; James & Pedder, 2006; Pastore & Andrade, 2019; Smith et al., 2014; Xu & Brown, 2016). For example, recent studies have linked AL with assessment identity (e.g., Cowie et al., 2014; Looney et al., 2018; Xu & Brown, 2016) which highlighted the importance of developing teachers' understanding of assessment practices and themselves as assessors. This concept was especially emphasized in Looney et al.'s (2018) dynamic framework of teacher assessment identity, which consists of a series of assessment knowledge and skills, teachers' confidence in implementing assessment, and feelings about assessment. This framework stated that the core of teacher assessment identity should be teachers' awareness of who they are. The AL frameworks of Pastore and Andrade (2019) and Chan and Luk (2022) incorporated other socio-emotional aspects such as ethical issues (e.g., assessment malpractices, fairness, and equity) into the AL concept. Researchers also suggested that AL is not stable and

changes over time because of a variety of internal and external factors (Mockler, 2011; Willis et al., 2013). For instance, in Xu and Brown's AL framework (2016), AL was conceptualized as an iterative and dynamic system.

The evolvement of the AL concept has become evident in recent teacher standards across diverse educational contexts, which have developed from a rough and limited perspective to a wide and thorough range of activities in the assessment field. Such change reflects more emphasis than in the past on assessment for ethical aspects such as fairness, teacher support, and communicating or reporting assessment results (DeLuca et al., 2016a; Pastore & Andrade, 2019). These teacher standards have been revised to cover a variety of topics regarding assessment practices (e.g., administration and use of assessment results, the use of digital assessment) and suggested to define AL broadly and pay attention to effectively integrate educational policies and teacher education paths. It can be seen in policy documents and assessment standards in North America, Europe, China, and elsewhere (DeLuca et al., 2016a). Furthermore, the use of digital tools in assessment was also emphasized in certain official documents. For instance, the guidelines issued by the Ministry of Education (MOE) in China reflect the evolution of assessment expectations. From 1999 to 2022, MOE has continuously increased new assessment needs, such as making use of technologies of artificial intelligence and big data in assessment.

Over the past decade, in order to examine whether teachers keep pace with increasing assessment expectations in assessment standards, researchers have conducted numerous studies to qualitatively evaluate teacher AL using methods such as interviews, written diaries, and classroom observations (e.g., Atjonen et al., 2022; Dinan Thompson, & Penney, 2015), or quantitatively develop and validate instruments to investigate teacher AL levels (e.g., Plake et al., 1993; Mertler, 2003). Regarding the quantitative measurement tools, various instruments

have been developed and validated in different contexts, such as the Teacher Assessment Literacy Questionnaire (Plake et al., 1993) and the Classroom Assessment Literacy Inventory (Mertler, 2005). Some previous studies on developing instruments have used the 1990 Standards as the guiding framework. Brookhart (2011) argued that the 1990 Standards have become outdated because the current conceptions of formative assessment and technical and social issues involved in standards-based educational reforms were not included (Brookhart, 2011). Therefore, it is suggested that measures of AL need to be constructed and analyzed based on contemporary assessment standards (Brookhart, 2011; Gotch & French, 2014).

Research Gaps

Despite these efforts in the measurement of AL, Gotch and French (2014) argued that the psychometric evidence in measuring AL is still weak and suggested researchers increase the validity of AL measures by examining whether assessment content involved in AL measures is consistent with changes in contemporary assessment standards, such as the use of formative assessment and digital assessment. The advent of technology has brought new opportunities and challenges to teachers' assessment practices. The use of web-based tools in assessment practices has fundamentally changed the landscape of educational assessment, offering opportunities to tailor assessments based on individual learning styles, provide timely feedback, and facilitate data-driven decision-making. For instance, reports from computer-based testing and online learning platforms facilitate teachers' evaluations of student learning, thus enabling them to give timely feedback to students that can be recorded online. Especially during the coronavirus disease (COVID-19) pandemic, teachers have increasingly adopted mobile apps, software programs, and web-based learning platforms to support students' virtual, remote, or blended learning through synchronous, asynchronous, or hybrid modes. Although teaching and learning

have returned to normal, the use of web-based tools in assessment has continued because of the availability and convenience of online tools.

As assessment is subject-specific and context-dependent in most educational practices (Abell & Siegel, 2011; Taylor, 2013), the knowledge and skills within AL do not apply to all subject areas and learning contexts. Therefore, some previous studies have taken a more integrated approach to AL, situating teachers' conceptions and practices of assessment in a specific context rather than mastering general assessment rules (Frey & Fisher, 2009; Wyatt-Smith et al., 2010). Xu and Brown (2016) suggested that placing teachers' assessment decision-making and action-taking processes among the different dynamics in specific contexts appears essential to understanding and developing solutions to improve teacher AL. Taking higher education as an example, teachers in higher education tend to undertake more targeted responsibilities apart from the general assessment dimensions that can be applied to all stages of education. For instance, teachers are expected to conceptualize assessment from an institutional perspective and treat assessment as a long-term, developmental process that supports students' capacity for academic integrity, self-regulated learning, and lifelong learning (e.g., Brown & Race, 2013; Kaslow, et al., 2007). Teachers are also required to use the clearly defined institution-wide grade scale system and apply it consistently across individual programs and courses (e.g., James et al., 2002; Luth, 2010).

Given the transformations in contemporary assessment practices, the increasing proportion of digital assessment, the critiques directed at the previous AL measures, and the unique characteristics of assessment in higher education, it is of great importance to develop a new AL measurement scale that more accurately reflects teachers' current assessment needs in higher education. Such a study has the potential to uncover the current AL level of higher

education teachers and provide insights for future teacher professional development in assessment.

Research Purpose

Due to the outbreak of COVID-19 in late 2019, China implemented a large-scale transition from traditional face-to-face teaching to fully online courses in 1,291 universities across the country, driven by the government's policy of "non-stop teaching and learning" at the beginning of 2020 (Bao, 2020; Sun et al., 2020). Therefore, teachers need to deliver courses and assess student learning through Learning Management Systems (LMSs), online video platforms (e.g., DingDing), or social media platforms (e.g., WeChat). Thus, it has brought new challenges to teachers' assessment work during and even after the pandemic. Furthermore, so far, some higher education institutions still lack a comprehensive assessment system that reflects contemporary assessment standards or provide insufficient training for teachers, thus making them clueless about the "what" and "how" of assessing student learning. Therefore, developing a measurement scale on AL in both classroom and digital settings is necessary.

Based on this context, the current study aims to develop and validate a measurement scale to measure higher education teachers' AL in both classroom and digital contexts. More specifically, this study used Onwuegbuzie et al.'s (2010) Instrument Development and Construct Validation (IDCV) process for optimizing the development and validation of the quantitative instrument. This process contains 10 phases, including 1) conceptualizing the construct of interest; 2) identifying and describing behaviors that underlie the construct; 3) developing the initial instrument; 4) pilot-testing the initial instrument; 5) designing and field-testing revised instrument: instrument fidelity; 6) quantitative analysis phase; 7) qualitative analysis phase; 8) mixed analysis phase: qualitative-dominant crossover analyses; 9) mixed analysis phase:

quantitative-dominant crossover analyses; 10) evaluating the process and product. A mix of qualitative and quantitative methods was adopted in the IDCV process.

Organization of the Thesis

The rest of the document is organized as follows. In Chapter Two, the existing theories and theoretical frameworks of AL and digital assessment literacy (DAL) were reviewed. The key components of the main theoretical frameworks were compared to identify the potential theoretical frameworks that may be appropriate for guiding the integration of DAL into teacher AL. Then, the professional documents related to educational assessment were reviewed and analyzed using document analysis. The specific content categories were identified. Previous studies on teacher AL measures and frameworks of instrument development and validation were also reviewed. The potential research gaps were identified. In Chapter Three, the IDCV process that was used for developing and validating the measurement scale was introduced. The methods used in each phase were explained. In Chapter Four, how this study implemented each step of the IDCV process, including data collection and data analysis, the interpretation of the results, and evaluations of the IDCV process were introduced and discussed. Chapter Five provides a summary of the study, including the implications, limitations, and future research.

Chapter Summary

Chapter One introduced the context of the problem addressed in this research. First, the context of AL and conceptualizations of AL were introduced. Then, the state-of-the-art studies on international assessment standards and teacher AL measurement instruments were reviewed. The research significance and purpose were identified. Lastly, this chapter outlined the organization of the remaining document.

CHAPTER TWO: LITERATURE REVIEW

This chapter provides an overview of the definition of AL and existing AL and DAL models. Specifically, four AL models with different foci were introduced. The core components of AL in each model were discussed. In addition, the main issues involved in DAL were discussed. This chapter also discussed policy considerations in AL and reviewed several international teacher standards for assessment using document analysis. Then, the measurement of AL was reviewed, specifically including the commonly used AL measures and their psychometric properties. It further highlighted the contextual factors associated with AL measurement. It also compared several frameworks of instrument development and validation and highlighted the advantages of Onwuegbuzie et al.'s (2010) IDCV process. Finally, the existing research gaps were identified.

The Definition of AL

AL is a concept first introduced by Stiggins (1991a) to refer to teachers' capabilities to apply assessment knowledge and skills to implement quality assessment tasks and to use the accurate interpretation of assessment results to inform instruction and guide student learning (Fullan & Watson, 2000; Stiggins & Duke, 2008). According to the American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), and National Education Association (NEA) (1990), AL represents "the process of obtaining information that is used to make educational decisions about students, to give feedback to students about their progress, strengths, and weaknesses, to judge instructional effectiveness and curricular adequacy, and to inform policy." Popham (2011) believes that AL is teachers' understanding of the basic concepts and procedures involved in assessment. To date, no consensus on the definition of AL exists. Researchers have proposed various models and frameworks of teacher AL with different

foci, such as AL in the classroom context (Popham, 2011), and connecting educational assessment and teacher education (Xu & Brown, 2016).

Studies of teacher AL can be traced back to several stages. Initially, it was discussed within the single dimension of assessment knowledge and skills, that is, the practical aspects of AL. In 1990, the AFT, NCME, and NEA cooperated to develop the Standards for Teacher Competence in Educational Assessment of Students. The 1990 *Standards* consists of seven assessment competencies relating to assessment knowledge and skills, including selecting and developing assessment methods and strategies appropriate for instructional decisions, implementing, grading, and interpreting the assessment results, making decisions about specific students and organizing instruction, curriculum and school improvement based on assessment results, communicating assessment results to stakeholders (e.g., students, parents, and other educators), and being aware of unethical, illegal, and otherwise inappropriate assessment practices and uses of assessment information. Since the 1990 *Standards* was proposed, subsequent studies have been conducted in discussing teacher AL, most of which framed AL around the "core" theme of knowledge and skills within assessment (e.g., Fullan & Watson, 2000; McMillan, 2001; Shepard, 2000; Stiggins, 1991a, 1991b, 2009). For instance, Stiggins (1991a; 1991b) listed the key actions the assessment literate teachers should do to ensure a sound assessment: 1) explaining and sharing with students the assessment purpose; 2) clearly stating assessment content; 3) designing an accurate assessment process; 4) reporting results in an effective way to all stakeholders of assessment. Stiggins (2009) further added the aspect of "involving students in their own assessment". The five key elements well state classroom assessment competencies for teachers. McMillan (2001) summarized a list of 11 principles for assessment, which described the assessment as a process of conducting professional judgment

built upon distinct but connected principles of measurement evidence and evaluation. The principles also involve the influences of assessment and the standards of a good assessment. Brookhart (2011) argued the *Standards* lacked consideration of formative assessment and standards-based assessment and pointed out the need to propose a more fine-grained definition of AL. According to Brookhart, teachers should be capable of 10 principles related to assessment practices (see details in Brookhart, 2011).

These standards discussed the AL knowledge base from different foci, such as targeting various subject areas (e.g., language, mathematics), assessment purposes (e.g., formative assessment or summative assessment), and different stakeholders (e.g., teachers, students, policymakers) (e.g., Abell & Siegel, 2011; Brookhart, 2011; Popham; 2009). They set goals for assessment education and lay a solid foundation for empirical studies on AL (Xu & Brown, 2016).

Existing Frameworks of AL

Previous research has revealed the complex nature of teacher AL. First, although any models or frameworks of AL contain predominantly the "core" knowledge dimension that is applicable to all the contexts (Xu & Brown, 2016), in most cases, assessment is implemented in the subject-specific and context-dependent circumstances (Abell & Siegel, 2011; Taylor, 2013; Willis et al., 2013). Second, with the advancement of educational practices, for instance, online education, the changing needs of teacher assessment practices make AL not static, which drives the continuous exploration of teacher AL forward. Third, assessment is also a sociocultural activity, as different stakeholders may be involved in social interactions related to assessment (Broadfoot, 1996; Gipps, 2002). Apart from a set of outlined universal skills that can be applied to all contexts, teacher assessment practices are also influenced by not only expected learning

(curriculum), and pedagogical guidance but also community expectations and national or state policies (Looney et al., 2018). Therefore, more recent studies have shifted to examine AL as a multidimensional concept beyond a one-dimensional perspective of knowledge and skills, covering a complicated interaction between different components connected to social, cultural, policy, professional, and experiential elements (e.g., Brown, 2004; Brown et al., 2011; James & Pedder, 2006; Smith et al., 2014; Xu & Brown, 2016). This shift is from the initial teachers' competencies on planning and implementing quality assessment tasks to interpreting assessment results and further engaging students as active participants in assessment (Looney et al., 2018; Pastore & Andrade, 2019). Furthermore, the emotional dimension of teacher assessment work, such as concepts of assessment, beliefs, and feelings on assessment, received some attention (e.g., Brown, 2004; Brown et al., 2011; Brown & Remesal, 2012; Giraldo Aristizábal, 2018; James & Pedder, 2006). Drawing on Thompson's (1992) earlier research on teachers' views and beliefs regarding the teaching and learning of mathematics, Brown et al. (2011) explored teachers' conceptions of assessment and further found teachers have multiple and conflicting concepts of assessment. Some researchers also explored teachers' values, beliefs, and perceptions of assessment (e.g., James & Pedder, 2006, Smith et al., 2014), which has been viewed as an important factor in shaping assessment practice and were influenced by sociological and psychological elements (Looney et al., 2018). In some relevant studies, these psychological aspects were integrated into teachers' identities as assessors (i.e., conceptions, beliefs, feelings) (Looney et al., 2018; Xu & Brown, 2016). Adie (2013) originally proposed assessment identity which refers to the perceptions that teachers had as assessors. Its importance has been highlighted for developing teachers' understanding of assessment practices and themselves as assessors (Looney et al., 2018; Scarino, 2013; Xu & Brown, 2016).

AL has also been investigated from a socio-cultural and contextualized perspective, which suggests that it is not stable and changes over time because of a series of internal and external factors (e.g., Beijaard et al., 2004; Mockler, 2011; Willis et al., 2013). Willis et al. (2013) argued that discussion about teacher AL needs to be theory-driven, and should consider it as a social, dynamic, and layered ethical practice. As stated by Xu and Brown (2016) and Looney et al. (2018), teacher AL development involves the accumulation of assessment knowledge and skills, as well as the development of a set of contextually appropriate, complex, and interrelated competencies. This results in recent attempts to focus on proposing more fine-grained and comprehensive AL frameworks. Based on a thorough review of related studies, Xu and Brown (2016), Looney et al. (2018), Pastore and Andrade (2019), and Chan and Luk (2022) reconceptualized teacher AL from a focus on knowledge base to covering various aspects related to assessment. Figure 1.1 shows these four conceptual frameworks of teacher AL. Teacher Assessment Literacy in Practice (TALiP) was a conceptual framework of AL proposed by Xu and Brown (2016). It introduced a pyramidal model of six components in which the knowledge base is located at the bottom. The knowledge base was revised based on the 1990 *Standards* and recent updates (e.g., Brookhart, 2011; Klinger et al., 2015), which includes knowledge about disciplines and pedagogical content, as well as knowledge about assessment procedures such as assessment methods, grading, giving feedback, communicating assessment results, and assessment ethics. It serves as the basis of all other components and serves as standards and criteria for evaluating assessment practices. The following components are teacher conceptions of assessment as an interpretive and guiding framework, macro socio-cultural and micro institutional contexts as the boundaries, teacher AL in practice as compromises made among tensions, teacher learning as the impetus for advancing TALiP, and assessor identity

(re)construction as the ultimate goal (the pinnacle of the framework). According to the proposed framework, teachers with a high level of AL are those who constantly reflect on their assessment practice, participate in professional activities related to community assessment, participate in professional communications related to assessment, self-interrogate their assessment concepts, and seek resources to understand assessment and their role as assessors. The components in the TALiP framework are dynamic and interrelated. It also follows a cyclical pattern, displaying a multidirectional flow between the bottom and top. Similarly, in Looney et al. (2018)'s review study, they argued that assessment identity is more sufficient than AL to reflect the scope and complexity of the dimensions involved in teacher assessment work, given the significance of teachers' feelings and beliefs about their assessment practices, and especially, their role of being assessors. Following this argument, they proposed a dynamic framework of teacher assessment identity encompassing "*I know*" (teachers' knowledge and skills), "*I feel*" (teacher's feelings about assessment), "*I believe*" (teacher's beliefs about assessment), "*I am confident*" (teacher's confidence and self-efficacy in undertaking assessment), and "*My role*" (who teachers are). In this framework, teacher assessment identity is constituted of a number of assessment knowledge and skills, teachers' confidence in implementing assessment, and their beliefs and feelings about assessment. This framework emphasizes not just what teachers know and do, but also who they are, which is the core of teacher assessment identity. However, despite providing significant insights into the current conceptualization of AL, these frameworks failed to provide an accurate assortment of multiple dimensions. Based on the dialectical analysis of these competing models, Pastore and Andrade (2019) proposed a three-dimensional framework encompassing conceptual, praxeological, and socio-emotional aspects by taking a holistic and adaptive perspective on competencies. The three-dimensional model includes:

1. Conceptual knowledge dimension

   Know what assessment is in terms of different models and methods (e.g., basic knowledge of assessment, data analysis)

2. Praxeological dimension

   Integrate the assessment process with other teaching practices (e.g., choosing assessment strategies to gather information about student learning)

3. Socio-emotional dimension

   Deal with the social and emotional aspects of assessment (e.g., effectively work with stakeholders; understand their role as assessors; ethical aspects)

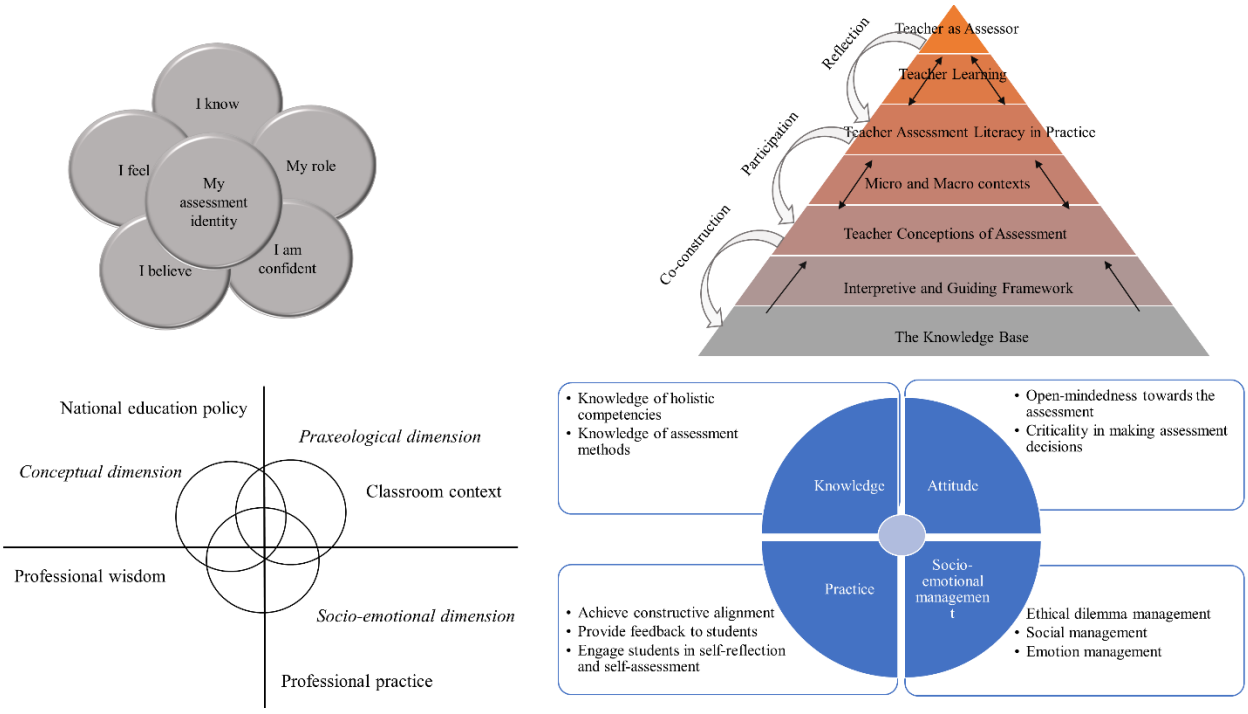
According to this model, AL is a dynamic competence, in which teachers should have the ability to use knowledge and skills flexibly and creatively across a range of educational contexts. In this model, AL was considered not only in terms of knowledge and skills about assessment but also in its social, emotional, and ethical aspects. Within the hierarchical and horizontally stacked cultural contexts of a classroom, school, system, and country, these dimensions might be emphasized in different ways. Moreover, the model can be used to support teacher education programs by examining the components of AL and the developmental paths to follow in order to reduce the distance between theories, policies, and assessment practice.

Chan and Luk (2022) argued that little attention was paid to the holistic competency development of teacher AL in previous models. Thus, they conducted interviews with academics in this field and proposed a framework of teacher AL in holistic competencies with four dimensions and 10 sub-dimensions. This framework includes 1) knowledge of holistic competencies and assessment methods (knowledge); 2) teachers' mindset toward assessment (attitude); 3) assessment-related capabilities (practice); 4) teachers' management of ethical

dilemmas, social relationships, and emotions (socio-emotional management). The dimensions of knowledge, practice, and socio-emotional management were similar to Pastore and Andrade's (2019) three-dimensional AL framework. Teacher assessment identity, which was included in the socio-emotional dimension in Pastore and Andrade's (2019) AL framework, was rephrased to the attitude dimension which focuses on teachers' open-mindedness toward assessment and criticality in making decisions.

Figure 1.1

A Sample of AL Conceptual Models



Note. Top left: teacher assessment identity (Looney et al., 2018). Top right: TALiP (Xu & Brown, 2016). Bottom left: three-dimensional AL framework (Pastore & Andrade, 2019). Bottom right: four-dimensional AL framework (Chan & Luk, 2022).

Digital Assessment Literacy

Digital assessment has been an integral part of teacher assessment practices due to the advancement of online education and technology-based teaching and learning. Digital assessment can enable teachers to better use the assessment resources and information (Oldfield et al., 2012), and show great potential in breaking through the limitations of traditional paper-based assessments to assess more complex knowledge and cognitive skills (Jamil et al., 2012). Therefore, to fully utilize the advantages of digital assessment, teachers need to be proficient in designing and implementing assessments by using digital tools or in digital contexts. The term, DAL, was first proposed by Eyal (2012) to refer to the role of a teacher as an assessor in the digital environment. According to Eyal (2012), teacher DAL is not stable but a continuum, moving from a basic level to an intermediate and advanced level. The basic level involves the use of digital tools. The intermediate level is a higher literacy level that involves the practical aspect of assessment, for example, implementing assessment tasks in a digital environment. The advanced level requires teachers to implement advanced assessment methods based on constructivist-social learning and self-regulated learning. Eyal (2012) concluded that a teacher with higher levels of DAL would be able to strategically use diverse applications and technological platforms or systems to drive students forward by adjusting various assessment strategies. The statements of DAL provide guidance for researchers to investigate teacher AL in the digital world and for schools to improve professional development programs aimed at creating digital assessment–literate teachers.

As “an entirely different kind of assessment – one that incorporates the assessment skills and tailors them to the digital environment” (Eyal, 2012), DAL was not especially emphasized or involved in most AL frameworks, but in certain AL frameworks for specific stakeholders (e.g., language teachers). For instance, the prevalence of computer-assisted language testing requires

teachers to be capable of using technological applications and systems in assessment practices. ACER (2021) developed an AL framework to support the International Baccalaureate, a platform that offers continuous international education through four challenging, high-quality educational programs to students aged from 3 years old to 19 years old. Digital assessment was involved as one of the seven elements of this framework. They argued that teacher professional development needs to constantly evolve as digital technologies evolve.

Schmidt and DeSchryver (2022) proposed another similar concept, digital application literacy (DAppLit), which refers to the flexible use of web-based tools in education, including assessment. As a form of media literacy, it requires educators and students to be capable of several skills in applying digital tools for moving from traditional assessments to assessments in a virtual learning environment, such as managing LMSs and audio or video-type products. Schmidt and DeSchryver (2022) argued that proposing and illustrating the DAppLit model could help promote valid online assessments that assess students' content knowledge and avoid the unexpected consequences of technology use on assessment results. In the DAppLit model, they introduced several issues related to online assessment and proposed approaches to address these issues, including the adaptability of online platforms such as better web navigation, and the emotional aspects such as reducing test anxiety.

Policy Considerations in AL

Apart from a set of outlined universal skills that can be implemented in all educational contexts, teacher assessment practices are also influenced by not only expected learning (curriculum) and pedagogical guidance but also community expectations and national or state policies (Looney et al., 2018; Xu & Brown, 2019). It has led to the construction of more fine-grained professional standards across countries that involve multiple assessment-related factors.

Apart from the 1990 Standards in the US, countries such as Australia, Canada, New Zealand, and the UK also released their own professional standards which reflect their educational policies. The accountability trend and the new emphasis on the compositions of AL have led to a revision of these standards for better covering a variety of topics and further informed teacher AL development. For example, the National Board for Professional Teaching Standards (NBPTS, 2012) recognized that teachers need to encourage student engagement in assessment and communicate feedback with various stakeholders. Similarly, InTASC standards state that it is necessary to moderate teachers' judgments to guarantee that assessment criteria are interpreted consistently both within and between schools.

Document Analysis

To build on DeLuca et al. (2016a) and the literature review of Pastore and Andrade (2019), I collected assessment standards from seven regions (i.e., China, Canada, Australia, New Zealand, the UK, the USA, and mainland Europe). In DeLuca et al.'s (2016a) review study, assessment standards of six regions (except China) were selected. DeLuca et al. (2016a) explained the reason for choosing these regions: 1) they have made efforts to promote classroom assessment practice, research, and policy; 2) their published standards shape and guide teacher practice of classroom assessment. This review study also suggested future research to investigate assessment standards in more regions.

These standards were analyzed using document analysis. Document analysis is a qualitative research method that involves reviewing and interpreting documents such as policy documents, business reports, and social media posts (Shaw et al., 2004; Snelson, 2016). It is an invaluable part of most schemes of triangulation and a combination of methods for studying the same phenomenon (Bowen, 2009). Researchers can adopt document analysis as a way to (1)

provide supplementary data, through which background information can be explored to help contextualize the research concepts within its subject or field; (2) track changes and development of the information; 3) ensure the research critical and comprehensive (Bowen, 2009). The READ approach was adopted for document analysis. It includes four steps: (1) ready the materials, (2) extract data, (3) analyze data, and (4) distil the findings (Dalglish et al., 2021).

Step 1: Ready the Documents. Assessment standards in these regions were obtained from public websites for professional organizations or the governmental department of education (e.g., Australian Department of Education, MOE of P. R. China, Association for Educational Assessment-Europe).

Step 2: Extract Data. Because the review intends to provide guidance for the study of measuring AL in the context of higher education, documents that are specific to K-12 education were excluded. Not all documents are targeted for assessment. Some of them are overall standards for education or teacher standards, such as China 1999, 2010, 2020, and Australia 2011. For these documents, only assessment-related areas were included in the document analysis. Table 2.1 shows the total of 13 documents and the included assessment standards.

Step 3: Analyze Data. The specific technique adopted was noting occurrences, or content analysis, where the use of particular words, phrases, and concepts was quantified (O'Leary, 2014). The texts were coded into manageable content categories. This is a process of selective reduction. By reducing the text to categories, I focused on and coded for specific words or patterns that inform the research question.

All 13 documents were divided into governmental and research-based assessment standards and teacher accreditation- and certification-based assessment standards (DeLuca et al., 2016a). They were first coded by region and publication date. In document analysis, phrases

were analyzed for identifying categories and sub-categories (see Table 2.2). The assessment-related phrases were coded for existence (see Table 2.3 and 2.4).

Categories. In DeLuca et al. (2016a), eight themes were extracted from assessment standards. Similar sub-categories were also identified in this study. These 13 sub-categories can be classified into three categories, i.e., assessment knowledge, assessment practice, and social and emotional aspects of assessment. The first category is assessment knowledge with four sub-categories (i.e., know what assessment is, what and why to assess, how to assess, and how to communicate assessment information to stakeholders). The associated codes identified from documents include knowledge of what assessment is and "what to assess" (Europe 2012), theories, principles, and purpose of assessment (NZ 2008), how to assess (UK 2011), and "know how to communicate assessment information appropriately to learners, their parents/caregivers and staff" (NZ 2008). In the USA 2011, essential knowledge about assessment is listed as an independent part of *Standard 6: Assessment*. Australia 2011 divided the teacher career stage into four levels (i.e., graduate, proficient, highly accomplished, and lead), in which the assessment knowledge is required at the graduate level.

The second category relates to assessment practice, which includes five sub-categories, i.e., design or select appropriate assessment methods, use assessment methods to gather data on student learning, interpret results and provide feedback, use data to make decisions and adjust instruction, and engage with stakeholders about assessment information. Teachers need to choose assessment methods and design assessment strategies such as integrating summative and formative assessments, and diagnostic assessments. For instance, in the USA 1990, 1999, 2011, the standards list the requirement of designing assessments that match learning objectives. Teachers also need to collect, analyze, and use assessment information to promote student

learning and inform planning (NZ 2008), interpret the results and provide targeted feedback (UK 2008, 2011), and use assessment information to modify teaching practice (Canada 1993, NZ 2019, USA 2011). Assessment practice also involves "consultation with experts or stakeholders to review reports" (Europe, 2012), or "reporting to students/parents/carers" (Australia 2011).

The last category focuses on the social and emotional aspects of assessment, which include working with stakeholders to optimize the assessment system, teacher as assessor, ethical aspects, and impact on student engagement. Australia 2011 requires teachers to collaborate with colleagues to evaluate the effectiveness of their assessment methods. Europe 2012 requires teachers to consult stakeholders. China 2020 also dictates to "build an assessment system with the participation of the government, schools, and society". Teacher as assessor refers to teachers being conscious of their own role as assessor, for instance, carefully collecting and storing assessment data (Europe 2012). Regarding ethical aspects, teachers are required to "systematically and critically engage with evidence to reflect on and refine their practice" (NZ 2008), manage cheating (China 2010), and avoid "teaching to test" (China 1999, 2010, 2020). The last sub-category is the impact of assessment on students, such as giving students time to reflect on and assess their own work (UK 2008), and engaging learners actively in assessment processes (USA 2011).

Existence of Categories. Table 2.3 and 2.4 show that the documents do not cover all the categories and sub-categories. Most documents concentrate on assessment practice. Several documents cover assessment knowledge or social aspects of assessment. It might be due to the attributes of the documents. Professional standards of Europe 2012 and USA 2011 are comprehensive, covering assessment knowledge, practice, and social aspects. Europe 2012 provides a framework of standards for educational assessment across Europe which is a

professional document specifically for assessment. USA 2011 is a professional document of teaching standards that provides a guiding framework for teacher development in teacher education programs, professional organizations, districts, and states.

USA 1990 (i.e., the 1990 Standards) was proposed to inform teacher educators and teachers in developing assessment competence. It consists of seven standards that have been widely represented and used as a reference in assessment textbooks, teacher education, or educational research. In 1993, the Joint Advisory Committee in Canada developed *Principles for Fair Student Assessment Practices for Education in Canada*, which covers the main assessment practice and social aspects of assessment. In 1999, the Standards for Educational and Psychological Testing was published by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. The goal is to offer standards and criteria for evaluating tests, testing procedures, and test-related consequences. These three documents are governmental or research-based standards. Twelve years later, a comprehensive and detailed document of teaching standards was developed by the Interstate Teacher Assessment and Support Consortium (InTASC) as a way to probe the complexity of the teacher's practice, in which assessment is one of the standards. Each standard delineates the knowledge, dispositions, and performances.

In 2008, the Assessment Reform Group in the UK developed *Changing Assessment Practices: Process, Principles and Standards*, aiming to guide and support change in assessment practice. This document of assessment standards covers comprehensive aspects involved in assessment practice and social aspects of assessment. Subsequently, in 2011, the UK Department for Education issued *Revised Teacher Standards*. This document lists 8 standards for teaching, of

which *Standard 6* lists detailed requirements for making accurate and productive use of assessment.

New Zealand Teachers Council issued *Graduating Teacher Standards* in 2008. The purpose is to provide a guiding framework for the qualification of new teachers entering this profession. It is organized into three dimensions: *Professional Knowledge, Professional Practice*, and *Professional Values and Relationships*. It includes a total of seven standards, each with three to six guidelines. Some of the guidelines are specific to assessment, for example, know how to communicate assessment results to relevant stakeholders. In 2019, the Teachers Council of Aotearoa New Zealand published *the Code of Professional Responsibility and Standards for the Teaching Profession*. This document articulates the expectations and aspirations of the teacher profession. The requirements for assessment within it mostly focus on assessment practice. In 2011, the Australia Department of Education issued *Australian Professional Standards for Teachers*. The requirements listed in *Standard 5* cover assessment knowledge, practice, and some social aspects of assessment.

In China, the Ministry of Education issued three governmental documents for education in 1999, 2010, and 2020. Various aspects of assessment are gradually emphasized with fine-grained requirements. These documents all aim to provide policy guidance for educational reform. For a long time, China has adopted high-stake examinations to select candidates (Stobart, 2008; Wang & Brown, 2014). Thus, teachers were directed to "teach to the test" in the context of high-stake assessment-oriented education (Cheng, 1997). Drawn on the development of contemporary educational theory and practice and the governmental priority of ensuring quality education, since the late 1970s, the educational reform trend in China has been focusing on eliminating such potential negative washback effect of high-stake assessment-oriented education

on student learning (Coombs et al., 2022). This can be reflected in the transformations of emphasis in the guidelines released by the Ministry of Education (MOE). In the 21st century, the educational reform started from the *Action Plan for Revitalizing Education for the 21st Century* (MOE, 1999) which emphasized education for all-around development rather than high-stake assessment-oriented teaching and learning. The subsequent reform continued to state quality education and emphasize student engagement (MOE, 2001) and the establishment of a comprehensive assessment system (MOE, 2010, 2020). Especially, in China 2020, the plan for the educational reform focused on the practical and social aspects of assessment and mentioned making use of technologies in artificial intelligence and big data in assessment, which implies the urgency of involving digital assessment in AL. Some studies have been conducted to investigate the Chinese teachers' AL in the context of this reform trend (e.g., Chen & Brown, 2016; Brown et al., 2009, 2011) and found various degrees of AL deficiencies in assessment conceptions and practices.

Step 4: Distil the findings. Results show most professional standards do not cover all the categories identified in all the documents. The different emphasis of these professional documents is mainly due to their different purposes. For instance, the purpose of publishing USA 1999 was to provide standards for evaluating tests, test practices, and the effects of test use. Therefore, the focus of this type of standards is mainly on the assessment practice and social aspects of assessment. Similar to DeLuca et al.'s (2016a) review, this paper also found a gradual evolvement of AL concept. Early assessment standards did not involve the knowledge dimension of AL and mainly focused on assessment practice and some social aspects of assessment. Regarding assessment practice, the emphasis was put on teachers' competence to design and implement summative forms of assessment. Since 2000, formative assessment has been

gradually emphasized in most documents (e.g., China 1999, 2010, 2020, Australia 2011, USA 2011) which implies that teachers need to monitor students learning progress through periodic assessments. Furthermore, data types of students' assessment information gathered by the teacher were also extended to both quantitative (e.g., exam scores) and qualitative (e.g., written examination scripts, recordings of speech). Regarding social aspects of assessment, a gradual emphasis was on the impact of assessment on student engagement in their own learning, for instance, "reflect on and assess their own work" (UK 2008).

Table 2.1

Assessment-related Professional Standards from National or International Organizations

Region (Year)	Document	Description	Standards/Principles/Elements about Assessment
Australia (2011, revised 2018)	Australian Professional Standards for Teachers (Australia Department of Education, 2011)	7 standards and related guidelines	Standard 5: Assess, provide feedback and report on student learning. • Focus area 5.1: Assess student learning • Focus area 5.2: Provide feedback to students on their learning • Focus area 5.3: Make consistent and comparable judgements • Focus area 5.4: Interpret student data • Focus area 5.5: Report on student achievement
Canada (1993)	Principles for Fair Student Assessment Practices for Education in Canada (JAC, 1993)	9 principles	Classroom assessment • Developing and choosing methods for assessment • Collecting assessment information • Judging and scoring student performance • Summarizing and interpreting results • Reporting assessment findings Assessments produced external to the classroom • Developing and selecting methods for assessment • Collecting and interpreting assessment information • Informing students being assessed • Implementing mandated assessment programs

Region (Year)	Document	Description	Standards/Principles/Elements about Assessment
New Zealand 2008	Graduating Teacher Standards (New Zealand Teachers Council, 2008)	7 standards and related guidelines	Standard 2: Graduating Teachers know about learners and how they learn • have knowledge of a range of relevant theories, principles and purposes of assessment and evaluation. Standard 5: Graduating Teachers use evidence to promote learning • gather, analyze, and use assessment information to improve learning and inform planning. • know how to communicate assessment information appropriately to learners, their parents/caregivers and staff.
New Zealand 2019	Standards for the Teaching Profession (Teaching Council of Aotearoa New Zealand, 2019)	6 standards	Standard 3: Professional relationships • communicate clear and accurate assessment for learning and achievement information. Standard 5: Design for learning • select teaching approaches, resources, and learning and assessment activities based on a thorough knowledge of curriculum content, pedagogy, progressions in learning and the learners. • gather, analyze and use appropriate assessment information, identifying progress and needs of learners to design clear next steps in learning and to identify additional supports or adaptations that may be required. Standard 6: Teaching • use an increasing repertoire of teaching strategies, approaches, learning activities, technologies and assessment for learning strategies and modify these in response to the needs of individuals and groups of learners. • ensure learners receive ongoing feedback and assessment information and support them to use this information to guide further learning.
UK 2008	Changing Assessment Practices: Process,	3 standards and related guidelines	Standard 1: Assessment generally • Policies require schools and local advisers to show how all assessment is being used to help students' learning.

Region (Year)	Document	Description	Standards/Principles/Elements about Assessment
	Principles and Standards (ARG, 2008)		• Introduction of new practices in assessment is accompanied by changes in teacher education and evaluation criteria necessary for their sustainability. • Schools are accountable for using formative and summative assessment to maximize the achievement of goals. • National standards of students' achievement are reported as a range of qualitative and quantitative data from surveys of representative samples. Standard 2: Formative use of assessment • Assessment to support learning is at the heart of government programs for raising standards of achievement. • Initial teacher education and professional development courses ensure that teachers have the skills to use assessment to support learning. • School inspection frameworks give prominence to the use of assessment to support learning. • Schools are encouraged to evaluate and develop their formative use of assessment. Standard 3: Summative use of assessment • Moderated assessment by teachers is used to report students' performance throughout the compulsory years of school. • Moderation of teachers' judgments is required to ensure common interpretation of criteria within and across schools. • Regulations ensure that arrangements for the summative use of assessment are compatible with the practice of using assessment to help learning. • Targets for school improvement are based on a range of indicators and are agreed through a process combining external evaluation and internal self-evaluation.
UK 2011	Revised Teacher Standards (Department for	8 standards for teaching	Standard 6: Make accurate and productive use of assessment

Region (Year)	Document	Description	Standards/Principles/Elements about Assessment
	Education-United Kingdom, 2011)		• know and understand how to assess the relevant subject and curriculum areas, including statutory assessment requirements • make use of formative and summative assessment to secure pupils' progress • use relevant data to monitor progress, set targets, and plan subsequent lessons • give pupils regular feedback, both orally and through accurate marking, and encourage pupils to respond to the feedback.
USA 1990	The Standards for Teacher Competence in Educational Assessment of Students (AFT et al., 1990)	7 standards	Standard 1: Choosing assessment methods appropriate to instructional decisions. Standard 2: Developing assessment methods appropriate for instructional decisions. Standard 3: Administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods. Standard 4: Using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement. Standard 5: Developing valid pupil grading procedures. Standard 6: Communicating assessment results to various stakeholders. Standard 7: Recognizing unethical, illegal, and inappropriate assessment methods and uses of assessment information.
USA 1999	Standards for Educational and Psychological Testing (AERA et al., 1999)	15 standards	Part 1: Test construction, evaluation, and documentation Standard 1: Validity Standard 2: Reliability and errors of measurement Standard 3: Test development and revision Standard 4: Scales, norms, and score comparability Standard 5: Test administration, scoring, and reporting Standard 6: Supporting documentation for tests

Region (Year)	Document	Description	Standards/Principles/Elements about Assessment
			Part 2: Fairness in testing Standard 7: Fairness in testing and test use Standard 8: The rights and responsibilities of test takers Standard 9: Testing individuals of diverse linguistic backgrounds Standard 10: Testing individuals with disabilities Part 3: Testing applications Standard 11: The responsibilities of test users Standard 12: Psychological testing and assessment Standard 13: Educational testing and assessment Standard 14: Testing in employment and credentialing Standard 15: Testing in program evaluation and public policy
USA 2011	InTASC Model Core Teaching Standards: A Resource for State Dialogue (InTASC, 2011)	10 standards	Standard 6: Assessment (9 points for performances, 7 for essential knowledge, 6 for critical dispositions) The teacher understands and uses multiple methods of assessment to engage learners in their own growth, to monitor learner progress, and to guide the teacher's and learner's decision making.
Europe 2012	European Framework of Standards for Educational Assessment 1.0 (AEA-Europe, 2012)	7 core elements	Element 1: Defining the goal (construct, target group, function) Element 2: Identifying the Nature of evidence and tasks Element 3: Gathering evidence (admin and logistics) Element 4: Capturing outcomes (scoring, rating) Element 5: Decision Making (aggregating, norms, grades, cut off scores) Element 6: Interpreting and reporting results Element 7: Evaluation and next iteration
China 1999	Action Plan for Revitalizing Education for the 21st	12 core elements	Element 1: Implement the "trans-century quality education project" to improve the national quality

Region (Year)	Document	Description	Standards/Principles/Elements about Assessment
	Century (MOE, 1999)		• Reform the curriculum system and assessment system and implement new assessment system Element 8: Higher education • Increase the assessment of students' capacities and comprehensive quality and explore diverse assessment methods and systems
China 2010	Outline of the National Medium and Long-term Education Reform and Development Plan (2010-2020) (MOE, 2010)	22 Chapters	Chapter 11: Reform of cultivating talents • Reform the education quality evaluation and talent evaluation system Chapter 12: Reform of examination and enrollment system • Promote the reform of the examination and enrollment system • Improve the examination and enrollment system of colleges and universities • Strengthen information disclosure and social supervision • Strengthen the responsibility for examination safety and construction of the integrity system, and resolutely prevent and seriously investigate cheating in examination enrollment.
China 2020	Overall plan for deepening the reform of education evaluation in the new era (MOE, 2020)	11 core elements	General requirements Element 2: Main principles • Improve result evaluation, process evaluation, comprehensive evaluation and explore value-added evaluation. • Make full use of information technology, and improve the scientific, professional, and objective nature of education evaluation. Element 3: Reform objectives • Promote more complete assessment system to guide teachers to devote themselves to educating people and more diversified assessment methods to advance students' all-around development. Main tasks Element 7: Reform student assessment and promote the all-round development of morality, intelligence, physical fitness, beauty and labor

Region (Year)	Document	Description	Standards/Principles/Elements about Assessment
			• Change the practice of labeling students with scores, innovate the process evaluation method of moral, intellectual, physical, aesthetic and labor, and improve the comprehensive quality evaluation system. • Improve the academic assessment system that organically combines formative assessment and summative assessment, promote classroom participation and classroom discipline, and guide students to establish a good style of study. • Change the fixed form of test items in entrance exams, enhance the openness of test items, and reduce the phenomenon of rote memorization and "mechanical brushing". Organization and implementation Element 10: Strengthen professional construction • Build an assessment system with the participation of the government, schools, and society. • Innovate assessment tools, explore, and carry out the longitudinal assessment of the whole process of students' learning in each grade and the horizontal assessment of all-around development.

Table 2.2

Categories, Sub-categories, and Codes Related to Assessment

Categories	Sub-categories	Codes
Assessment knowledge	Know what assessment is	• what an assessment • have knowledge of a range of relevant theories
	Know what and why to assess	• what is going to measure • the content covered in the assessment should include • purpose
	Know how to assess	• principles • question and task formats

		• demonstrate understanding of assessment strategies • know how to assess
	Know how to communicate assessment information to stakeholders	• know how to communicate assessment information to stakeholders • Demonstrate understanding of providing feedback to students
Assessment practice	Design or select appropriate assessment methods	• choose a method • develop, select and use assessment strategies • formative & summative, formal, diagnostic • assessments match learning objectives
	Use assessment methods to gather data on student learning	• forms of the evidence to be gathered • collect assessment information
	Interpret results and provide feedback	• assessment reports • provide targeted feedback • interpret student assessment data
	Use data to make decisions and adjust instruction	• Using assessments to make decisions • diagnose learning needs • modify teaching practice
	Engage with stakeholders about assessment information	• consultation with experts or stakeholders to review reports • report to students/parents/carers • work with colleagues to use data
Social and emotional aspects of assessment	Work with stakeholders to optimize assessment system	• consult stakeholders • support colleagues • work with colleagues to construct
	Teacher as assessor	• assessment data should be carefully collected and stored
	Ethical aspects	• reflect on and refine practice • cheating • teaching to test
	Impact on student engagement	• impact of the assessment on students • affect the candidates' learning • support students to use this information to guide further learning • students to reflect on and assess their own work.

Table 2.3

The Existence of Sub-Categories for Professional Documents from Canada, the USA, and China

Categories	Sub-categories	Government and research-based standards								
		USA (1990)	Canada (1993)	USA (1999)	UK (2008)	UK (2011)	Europe (2012)	China (1999)	China (2010)	China (2020)
Assessment knowledge	Know what assessment is						√			
	Know what and why to assess					√	√			
	Know how to assess					√	√			
	know how to communicate assessment information to stakeholders									
Assessment practice	Design or select appropriate assessment methods	√	√	√	√	√	√	√	√	√
	Use assessment methods to gather data on student learning	√	√	√	√	√	√		√	√

	Interpret results and provide feedback	√	√	√	√	√	√		√	√
	Use data to make decisions and adjust instruction	√	√	√	√	√	√		√	√
	Engage with stakeholders about assessment information	√	√	√	√		√		√	√
Social aspects of assessment	Work with stakeholders to optimize assessment system		√	√	√		√	√	√	√
	Teacher as assessor	√	√	√	√		√			√
	Ethical aspects	√	√	√					√	√
	Impact on student engagement		√	√	√		√		√	√

Table 2.4

The Existence of Sub-Categories for Professional Documents from New Zealand, Australia, and the UK

Categories	Sub-categories	Teacher accreditation and certification-based standards			
		New Zealand (2008)	New Zealand (2019)	Australia (2011)	USA (2011)
Assessment knowledge	Know what assessment is	√		√	√
	Know what and why to assess	√		√	√
	Know how to assess	√		√	√
	know how to communicate assessment information to stakeholders	√		√	√
Assessment skills and practice	Design or select appropriate assessment methods		√	√	√
	Use assessment methods to gather data on student learning	√	√	√	√
	Interpret results and provide feedback	√	√	√	√
	Use data to make	√	√	√	√

	decisions and adjust instruction				
	Engage with stakeholders about assessment information		√	√	√
Social aspects of assessment	Work with stakeholders to optimize assessment system			√	√
	Teacher as assessor	√			√
	Ethical aspects				√
	Impact on student engagement		√		√

Measurement of AL

Despite previous empirical studies have shown significant gains in student learning achievement when teachers facilitate teaching activities by using assessment (Gardner, 2006; Willis, 2010), researchers found that teachers still struggle to do assessment work in alignment with assessment theories and national policies regarding assessment (Deluca & Klinger, 2010). Over the past decade, to examine whether teachers keep pace with increasing assessment expectations, researchers have conducted numerous studies to qualitatively evaluate teacher AL using methods such as interviews, written diaries, and classroom observations (e.g., Atjonen et al., 2022; Dinan Thompson, & Penney, 2015), or quantitatively develop and validate instruments to investigate teacher AL levels (e.g., Plake et al., 1993; Mertler, 2003).

A thorough review of commonly used instruments was conducted to provide an overview of them and identify any possible gaps. Seven instruments were identified as representative instruments for measuring AL and have been applied to various contexts (see Table 2.5). The 1990 *Standards* serves as the guiding framework of most instruments (n = 5). Within them, the earliest instrument was TALQ (Teacher Assessment Literacy Questionnaire) developed by Plake et al. (1993). The TALQ is a 35-item (five items per standard) content-based instrument used to measure in-service teachers' assessment competence. The items used a binary scoring system (0 = incorrect, 1 = correct), and a high total score indicated a high AL level. The instrument was validated by a sample of 555 elementary and secondary school teachers around the United States. The reliability estimate of Cronbach's α was 0.54. Participants had an average score of 23.2 (*SD* = 3.3) (Plake et al. 1993). It was also validated in other contexts, such as pre-service and in-service teachers in Oman (Alkharusi, 2011; Alkharusi et al., 2011), and Chinese university English teachers (Xu & Brown, 2017).

In later years, Campbell et al. (2002) renamed TALQ to ALI (Assessment Literacy Inventory). It was applied to a sample of 220 pre-service teachers. The average score of this sample was 21 and the reliability was 0.74. Mertler and Campbell then slightly revised TALQ into CALI (Classroom Assessment Literacy Inventory) (Mertler, 2003) and later to Revised ALI (Mertler & Campbell, 2005). Mertler (2003) obtained similar results to those of Plake et al. (1993) and Campbell et al. (2002). The study showed a reliability estimate of 0.57 for the in-service teachers and 0.74 for the pre-service teachers. This study also made a comparison of the AL levels between in-service and pre-service teachers. Campbell et al. (2002) and Mertler (2003) argued that the psychometric performance of the original instrument was poor and the language was obscure. Therefore, they redeveloped the AL scale based on TALQ, which is called Revised ALI. The Revised ALI differs in structure and items compared to the earlier instruments. It includes five scenarios, each with seven questions. Some of the items measure general assessment concepts and knowledge of standardized testing. Other items were developed in the context of classroom assessment. The pilot data analysis showed a reliability of 0.74.

API, by Zhang and Burry-stock (1997), was also developed based on the 1990 Standards, aiming to measure in-service teachers' self-perceived assessment skills (e.g., perceived skillfulness in applying paper-based tests, perceived skillfulness in communicating assessment results). It includes 67 items on a 7-point Likert scale (1 = not confident, 7 = very confident). On a sample of 297 in-service teachers, the entire instrument obtained a reliability estimate of 0.97, with a range of 0.79 to 0.93 for sub-scales.

A significant advance in research into teacher's conceptions of assessment can be seen in the work conducted by Brown and his colleagues. TCoA (Teacher Conceptions of Assessment) inventory was initially developed by Brown (2002) and validated on a relatively small sample of

interns and pre-service teachers. The new version, TCoA-III, was validated by a large national survey of New Zealand primary teachers (Brown, 2004). The abridged version (TCoA-IIIA), which has the same structure as the full version, was validated with large samples of New Zealand and Queensland primary teachers (Brown, 2006), Queensland primary and secondary teachers (Brown, 2011), and a sample of Hong Kong teachers in Chinese version (Brown et al., 2009). TCoA-IIIA is a self-reported survey with 27 statements that adopts a positively packed rating scale, that is, including four positive response points from slightly agree to strongly agree and two negative points, i.e., mostly and strongly disagree. Teachers can rate how much they agree with statements related to the four main purposes of assessment (i.e., assessment makes schools accountable; assessment makes students accountable; assessment improves education; and assessment is irrelevant). It was reported to have good fit characteristics for both New Zealand and Queensland primary teachers (see Table 2.5). The abridged version was found more efficient than the full scale.

The ACAI (Approaches to Classroom Assessment Inventory) was developed based on a thorough review of 15 national or professional assessment standards (i.e., 1990–present) from five regions (DeLuca et al., 2016b). This instrument consists of three parts. Part one (scenario-based) has five scenarios relating to contemporary assessment dilemmas faced by teachers. Part two (Likert) was developed based on Classroom Assessment Standards (JCSEE, 2015), aiming to determine teachers' perceived skills about contemporary classroom assessment practices. Part three (Likert) was developed to investigate teachers' professional priorities and preferences in assessment. The results of pilot testing on a sample of about 400 pre-service and in-service teachers showed that the estimated reliability values ranged from 0.74 to 0.92.

Table 2.5

Commonly Used AL instruments

Instruments	Authors	Sample	Item characteristics	Guiding frameworks	Psychometric properties
Assessment Literacy Inventory (ALI)	Campbell et al. (2002)	220 undergraduate prospective teachers	35 content-based items (5 items per standard)	1990 Standards	α = 0.74; *M* = 21
Classroom Assessment Literacy Inventory (CALI)	Mertler (2003)	197 in-service teachers 67 pre-service teachers	35 content-based items (5 items per standard)	1990 Standards	1) α = 0.57; *M* = 22.0, *SD* = 3.4 2) α = 0.74; *M* = 19.0, *SD* = 4.7
Revised Assessment Literacy Inventory (revised ALI)	Mertler & Campbell (2005)	250 undergraduate prospective teachers in the U.S.	35 scenario-based items (5 scenarios; 5 items per standard)	1990 Standards	α = 0.74; *M* = 23.9, *SD* = 4.6
Approaches to Classroom Assessment Inventory (ACAI)	DeLuca et al. (2016b)		Part 1: 20 scenario-based items (5 scenarios; 4 items per theme) Part 2: 12 Likert-type items Part 3: 21 Likert-type items	Classroom Assessment Standards (JCSEE, 2015)	α ranged from 0.74 to 0.92
Teacher Assessment Literacy Questionnaire (TALQ)	Plake et al. (1993)	555 elementary and secondary school teachers in U.S.	35 content-based items (5 items per standard)	1990 Standards	α = 0.54; *M* = 23.2, *SD* = 3.3
Assessment Practices Inventory (API)	Zhang & Burry-stock (1997)		67 Likert-type items	1990 Standards	α ranged from 0.79 to 0.93, α = 0.97 for total instrument

Teacher Conceptions of Assessment Version III abridged inventory (TCoA-IIIA)	Brown (2006)	525 New Zealand and 692 Queensland primary school teachers	27 Likert-type items (3 items per factor)	four main purposes of educational assessment (i.e., assessment makes schools accountable, assessment makes students accountable, assessment improves education, and assessment is irrelevant)	New Zealand (χ_{311}^2 = 841.02; RMSEA = 0.057; TLI = 0.87) and Queensland (χ_{311}^2= 1492.61; RMSEA = 0.074; TLI = 0.80)

Contextual Factors Associated with AL Measurement

Previous studies found the AL level of in-service teachers tended to be higher than that of pre-service teachers (Mertler, 2003, 2004; Mertler & Campell, 2005). Mertler (2004) pointed out that it was possibly due to the lack of sufficient teacher preparation courses in assessment for pre-service teachers, especially the targeted training to secondary level classrooms presented by the measurement specialists. In their later studies, Mertler and Campbell (2005) also reported such discrepancies and suggested examining the link between teaching experience and assessment competency. Most AL measurement instruments that were developed in previous studies were situated in the K-12 context. They recruited pre-service or in-service teachers for validation, for example, primary or secondary school teachers (e.g., Brown, 2006; Mertler & Campell, 2005). The scenarios or items described in these instruments were applied to K-12 classrooms. Therefore, when examining the AL level of teachers in post-secondary education, some researchers made revisions to these scenarios or items. For example, in order to investigate the AL level of Chinese English teachers, Xu and Brown (2017) adjusted the education context

of TALQ items. The context of certain items was modified from K-12 education to higher education settings. Several items that were specific to the U.S. policy and practice were deleted. They argued that TALQ mainly focused on the basic level of the AL dimension, which was not sufficient to examine the enormous assessment responsibility of university English teachers, nor can it reflect AL changes in contexts, policies, and culture.

AL in Postsecondary Education

Apart from the general assessment dimensions that can be applied to all stages of education, teachers in postsecondary education tend to undertake more targeted responsibilities. A literature review conducted by Lindstrom et al. (2017) provided a summary of major themes related to assessment principles in postsecondary education. It involves four aspects, i.e., conceptualizing assessment, assessment practice, support for assessors, and institutional support. These themes of assessment principles reveal multi-facets of AL, which largely align with Pastore and Andrade's three-dimensional framework, i.e., conceptual, praxeological, and socio-emotional dimensions. Furthermore, these themes cover some responsibilities targeted to postsecondary institutions. For instance, when conceptualizing assessment from an institutional perspective, assessment is thought of as a long-term, developmental process that supports students' capacity for academic integrity, self-regulated learning, and lifelong learning (e.g., Brown & Race, 2013; Kaslow, et al., 2007). It is also a continuous activity rooted in the culture of the institutions and curriculums, rather than just a course component aimed at providing final scores for specific learning modules (e.g., Ndoye & Parker, 2010; Stassen, 2012). Regarding institutional support, teachers are required to use the clearly defined institution-wide grade scale system and apply it consistently across individual programs and courses (e.g., James et al., 2002; Luth, 2010). Furthermore, higher education institutions attempt to establish an assessment

culture that can direct change processes, integrate assessment into policy frameworks, and improve the sustainability of organizations to support student learning (e.g., Heinrichs et al., 2015; Stassen, 2012).

Frameworks of Instrument Development and Validation

For the Mixed Methods Research (MMR) in instrument development and validation, some frameworks have been proposed for guiding researchers with designs and approaches to construct reliable and valid instruments. These frameworks/principles/procedures are beneficial for the conceptualization of the conduct of mixed-methods instrument development design and have been used as guidelines for measuring various concepts of interest (e.g., Hoehle & Venkatesh, 2015; Koskey et al., 2018). Several frameworks or procedures for developing and validating instruments are listed in Table 3.6. Generally speaking, instrument development involves five main phases, including defining the construct, generating items, pilot testing, revising, and finalizing the scale (Burton & Mazerolle, 2011). For example, Yin (2006) proposed five procedures for scale development and validation: 1) research questions; 2) units of analysis; 3) samples for study; 4) instrumentation and data collection; 5) analytic strategies. Even though the author suggested using more approaches in each procedure, detailed approaches were not discussed in depth due to the scope of the study. Onwuegbuzie et al. (2010) proposed a more detailed framework, the Instrument Development and Construct Validation (IDCV) process. It entails 10 steps from construct identification and conceptualization to the evaluation of the process and product. Based on the synthesis of previous studies on instrument development (DeVellis, 2012; Straub, 1989; Straub et al., 2004), MacKenzie et al. (2011) presented a comprehensive procedure of conceptualizing construct, developing and validating a measurement instrument. In total, the procedure consists of 10 steps. In DeVellis's (2012)

framework, the process of instrument development could be broken down into nine steps. Carpenter (2018) also broke down the scale development process into ten manageable steps. Zhou (2019) simplified this process into five steps. Some new procedures, for example, the MEASURE approach (e.g., Kalkbrenner, 2021), were not listed and discussed here because of less validation in empirical studies.

Most of them (e.g., Carpenter, 2018; MacKenzie et al., 2011; Zhou, 2019) adopted a sequential mixed methods research design for scale development (Creswell & Plano Clark, 2011). It usually follows a sequence of three stages: a qualitative stage of defining the construct of an instrument, a stage of developing the instrument including item design and revision, and a quantitative stage of validating the instrument. Although MacKenzie et al. (2011) provided a great detailed guide for validating instruments which enables cross-validation of instruments, there was no in-depth discussion about the process of generating items. DeVellis (2011) is quantitative-oriented and provides a detailed explanation of the quantitative phase. In Carpenter (2018), more emphasis was put on factor analysis (e.g., factor extraction, and rotation). Zhou (2019) mentioned that qualitative codes could be converted into instrument items but did not provide a detailed description of the process.

Although numerous quantitative and qualitative approaches have been introduced and used in previous related studies, they are often employed in isolation instead of being completely integrated to guide instrument validation (Koskey et al., 2018). This limitation can be addressed by crossover analyses in Onwuegbuzie et al.'s (2010) framework (see Figure 1). Crossover analyses enable researchers to make Gestalt switches (Kuhn, 1962) between qualitative and quantitative perspectives, traveling back and forth multiple times to mine the deep meaning

underlying the data (Onwuegbuzie et al., 2010). Thus, multiple forms of evidence can be collected to inform criterion, content, and construct validity.

The Current Study

Concerning commonly used instruments for measuring AL, the psychometric properties of these instruments revised based on TALQ showed greater reliability than the original scale. Three out of seven instruments were content-based type. Two were Likert-scale. One was scenario-based, and one was a combination of both scenario-based and Likert-scale. Most of them were developed based on the 1990 Standards. Many scholars have adopted them in various national contexts. Despite these efforts in the measurement of AL, Gotch and French (2014) argued that the psychometric qualities of previous AL measures were still weak and suggested that researchers need to increase the validity of AL measurements by examining whether they can represent and reflect the current transformations in assessment, such as the increased trend of accountability in schools and the use of formative assessments to continuously monitor student learning. Current AL instruments do not fully reflect the transformations in the assessment field and are still based on outdated standards for teachers' classroom assessment practice (DeLuca et al., 2016a). Brookhart (2011) pointed out that the 1990 Standards have become outdated because the current conceptions of formative assessment and technical and social issues involved in standards-based educational reforms were not included. Thus, AL measures that set the 1990 Standards as the dominant reference might be out of step with contemporary assessment needs. Furthermore, even though digital tools and technologies have provided new opportunities for assessment, how teachers can maximize the advantages of digital resources in their assessment work has received very little attention (Girgla et al., 2021). Given significant changes in assessment practices over the past decade and the rising status of technology-based assessment-

related activities, there is a need to investigate teacher AL in various contexts based on contemporary educational reforms and recent work on the conceptualization of AL. The measurement of AL could further consider contextual factors such as different professional careers (e.g., novice, expert) and different educational levels (e.g., university, kindergarten) because AL is culturally situated and contextually sensitive (Xu & Brown, 2016). Furthermore, considering regional policies and priorities in measurement would benefit geographic validity (DeLuca et al., 2016a; Messick, 1989).

To address the above issues identified in previous studies, this study developed a measurement instrument that aims to measure the AL level of teachers in higher education. Specifically, this study adopted Onwuegbuzie et al.'s (2010) framework of IDCV to: 1) develop a measurement instrument that reflects the current assessment standards; and 2) validate the instrument by both qualitative and quantitative methods.

Chapter Summary

Chapter 2 provided a comprehensive review of AL-related issues in the research context. Firstly, it reviewed the definition of AL and existing AL and DAL models. Specifically, four AL theoretical models with different foci were presented. The core components of AL in each model were discussed. In addition, this chapter also discussed the main issues involved in DAL. Then, this chapter presents a document analysis of international teacher standards for assessment. It also reviewed the measurement of AL, specifically including the commonly used AL measures and their psychometric properties. It further highlighted the contextual factors associated with AL measurement. It also compared several frameworks of instrument development and validation and highlighted the advantages of Onwuegbuzie et al.'s (2010) IDCV process. Finally, several research gaps were identified based on the literature review. The next chapter specifically

introduces the IDCV process and the methods used for developing and validating the AL measurement instrument.

Table 2.6

A Sample of Frameworks/Principles/Procedures for Instrument Development and Validation

Study	Frameworks/Principles/Procedures	Specific steps	Related Research (e.g.)
Yin (2006)	Five procedures for scale development and validation	1. Research questions 2. Units of analysis 3. Samples for study 4. Instrumentation and data collection 5. Analytic strategies	Taghipoorreyneh et al. (2019)
Onwuegbuzie et al. (2010)	IDCV	1. Conceptualize the construct of interest 2. Identify and describe behaviors that underlie the construct 3. Develop initial instrument 4. Pilot-test initial instrument 5. Design and field-test revised instrument 6. Validate revised instrument: Quantitative analysis phase 7. Validate revised instrument: Qualitative analysis phase 8. Validate revised instrument: Mixed analysis phase: Qualitative-dominant crossover analyses 9. Validate revised instrument: Mixed analysis phase: Quantitative-dominant crossover analyses 10. Evaluate the instrument development/construct evaluation process and product	Koskey et al. (2018), Mohamed (2019), Shiyanbola et al. (2021)
MacKenzie et al. (2011)	Ten Steps for scale development and validation	1. Construct definition 2. Measure development 3. Content validity assessment 4. Measurement model specification	Hoehle & Venkatesh (2015)

		5. Pre-test of the scales 6. Scale purification 7. Assess scale validity 8. New sample data collection 9. Cross validation 10. Norm development	
DeVellis (2012)	Guidelines in scale development	1. Determine clearly what you want to measure 2. Generate an item pool 3. Determine the format of the measure 4. Have experts review the initial item pool 5. Cognitive interviewing 6. Consider the inclusion of validation items 7. Administer items to development sample 8. Evaluate the items 9. Optimize scale length	Howell Smith et al. (2020), Hung et al. (2016)
Carpenter (2018)	Ten steps in scale development and reporting	1. Research the intended meaning and breadth of the theoretical concept 2. Determine sampling procedure 3. Examine data quality 4. Verify the factorability of the data 5. Conduct common factor analysis 6. Select factor extraction method 7. Determine number of factors 8. Rotate factors 9. Evaluate items based on a priori criteria 10. Present results	Weretecki et al. (2021), Lin et al. (2022)
Zhou (2019)	Five steps	1. Qualitative investigation to explore the construct 2. Conversion of qualitative findings to develop items 3. Conducting mixed-methods validation to determine content validity 4. Administration of the developed scale to the target population	Sreedharan et al. (2022)

	5. Quantitative validation to examine items' construct validity

CHAPTER THREE: METHODOLOGY

This chapter introduces the mixed research design used for the current study. First, the reliability and validity of instrument development were discussed. Then, Onwuegbuzie et al. (2010)'s IDCV process for optimizing the development of the quantitative instrument was introduced.

Methods

This study adopted the IDCV process as a mixed-methods meta-framework for optimizing the development and validation of the quantitative instrument. This process contains 10 phases for instrument development/fidelity and construct validation. The 10 phases are shown in Figure 3.1. This meta-framework has two advantages. It allows crossover analyses to integrate qualitative and quantitative data analyses. It also incorporates a debriefing system for researchers (Onwuegbuzie et al., 2008), whereby researchers are interviewed by outside experts in order to impartially evaluate the process fidelity and the decision-making in the study. Under this framework, various methods were used to ensure the reliability and validity of the instrument.

Reliability

One of the first principles regarding instrument development is that the developed instrument should be reliable. Reliability is defined as the reproducibility of measurements and reflects the degree to which a measure produces the same results when applied repeatedly to a person or process that has not changed (Shrout & Lane, 2012). If reliability is compromised, then the validity of the measure will also be compromised. The greater the influence of measurement error on a measure, the less useful it is. Therefore, establishing reliability is usually considered the first step in determining the measurement quality (Shrout & Lane, 2012).

Validity

Onwuegbuzie et al. (2009) extended Messick's (1989, 1995) theory of validity and provided a useful meta-framework for IDCV (see Onwuegbuzie et al., 2009). Onwuegbuzie et al. (2009) listed three main forms of validity – content-related validity, criterion-related validity, and construct-related validity. Content validity refers to the extent to which the instrument completely assesses or measures the construct of interest (Haynes et al., 1995), which contains face validity, item validity, and sampling validity. Face validity concerns how clear and meaningful the items are to the respondents. Item validity refers to the degree to which a specific item represents the measurement of the expected domain. Sampling validity concerns whether the items can sample the whole concept.

Criterion validity indicates how well the responses of a scale converge with criterion variables with which the scale is supposed to converge (Cronbach & Meehl, 1955). As one of criterion-related validity, concurrent validity refers to obtaining criterion variables and the responses of a measurement scale simultaneously, while predictive/postdictive validity refers to measuring criterion variables after or before the current measurement (Grimm & Widaman, 2012).

Construct validity determines the degree to which a test measures the construct it is designed to measure (Grimm & Widaman, 2012), which has been regarded as the most fundamental and important aspect of psychometrics. Construct validity also contains several subtypes, such as convergent and divergent validity, discriminant validity, and outcome validity. Convergent validity focuses on whether there is a strong correlation between the scores from the developed instrument and other instruments measuring the same construct, while when validating divergent validity, researchers should confirm that scores from the developed instrument are not highly correlated with the antithetical instruments. Discriminant validity can

be tested by demonstrating the extent of correlation between the measures of unrelated constructs. Outcome validity focuses on the meaning of scores and the expected and unexpected effects of the instrument.

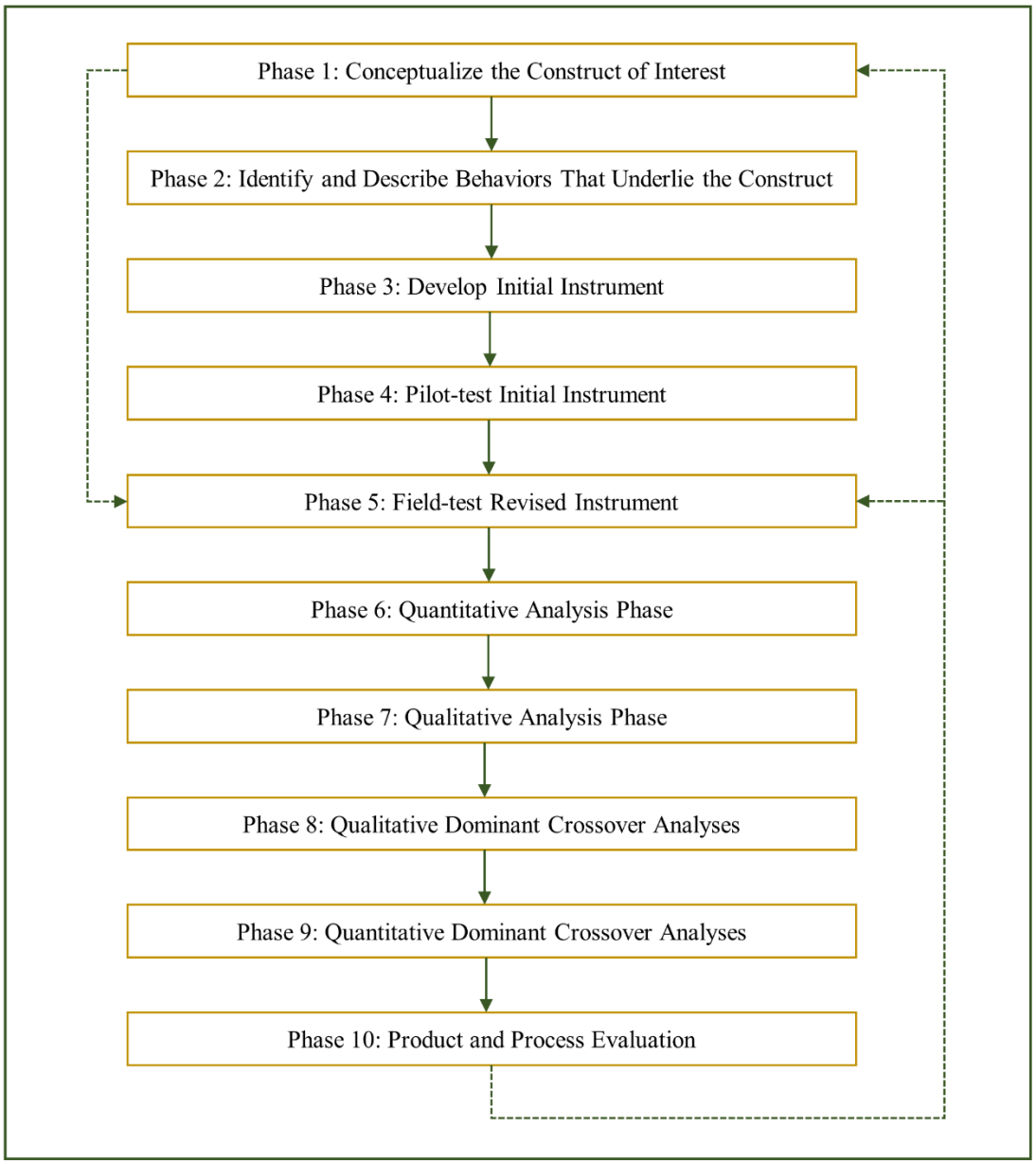
Based on this framework, Onwuegbuzie et al. (2010) recommended eight types of crossover analyses (e.g., data comparison, data transformation, data correlation, integrated data display, and integrated data reduction) that can be conducted for each validity type. They suggested that qualitative methods can be used to address most of these validity types, although quantitative methods are the common choices. For example, structural validity has been typically addressed entirely quantitatively (e.g., EFA). It also can be assessed by qualitative analysis techniques, for example, constant comparison of open-ended responses.

Procedures

This study was approved by the Human Subjects Ethics Committee of the author's university in September 2023. After approval, this study conducted data collection and analysis in the order of the 10 phases in the IDCV process. Data collection started in September 2023 and ended in March 2024. In this section, the approaches and steps based on the IDCV process were introduced, and the role that mixed methods can play in developing an instrument for assessing teacher AL was explained.

Figure 3.1

Instrument Development and Construct Validation (IDCV) Process by Onwuegbuzie et al. (2010)



Phase 1: Conceptualize the Construct of Interest

The main goal in Phase 1 is for instrument developers to be conscious of their own three-dimensional belief systems, namely, overall worldview, research philosophy, and discipline-specific philosophy (Combs et al., 2010).

The first dimension of belief systems requires researchers to be aware of their perspectives on seeing and interpreting the world (Johnson et al., 2004). That is to say, the first

step is to clarify the research paradigm. The research paradigm involves the abstract beliefs and principles that influence the way researchers see, interpret, and act with the world (Kivunja & Kuyini, 2017). It is related to how researchers locate their research (e.g., the choice of methodologies) and can guide researchers in conducting their research. A research paradigm contains four elements, including epistemology, ontology, methodology, and axiology (Lincoln & Guba, 1985). Epistemology is used to describe how we come to know the reality within this world. It focuses on the nature and forms of knowledge, and how it can be acquired and transmitted. It also concerns the nature of the relationship between the knower and what can be known (Kivunja & Kuyini, 2017). Researchers should address a core question for understanding the epistemological element: *How do we know what we know*? The second element, ontology, involves the study of what exists or what is real, how things are grouped into categories, and how they are related to one another. It makes researchers consider the nature of what they are investigating. It is also a crucial element for researchers to understand how to make the collected data meaningful. As a broad term, methodology covers the research design, methods, procedures, and approaches used to investigate something (Keeves, 1997). It is concerned with the overall flow and logic of the research processes in a research project which aims to acquire knowledge about a research question (Kivunja & Kuyini, 2017). Even when investigating the same social phenomenon, different epistemological and ontological views can lead to different methodologies (Grix, 2004). The last element is axiology, which involves the ethical issues that need to be considered in a research project. Specifically, researchers need to define, understand, and evaluate right or wrong behaviors when conducting research, for instance, respecting all participants' rights and avoiding or minimizing risk or harm to participants (Kivunja & Kuyini, 2017).

Researchers have proposed a number of paradigms that can be arranged into three main taxonomies, including positivist, interpretivist, and critical paradigms (Candy, 1989). Positivism is a paradigm that relies on experimentation, observation, and reason as the basis for investigating human behavior. It believes that knowledge is discovered from measurable observations of activities or actions. The epistemological element of positivism points towards objectivism, ontology towards realism, methodology towards experimentation, and axiology towards beneficence (Kivunja & Kuyini, 2017). The nature of the positivist paradigm lies in explaining relationships using scientific methods. The interpretivist paradigm, also called the constructivist paradigm, focuses on the subjective world of human experience (Guba & Lincoln, 1989). Its epistemology is subjectivist, ontology is relativist, methodology is naturalistic, and axiology is balanced (Kivunja & Kuyini, 2017). This paradigm advocates qualitative methods. For instance, data about individuals' experiences and their reflections on these experiences can be collected and analyzed under the guidance of grounded theory. The critical paradigm, sometimes called the transformative paradigm, focuses on social, political, and economic issues, such as social change, power, and inequality. This paradigm advocates a transactional epistemology, an ontology of historical realism, a dialogic methodology, and an axiology of respecting cultural norms. The methods involved in critical paradigm are similar to interpretivism, which also gathers and analyzes qualitative data from methods such as focus groups and interviews. Furthermore, a fourth paradigm called the pragmatic paradigm was proposed by later researchers (e.g., Tashakkori & Teddlie, 2003), which is based on the concept of using the most appropriate methods to answer research questions. It assumes a relational epistemology, an ontology of non-singular reality, a mixed methods methodology, and an axiology that advocates benefiting people in research (Kivunja & Kuyini, 2017). In sum, when

following a research paradigm, researchers can locate their research in specific assumptions of epistemology, ontology, methodology, and axiology. The influence of the chosen paradigm on research permeates all aspects, including identifying research questions, selecting participants, data collection, analysis, and interpretation.

Second, instrument developers need to identify the paradigm of their research. The paradigmatic differences between quantitative and qualitative research have been discussed by many scholars. The qualitative versus quantitative debate may make some individuals confused about the logic of justification with research methods. However, as two separate elements of paradigm, differences in epistemology should not absolutize the choice of methodologies (Johnson & Onwuegbuzie, 2004). A quantitative researcher could use qualitative data collection methods, and vice versa. Moreover, due to the increasing trend of interdisciplinary research nowadays, many researchers seek to break the single research tradition under certain paradigms and attempt to combine multiple methods to answer complex research questions (Johnson & Onwuegbuzie, 2004). In the field of MMR, the incompatibility of different paradigms has been broken. Therefore, researchers can freely choose multiple paradigms instead of debating which one is superior (Morgan, 2022). For example, researchers might mix or combine the use of quantitative and qualitative approaches under the research philosophy of blending the postpositivist stance with the constructivist stance. Constructivist researchers seek to understand the experience of research participants to discover the participants' subjective truth or perceptions, while the post-positivism holds the view that social science research should be objective. For example, to blend these metatheoretical viewpoints, researchers could use factor analysis to examine the structure of themes that were extracted from qualitative data, which can further validate the construct (i.e., sequential qualitative-quantitative mixed analysis)

(Onwuegbuzie et al., 2010). Alternatively, the themes obtained from a thematic analysis can be used to compare to factors derived from a factor analysis, i.e., concurrent mixed analysis, which compares findings from quantitative analysis with qualitative analysis to achieve convergence (Greene et al., 1989, Onwuegbuzie et al., 2010).

Onwuegbuzie et al.'s (2010) IDCV proposed a form of comprehensively combining qualitative and quantitative data analysis methods, i.e., crossover analyses, which involves alternating between inductive and deductive reasoning (Onwuegbuzie et al., 2010). These analyses enable the incorporation of emic (i.e., derived from the researchers of the ongoing study) views and etic (e.g., derived from previous theories and hypotheses) views for instrument development and construct validation. Such analyses are appropriate for ensuring the quality of meta-inferences by integrating interpretations of the qualitative results with quantitative findings to draw coherent and reasonable conclusions (Onwuegbuzie et al., 2010).

Third, instrumental developers should identify their own construct-specific philosophies (Onwuegbuzie et al., 2010), which include the thoughts and views surrounding the research of interest. In Phase 1, studies related to teacher AL were reviewed to identify the core components of this construct and find theoretical frameworks to guide the study. Moreover, focus groups of key informants (i.e., instructors of universities) were set up to obtain information on their assessment practices and the use of technology in assessment. Guest et al. (2017) suggested that two to three focus groups can extract most themes that reflect the experiences shared by key informants. Focus groups play an important role in ensuring the voices of key informants are covered and understanding their cultural backgrounds. Qualitative data collected in focus groups can help identify the key components in each dimension of the AL.

Phase 2: Identify and Describe Behaviors That Underlie the Construct

The information from the literature review and the qualitative data from focus groups were used to prepare the framework of the instrument. In Phase 2, open coding based on grounded theory was used to turn qualitative data into small and discrete components of data with a descriptive label, followed by axial coding which was used for exploring connections between codes, and aggregating and summarizing codes into broader categories. Finally, selective coding was used to identify core concepts. Then, a preliminary framework was developed.

A panel of experts was invited to evaluate and provide their feedback and suggestions on the preliminary framework. Experts were selected based on the following criteria: (1) studying topics relating to AL; or (2) having published at least one journal article on this topic. The framework was revised based on the results of the expert review.

Phase 3: Develop Initial Instrument

In Phase 3, the items in Likert scaling (Likert, 1932) were generated based on the revised framework. To write good items, I reviewed multiple published sources that presented AL questionnaires specifically for teachers (e.g., TALQ), and brainstormed the items. AL and DAL were integrated when generating items. For assessment knowledge, the items were generated and revised based on Al-Bahlani and Ecke (2023)'s questionnaire and the framework yielded in Phase 2. The items in this part were used to investigate the level of assessment knowledge. For assessment practices and socio-emotional aspects of assessment, the items were generated in the form of subjective statements to measure teachers' self-perceived competence in assessment. At least three items were generated for each concept.

I also asked the expert group for feedback on the items to ensure that the items were informed by both emic and etic perspectives. This was a process of back-and-forth discussion

and revisions of the items. During this process, both quantitative items and a qualitative question in short and clear language were developed. The first full version of the instrument was evaluated in Phase 4.

Phase 4: Pilot-Test Initial Instrument

Mixed methods were used to justify the quality of the instrument, that is, to ensure the instrument is clearly worded, well structured, and covers all the concepts that should be measured. More specifically, in Phase 4, content validity was evaluated. The instrument was evaluated by the content validity index (CVI) and the content validity ratio (CVR) based on the results of the expert review. A sample of key informants was also recruited to obtain their feedback on item quality.

This phase involves two steps. The first step is the assessment of items through the expert review. The item review survey was sent to the experts recruited in Phase 2 with a cover letter explaining the study and their willingness to review the items. It also included clear guidelines on how to rate each item.

Each item was evaluated by experts to rate its 1) relevance to the instrument's aim, 2) clarity of each item; 3) essentiality of each item. CVI (Lawshe, 1975; Lynn, 1986; Polit & Beck, 2006; Polit et al., 2007; Zamanzadeh et al., 2015) was calculated based on the results of experts' evaluations. CVI is a quantitative indicator of content validity and the most commonly used method. It includes the Item-CVI (I-CVI) and the Scale-level-CVI (S-CVI) (Zamanzadeh et al., 2015). I-CVI is calculated as dividing the total number of experts by the number of experts who rate each item as relevant. Table 3.1 presents the cut-off values of I-CVI according to different numbers of experts. There are two methods to calculate S-CVI. One is the Universal Agreement (UA) among experts (S-CVI/UA), and the other is the Average CVI (S-CVI/Ave) which is a less

conservative method (Zamanzadeh et al., 2015). S-CVI/UA is calculated as dividing the total number of items by adding all items with I-CVI equal to 1, while S-CVI/Ave is calculated as dividing the total number of items by taking the sum of the I-CVIs (Rodrigues et al., 2017; Zamanzadeh et al., 2015). A value of 0.80 for S-CVI/UA and the S-CVI/Ave was acceptable (Polit & Beck, 2006; Waltz et al, 2005); values up to 0.90 indicate an excellent average (Holle et al., 2014).

Despite that CVI was widely employed by researchers to measure content validity, it does not consider the risk of inflated values due to chance agreement. Therefore, Wynd et al. (2003) suggested calculating the kappa statistic that can provide information on interrater agreement beyond chance (coded as K*). The probability of chance agreement should be first determined for each item using the following formula, in order to get the modified kappa statistic:

$$P_c = \left[\frac{n!}{A!\,(n-A)!}\right] \times 0.5^n$$

Where *n* is the number of experts, *A* is the number of experts rating each item as relevant. Next, K* can be calculated by the following formula:

$$K^* = \frac{I\text{-}CVI - P_c}{1 - P_c}$$

K* levels were classified as excellent, good, and fair, respectively, if they were above 0.74, between 0.60 and 0.74, and between 0.40 and 0.59 (Cicchetti & Sparrow, 1981). The items with fair K* values should be revised or deleted.

The second quantitative indicator is CVR, which is a measure of the essentiality of an item (Waltz et al, 2005). CVR has a value between -1 and 1. The higher value indicates greater agreement among experts (Zamanzadeh et al., 2015). CVR can be computed by the following equation:

$$CVR = \frac{(N_e - \frac{N}{2})}{\frac{N}{2}}$$

where N_e is the number of experts rating an item as essential and *N* is the total number of experts.

Table 3.1

The Number of Experts and the Acceptable Cut-off Value of I-CVI

Number of experts	Acceptable CVI values	Sources
2	⩾ 0.80	Davis (1992)
3 - 5	1	Polit & Beck (2006), Polit et al. (2007)
⩾ 6	⩾ 0.83	Polit & Beck (2006), Polit et al. (2007)
6 - 8	⩾ 0.83	Lynn (1986)
⩾ 9	⩾ 0.78	Lynn (1986)

The second step is pilot testing of the instrument. This step can test the face validity of the revised instrument. A group of participants were recruited to participate in the pilot testing. They were provided consent forms and the proposed instrument in the review form. The form contains instructions for rating each item 1, 2, 3, or 4 on clarity and understandability. Similar to CVI, face validity index (FVI), including I-FVI and S-FVI (S-FVI/Ave and S-FVI/UA), were calculated. The acceptable FVI value for about 30 raters was at least 0.80 in most previous studies (Yusoff, 2019).

Phase 5: Design and Field-Test Revised Instrument

The problematic items were revised according to the results of the expert review and pilot testing in Phase 4. Then, the revised measurement instrument was sent out for a widespread field test. In the field test, the sample size must be large enough to conduct the factor analysis and yield adequate reliabilities with relatively narrow confidence intervals.

Phase 6: Quantitative Analysis Phase

In Phase 6, quantitative data collected from the field test were analyzed. Factor analysis and reliability testing were conducted for the instrument. Specifically, a conceptual model was built. Confirmative factor analysis (CFA) was performed to verify the factor structure in the instrument.

Phase 7: Qualitative Analysis Phase

Phase 7 involves the analysis of the qualitative data collected from the open-ended question in field-testing. The major goal of mixing quantitative and qualitative data is to address triangulation, complementarity, development, initiation, or expansion (Greene et al., 1989). The open-ended responses were preprocessed and analyzed using the thematic analysis method. Thematic analysis is a data-analytic process that helps identify, analyze, and interpret patterns or themes in qualitative data.

Phase 8: Qualitative-Dominant Crossover Analyses

Crossover analyses were used to integrate the interpretation of qualitative and quantitative data analysis results, in order to further address one or more of the five goals for mixing qualitative and quantitative data identified by Greene et al. (1989). In this phase, a quantitative analysis of the qualitative data extracted in Phase 7 was conducted. Specifically, the themes that emerged from the thematic analysis were quantified, in which the themes were identified and transformed into numerical values that can be analyzed quantitatively (i.e., the frequency of each theme).

Phase 9: Quantitative-Dominant Crossover Analyses

Phase 9 involves conducting a quantitative-dominant crossover analysis. The quantified themes that emerged in Phase 7 were analyzed to find correlations to the factors that were confirmed in Phase 6. Specifically, correlation analysis was conducted on the relationships

between the frequencies of the themes and the factor loadings of the factors. The assumption of this analysis is that the more the themes teachers shared in their assessment experiences, the higher factor loadings they might have. This indicates that they are important elements in AL.

Phase 10: Process and Product Evaluation

Phase 10 involves a comprehensive evaluation of both the product and the process. Analysis of the product aims to evaluate the validity of the instrument. If the quantitative and qualitative analyses show inconsistent results, the IDCV process allows for reexamining instrument fidelity for continuous improvement.

Chapter Summary

Chapter 3 mainly describes methods used for developing and validating an instrument. It first discussed the reliability and validity that a measurement instrument requires. Then, it detailed the logic and methods in each phase of the IDCV process, from conceptualizing the construct of interest to evaluating the instrument development. The next chapter presents the results and discussion.

CHAPTER FOUR: RESULTS AND DISCUSSION

This chapter reported and discussed the results generated from each phase in the IDCV process.

Phase 1: Conceptualize the Construct of Interest

The research interest of the current study is higher education teacher AL in both classroom and digital settings. This phase involved a systematic review of related literature and examined my belief systems. In the first step of the systematic review, the concept of teacher AL was explored by reviewing the relevant studies that focused on how this concept evolved and how it was conceptualized (e.g., Chan & Luk, 2022; Looney et al., 2018; Pastore & Andrade, 2019; Xu & Brown, 2016). The concept of DAL and digital components in the assessment were also reviewed (e.g., Eyal, 2012; Schmidt & DeSchryver, 2022). The second step reviewed the governmental or institutional documents on assessment in education to discuss policy considerations in teacher AL, especially the policies released by the Ministry of Education of China. In the third step, the existing measures of teacher AL were reviewed, including how they were designed and how they were validated (e.g., DeLuca et al., 2016b; Plake et al., 1993).

Additionally, the researcher reviewed the methods used for developing and validating measurement instruments and determined the methods for the current study (e.g., Onwuegbuzie et al., 2010; Younas, 2020). A mix of quantitative and qualitative assumptions under the research philosophy of blending the postpositivist stance with the constructivist stance was identified to guide the current study. The following phases reported on why and how each qualitative and quantitative method was used, as well as the results and discussion of them.

Phase 2: Identify and Describe Behaviors That Underlie the Construct

This phase involves two steps, including the focus group interviews and the expert review of the preliminary framework.

Focus Group

To integrate teacher AL and DAL, namely, to identify key features of teacher AL in both classroom and digital environments, focus group interviews were set up to obtain information on their assessment practices and the use of technology in assessment. The specific questions involved in the focus groups are listed in Appendix A.

Consent forms and a brief introduction to this research were sent to some teachers who work at universities in China. A total of nine higher education teachers from six disciplines (i.e., English, Education, Electronic Information, Engineering, Journalism, and Anthropology) participated in three focus group interviews (three teachers per group). All teachers were between the ages of 30 and 40. Six were female and three were male. Four teachers had more than five years of working experience. Four teachers had two to five years of working experience. Only one teacher had less than two years of working experience.

Firstly, three teachers who agreed to participate were assigned to the first focus group. The interview data of the first focus group were transcribed. First, by open coding, qualitative data were broken into discrete parts, and codes were created to label them. During this process, all discrete parts that were labeled with a particular code were collated. Similar events described by interviewees were continuously compared and contrasted to make sure meaningful and distinct codes. With axial coding, the codes and their underlying data were checked to find how the codes could be grouped into categories. Then, three categories were created. Each included a number of different codes. For instance, codes related to conducting assessment activities were grouped into the category of “assessment practices”. The last step was selective coding, where

the categories were connected around one core concept, namely, higher education teacher AL in both classroom and digital settings. Then, six higher education teachers were assigned to the second and third focus groups. The interview data from the second and third focus groups were also analyzed. However, no new information emerged. Therefore, this study did not invite more teachers to conduct new focus group interviews.

Table 4.1 shows the codes and categories yielded from the focus group interviews. It can be seen that the codes and categories correspond to the categories extracted from professional assessment standards (see Table 2.2). Moreover, these codes and categories align with the AL frameworks of Pastore and Andrade (2019) and Chan and Luk (2022), in which AL consists of three dimensions, i.e., assessment knowledge, assessment practices, and socio-emotional aspects of assessment. For example, about assessment knowledge, one teacher reported that "…*This university also offers some new teacher training on assessment knowledge*…". Regarding assessment practices, some of them mentioned that "…*If there are regulations in the department, follow them*…" (assessment criteria), "…*Well, for classroom teaching, regular inspections, individual group presentations, and sharing, and finally final exams*…" (select and implement assessment strategies), or "…*So we will use this as an indicator, and in our next semester's teaching, we will focus on strengthening our training and teaching in collaboration*…" (use assessment data to adjust teaching). Regarding the socio-emotional aspects of assessment, some of them reported that "…*So we collaborated with Xinhua News Agency this year…in this section, which is the practical evaluation, we contacted the teaching staff from Xinhua News Agency to conduct the assessment together*…" (work with stakeholders), or "…*If there is such a situation of suspected plagiarism, I will definitely not tolerate it*…" (ethical aspects). Teachers in the focus groups also shared their experiences and feelings about using technologies, for instance, "…*We*

can directly use the speed grader on Canvas to score, and then give the students a final score directly...". Furthermore, they reported that they had taken on some responsibilities in higher education, such as cooperating with relevant enterprises or institutions to enhance assessment systems for the purpose of improving students' job skills (e.g., the abovementioned *Xinhua News Agency*).

Then, a preliminary framework was developed based on coding analysis results and the previous literature (see Table 4.2).

Table 4.1

Codes and Categories

Categories	Code name	Code description
Assessment knowledge	Assessment knowledge	The knowledge and understanding of assessment.
Assessment practices	Assessment criteria	1) Define learning targets and assessment criteria and align them with the assessment aims. 2) Implement the assessment requirements stipulated by the country, school, and department.
	Select and implement assessment strategies	Select assessment tools and strategies and implement assessment activities, for example, formative assessment and summative assessment.
	Interpret evidence of student learning	Interpret assessment results.
	Use assessment data to adjust teaching	Adjust instruction, curriculum, teaching methods, or assessment strategies based on evidence of student learning, for example, results from formative and summative assessments.
	Provide feedback	Communicate feedback for students, organize and document all the mutual feedback.
	Engage with colleagues	Work with other colleagues on assessments
	Regulate students' learning	Teach and support students in using assessment information to regulate their learning.
	Assessment reports	Produce relevant assessment reports from within the LMS, report and communicate assessment results to students, administrators, and other major users.
Socio-emotional aspects	Work with stakeholders	Work with colleagues, and other stakeholders (for example, institutions, platform developers, and enterprises) to enhance assessment systems.
	Teacher as assessor	Are conscious of their role as assessors and of issues of responsibilities, and rights.

Ethical aspects	Ethical aspects such as: 1) online cheating, teaching to the test, and other assessment malpractices; 2) Fairness and equity.
Engagement and Relationships	1) Students' involvement/engagement (e.g., providing choices for learners about goals, tasks, information sources, and products); 2) Teacher-student relationships.
Emotional aspects	Emotional dynamics of assessment from the student's point of view and to the dispositions that can influence learning (e.g., excessive emphasis on GPA, persistence, test anxiety, and resistance to evaluation) from online activities (e.g., social media).

Expert Review

Four experts studying in AL who are also teaching and research personnel working at four world-renowned universities agreed to participate in this study. All experts have published academic articles related to AL and educational measurement. They were sent consent forms and the link to the framework evaluation survey via email (see Appendix B). The framework evaluation consists of two parts, including the ratings for the validity questions to evaluate the framework and the open-ended questions for identifying any suggestions and comments. All the experts agreed that the preliminary framework covers the AL dimensions of assessment knowledge, assessment practices, and socio-emotional aspects of assessment. However, most of them suggested adding more elements of digital assessment and assessment in the context of higher education. Then, the framework was revised based on the experts' feedback. The revised framework is presented in Table 4.3.

Table 4.2

The Preliminary Framework

Dimensions	AL Components	
Conceptual Knowledge Dimension		Knowledge of assessment purpose, content, and methods Knowledge of grading and data analysis (e.g., gathering information and statistical analyses) Knowledge of assessment interpretation and communication (e.g., to students, parents, and administrators)

		Knowledge of assessment ethics (e.g., cheating, teaching to the test) Knowledge of digital assessment (i.e., how to complement valid assessments in digital environments)
Praxeological Dimension	Define assessment criteria	1) Define assessment criteria and align them with the assessment aims. 2) Implement the assessment requirements stipulated by the department, school, and country.
	Collect evidence of student learning	1) Select assessment strategies (e.g., formative and summative assessments) and implement valid assessments. 2) Collect assessment data from multiple sources, including LMS, apps, online learning platforms, and classrooms.
	Interpretation and providing feedback	1) Make accurate interpretations of assessment results. 2) Communicate feedback face-to-face and through LMS to students. 3) Organize and document all the mutual feedback.
	Adjust teaching	Adjust instruction, curriculum, teaching methods, or assessment strategies based on evidence of student learning, for example, results from formative and summative assessments.
	Engage with stakeholders	Engage with other stakeholders (e.g., other teachers, administrators) about assessment information and assessment reports, by disseminating information, and providing access and permissions to various digital environments.
	Regulate students' learning	Teach and support students in using assessment information to regulate their learning: 1) Manage student-involved assessment practices and encourage students to use advanced technologies (e.g., blogs and computerized practice tests) to implement self-assessment and reflection. 2) Scaffold student understanding of self- and peer-assessment practice through integration of collaborative technologies that enable comment and discussion.
Socio-emotional Dimension	Work with stakeholders	Are effective in working with colleagues, and other stakeholders (for example, institutions, platform developers, enterprises) to create a shared sense-making of assessment practices and enhance assessment systems in the service of student learning
	Teacher as assessor	Their feelings and beliefs about being an assessor and issues of trust, responsibilities, and rights 1) Open-mindedness towards the assessment. 2) Critically in making assessment decisions.
	Ethical aspects	Attend to ethical aspects such as: 1) Unintended consequences, for instance, students should be assessed on their content knowledge and not on their savvy tech usage and adaptability, or test-taken behavior. 2) Cheating, teaching to the test, and other assessment malpractices. 3) Fairness and equity, for instance, some students' older desktop technology may affect their speed.

	Awareness of power	Have awareness of power, and the impact assessment has on: 1) Students' involvement/engagement (e.g., providing choices for learners with regard to goals, tasks, information sources and products). 2) Teacher-student relationships
	Emotional aspects	Attend to the emotional dynamics of assessment from the student's point of view and to the dispositions that can influence learning (e.g., excessive emphasis on GPA, persistence, test anxiety because of their lack of familiarity with the assessment software) from online activities.

Table 4.3

The Revised Framework

Dimensions	Descriptions
Assessment Knowledge	
Knowledge of assessment purpose, content, and methods	**What is assessment? Why assess? What to assess? How to assess?**
Knowledge of grading and data analysis	**1) know how to gather multiple sources of information (e.g. ranging from big data to quantitative and qualitative evidence gathered from both classroom and digital settings)** **2) Relevant statistical and psychometrical concepts, procedures, Techniques** **3) Know how to use digital tools to do data analyses**
Knowledge of assessment interpretation and communication	**1) Know how to interpret assessment results.** **2) Know how to communicate assessment results to relevant stakeholders via different digital tools.**
Knowledge of assessment ethics related to both classroom and remote settings	**Know ethics issues such as academic integrity and AI writing.**
Knowledge of implementing digital assessments	**Know how to implement valid digital assessments.**
Assessment Practice	
Assessment criteria	1) Define assessment criteria and align them with the assessment aims. **2) Use the clearly defined institution-wide grade scale system and apply it consistently across individual programs and courses.** **3) Embed assessment in the culture of the institution and curriculum.**
Assess student learning	1) Select assessment strategies (e.g., formative and summative assessments) and **design an effective assessment plan.**

	2) Implement valid classroom and digital assessments. 2) Collect assessment data from multiple sources, including LMS, apps, online learning platforms, and classrooms.
Interpretating results and providing feedback	1) Make accurate interpretations of **both classroom and digital assessment results.** 2) Communicate feedback to students. 3) Organize and document all the mutual feedback.
Adjust teaching	Adjust instruction, curriculum, teaching methods, or assessment strategies based on evidence of student learning, for example, results from formative and summative assessments.
Engage with colleagues	1) Engage with colleagues about assessment information and assessment reports **from both classroom and digital settings**, such as by disseminating information, and providing access and permissions to various digital environments. **2) Collaborate with colleagues to contribute to the work of determining the areas where course-based and program-based assessment can be integrated and inform each other.**
Regulation students' learning	1) Teach and support students in using assessment information to regulate their learning: a) Manage student-involved assessment practices and encourage students to use advanced technologies (e.g., blogs and computerized practice tests) to implement self-assessment and reflection. b) Scaffold student understanding of self- and peer-assessment practice through the integration of **online collaborative technologies** that enable comment and discussion. **2) Support students to become self-regulated learners with learning motivation and self-esteem.**
Social and Emotional Management	
Work with stakeholders	Work with colleagues, and other stakeholders (e.g., institutions, platform developers, and enterprises) to create a shared sense-making of assessment practices and improve assessment systems **at the school, college, or institutional level that best support student learning.**
Teacher as assessor	Their feelings and beliefs about being an assessor and issues of trust, responsibilities, and rights 1) Open-mindedness towards the assessment, **for instance, learning how to use new digital tools for assessment.** **2) Critical thinking in assessment, for instance, reflecting on the alignment between authentic learning experiences in or across disciplines to understand challenges around assessment.**

	3) Sustainable assessment beliefs for developing students' skills throughout their work and lives.
Ethical aspects	Attend to ethical aspects such as: 1) Unintended consequences, for instance, students should be assessed on their content knowledge and not on their savvy tech usage and adaptability, or test-taken behavior. 2) **Academic integrity, AI writing,** and other assessment malpractices. 3) Equity and justice, for instance, some students' older desktop technology may affect their speed.
Awareness of power	Have awareness of power, and the impact assessment has on: 1) Students' involvement/engagement (e.g., providing choices for learners with regard to assessment goals, tasks, information sources and products). 2) Teacher-student relationships
Emotional aspects	Attend to the emotional dynamics of assessment from the student's point of view and to the dispositions that can influence learning (e.g., excessive emphasis on GPA, persistence, test anxiety because of their lack of familiarity with the assessment software) from online activities.

Note. The bold part is the revisions.

Phase 3: Develop Initial Instrument

The initial measurement scale was developed based on the revised framework and relevant studies. It contains three parts (see Appendix C). The first part asked respondents to rate how knowledgeable they are with 30 statements based on a 5-point Likert scale (from not knowledgeable at all to extremely knowledgeable), for example, "*define rating scales and rubrics for an assessment*". These statements are related to assessment knowledge. Items were developed based on the revised framework and also other AL measurement instruments as references (e.g., Al-Bahlani & Ecke, 2023).

The second part contains 44 statements about university teachers' assessment practices and social and emotional aspects of assessment. Respondents can rate how much they agree with the statements on a 6-point Likert scale ranging from 1 strongly disagree to 6 strongly agree, for example, "*I can collect effective formative assessment results through digital tools such as e-learning platforms or apps*". Six negatively worded items were generated to reduce participants' acquiescence bias.

The third part contains one open-ended question asking respondents to introduce their assessment work and the use of digital assessment. The qualitative data were integrated and compared with the quantitative responses to enhance instrument development and validation.

Phase 4: Pilot-Test Initial Instrument

This phase involves two steps. The first step is the assessment of content validity through the expert review. The second step is the assessment of face validity through the pilot test.

Expert Review

Four experts in Phase 2 evaluated the quality of the instrument. They reviewed each item and rated its 1) relevance to the aim of the scale, 2) clarity of each item; 3) essentiality of each item.

CVI Results. Table 4.4 shows the calculated I-CVI and K* for each item. The I-CVI and K* of 54 items are equal to 1. According to the cut-off values of K*, they were marked as excellent, which means they are relevant items and can be retained. Therefore, it indicates that the rest 21 items should be revised or deleted to improve the quality of the instrument. The S-CVI/UA was 0.730 and the S-CVI/Ave was 0.922. The universal agreement method shows moderate content validity while the average approach demonstrates high content validity of this instrument.

Table 4.4

I-CVI and K for Each Item*

Items	The number of experts rating 3 or 4	I-CVI	P_c	K*	Evaluation results
Part 1					
1	4	1	0.063	1	Excellent
2	4	1	0.063	1	Excellent
3	4	1	0.063	1	Excellent
4	4	1	0.063	1	Excellent

5	4	1	0.063	1	Excellent
6	4	1	0.063	1	Excellent
7	4	1	0.063	1	Excellent
8	4	1	0.063	1	Excellent
9	4	1	0.063	1	Excellent
10	3	0.750	0.250	0.667	Good
11	4	1	0.063	1	Excellent
12	4	1	0.063	1	Excellent
13	4	1	0.063	1	Excellent
14	2	0.500	0.375	0.200	No agreement
15	3	0.750	0.250	0.667	Good
16	4	1	0.063	1	Excellent
17	3	0.750	0.250	0.667	Good
18	3	0.750	0.250	0.667	Good
19	4	1	0.063	1	Excellent
20	4	1	0.063	1	Excellent
21	4	1	0.063	1	Excellent
22	4	1	0.063	1	Excellent
23	4	1	0.063	1	Excellent
24	4	1	0.063	1	Excellent
25	4	1	0.063	1	Excellent
26	2	0.500	0.375	0.200	No agreement
27	4	1	0.063	1	Excellent
28	3	0.750	0.250	0.667	Good
29	4	1	0.063	1	Excellent
30	4	1	0.063	1	Excellent
Part 2					
1	4	1	0.063	1	Excellent
2	4	1	0.063	1	Excellent
3	4	1	0.063	1	Excellent
4	3	0.750	0.250	0.667	Good
5	4	1	0.063	1	Excellent
6	4	1	0.063	1	Excellent
7	4	1	0.063	1	Excellent
8	3	0.750	0.250	0.667	Good
9	4	1	0.063	1	Excellent
10	4	1	0.063	1	Excellent
11	4	1	0.063	1	Excellent
12	4	1	0.063	1	Excellent
13	4	1	0.063	1	Excellent
14	4	1	0.063	1	Excellent
15	3	0.750	0.250	0.667	Good
16	4	1	0.063	1	Excellent
17	4	1	0.063	1	Excellent

18	3	0.750	0.250	0.667	Good
19	4	1	0.063	1	Excellent
20	4	1	0.063	1	Excellent
21	4	1	0.063	1	Excellent
22	4	1	0.063	1	Excellent
23	4	1	0.063	1	Excellent
24	3	0.750	0.250	0.667	Good
25	4	1	0.063	1	Excellent
26	4	1	0.063	1	Excellent
27	3	0.750	0.250	0.667	Good
28	3	0.750	0.250	0.667	Good
29	2	0.500	0.375	0.200	No agreement
30	3	0.750	0.250	0.667	Good
31	3	0.750	0.250	0.667	Good
32	4	1	0.063	1	Excellent
33	4	1	0.063	1	Excellent
34	4	1	0.063	1	Excellent
35	3	0.750	0.250	0.667	Good
36	3	0.750	0.250	0.667	Good
37	4	1	0.063	1	Excellent
38	4	1	0.063	1	Excellent
39	4	1	0.063	1	Excellent
40	4	1	0.063	1	Excellent
41	4	1	0.063	1	Excellent
42	4	1	0.063	1	Excellent
43	4	1	0.063	1	Excellent
44	3	0.750	0.250	0.667	Good

CVR Results. The CVR was calculated for each item (see Table 4.5). It shows that 53 out of 74 items had a CVR of 1, which means they were essential. Seventeen items had a CVR of 0.500 and four items had a CVR of 0. The average CVR was 0.831.

Table 4.5

CVR for Each Item

Items	The number of experts rating 3 or 4	CVR
Part 1		
1	4	1
2	4	1
3	4	1

4	4	1
5	4	1
6	4	1
7	4	1
8	4	1
9	4	1
10	3	0.500
11	4	1
12	4	1
13	4	1
14	2	0
15	3	0.500
16	3	0.500
17	2	0
18	3	0.500
19	4	1
20	4	1
21	4	1
22	4	1
23	4	1
24	4	1
25	4	1
26	2	0
27	4	1
28	3	0.500
29	4	1
30	4	1
Part 2		
1	4	1
2	4	1
3	4	1
4	3	0.500
5	4	1
6	4	1
7	4	1
8	3	0.500
9	4	1
10	4	1
11	4	1
12	4	1
13	4	1
14	4	1
15	3	0.500
16	4	1

17	4	1
18	3	0.500
19	4	1
20	4	1
21	4	1
22	4	1
23	4	1
24	3	0.500
25	4	1
26	4	1
27	3	0.500
28	3	0.500
29	2	0
30	3	0.500
31	3	0.500
32	4	1
33	4	1
34	4	1
35	3	0.500
36	3	0.500
37	4	1
38	4	1
39	4	1
40	4	1
41	4	1
42	4	1
43	4	1
44	3	0.500

Instrument Revision Results. The developed instrument was revised based on the CVI and CVR results. Table 4.6 lists 22 items that were revised, retained, or deleted. Apart from the 21 items with a CVI below 1, item 21 in Part 2 was also revised based on the experts' feedback. In Part 1, items 14, 26, and 28 were deleted. Item 15 was retained because interpreting measurement errors was critical in the assessment. Furthermore, one expert suggested dividing item 13 in Part 1 and items 21 and 24 in Part 2 into two separate items, that is, separating qualitative and quantitative, as well as student self-assessment and peer assessment. Other items were revised based on the previous literature on AL and features that have emerged from

assessment in postsecondary education, and the qualitative analysis of focus groups. For example, items 17 and 18 were revised because the respondents from focus groups reported that they needed to consider the requirements of the major and university in assessment. Finally, the revised instrument was translated into Chinese for the subsequent item review.

Table 4.6

Instrument Revision

Item	Original	Revision
Part 1		
10	avoid bias (personal preferences) in grading.	avoid personal preferences and stereotypes in grading.
13	analyze both qualitative and quantitative learning evidence.	13. analyze qualitative assessment data. 14. analyze quantitative assessment data.
14	know relevant statistical techniques.	deleted
15	interpret measurement error.	retained
17	determine if an assessment aligns with a local system of accreditation.	determine if an assessment aligns with the requirements of the professional field.
18	determine if an assessment aligns with a local educational system.	determine if an assessment aligns with the university's requirements.
26	use LMS to design tests (e.g., online quizzes, online projects).	deleted
28	give students online feedback on assignments through LMS or other apps and websites.	deleted
Part 2		
4	When the assessment criteria do not match the actual assessment situation, I still use the previous assessment criteria.	Retained
8	When I need to use online learning platforms, apps, or other digital tools to assess students' learning progress, I feel overwhelmed and want to give up.	I feel overwhelmed by using digital tools for assessment and refuse to use them.
15	When I find that the content that needs to be adjusted is complex, I will give up the adjustment.	When I find that the teaching content or methods that need to be adjusted are complex, I will give up.
18	I am willing to share assessment information online with other teachers, such as viewing my online test design.	When necessary, I am willing to share digital assessment information with other teachers of the same course.
21	I am willing to adopt peer assessment and self-assessment of students.	21. I am willing to adopt student peer-assessment. 22. I am willing to adopt student self-assessment.
24	I am not very willing to use student self-assessment and peer assessment because I	25. I am not very willing to use student self-assessment because I believe their assessments are not fair enough.

	believe their assessments are not fair enough.	26. I am not very willing to use student peer-assessment because I believe their assessments are not fair enough.
27	I can collaborate with enterprises or organizations to develop a student professional quality assessment system that meets job market demand.	I can collaborate with relevant institutions or organizations to improve assessment systems.
28	I don't think there is a need to consider job market needs in planning assessment.	I don't think there is a need to collaborate with relevant institutions or organizations when improving assessment systems.
29	I understand my identity as the subject of assessment and the responsibilities I bear for it.	deleted and added a new item: I always keep open-minded towards assessment.
30	I believe that I have taken on my due responsibility in the assessment work.	I believe that I have fulfilled my due responsibility in the assessment.
31	I believe that I have made the right decisions in the assessment work.	I believe that I have applied critical thinking in making assessment decisions.
35	In order to cope with the pressure of teaching performance, I often teach to the test.	I try my best to avoid teaching to the test.
36	When some students' assessment performance is affected by outdated equipment or the internet, I will take appropriate measures to ensure fairness and equity in the assessment.	When some students' assessment performance is affected by emergent situations (e.g., sudden internet disconnection, illness), I will take appropriate measures to ensure fairness and equity in the assessment.
44	I think it's normal for students to be under pressure and not to be taken seriously.	When providing assessment feedback to students, I can notice their negative emotions and take appropriate actions to improve them.

Pilot-Test

The web link to the revised instrument was sent to a sample of participants. It includes not only people who actually take this instrument (i.e., higher education teachers), but also the general public (i.e., Chinese native speakers). The purpose was to avoid any obscureness of language, such as acronyms and informal expressions. They were recruited to judge whether the item was clear and understandable on a four-point Likert scale. Thirty-six raters completed the pilot test of the instrument. Finally, 30 valid responses were included in the analysis.

Table 4.7 presents the I-FVI of each item. Fifty-nine items had an I-FVI of 1. The other 15 items had an I-FVI ranging from 0.866 to 0.966. It indicates that all the items are clear and

understandable. The S-FVI/Ave was 0.967 and the S-FVI/UA was 0.784, which means that the whole instrument has achieved a satisfactory level of face validity. Finally, revisions were made to items 1, 2, 27, and 28 of Part 1 and items 4, 8, 30, and 31 of Part 2 to ensure that native Chinese speakers can understand these contents. The final survey is presented in Appendix D (English version) and Appendix E (Chinese version), which includes four parts, 1) background information; 2) self-perceived assessment knowledge; 3) self-perceived assessment competence; 4) an open-ended question. Table 4.8 lists items included in each subscale for Part II and Part III.

Table 4.7

Face Validity Index

Items	The number of experts rating 3 or 4	I-FVI
Part 1		
1	26	0.866
2	28	0.933
3	30	1
4	30	1
5	30	1
6	30	1
7	30	1
8	30	1
9	30	1
10	30	1
11	30	1
12	30	1
13	27	0.900
14	30	1
15	27	0.900
16	30	1
17	30	1
18	30	1
19	30	1
20	30	1
21	30	1
22	30	1
23	28	1
24	30	1

25	28	0.933
26	30	1
27	27	0.900
28	29	0.966
Part 2		
1	29	0.966
2	30	1
3	30	1
4	28	0.933
5	30	1
6	30	1
7	30	1
8	26	0.866
9	30	1
10	30	1
11	30	1
12	30	1
13	30	1
14	30	1
15	26	0.866
16	30	1
17	30	1
18	30	1
19	30	1
20	30	1
21	30	1
22	30	1
23	30	1
24	30	1
25	26	0.866
26	26	0.866
27	30	1
28	30	1
29	30	1
30	26	0.866
31	29	0.966
32	30	1
33	30	1
34	30	1
35	30	1
36	30	1
37	30	1
38	30	1
39	30	1

40	30	1
41	30	1
42	30	1
43	30	1
44	30	1
45	30	1
46	30	1

Table 4.8

Items Included in Each Subscale for Part II and Part III

	Constructs	Item
Part II		
	Assessment Concepts, Purpose, Content, Methods	1-8
	Grading and Data Analysis	9-14
	Interpreting Assessment Results	15-18
	Communicating Assessment Results	19-22
	Assessment Ethics	23-25
	Developing Digital Assessment	26-28
Part III		
	Assessment Criteria	1-4
	Assessing Student Learning	5-8
	Interpreting Results and Providing Feedback	9-12
	Adjusting Teaching	13-16
	Engaging with Colleagues	17-20
	Critical Reflection	21-26
	Working with Stakeholders	27-30
	Mindset	31-34
	Ethical Management	35-38
	Engagement and Relationships	39-41
	Emotional Management	42-46

Phase 5: Field-Test Revised Instrument

The web link to the instrument was sent on an internal social media platform for higher education teachers. A total number of 265 university teachers completed the online survey. Table 4.9 lists the basic demographic information of all the respondents, including their gender, age, department, years of working experience, university level, and training experience. All the participants are volunteers from different disciplines, except for 18 participants who did not report their department. Most of them work in the School of Arts and Humanities and the School

of Engineering. They are aged between 22 and 65 (*M* = 37.098, *SD* = 7.548); 48.7% are male (*n* = 129) and 49.8% are female (*n* = 132). They reported an average of 9.390 years of professional educators (*SD* = 7.735). The universities they work in are mostly from the 985 Project and ordinary second-tier (*n* = 148, 56.1%). Furthermore, about one-fifth of participants reported that they have had 2+ full-semester courses in assessment. One-fifth of participants reported they have participated in the teacher preparation program with assessment topics.

Table 4.9

Demographic Information of the Participants

Demographic information		n (%)
Gender	Male	129 (48.7%)
	Female	132 (49.8%)
	Others	1 (0.4%)
	Prefer not to respond	3 (1.1%)
Age	M = 37.098, SD = 7.548	
Department	Arts & Humanities	73 (27.7%)
	Media & Journalism	20 (7.6%)
	Psychology	5 (1.9%)
	Sociology & Law	7 (2.6%)
	Business	6 (2.3%)
	Kinesiology	1 (0.4%)
	Education	10 (3.8%)
	Economy & Management	20 (7.6%)
	Engineering	65 (24.5%)
	Science (Mathematics, Physics, etc.)	10 (3.8%)
	Life Sciences	6 (2.3%)
	Agriculture, Botany & Zoology	8 (3.0%)
	Ocean, Earth Science, Geography & Geology	7 (2.6%)
	Environment & Sustainability	5 (1.9%)
	Medicine & Pharmacy	4 (1.5%)
	Missing	18 (6.8%)
Years of working experience	M = 9.390, SD = 7.735	
University level	985 Project	77 (29.1%)
	211 Project	47 (17.7%)
	Ordinary first-tier universities	71 (26.8%)

	Ordinary second-tier universities	45 (17.0%)
	Vocational and technical college	18 (6.8%)
	Prefer not to respond	7 (2.6%)
Training experience	2+ full-semester courses in assessment	53 (20.0%)
	1 semester course in assessment	32 (12.1%)
	A short or mini course in assessment	18 (6.8%)
	A module in assessment (e.g., one face-to-face workshop, one online session)	26 (9.8%)
	Teacher preparation program with assessment topics	55 (20.8%)
	No formal preparation for assessment	43 (16.2%)
	I don't remember	38 (14.3%)

Phase 6: Quantitative Analysis Phase

Part II: Self-perceived Assessment Knowledge

Table 4.10 shows descriptive statistics (mean and standard deviation) and Cronbach's α on the self-perceived assessment knowledge of university teachers. A high degree of internal consistency among the items in the scale is indicated by high Cronbach's alpha values. Participants reported they were between *moderately knowledgeable* and *very knowledgeable* in assessment knowledge (M = 3.427, SD = 0.846). Specifically, they reported that they were more knowledgeable on grading and data analysis, and assessment ethics.

Table 4.10

Descriptive Statistics for Part II

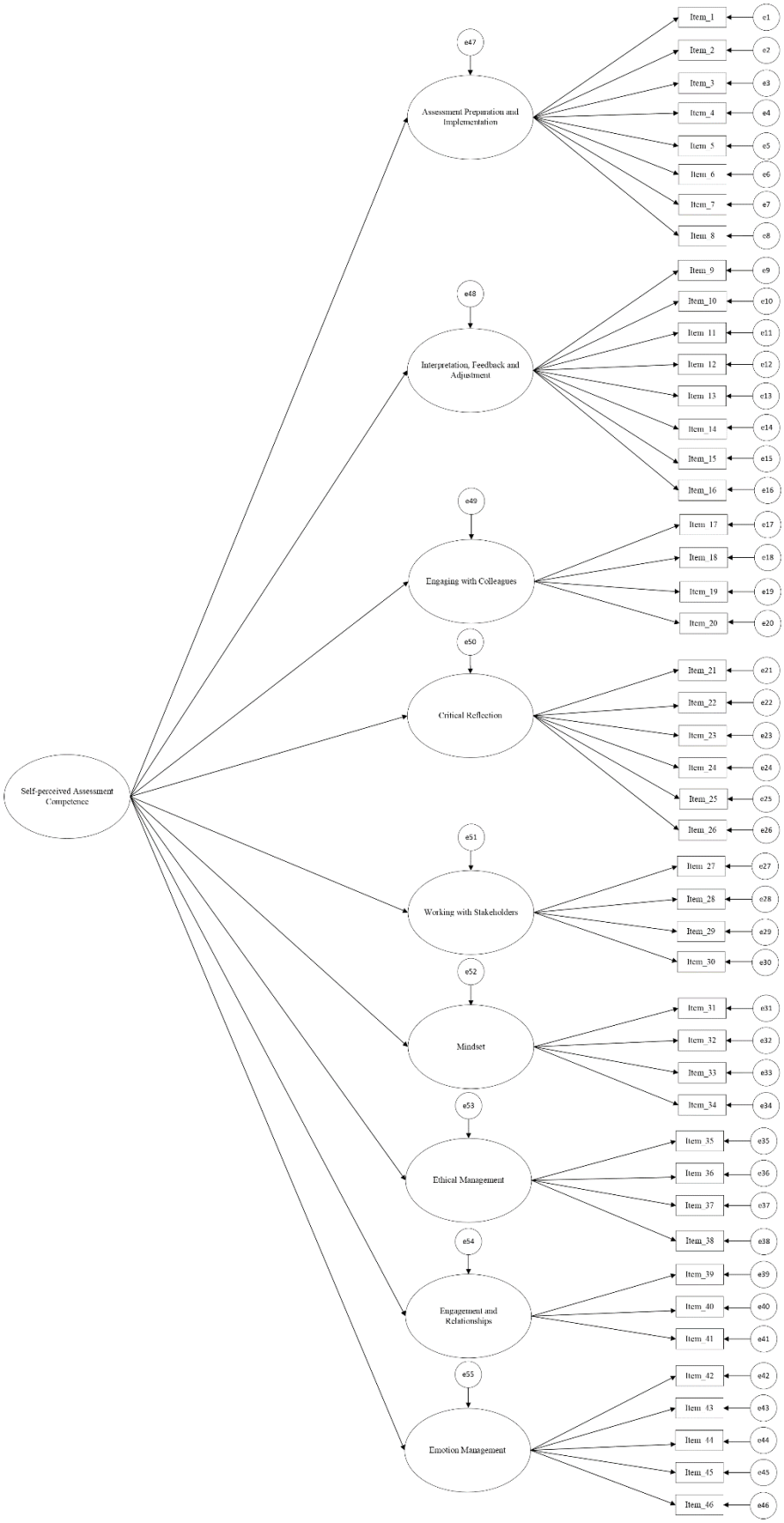
		Item	M	SD	Cronbach'α
Subscale	Assessment Concepts, Purpose, Content, Methods	1-8	3.266	0.922	0.914
	Grading and Data Analysis	9-14	3.609	0.929	0.913
	Interpreting Assessment Results	15-18	3.399	0.938	0.882
	Communicating Assessment Results	19-22	3.343	0.927	0.823
	Assessment Ethics	23-25	3.722	1.019	0.853
	Developing Digital Assessment	26-28	3.346	1.082	0.927
Total scale			3.427	0.846	0.970

Part III: Self-perceived Assessment Competence

Conceptual Model. A conceptual model for university teachers' self-perceived assessment competence was built based on a thorough literature review. Assessment competence was hypothesized covering assessment practice and socio-emotional aspects of assessment. According to the framework of the instrument, assessment practice contains 6 components, including defining assessment criteria, assessing student learning, interpreting results and providing feedback, adjusting teaching, engaging with colleagues, and regulating students' learning. Assessment criteria and assessing student learning were integrated into a latent variable, labeled as *Assessment Preparation and Implementation*. Interpreting results, providing feedback, and adjusting teaching were integrated into a latent variable, labeled *Interpretation, Feedback and Adjustment.* Finally, 9 latent variables were identified as first-order factors and hypothesized loading onto one second-factor, *Self-perceived Assessment Competence* (see Figure 4.1).

Figure 4.1

The Conceptual Model


Self-perceived Assessment Competence
Assessment Preparation and Implementation
Interpretation, Feedback and Adjustment
Engaging with Colleagues
Critical Reflection
Working with Stakeholders
Mindset
Ethical Management
Engagement and Relationships
Emotion Management
e47
e48
e49
e50
e51
e52
e53
e54
e55
Item_1
Item_2
Item_3
Item_4
Item_5
Item_6
Item_7
Item_8
Item_9
Item_10
Item_11
Item_12
Item_13
Item_14
Item_15
Item_16
Item_17
Item_18
Item_19
Item_20
Item_21
Item_22
Item_23
Item_24
Item_25
Item_26
Item_27
Item_28
Item_29
Item_30
Item_31
Item_32
Item_33
Item_34
Item_35
Item_36
Item_37
Item_38
Item_39
Item_40
Item_41
Item_42
Item_43
Item_44
Item_45
Item_46
e1
e2
e3
e4
e5
e6
e7
e8
e9
e10
e11
e12
e13
e14
e15
e16
e17
e18
e19
e20
e21
e22
e23
e24
e25
e26
e27
e28
e29
e30
e31
e32
e33
e34
e35
e36
e37
e38
e39
e40
e41
e42
e43
e44
e45
e46

Data Cleaning. Data cleaning was first performed to delete the extreme responses and neutral responses, which could avoid response bias. Moreover, some participants may respond to agree with all the statements regardless of the content (acquiescence bias; Cronbach, 1942). Such responses would threaten the validity of the survey instrument (Cronbach, 1950). After cleaning these invalid responses, 253 samples were retained for the following analysis.

Descriptive Statistics and Reliability. Table 4.11 presents descriptive statistics and reliability estimates of 46 items in Part III which asked participants to rate how they perceived their assessment competence, including the mean, standard deviation, internal consistency (Cronbach's α) of each factor, and the total scale. It shows that their responses were between *somewhat agree* and *agree* on all the statements.

Regarding Cronbach's α, although a value of greater than 0.700 was widely used to indicate adequate reliability, researchers also recommended a reliability estimate of greater than 0.800 as good and 0.700 as modest (Carmines & Zeller, 1979; Nunally, 1978). Table 4.11 shows that Cronbach's α of some factors before deleting negatively worded items were not all above 0.800. When including 6 negatively worded items in the analysis (item 4, 8, 15, 25, 26, and 30), the Cronbach's α of F1, F4, and F5 were between 0.671 and 0.729. After deleting these items in the analysis, Cronbach's α values of all factors were above 0.800 and Cronbach's α of the total scale reached 0.978. It indicates that the negatively worded items have threatened the reliability of the scale.

Table 4.11

Descriptive Statistics and Cronbach's α for Part III

		Item	M	SD	α
F1	Assessment Preparation and Implementation	1-8	4.094	0.719	0.712
		1-8 (excluded item 4 and 8)	4.194	0.965	0.899

F2	Interpretation, Feedback, and Adjustment	9-16	4.159	0.819	0.844
		9-16 (excluded item 15)	4.242	0.958	0.926
F3	Engaging with Colleagues	17-20	4.441	1.063	0.916
F4	Critical Reflection	21-26	4.010	0.850	0.729
		21-24 (excluded item 25 and 26)	4.209	1.083	0.887
F5	Working with Stakeholders	27-30	4.007	0.861	0.671
		27-29 (excluded item 30)	4.327	1.010	0.885
F6	Mindset	31-34	4.537	0.977	0.892
F7	Ethical Management	35-38	4.269	0.936	0.799
F8	Engagement and Relationships	39-41	4.376	1.016	0.872
F9	Emotional Management	42-46	4.385	0.987	0.912
Total scale			4.219	0.732	0.962
Total scale (excluded item 4, 8, 15, 25, 26, 30)			4.316	0.859	0.978

Note. Item 4, 8, 15, 25, 26, and 30 are negatively worded items. F means factor.

CFA. CFA was used to examine the validity of 9 key factors in self-perceived assessment competence. The CFA analysis was conducted using the *cfa* function of the R *lavaan* package (Rosseel, 2012). First, a one-factor second-order model (Model 1) was built with a total of 46 items, using maximum likelihood estimation. The results show unacceptable model fitness indices ($x^2/df = 3.237$, CFI = 0.780, TLI = 0.768, RMSEA = 0.094, SRMR = 0.081), according to the general criteria (Hu & Bentler, 1999), i.e., CFI, TLI > 0.900, RMSEA < 0.080, SRMR < 0.100. Furthermore, the standardized factor loadings of all 6 negatively worded items are not acceptable (see Table 4.12), according to the minimum accepted range for standardized factor loadings of 0.400 (Hinton et al., 2014; Steven, 2002). Their communality values are also below 0.100. It suggests that negatively worded items do not load onto the target factors. Research evidence has shown that positively worded items and negatively worded items may load onto separate factors (i.e., a method effect), and thus threaten the construct validity of the scale

(Greenberger et al., 2003; Ibrahim, 2001). For model modification, the modification indices that are greater than 20 were checked. It was found that the residual covariances of the combination of all 6 negatively worded items were large. Therefore, 6 negatively worded items were deleted in Model 2.

The results of Model 2 show an acceptable model fit ($x^2/df = 2.173$, CFI = 0.911, TLI = 0.900, RMSEA = 0.068, SRMR = 0.046). The nine coefficients range from 0.851 (F9) to 0.960 (F7), indicating that the first-order factors fall to the second-order factor in high effect sizes (see Figure 5.12). All the items load onto the target first-order factors with standardized factor loadings of greater than 0.600. It suggests that all the items converge well on the target construct.

Examining the correlation between two constructs is one of the most commonly used criteria for testing discriminant validity. Therefore, the inter-correlations of 9 factors were also examined (see Table 4.13). According to the criterion proposed by John and Benet-Martínez (2000), Rönkkö and Cho (2022), i.e., r_{xy} < 0.900, all 9 factors are correlated but also distinct constructs.

Table 4.12

Standardized Factor Loading and Communality of All Items in Model 1

Factor	Item	Standardized factor loading	Communalities
F1	1	0.765	0.585
	2	0.775	0.600
	3	0.834	0.696
	4	0.301	0.091
	5	0.808	0.653
	6	0.720	0.518
	7	0.760	0.577
	8	0.162	0.026
F2	9	0.769	0.592
	10	0.836	0.699
	11	0.810	0.656
	12	0.766	0.586
	13	0.852	0.726
	14	0.812	0.660

	15	0.206	0.042
	16	0.780	0.608
F3	17	0.844	0.712
	18	0.895	0.801
	19	0.837	0.700
	20	0.856	0.733
F4	21	0.773	0.597
	22	0.782	0.612
	23	0.849	0.721
	24	0.844	0.712
	25	0.069	0.005
	26	0.005	0.000
F5	27	0.854	0.729
	28	0.830	0.688
	29	0.861	0.742
	30	0.072	0.005
F6	31	0.799	0.638
	32	0.874	0.763
	33	0.811	0.658
	34	0.805	0.648
F7	35	0.743	0.553
	36	0.661	0.437
	37	0.669	0.448
	38	0.744	0.553
F8	39	0.762	0.581
	40	0.886	0.785
	41	0.864	0.746
F9	42	0.798	0.637
	43	0.817	0.668
	44	0.805	0.648
	45	0.850	0.723
	46	0.837	0.700
AC	F1	0.932	0.837
	F2	0.923	0.868
	F3	0.881	0.728
	F4	0.854	0.721
	F5	0.879	0.755
	F6	0.904	0.818
	F7	0.960	0.868
	F8	0.863	0.792
	F9	0.851	0.728

Table 4.13

Inter-correlations of 9 factors in Model 2

Factors	F1	F2	F3	F4	F5	F6	F7	F8	F9
F1	1								
F2	0.868**	1							

F3	0.710**	0.734**	1						
F4	0.698**	0.712**	0.678**	1					
F5	0.688**	0.737**	0.752**	0.686**	1				
F6	0.742**	0.748**	0.729**	0.680**	0.726**	1			
F7	0.705**	0.719**	0.594**	0.660**	0.628**	0.714**	1		
F8	0.680**	0.747**	0.666**	0.638**	0.641**	0.702**	0.804**	1	
F9	0.676**	0.680***	0.713**	0.632**	0.652**	0.701**	0.723**	0.760**	1

Figure 4.2

The Measurement Model (Model 2)


Self-perceived Assessment Competence
.932
.923
.881
.854
.879
.904
.960
.863
.851
Assessment Preparation and Implementation
Interpretation, Feedback and Adjustment
Engaging with Colleagues
Critical Reflection
Working with Stakeholders
Mindset
Ethical Management
Engagement and Relationships
Emotion Management
e41
e42
e43
e44
e45
e46
e47
e48
e49
.131
.147
.224
.270
.227
.183
.079
.255
.275
Item_1
Item_2
Item_3
Item_5
Item_6
Item_7
Item_9
Item_10
Item_11
Item_12
Item_13
Item_14
Item_16
Item_17
Item_18
Item_19
Item_20
Item_21
Item_22
Item_23
Item_24
Item_27
Item_28
Item_29
Item_31
Item_32
Item_33
Item_34
Item_35
Item_36
Item_37
Item_38
Item_39
Item_40
Item_41
Item_42
Item_43
Item_44
Item_45
Item_46
.745
.740
.821
.818
.706
.746
.803
.823
.800
.731
.858
.832
.777
.815
.869
.837
.869
.729
.740
.865
.858
.849
.825
.865
.797
.876
.808
.809
.734
.643
.653
.696
.754
.887
.868
.824
.846
.803
.793
.769
e1
e2
e3
e4
e5
e6
e7
e8
e9
e10
e11
e12
e13
e14
e15
e16
e17
e18
e19
e20
e21
e22
e23
e24
e25
e26
e27
e28
e29
e30
e31
e32
e33
e34
e35
e36
e37
e38
e39
e40
.445
.453
.327
.331
.502
.443
.355
.323
.361
.466
.264
.308
.396
.336
.244
.299
.245
.469
.452
.251
.264
.279
.319
.252
.365
.232
.348
.345
.462
.587
.574
.515
.432
.213
.247
.321
.284
.356
.371
.409

Phase 7: Qualitative Analysis Phase

In addition to completing both 74 items Part II and Part III, participants were asked to introduce all their assessment work and the use of digital tools in assessment (Qualitative phase). Some responses were deleted because the participants only listed the digital tools they used for assessment. Out of 265 responses, 152 were valid and analyzed using thematic analysis. Table 4.14 lists 10 themes that emerged from this analysis process, which further validated the AL measurement instrument: (1) assessment preparation and implementation, (2) interpretation, feedback, and adjustment, (3) Reflection on learning, (4) engaging with colleagues, (5) working with stakeholders, (6) mindset, (7) ethical management, (8) student engagement, (9) emotional management, (10) teacher-student relationships. For assessment preparation and implementation, some participants mentioned their assessment practices in choosing assessment strategies, defining assessment criteria, implementing assessment, and grading. For example, one participant reported that "…*In course teaching, it is not only about the end of the term, but also about using regular exercises to assess what students have learned, and enriching various forms, not just simple answering questions*…". The second theme is interpretation, feedback, and adjustment, which includes analyzing and interpreting assessment results, learning outcomes, providing feedback, and making adjustments. For example, a teacher reported that "…*summarize and evaluate (assessments) using data mining methods and guide work for the next semester*…". The third theme is reflection on learning, reflecting teachers using self and peer assessment to ask students to critically reflect on their own learning and learning of their peers. It was extracted from the shared experiences of teachers in implementing self and peer assessments, for example, "…*There are two methods for group assignments: self-assessment and peer assessment*…". The fourth theme is engaging with colleagues. Some teachers reported that "…*I would also have*

meetings with colleagues to discuss assessment work…", "*…prepare and design assessment tools together…*", or "*…I often communicate with other teachers, especially those who have the same students or subjects as me, which can provide me with more information and be very beneficial…*". The fifth theme is working with stakeholders, for example, a teacher reported that "*Previously, assessment did not rely heavily on quantitative analysis techniques using digital tools, and we look forward to collaboration with platforms and organizations…*". The sixth theme is mindset, which covers teachers' feelings and beliefs about assessment, and their awareness of being assessors. For example, some teachers reported that "*…The assessment system is traditional and lacks room for individual teachers to make adjustments…*", "*…We cannot rely solely on assessment…We must combine assessment with actual work to achieve rationalization of assessment results…*", or "*…I did not use digital tools very much and am willing to participate in similar training to improve my abilities and level in this area…*". The seventh theme is ethical management, which covers issues of fairness and academic integrity in assessment. For example, a teacher reported that "*…There is a lot of room for dishonesty in online assessment…*". The eighth theme is student engagement, which involves students' learning engagement and classroom interaction, for example, "*…I value classroom interaction more…*". The ninth theme is emotional management. For example, a teacher reported that "*…At the beginning of the semester, I conduct research on students' expectations for the course, and at the end of the semester, I conduct research on whether students are satisfied with the course…*". The last theme is teacher-student relationships. For example, a teacher reported that "*…Students may not be satisfied with assessment results and gave low scores on my teaching. This undermined my teaching enthusiasm…*".

Table 4.14

Codes and Themes Emerged from the Open-ended Responses

Theme	Codes	Frequency *n* (%)
Assessment Preparation and Implementation	Assessment strategies	48 (31.6%)
(*n* = 74, 48.7%)	Assessment criteria	16 (10.5%)
	Implementing assessment	7 (4.6%)
	Grading	3 (2.0%)
Interpretation, Feedback and Adjustment	Analyzing and interpreting assessment results	7 (4.6%)
(*n* = 27, 17.8%)	Learning outcomes	6 (3.9%)
	Providing feedback	9 (5.9%)
	Making adjustment	5 (3.3%)
Reflection on Learning	Peer assessment	5 (3.3%)
(*n* = 9, 5.9%)	Self-assessment and peer assessment	4 (2.6%)
Engaging with Colleagues	Communicating with colleagues	3 (2.0%)
(*n* = 10, 6.6%)	Collaborating with colleagues	3 (2.0%)
	Discussing with colleagues	4 (2.6%)
Working with Stakeholders	Work with institutions	1 (0.7%)
(*n* = 2, 1.3%)	Work with platform staff	1 (0.7%)
Mindset	Feelings about digital assessment tools	18 (11.8%)
(*n* = 41, 27.0%)	Feelings about assessment	16 (10.5%)
	Self-evaluation	7 (4.6%)
Ethical Management	Fairness	9 (5.9%)
(*n* = 11, 7.2%)	Academic integrity	2 (1.3%)
Student Engagement	Learning engagement	5 (3.3%)
(*n* = 6, 3.9%)	Classroom interaction	1 (0.7%)
Emotional Management	Students' emotion	2 (1.3%)
(*n* = 4, 2.6%)	Satisfaction	1 (0.7%)
	Learning motivation	1 (0.7%)
Teacher-student relationships	Teacher-student relationships	1 (0.7%)

(*n* = 1, 0.7%)

Qualitative-Dominant Crossover Analyses

In this phase, qualitative responses were quantified. Specifically, the frequencies of the themes that were extracted from Phase 7 were calculated (see Table 4.14). It shows that *Assessment Preparation and Implementation* was the most (*n* = 74, 48.7%), followed by teachers' *Mindset* (*n* = 41, 27.0%). It suggests that teachers mainly concentrated on assessment strategies, assessment criteria, and assessment implementation in their assessment practice. They kept critical thinking in their assessment work and were willing to share their attitudes and feelings toward assessment. *Working with stakeholders* and *Teacher-student Relationships* were the least topics that teachers introduced (*n* = 2, 1.3%; *n* = 1, 0.7%).

Phase 9: Quantitative-Dominant Crossover Analyses

In this phase, the frequency of each factor calculated from Phase 7 was correlated to the factor loadings stemming from the CFA of Model 2 in Phase 6. Shapiro-Wilk's test of normality shows that the data was not normally distributed ($p > 0.05$). Because of the small sample size, Kendall's correlation coefficient was obtained ($t = 0.564$, $p < 0.05$). It indicates that the factors extracted from open-ended responses were significantly correlated with their factor loadings, suggesting that the more they mentioned this theme in their assessment work, the higher factor loading it has on the core construct, assessment competence. For example, among all open-ended responses, 48.7% mentioned the factor *Assessment Preparation and Implementation*, with a relatively high factor loading of 0.932, which implies that this theme is a vital factor in teacher self-reported assessment competence.

Phase 10: Product and Process Evaluation

Product Evaluation

Both quantitative and qualitative data were collected and analyzed to draw inferences regarding the reliability and validity of the measurement instrument. Specifically, the CFA results and open-ended responses, as well as the crossover analyses, indicate that this instrument have adequate face validity, item validity, construct validity, and discriminant validity.

First, in this study, higher education teachers' AL in both classroom and digital contexts was assessed in two parts, self-reported assessment knowledge and assessment competence. CFA confirmed 9 first-order factors within one second-order factor, assessment competence, including 1) *Assessment Preparation and Implementation*; 2) *Interpretation, Feedback, and Adjustment*; 3) *Engaging with Colleagues*; 4) *Critical Reflection*; 5) *Working with Stakeholders*; 6) *Mindset*; 7) *Ethical Management*; 8) *Engagement and Relationships*; 9) *Emotional Management.* These factors reflect the assessment practices and socio-emotional aspects of assessment, as indicated in the revised framework. The inter-correlation coefficients provided a clear picture of how the 9 factors correlated with each other in the assessment. It indicates that the 9 factors are distinct but positively interrelated and interwoven during the assessment process.

The *Assessment Preparation and Implementation* factor (6 items) is comprised of defining assessment criteria and differentiating assessment strategies as well as implementing assessments. The *Interpretation, Feedback, and Adjustment* factor (7 items) involves interpreting assessment results, providing feedback, and adjusting their teaching. These activities typically occur after the implementation of assessments. The first two factors constitute the main assessment practices, which can also be reflected in the open-ended responses, with 66.4% of respondents reporting these activities in their assessment work. The *Engage with Colleagues* factor (4 items) involves working with colleagues about assessment information. For example, some respondents reported that they discussed assessments with colleagues and obtained

valuable information and experience from them. The *Critical Reflection* factor (4 items) includes using self- and peer-assessment to support students to reflect on their own learning and the learning of their peers and finally become self-regulated learners. Self-regulated learning is one key characteristic of assessment in postsecondary education when conceptualizing assessment from an institutional perspective. In open-ended responses, these two kinds of assessments were also mentioned by several teachers. The *Work with Stakeholders* factor (3 items) includes working with stakeholders to improve the assessment plan and system. Although it has been revealed in the literature and focus group interviews, it was the least reported factor among respondents in the open-ended question. The *Mindset* factor (4 items) involves teachers' attitude towards assessment, for example, being conscious of their own role as assessor. This is one key emotional construct in AL frameworks (e.g., Chan & Luk, 2022; Pastore & Andrade, 2019). The open-ended responses also confirmed its significance with 27.0% occurrence. The *Ethical Management* factor (4 items) includes fairness, equity, and academic integrity in assessment, for example, cheating and teaching to test. This is also one important component of assessment, according to the results of crossover analyses. The *Engagement and Relationships* factor (3 items) includes student engagement in assessment and learning and student-teacher relationships. This factor was reflected by two themes identified in open-ended responses with 4.6% occurrence. The last factor, *Emotional Management* (5 items), involves students' emotions and feelings of assessment and their dispositions that influence learning such as test anxiety and learning enthusiasm. Despite a relatively lower occurrence in open-ended responses, it has been recognized as one important component in AL frameworks.

Second, negatively worded items were adopted in designing the measurement instrument to alleviate response bias. However, CFA results show that negatively worded items damaged

the reliability and construct validity of the measurement instrument. It indicates that including negatively worded items in the scale needs caution, as suggested by researchers (e.g., Ibrahim, 2001; Salazar, 2015). Some researchers also suggested better not to mix negative and positive worded statements, but if mixing, use strategies, such as warning respondents to look out for negative wording (Mathews & Shepherd, 2002) and administering the survey when respondents are not fatigued (Merritt, 2012). In this study, given the length of the scale, the negatively worded items were deleted. CFA without these items shows adequate model fit. In future research regarding developing a scale of AL, researchers can try to design negatively worded items of high quality and examine the possible latent variable they reflect.

During the process of evaluating the product, one researcher who is both a higher education teacher and an expert in educational measurement served as the debriefer. We discussed the findings of this study and the scale developed in this study. I learned that the results support the AL frameworks of Pastore and Andrade (2019) and Chan and Luk (2022), which conceptualized AL as a multidimensional construct. The three dimensions in Pastore and Andrade's (2019) AL framework and the four dimensions in Chan and Luk's (2022) framework were assessed not in a single part but in separate two parts with different forms, where assessment knowledge was measured separately. The reason for this separation is to distinguish between the assessment knowledge and a series of actions generated in the assessment work, more specifically, between "I know" and "I can", "I am willing to" or "I have confidence in". Therefore, assessment knowledge was measured by their ratings of how much they knew about it, while other dimensions were measured by how they rated their ability and willingness to do an assessment activity. The quantitative and qualitative data analysis results confirmed that the items in this measurement scale can measure what it was designed to measure. The reflection on

how findings are consistent or inconsistent with the original conceptual framework of the studied construct helps us understand how to ensure the construct validity of a measurement scale.

Process Evaluation

Debriefing occurred during the instrument development and validation phase, mainly including the length of the scale, sampling, verifying item wording, and qualitative and quantitative data analysis. The areas of expertise determine if a debriefing is necessary.

First, for instrument development, debriefing occurs when experts on this construct discuss whether the items can measure the original theory of the construct. This is of great importance to obtain valuable information from experts and increase my understanding of the instrument development process. Furthermore, Chinese native speakers served as debriefing interviewers for modifying the item wording. For example, item wording should avoid abbreviations and double negative expressions that may confuse respondents. This debriefing process helps increase the face validity of the instrument.

Second, psychometric inquiry with an expert in this area provided valuable insights for instrument development and sampling. Specifically, we discussed what the target sample should be in each phase and how the length of a scale affects the response rate and causes fatigued. For instance, we found that the information obtained from the first focus group interview covers almost all the aspects of the original AL theoretical frameworks. No new information has been found since the second focus group interview. Therefore, a total of three focus groups with nine interviewees were conducted. Furthermore, the sample size for field testing should be large enough according to the number of indicators and factors as well as covariances between factors. The sample also needs to comprehensively cover the target audience. For example, the experts suggested sampling by department and years of working experience to avoid sampling a single

group. The debriefing process benefits instrument validation based on psychometric considerations.

Third, this study shows that inquiring with experts in qualitative, quantitative, and mixed-method research is necessary for the researcher to implement the IDCV process. In this study, a series of different qualitative and quantitative methods were used to finally conduct concurrent mixed research with a rationale of triangulation. As one example, CFA was employed to analyze the quantitative data instead of EFA because we discussed the research logic and identified the confirmatory nature of the research, namely, finding evidence for the hypothesis generated from the literature. The subsequent crossover analyses also yielded strong meta-inferences for this study. We recognized that only qualitative or quantitative findings would lead to insufficient evidence. Even the sequential mixed methods would not inform better instrument quality than concurrent mixed methods. The strengths of analyzing qualitative and quantitative data concurrently have also been demonstrated in previous studies on instrument development and validation (e.g., Koskey et al., 2018).

Chapter Summary

Chapter 4 reported and discussed the main findings in each phase of the IDCV process. Phase 1 to 4 involve how the instrument was developed and revised. In Phase 5, the instrument was sent to participants to collect both qualitative and quantitative data. Phase 6 to 9 involve the validation of the instrument. Specifically, qualitative and quantitative data were analyzed separately. Then, crossover analyses were adopted by combining two research approaches. In Phase 10, the researcher evaluated and reflected on the research process and product. The next chapter provides a summary of this study, followed by practical implications as well as limitations and future research.

CHAPTER FIVE: CONCLUSIONS

This study implemented Onwuegbuzie et al.'s (2010) IDCV process to develop and validate a measurement scale for higher education teachers' AL in both classroom and digital contexts. The IDCV process serves as a mixed-methods meta-framework for developing and validating measurement scales. It contains 10 phases from identifying and conceptualizing the construct of the interest to validating and evaluating the process and product. In Phase 1, the belief systems of the researcher, including the researcher's overall worldview, research philosophy, and discipline-specific philosophy, were established. Specifically, a mix of quantitative and qualitative assumptions under the research philosophy of blending the postpositivist stance with the constructivist stance was identified to guide this study. A systematic review of previous AL theoretical frameworks and governmental or institutional documents on educational assessment as well as the existing measures of teacher AL was conducted to gain valuable information on the construct of interest. In Phase 2, the qualitative data from focus groups were analyzed to identify the behaviors underlying the construct of AL and generate a framework for developing the initial instrument. Then, the framework was revised based on feedback from expert review. In Phase 3, The initial measurement scale was developed based on the literature and revised framework. In Phase 4, the initial measurement scale was revised based on the results from the expert review and pilot testing. Then, the revised measurement scale was administered to the target audience in Phase 5. Phase 6 to Phase 9 involves a series of data analyses. Specifically, quantitative data were analyzed using descriptive statistics and CFA. Qualitative data, namely, responses to the open-ended question, were analyzed using thematic analysis. Then, crossover analyses were conducted to combine interpretations of the qualitative and quantitative data analysis results for the purpose of

informing the validation of the measurement scale. The last phase involves a thorough evaluation of the measurement scale and the IDCV process.

Research Significance

The current study developed a framework of AL in higher education for blended learning that combines digital and face-to-face contexts. This framework draws on the assessment experience of teachers and opinions from the AL experts, as well as previous theoretical frameworks of AL. It reflects the multidimensionality of AL and the characteristics of assessment in higher education and using digital tools in assessment. It may contribute to the conceptualization of AL in specific contexts, which enriches the understanding of the interplay of various practical, social, and emotional factors involved in assessment. Furthermore, as a new scale aiming to measure higher education teachers' AL in the blended learning context, the instrument developed in this study that shows adequate psychometric qualities can also contribute to the field of AL measurement.

As regional policies and priorities should be considered to ensure geographic validity, this study reviewed professional standards for assessment from several regions and developed items that are applicable to higher education in China. This consideration based on regional features enables researchers to compare and contrast AL measurements that are grounded in different regional contexts, thus deepening our understanding of the concept of AL and its practical representation in various contexts.

Practical Implications

This study developed and validated a measurement instrument to measure the AL of higher education teachers in both classroom and digital environments.

Regarding the methodology, mixed methods in the IDCV process were used. Earlier research on instrument development and validation focused on solely quantitative or qualitative approaches. For example, regarding quantitative instrument development, Campbell and Fiske (1959) introduced multitrait-multimethod matrix (MTMM) in their seminal article as a comprehensive quantitative framework. This framework adopted a matrix of intercorrelations among tests representing at least two traits, each measured by at least two methods, for assessing construct-related validity using two validity types: convergent validity and discriminant validity. Some scholars regarded this study as a start of combining quantitative and qualitative approaches (Powell et al., 2008), as they formalized the idea of using multiple research methods and introduced the idea of triangulation in methodology. Although MTMM has been regarded as an innovative and useful tool for quantitative cross-validation, it solely relied on quantitative techniques (Brewer & Hunter, 2006; Teddlie & Johnson, 2009). Collins et al. (2006) suggested that qualitative methods can be used to promote the development of quantitative instruments and vice versa. For qualitative instrument development, Hak et al. (2008) proposed a qualitative method, TSTI, for pretesting a self-completion questionnaire, which includes three steps: 1) (respondent-driven) observing response behavior; 2) (interviewer-driven) follow-up investigation to fill gaps in observational data; 3) (interviewer-driven) debriefing to gain experiences and opinions. This qualitative-dominated method was proved by the authors to be effective in identifying problems that stem from the gap between the ‘theory’ underlying the research questions and the actual behavior and biographical characteristics of the respondents. They also suggested that different methods can be applied to the same instrument. Under the methodological wave of mixed research, TSTI has also been used as a validation method for quantitative instruments (see Van der Veer et al., 2013).

This methodological wave has gained much attention and become popularized in recent decades (Denscombe, 2008). It advocated researchers to employ qualitative and quantitative methods in the same study (Teddlie & Johnson, 2009; Johnson & Onwuegbuzie, 2004), "for the broad purposes of breadth and depth of understanding and corroboration'' (Johnson et al., 2007, p. 123). Such MMR allows researchers to integrate the strengths of qualitative and quantitative worldviews and approaches and comprehensively incorporate the insiders' (i.e., emic) and researcher-observers' (i.e., etic) perspectives on the research phenomenon (Johnson et al., 2007; Younas et al., 2020). More specifically, MMR can ensure that the discrepancy reflected refers to the trait assessed and not to the methodology (i.e., multiple operationalism) (Campbell & Fiske, 1959), maximize the advantages of triangulation between and within methods (Denzin, 1978; Jick, 1979; Morse, 1991), and show strengths on four dimensions, i.e., participant enrichment, instrument fidelity, treatment integrity, and significance enhancement (Collins et al., 2006). Onwuegbuzie et al.'s (2010) framework has been validated as an effective mixed methods research model in developing and validating instruments due to the integration of qualitative and quantitative methods concurrently rather than sequentially. Crossover analyses in this model enable researchers to make better decisions in instrument quality. Debriefing in this model also helps researchers to evaluate the decisions made in the research process. Guided by Onwuegbuzie et al.'s (2010) framework, this study adopted qualitative methods such as focus groups and open-ended questions, and quantitative methods such as the Likert scale and CFA for data analysis. It provides an illustrative case for applying IDCV. Under this framework, more quantitative and qualitative methods could be introduced, such as Rasch modeling for the quantitative phase and think-aloud interviews for the qualitative phase (e.g., Koskey et al., 2018).

Concerning AL, this study examined it in the contexts of higher education and blended learning environments. The last decade has witnessed a great number of studies that investigate what AL is, how it is conceptually defined, how it can be learned, and in what ways it can be measured and explored among teachers. The conceptualization of AL has expanded from a set of assessment knowledge and skills to a multidimensional construct that involves sociocultural and socio-emotional aspects, such as the AL frameworks of Pastore and Andrade (2019) and Chan and Luk (2022). Accordingly, the professional standards of AL have evolved from a rough and limited perspective to a wide range of diverse activities in the assessment domain. This study extracted both the general requirements of AL and the specific characteristics of digital assessments and assessments in higher education from literature and related documents. Information from the focus group interviews was also used to complement and improve the proposed AL framework. This framework could be used as a guide for policy and practice of teacher training programs in higher education. The AL components may be helpful for policymakers to devise assessment standards for teachers.

Regarding AL measurement, the instrument developed and validated in this study has adequate reliability and validity. It can be used by researchers and educators to support and promote assessment work in higher education. This instrument has two main parts, self-perceived assessment knowledge and assessment competence. The part of assessment knowledge can be used to evaluate the knowledge level of teachers. The part of assessment competence can be used to evaluate the level of teachers' abilities to carry out assessment work, such as, what they can do and what they need to improve. Thus, it can help promote interventions or diagnose problems in training pre- or in-service teachers on assessment.

Limitations and Future Research

This study has some limitations that need consideration. First, the AL scale developed in the present study has been validated and can be used to measure higher education teachers' AL in blended contexts. However, this scale is designed based on China's educational background. Its generalizability to other countries needs caution. Future research can be conducted to compare it with other scales from a cross-cultural perspective. Second, CFA was employed to confirm the factor structure of a collection of observed variables. Future studies can adopt other statistical techniques as complementary methods for scale validation. For instance, Rasch modeling has been used in validating measurement scales (e.g., David et al., 2018; Koskey et al., 2018). For instance, when applying the IDCV process to develop the transformative experience questionnaire, Koskey et al. (2018) used the Rasch model to assess the item fit and person fit indices, which can be utilized to provide information about the content-related and construct-related validity. Third, although negatively worded items did not perform well in the psychometric properties of the developed AL scale, this does not mean that they are useless. Future studies can generate high-quality negatively worded items to contribute to the development and validation of the measurement scale.

References

Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205–221). Dordrecht: Springer.

ACER. 2021. *Developing a teachers' assessment literacy and design competence framework*. International Baccalaureate Organization.

Adie, L. (2013). The development of teacher assessment identity through participation in online moderation. *Assessment in Education: Principles, Policy & Practice*, *20*, 91–106. https://doi.org/10.1080/0969594X.2011.650150

Al-Bahlani, S. M., & Ecke, P. (2023). Assessment competence and practices including digital assessment literacy of postsecondary English language teachers in Oman. *Cogent Education*, *10*(2). https://doi.org/10.1080/2331186X.2023.2239535

Alkharusi, H. (2011). Psychometric properties of the teacher assessment literacy questionnaire for preservice teachers in Oman. *Procedia, Social and Behavioral Sciences, 29*, 1614–1624. https://doi.org/10.1016/j.sbspro.2011.11.404

Alkharusi, H., Kazem, A. M., & Al-Musawai, A. (2011). Knowledge, skills, and attitudes of preservice and inservice teachers in educational measurement. *Asia-Pacific Journal of Teacher Education, 39*(2), 113–123. https://doi.org/10.1080/1359866X.2011.560649

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA). (1999). *Standards for Educational and Psychological Testing*. Retrieved from http://www.teststandards.org/

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.

Assessment Reform Group (ARG). (2008). *Changing assessment practices: Process, principles and standards*. Retrieved from 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.

Atjonen, P., Pöntinen, S., Kontkanen, S., & Ruotsalainen, P. (2022). In enhancing preservice teachers' assessment literacy: Focus on knowledge base, conceptions of assessment, and teacher learning. *Frontiers in Education (Lausanne), 7*. https://doi.org/10.3389/feduc.2022.891391

Australian Institute for Teaching and School Leadership. (2011). *Australian Professional Standards for Teachers*. AITSL. Melbourne.

Bao, W. (2020). COVID‐19 and online teaching in higher education: A case study of Peking University. *Human Behavior and Emerging Technologies, 2*(2), 113–115. https://doi.org/10.1002/hbe2.191

Beijaard, D., Meijer, P. C., & Verloop, N. (2004). Reconsidering research on teachers' professional identity. *Teaching and Teacher Education, 20*(2), 107–128. https://doi.org/10.1016/j.tate.2003.07.001

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Brewer, J., & Hunter, A. (2006). *Foundations of multimethod research: Synthesizing styles*. Thousand Oaks, CA: SAGE.

Broadfoot, P. (1996). *Education, assessment and society: a sociological analysis*. Open University Press.

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

Brown, G. T. L. (2002). *Teachers' conceptions of assessment* [Unpublished doctoral dissertation]. University of Auckland, Auckland. Available online at: http://researchspace.auckland.ac.nz/handle/2292/63

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice, 11*(3), 301–318.

Brown, G. T. L. (2006). Teachers' conceptions of assessment: Validation of an abridged instrument. *Psychological Reports, 99*, 166–170.

Brown, G. T. (2011). Teachers' conceptions of assessment: Comparing primary and secondary teachers in New Zealand. *Assessment Matters, 3*(3), 45–70. https://doi.org/10.18296/am.0097

Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for student improvement: Understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy & Practice, 16*(3), 347–363. https://doi.org/10.1080/09695940903319737

Brown, G. T. L., Hui, S. K. F., Yu, F. W. M., & Kennedy, K. J. (2011). Teachers' conceptions of assessment in Chinese contexts: A tripartite model of accountability, improvement, and

irrelevance. *International Journal of Educational Research*, *50*(5), 307–320. https://doi.org/10.1016/j.ijer.2011.10.003

Brown, S., & Race, P. (2013). Using effective assessment to promote learning. In L. Hunt & D. Chalmers (Eds.), *University teaching in focus: A learning-centred approach* (pp. 74–91). New York: Routledge.

Brown, G. T. L., & Remesal, A. (2012). Prospective teachers' conceptions of assessment: A cross-cultural comparison. *Spanish Journal of Psychology*, *15*(1), 75–89. https://doi.org/10.5209/rev_SJOP.2012.v15.n1.37286

Burton, L. J., & Mazerolle, S. M. (2011). Survey instrument validity Part I: Principles of survey instrument development and validation in Athletic training education research. *Athletic Training Education Journal*, *6*(1), 27–35. https://doi.org/10.4085/1947-380X-6.1.27

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Campbell, C., Murphy, J. A., & Holt, J. K. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to pre-service teachers.* Columbus, OH: Paper presented at the Mid-Western Educational Research Association.

Candy. P.C. (1989). Constructivism and the study of'self-direction in adult learning. *Studies in the Education of Adults*, *21*(2), 95–116. https://doi.org/10.1080/02660830.1989.11730524

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage.

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures, 12*(1), 25–44, https://doi.org/10.1080/19312458.2017.1396583

Chan, C. K. Y., & Luk, L. Y. Y. (2022). A four-dimensional framework for teacher assessment literacy in holistic competencies. *Assessment and Evaluation in Higher Education, 47*(5), 755–769. https://doi.org/10.1080/02602938.2021.1962806

Chen, J., & Brown, G. T. L. (2016). Tensions between knowledge transmission and student-focused teaching approaches to assessment purposes: helping students improve through transmission. *Teachers and Teaching, Theory and Practice, 22*(3), 350–367. https://doi.org/10.1080/13540602.2015.1058592

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*(1), 38–54. https://doi.org/10.1080/09500789708666717

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*(2), 127–137.

Collins, K. M. T., Onwuegbuzie, A. J. & Sutton, I. L. (2006). A model incorporating the rationale and purpose for conducting mixed methods research in special education and beyond. *Learning Disabilities: A Contemporary Journal, 4*, 67–100.

Combs, J. P., Bustamante, R. M., & Onwuegbuzie, A. J. (2010). A mixed methods approach to conducting literature reviews for stress and coping researchers: An interactive literature review process framework. In K. M. T. Collins, A. J. Onwuegbuzie, & Q. G. Jiao (Eds.), *Toward a broader understanding of stress and coping: Mixed methods approaches* (pp. 213-241). Greenway, CT: Information Age.

Coombs, A., Rickey, N., DeLuca, C., & Liu, S. (2022). Chinese teachers' approaches to classroom assessment. *Educational Research for Policy and Practice, 21*(1), 1–18. https://doi.org/10.1007/s10671-020-09289-z

Cowie, B., Cooper, B., & Ussher, B. (2014). Developing an identity as a teacher-assessor: Three student teacher case studies. *Assessment Matters, 7*. https://doi.org/10.18296/am.0128

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology, 33*(6), 401–415. https://doi.org/10.1037/h0054677

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*(1), 3–31. https://doi.org/10.1177/001316445001000101

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Dalglish, S. L., Khalid, H., & McMahon, S. A. (2021). Document analysis in health policy research: The READ approach. *Health Policy and Planning, 35*(10), 1424–1431. https://doi.org/10.1093/heapol/czaa064

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future*. New York: Teachers College Press.

David, S. L., Hitchcock, J. H., Ragan, B., Brooks, G., & Starkey, C. (2018). Mixing interviews and rasch modeling: Demonstrating a procedure used to develop an instrument that measures trust. *Journal of Mixed Methods Research, 12*(1), 75–94. https://doi.org/10.1177/1558689815624586

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*(4), 194–197. https://doi.org/10.1016/S0897-1897(05)80008-4

Denscombe, M. (2008). Communities of practice. *Journal of Mixed Methods Research, 2*(3), 270–283. https://doi.org/10.1177/1558689808316807

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 17*(4), 419–438. https://doi.org/10.1080/0969594X.2010.516643

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016a). Teacher assessment literacy: a review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251–272. https://doi.org/10.1007/s11092-015-9233-6

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016b). Approaches to Classroom Assessment Inventory: A New Instrument to Support Teacher Assessment Literacy. *Educational Assessment, 21*(4), 248–266. https://doi.org/10.1080/10627197.2016.1236677

DeLuca, C., Valiquette, A., Coombs, A., LaPointe-McEwan, D., & Luhanga, U. (2018). Teachers' approaches to classroom assessment: A large-scale survey. *Assessment in Education: Principles, Policy & Practice, 25*(4), 355–375. https://doi.org/10.1080/0969594X.2016.1244514

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods*. New York: Praeger.

Department for Education-United Kingdom. (2011). *Teachers' standards.* Retrieved from http://www.education.gov.uk/schools/teachingandlearning/reviewofstandards/a00205581/teachers-standards1-sep-2012

DeVellis R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: Sage.

Dinan Thompson, M., & Penney, D. (2015). Assessment literacy in primary physical education. *European Physical Education Review, 21*(4), 485–503. https://doi.org/10.1177/1356336X15584087

Eyal, L. (2012). Digital assessment literacy — the core role of the teacher in a digital environment. *Educational Technology & Society, 15*(2), 37–49.

Fulcher, G. (2012). Assessment Literacy for the Language Classroom. *Language Assessment Quarterly, 9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041

Fullan, M., & Watson, N. (2000). School-based management: Reconceptualizing to improve learning outcomes. *School Effectiveness and School Improvement, 11*, 453–473. https://doi.org/10.1076/sesi.11.4.453.3561

Gardner, J. (2006). Assessment for learning: A compelling conceptualization. In J. Gardner (Ed.), *Assessment and learning* (pp. 197–204). London: Sage.

Gipps, C. (2002). Sociocultural perspectives on assessment. In G. Wells, & G. Claxton (Eds.), *Learning for life in the 21st century: Sociocultural perspectives on the future of education* (pp.73–83). Malden, MA: Blackwell Publishers.

Giraldo Aristizábal, F. (2018). A diagnostic study on teachers' beliefs and practices in foreign language assessment. *Íkala : Revista De Lenguaje Y Cultura*, *23*(1), 25–44. https://doi.org/10.17533/udea.ikala.v23n01a04

Girgla, A., Good, L., Krstic, S., McGinley, B., Richardson, S., Sneidze-Gregory, S., & Star, J. (2021). *Developing a teachers' assessment literacy and design competence framework*.

https://www.ibo.org/research/ curriculum-research/cross-programme/developing-a-teachers-assessment-literacy-and-design-competenceframework-2021/

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement, Issues and Practice, 33*(2), 14–18. https://doi.org/10.1111/emip.12030

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and Individual Differences, 35*(6), 1241–1254. https://doi.org/10.1016/S0191-8869(02)00331-8

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*, 255–274. https://doi.org/10.2307/1163620

Grimm, K. J., & Widaman, K. F. (2012). Construct validity. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 621–642). American Psychological Association. https://doi.org/10.1037/13619-033

Grix, J. (2004). *The foundations of research*. London: Palgrave Macmillan.

Guba, E. G. & Lincoln. Y.S. (1989). What is this constructivist paradigm anyway? In E. G. Guba & Y. S. Lincoln (Eds.), *Fourth Generation Evaluation* (pp. 79–90). London: Sage Publications.

Hak, T., van der Veer, K., & Jansen, H. (2008). The Three-Step Test-Interview (TSTI): An observation-based method for pretesting self-completion questionnaires. *Survey Research Methods, 2*(3), 143–150. https://doi.org/10.18148/srm/2008.v2i3.1669

Haynes, S., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247.

Heinrichs, S., Bernotsky, R. L., & Danner, L. R. (2015). Guiding principles to impact an institution-wide assessment initiative. *Research & Practice in Assessment, 10*, 60–64.

Hinton, P. R., McMurray, I., & Brownlow, C. (2014). *SPSS explained*. London: Routledge.

Hoehle, H., & Venkatesh, V. (2015). Mobile application usability: Conceptualization and instrument development. *MIS Quarterly, 39*(2), 435–472. https://doi.org/10.25300/MISQ/2015/39.2.08

Howell Smith, M. C., Babchuk, W. A., Stevens, J., Garrett, A. L., Wang, S. C., & Guetterman, T. C. (2020). Modeling the use of mixed methods–grounded theory: Developing scales for a new measurement model. *Journal of Mixed Methods Research, 14*(2), 184–206. https://doi.org/10.1177/1558689819872599

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hung, C.-C., Liu, H.-C., Lin, C.-C., & Lee, B.-O. (2016). Development and validation of the simulation-based learning evaluation scale. *Nurse Education Today, 40*, 72–77. https://doi.org/10.1016/j.nedt.2016.02.016

Ibrahim, A. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88*(2), 497–500. https://doi.org/10.2466/PR0.88.2.497-500

Interstate Teacher Assessment and Support Consortium (InTASC). (2011). *InTASC model core teaching standards: a resource for state dialogue*. Washington: Council of Chief State School Officers. Retrieved from

http://www.ccsso.org/Resources/Publications/InTASC_Model_Core_Teaching_Standards_A_Resource_for_State_Dialogue_DApril_2011].html

James, R., McInnis, C., & Devlin, M. (2002). *Core principles effective assessment*. Retrieved from http://learnline.cdu.edu.au/commonunits/documents/Core%20principles%20of%20effective%20assessment.pdf.

James, M., & Pedder, D. (2006). Beyond method: Assessment and learning practices and values. *Curriculum Journal (London, England)*, *17*(2), 109–138. https://doi.org/10.1080/09585170600792712

Jamil, M., Tariq, R. H., & Shami, P. A. (2012). Computer-based vs paper-based examinations: Perceptions of university teachers. *Turkish Online Journal of Educational Technology - TOJET, 11*(4), 371–381.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, *24*, 602–611.

John, O. P., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339–369). Cambridge University Press.

Johnson, R. B., Meeker, K., & Onwuegbuzie, A. J. (2004, April). *Development and use of the Philosophical Beliefs Questionnaire*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26. https://doi.org/10.3102/0013189X033007014

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112–133. https://doi.org/10.1177/1558689806298224

Joint Advisory Committee (JAC). (1993). Principles for fair student assessment practices for education in Canada. Retrieved from https://www.wcdsb.ca/wp-content/uploads/sites/36/2017/03/fairstudent.pdf

Joint Committee on Standards for Education Evaluation (JCSEE). (2015). *Classroom assessment standards*. Retrieved from https://evaluationstandards.org/classroom/

Kalkbrenner, M. T. (2021). A practical guide to instrument development and score validation in the social sciences: The measure approach. *Practical Assessment, Research & Evaluation, 26*, 1–18. https://doi.org/10.7275/svg4-e671

Kaslow, N. J., Beneau, M. J., Lichtenberg, J. W., Portnoy, S. M., Rubin, N. J., Leigh, I. W., Nelson, P. D., & Smith, I. L. (2007). Guiding principles and recommendations for the assessment of competence. *Professional Psychology: Research and Practice, 38*(5), 441–451.

Keeves, J. P. (1997). *Educational research methodology and measurement*. Cambridge: Cambridge University Press.

Kivunja, C., & Kuyini, A. B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education, 6*(5), 26. https://doi.org/10.5430/ijhe.v6n5p26

Klinger, D. A., McDivitt, P. R., Howard, B. B., Munoz, M. A., Rogers, W. T., & Wylie, E. C. (2015). *The classroom assessment standards for PreK-12 teachers*. Kindle Direct Press.

Koskey, K. L. K., Sondergeld, T. A., Stewart, V. C., & Pugh, K. J. (2018). Applying the mixed methods instrument development and construct validation process: The transformative experience questionnaire. *Journal of Mixed Methods Research, 12*(1), 95–122. https://doi.org/10.1177/1558689816633310

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Likert, R. (1932). A technique for measurement of attitudes. *Archives of Psychology*, *140*, 5–55.

Lin, G.-Y., Tseng, T. H., Yeh, C.-H., Wang, Y.-M., Wang, Y.-Y., & Wang, Y.-S. (2022). Development and validation of an internet unethical behavior scale. *Library & Information Science Research, 44*(2), 101153. https://doi.org/10.1016/j.lisr.2022.101153

Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic inquiry.* Thousand Oaks: Sage.

Lindstrom, G., Taylor, L., Weleschuk, A. (2017). Guiding principles for assessment of student learning. *Taylor Institute for Teaching and Learning Guide Series*. Calgary, AB: Taylor Institute for Teaching and Learning at the University of Calgary.

Looney, A., Cumming, J., van Der Kleij, F., & Harris, K. (2018). Reconceptualising the role of teachers as assessors: Teacher assessment identity. *Assessment in Education : Principles, Policy & Practice, 25*(5), 442–467. https://doi.org/10.1080/0969594X.2016.1268090

Luth, R. W. (2010). Assessment and grading at the University of Alberta: Policies, practices, and possibilities. A report to the Provost and the University. Edmonton, AB.: University of Alberta.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research (New York), 35*(6), 382–386.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, *35*(2), 293–334. https://doi.org/10.2307/23044045

Mathews, B. P., & Shepherd, J. L. (2002). Dimensionality of Cook and Wall's (1980) British Organizational Commitment Scale revisited. *Journal of Occupational and Organizational Psychology*, *75*(3), 369–375. https://doi.org/10.1348/096317902320369767

McMillan, J. H. (2001). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research & Evaluation*, *7*(8). https://doi.org/10.7275/5kc4-jy05

Merritt, S. M. (2012). The Two-Factor Solution to Allen and Meyer's (1990) Affective Commitment Scale: Effects of Negatively Worded Items. *Journal of Business and Psychology*, *27*(4), 421–436. https://doi.org/10.1007/s10869-011-9252-3

Mertler, C.A. (2003). *Pre-service versus in-service teachers' assessment literacy: Does classroom experience make a difference*? In Annual meeting of the Mid-Western Educational Research Association, Columbus.

Mertler, C.A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education*, *33*(2), 76.

Mertler, C. A. (2005). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education*, *33*(2), 49–64.

Mertler, C.A., & Campbell, C. (2005). *Measuring teachers' knowledge & application of classroom assessment concepts: Development of the assessment literacy inventory.* In Annual meeting of the American Educational Research Association, Montreal.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Ministry of Education of P. R. China (MOE) (1999). *National professional standards for K-12 teachers*. Retrieved from http://www.moe.edu.cn/publicfiles/business/htmlfiles/moe/s6991/20121 2/xxgk_145603.html

Ministry of Education of P. R. China (MOE). (2001). *Curriculum reform guidelines for basic education*. Retrieved from http://www.moe.edu.cn/base/jckecheng/

Ministry of Education of P. R. China (MOE). (2010). *Outline of the national medium and long-term education reform and development plan (2010-2020).* Retrieved from http://www.v-gov.cn/newsview.asp?id=779

Ministry of Education of P. R. China (MOE) (2020). *Overall plan for deepening the reform of education evaluation in the new era*. Retrieved from http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202010/t20201013_494381.html

Mockler, N. (2011). Beyond 'what works': Understanding teacher identity as a practical and political tool. *Teachers and Teaching: Theory and Practice, 17*(5), 517–528. https://doi.org/10.1080/13540602.2011.602059

Mohamed, M., Aziz, M. S. A., & Ismail, K. (2019). The validation of assessment for learning audit instrument: A mixed methods approach. *3L, Language, Linguistics, Literature, 25*(4), 209–226. https://doi.org/10.17576/3L-2019-2504-13

Morgan, D. L. (2022). Paradigms in mixed methods. In J. H. Hitchcock & A. J. Onwuegbuzie (Eds.), *The Routledge handbook for advancing integration in mixed methods research* (pp. 97–112). Taylor & Francis Group.

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research (New York), 40*(2), 120–123. https://doi.org/10.1097/00006199-199103000-00014

National Board for Professional Teaching Standards (NBPTS). (2012). *What teachers should know and be able to do*. Arlington, VA. Retrieved from http://www.nbpts.org

Piascik, P., & Bird, E. (2008). Creating and sustaining a culture of assessment. *American Journal of Pharmaceutical Education, 72*(5), 97–97. https://doi.org/10.5688/aj720597

New Zealand Teachers Council. (2008). *Graduating teacher standards*. Retrieved from https://traintheteacher.me/graduating-teacher-standards-e-portfolio/new-zealand-graduating-teacher-standards/

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

O'Leary, Z. (2014). *The essential guide to doing your research project* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Oldfield, A., Broadfoot, P., Sutherland, R., & Timmis, S. (2012). *Assessment in a digital age: A research review*. University of Bristol.

Onwuegbuzie, A. J. (2003). Effect sizes in qualitative research: A prolegomenon. *Quality & Quantity: International Journal of Methodology, 37*, 393–409.

Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research, 4*(1), 56–78. https://doi.org/10.1177/1558689809355805

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity, 43*(2), 197–209. https://doi.org/10.1007/s11135-007-9112-4

Onwuegbuzie, A. J., Leech, N. L., & Collins, K. M. T. (2008). Interviewing the interpretive researcher: A method for addressing the crises of representation, legitimation, and praxis. *International Journal of Qualitative Methods, 7*(4), 1–17. https://doi.org/10.1177/160940690800700401

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education, 84*, 128–138. https://doi.org/10.1016/j.tate.2019.05.003

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement, Issues and Practice, 12*(4), 10–12. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Polit, D., & Beck, C. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497.

Polit, D. F., Beck, C. T., and Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459–467. https://doi.org/10.1002/nur.20199

Popham, J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice, 48*, 4–11. https://doi.org/10.1080/00405840802577536

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*, 265–273. https://doi.org/10.1080/08878730.2011.605048

Powell, H., Mihalas, S., Onwuegbuzie, A. J., Suldo, S., & Daley, C. E. (2008). Mixed methods research in school psychology: A mixed methods investigation of trends in the literature. *Psychology in the Schools, 45*, 291–309.

Rodrigues, I., Adachi, J., Beattie, K., & MacDermid, J. (2017). Development and validation of a new tool to measure the facilitators, barriers and preferences to exercise in people with osteoporosis. *BMC Musculoskeletal Disorders, 18*(1), 540.

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods, 25*(1), 6–14. https://doi.org/10.1177/1094428120968614

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309–327. https://doi.org/10.1177/0265532213480128

Schildkamp, K., & Lai, M. K. (2013). Introduction. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 1–7). Dordrecht, the Netherlands: Springer.

Schmidt, L. J., & DeSchryver, M. (2022). The role of digital application literacy in online assessment. *Journal of Educational Technology Systems, 50*(3), 356–378. https://doi.org/10.1177/00472395211052644

Scotland, J. (2012). Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and

critical research paradigms. *English Language Teaching (Toronto), 5*(9), 9–16. https://doi.org/10.5539/elt.v5n9p9

Shaw, S., Elston, J., & Abbott, S. (2004). Comparative analysis of health policy implementation. *Policy Studies, 25*(4), 259–266. https://doi.org/10.1080/0144287042000288451

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14. https://doi.org/10.3102/0013189X029007004

Shiyanbola, O. O., Rao, D., Bolt, D., Brown, C., Zhang, M., & Ward, E. (2021). Using an exploratory sequential mixed methods design to adapt an Illness Perception Questionnaire for African Americans with diabetes: The mixed data integration process. *Health Psychology & Behavioral Medicine, 9*(1), 796–817. https://doi.org/10.1080/21642850.2021.1976650

Shrout, P. E., & Lane, S. P. (2012). Reliability. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology (Vol. 1. Foundations, planning, measures, and psychometrics)* (pp. 643–660). American Psychological Association. https://doi.org/10.1037/13619-034

Smith, L. F., Hill, M. F., Cowie, B., & Gilmore, A. (2014). Preparing teachers to use the enabling power of assessment. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing assessment for quality learning: The enabling power of assessment* (pp. 303–323). Dordrecht, Netherlands: Springer.

Snelson, C. L. (2016). Qualitative and mixed methods social media research: A review of the literature. *International Journal of Qualitative Methods, 15*, 1–15. https://doi.org/10.1177/1609406915624574

Solís Salazar, M. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192–200. https://doi.org/10.7334/psicothema2014.266

Sreedharan, J. K., Rao, U. K., Al Ahmari, M., Kotian, S. M., & Mokshanatha, P. B. (2022). Validation of a structured questionnaire to assess the perception and satisfaction of respiratory therapy students toward career prospects and learning resources. *Canadian Journal of Respiratory Therapy: CJRT = Revue Canadienne de La Thérapie Respiratoire: RCTR*, *58*, 162–168. https://doi.org/10.29390/cjrt-2022-032

Stassen, M. L. A. (2012). Accountable for what? *Journal of Assessment and Institutional Effectiveness*, *2*(2), 137–142.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Erlbaum.

Stiggins, R. J. (1991a). Assessment literacy. *Phi Delta Kappan*, *72*, 534–539.

Stiggins, R. J. (1991b). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice*, *10*(1), 7–12.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238–245.

Stiggins, R. (2009). Assessment for learning in upper elementary grades. *Phi Delta Kappan*, *90*(6), 419–421. https://doi.org/10.1177/003172170909000608

Stiggins, R. J. (2017). *The perfect assessment system*. Alexandra, VA: ASCD.

Stiggins, R. J., & Duke, D. (2008). Effective instructional leadership requires assessment leadership. *Phi Delta Kappan*, *90*, 285–291. https://doi.org/10.1177/003172170809000410

Stobart, G. (2008). Testing times: The uses and abuses of assessment. New York: Routledge.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly*, *13*(2), 147–169. https://doi.org/10.2307/248922

Straub, D. W., Boudreau, M. -C., & Gefen, D. (2004). Validation guidelines for is positivist research. *Communications of the Association for Information Systems*, *13*, 24. https://doi.org/10.17705/1CAIS.01324

Sun, L., Tang, Y., & Zuo, W. (2020). Coronavirus pushes education online. *Nature Materials*, *19*(6), 687–687. https://doi.org/10.1038/s41563-020-0678-8

Taghipoorreyneh, M., & de Run, E. C. (2020). Using mixed methods research as a tool for developing an indigenous cultural values instrument in Malaysia. *Journal of Mixed Methods Research*, *14*(3), 403–424. https://doi.org/10.1177/1558689819857530

Tashakkori A & Teddlie C. (2003). *Handbook of mixed methods in social & behavioral research*, Sage, California.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing*, *30*(3), 403–412. https://doi.org/10.1177/0265532213480338

Teachers Council of Aotearoa New Zealand. (2019). *The code of professional responsibility and standards for the teaching profession*. Retrieved from https://teachingcouncil.nz/assets/Files/Code-and-Standards/Our-Code-Our-Standards-Nga-Tikanga-Matatika-Nga-Paerewa.pdf

Teddlie, C., & Johnson, R. B. (2009). Methodological thought since the 20th century. In C. Teddlie & A. Tashakkori (Eds.), *Foundations of mixed methods research: Integrating quantitative and qualitative techniques in the social and behavioral sciences* (pp. 62–82). Thousand Oaks, CA: SAGE.

Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 127–146). Macmillan Publishing Co, Inc.

Van der Veer, C., Ommundsen, R., Yakushko, O., Higler, L. E., Woelders, S., & Hagen, K. (2013). Psychometrically and qualitatively validating a cross-national cumulative measure of fear-based xenophobia. *Quality & Quantity*, *47*(3), 1429–1444. https://doi.org/10.1007/s11135-011-9599-6

Waltz, C., Strickland, O., & Lenz, E. (2005). Validity of measures. In: Waltz C., Stickland O., & Lenz E. (Eds.), *Measurement in nursing and health research (3rd ed.)* (pp. 154–189). New York: Springer Pub.

Wang, Z., & Brown, G. T. L. (2014). Hong Kong tertiary students' conceptions of assessment of academic ability. *Higher Education Research and Development*, *33*(5), 1063–1077. https://doi.org/10.1080/07294360.2014.890565

Weretecki, P., Greve, G., & Henseler, J. (2021). Experiential value in multi-actor service ecosystems: Scale development and its relation to inter-customer helping behavior. *Frontiers in Psychology*, *11*, 593390–593390. https://doi.org/10.3389/fpsyg.2020.593390

Williams, J. C. (2015). “Assessing without levels”: Preliminary research on assessment literacy in one primary school. *Educational Studies*, *41*(3), 341–346. https://doi.org/10.1080/03055698.2015.1007926

Willis, J. (2010). Assessment for learning as a participatory pedagogy. *Assessment Matters*, *2*, 65–84.

Willis, J., Adie, L., & Klenowski, V. (2013). Conceptualizing teachers' assessment literacies in an era of curriculum and assessment reform. *Australian Educational Researcher, 40*(2), 241–256. https://doi.org/10.1007/s13384-013-0089-9

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research, 25*(5), 508–518. https://doi.org/10.1177/0193945903252998

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

Xu, Y., & Brown, G. T. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment, 6*(1), 133–158.

Yin R. K. (2006). Mixed method research: Are the method genuinely integrated or merely parallel? *Research in the School, 13*(1), 41–47.

Younas, A., Rasheed, S. P., Zeb, H., & Inayat, S. (2020). Data integration using the building technique in mixed-methods instrument development: Methodological discussion. *Journal of Advanced Nursing*. https://doi.org/10.1111/jan.14415

Yusoff, M. S. B. (2019). ABC of response process validation and face validity index calculation. *Education in Medicine Journal, 11*(3), 55–61. https://doi.org/10.21315/eimj2019.11.3.6

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences, 4*(2), 165–178.

Zhang, Z., & Burry-stock, J.A. (1997). *Assessment practices inventory: a multivariate analysis of teachers' perceived assessment competency*. In Annual meeting of the American Educational Research Association, Chicago.

Zhou, Y. (2019). A mixed methods model of scale development and validation analysis. *Measurement (Mahwah, N.J.), 17*(1), 38–47. https://doi.org/10.1080/15366367.2018.1479088

Appendix A

Focus Group Plan
Date:

ID	Name	University	Department	Age	Gender	Years of teaching experiences
1				○ 30 to 40 ○ 41 to 50 ○ 51 to 60 ○ Over 60	○ Male ○ Female ○ Others	○ Less than 2 years ○ 2 to 5 years ○ More than 5 years
2				○ 30 to 40 ○ 41 to 50 ○ 51 to 60 ○ Over 60	○ Male ○ Female ○ Others	○ Less than 2 years ○ 2 to 5 years ○ More than 5 years
3				○ 30 to 40 ○ 41 to 50 ○ 51 to 60 ○ Over 60	○ Male ○ Female ○ Others	○ Less than 2 years ○ 2 to 5 years ○ More than 5 years

Focus Group Introduction

WELCOME
Thanks for agreeing to be part of the focus group. We appreciate your willingness to participate.
INTRODUCTIONS
Moderator; assistant moderator
PURPOSE OF FOCUS GROUPS
The focus group discussion is one part of the thesis of PhD candidate Qin Wang. The reason we are having these focus groups is to investigate the assessment literacy of university teachers. We need your input and want you to share your honest and open thoughts with us.
GROUND RULES
1. WE WANT YOU TO DO THE TALKING.
We would like everyone to participate.
I may call on you if I haven't heard from you in a while.
2. THERE ARE NO RIGHT OR WRONG ANSWERS
Every person’s experience and opinion are important.
Speak up whether you agree or disagree.
We want to hear a wide range of opinions.

3. WHAT IS SAID IN THIS ROOM STAYS HERE
We want folks to feel comfortable sharing when sensitive issues come up.
4. WE WILL BE TAPE RECORDING THE GROUP
We want to capture everything you have to say.
We don't identify anyone by name in our report. You will remain anonymous.

Questions

Engagement questions:
1. How do you know about assessment knowledge (e.g., teacher training, or personal experience)?
2. What assessment work do you usually do? What kind of assessment do you usually use?
Exploration Questions:
3. How do you usually prepare for an assessment?
4. How do you use the information from the assessment?
5. How do you feel about teacher-student relationships and the ethical aspects of assessment, such as fairness and cheating?
6. How do you feel about yourself when you work with others in assessment practices?
7. How do you feel about institutional support in assessment practices?
8. Do you use technology such as technological applications and online systems for assessment? If you do, how do you use them to assess student learning?
Exit question:
9. Is there anything else you would like to say about the assessment?

Appendix B

Framework Evaluation Survey

This is an evaluation survey to get information of how valid the survey framework is to develop an instrument to assess university teachers' assessment literacy in both classroom and digital settings. Please read through the framework carefully before answering the items below.

Part 1. Please check the extent to which you agree or disagree with each of the following statements about the framework.

Item	Evaluation questions	Ratings			
		Strongly agree	Agree	Disagree	Strongly disagree
1	The topics of the framework represent the constructs of assessment literacy.				
2	It provides a framework for measuring the conceptual knowledge dimension of teacher assessment literacy.				
3	It provides a framework for measuring the praxeological dimension of teacher assessment literacy.				
4	It provides a framework for measuring the socio-emotional dimension of teacher assessment literacy.				
5	It covers measuring teacher digital assessment literacy.				
6	The categories of the framework are well structured.				

Part 2. For the following questions, please describe your opinions about the framework.

1. For each item to which you responded "Strongly disagree" or "Disagree", please explain why you disagree and suggest how the framework might be improved.

2. What do you think may be missing from the content of the framework related to the constructs of teacher assessment literacy?
3. What parts of the framework may be extraneous or not as important for measuring the constructs of teacher assessment literacy?
4. Do you have any other suggestions for improving the survey framework? Please describe.

Appendix C

The Preliminary Instrument

Part I

You will be presented with 30 statements in this section. For each statement, please identify how knowledgeable you are (1=not knowledgeable at all; 2=slightly knowledgeable; 3=moderately knowledgeable; 4= very knowledgeable; 5=extremely knowledgeable).

Assessment Concepts, Purpose, Content, Methods

1. know the concepts of reliability and validity
2. differentiate assessment methods (e.g., formative assessment, summative assessment, diagonal assessment)
3. develop assessment methods (e.g., assigning projects, assignments, reports, presentations, etc.) based on clearly defined objectives
4. design tests that suit both low-achieving and high-achieving students
5. define rating scales and rubrics for an assessment
6. assess student content knowledge, skills, and development over time
7. use self-assessment
8. use peer assessment

Grading and Data Analysis

9. determine students' grades according to students' average performance
10. avoid bias (personal preferences) in grading
11. identify different factors to be considered when grading
12. teach, assess, and grade in correspondence to the main learning objectives
13. analyze both qualitative and quantitative learning evidence
14. know relevant statistical techniques

Interpreting Assessment Results

15. interpret measurement error
16. interpret what a particular score says about an individual's ability
17. determine if an assessment aligns with a local system of accreditation
18. determine if an assessment aligns with a local educational system

Communicating Assessment Results

19. communicate assessment results to students
20. provide written feedback to students
21. provide oral feedback to students
22. communicate assessment results to other stakeholders (e.g., colleagues, the department administrators)

Assessment Ethics

23. keep the assessment results of each student confidential
24. prevent students from cheating on tests
25. avoid teaching to the test when preparing students for tests

Developing Digital Assessment
26. use LMS to design tests (e.g., online quizzes, online projects).
27. assess student learning using digital tools (computerized practice tests, apps, discussion boards, blogs, etc.)
28. give students online feedback on assignments through LMS or other apps and websites.
29. vary digital assessment tools according to their effectiveness for classroom purposes
30. use online assessment data to plan future teaching

Part II

Below are 44 statements about assessment. Please indicate your level of agreement with each statement (1=strongly disagree; 2=disagree; 3=somewhat disagree; 4=somewhat agree; 5=agree; 6=strongly agree).

Assessment Preparation and Assessment Implementation
1. I can develop assessment criteria that meet the assessment aims.
2. The assessment criteria I have developed can be consistent with the relevant assessment policies and requirements of the department, school, and country.
3. When the assessment criteria do not match the actual assessment situation, I have confidence in modifying it.
4. When the assessment criteria do not match the actual assessment situation, I still use the previous assessment criteria.
5. I can choose different assessment strategies to assess students' learning outcomes.
6. I am willing to use e-learning platforms, apps, and other digital tools to assess students' learning.
7. I can collect effective formative assessment results through digital tools such as e-learning platforms or apps.
8. When I need to use online learning platforms, apps, or other digital tools to assess students' learning progress, I feel overwhelmed and want to give up.

Interpretation, Feedback and Adjustment
9. I can accurately interpret the assessment results.
10. Whether in person or online environments, I can provide timely feedback to students based on the assessment results.
11. I will provide feedback based on each student's own situation.
12. I can organize and document all the mutual feedback.
13. I can adjust teaching content and methods based on the assessment results.
14. When I find problems with the previous assessment methods, I can adjust the assessment strategies.
15. When I find that the content that needs to be adjusted is complex, I will give up the adjustment.
16. When I find that the content that needs to be adjusted is complex, I am confident in making corresponding adjustments.

Engaging with Colleagues

17. I am willing to discuss assessment information and results with other teachers.
18. I am willing to share assessment information online with other teachers, such as viewing my online test design.
19. I can obtain useful information from the assessment experience of others to improve my own assessment work.
20. When I find problems with the assessment, I am willing to collaborate with others to solve problems.

Self-regulation
21. I am willing to adopt peer assessment and self-assessment of students.
22. I can use digital tools to promote student self-assessment and reflection, such as computerized practice tests or writing reflection blogs.
23. I can use digital tools to promote students' understanding of peer assessment, such as online discussions or posting comments.
24. I am not very willing to use student self-assessment and peer assessment because I believe their assessments are not fair enough.

Working with Stakeholders

25. I can collaborate with colleagues to improve the assessment plan.
26. I can collaborate with platform staff to improve online assessment.
27. I can collaborate with enterprises or organizations to develop a student professional quality assessment system that meets job market demand.
28. I don't think there is a need to consider job market needs in planning assessment.

Mindset
29. I understand my identity as the subject of assessment and the responsibilities I bear for it.
30. I believe that I have taken on my due responsibility in the assessment work.
31. I believe that I have made the right decisions in the assessment work.
32. I view the assessment process as a valuable research process on teaching.

Ethical Management

33. When using digital tools (e.g., online tests) for assessment, I can try to avoid students taking shortcuts due to their proficiency in software technology or platform familiarity.
34. I can handle the issue of online cheating properly.
35. In order to cope with the pressure of teaching performance, I often teach to the test.
36. When some students' assessment performance is affected by outdated equipment or the internet, I will take appropriate measures to ensure fairness and equity in the assessment.

Engagement and Relationships
37. When designing assessments, I will provide students with multiple options to complete the assessment.
38. I can handle the relationship with students properly during the assessment process.
39. When designing assessments, I will consider the impact of assessment difficulty on subsequent student engagement in learning.

Emotional Management

40. When the assessment destroys students' learning enthusiasm, I will try to take action to solve this problem.
41. I can take action to alleviate students' exam anxiety.
42. I can pay attention to students' excessive emphasis on GPA and provide appropriate emotional support.
43. When students feel frustrated due to unfamiliarity with digital tools such as learning apps or platforms, I will assist them and provide emotional support.
44. I think it's normal for students to be under pressure and not to be taken seriously.

Part III: Open-ended Question

Below is an open-ended question about your assessment practices. We highly appreciate that you could share your experience and thoughts.

Question: Please briefly introduce all the assessment practices in your work.

Appendix D

The Final Instrument (English version)

Part I: Background information

1. My gender is:
   - Female
   - Male
   - Other
   - Prefer not to respond

2. My age is

3. My department is

4. How long have I been a professional educator in university?

5. The university I work for belongs to
   - 985 Project
   - 211 Project
   - Ordinary first tier universities
   - Ordinary second tier universities
   - Vocational and technical college
   - Prefer not to respond

6. I completed the following preparation in assessment:
   - 2+ full-semester courses in assessment
   - 1 semester course in assessment
   - A short or mini course in assessment
   - A module in assessment (e.g., one face-to-face workshop, one online session)
   - Teacher preparation program with assessment topics
   - No formal preparation in assessment
   - I don't remember

Part II

You will be presented with 28 statements in this section. For each statement, please identify how knowledgeable you are (1=not knowledgeable at all; 2=slightly knowledgeable; 3=moderately knowledgeable; 4= very knowledgeable; 5=extremely knowledgeable).

1. Understand the concepts of reliability and validity.
2. differentiate assessment methods (e.g., formative assessment, summative assessment, diagonal assessment)

3. develop assessment methods (e.g., assigning projects, assignments, reports, presentations, etc.) based on clearly defined objectives
4. design tests that suit both low-achieving and high-achieving students
5. define rating scales and rubrics for an assessment
6. assess student knowledge and skills over time
7. use self-assessment
8. use peer assessment
9. determine students' grades according to students' average performance
10. avoid personal preferences and stereotypes in grading
11. identify different factors to be considered when grading (e.g., participation, learning behavior)
12. teach, assess, and grade in correspondence to the main learning objectives
13. analyze qualitative assessment data
14. analyze quantitative assessment data
15. interpret measurement error
16. interpret what a particular score says about an individual's ability
17. determine if an assessment aligns with the requirements of the professional field
18. determine if an assessment aligns with the university's requirements
19. communicate assessment results via digital tools
20. provide written feedback to students
21. provide oral feedback to students
22. communicate assessment results to other stakeholders (e.g., colleagues, the department administrators)
23. keep the assessment results of each student confidential
24. prevent students from cheating on assessments
25. avoid teaching to the test when preparing students for tests
26. assess student learning using digital tools (computerized practice tests, apps, discussion boards, blogs, etc.)
27. vary digital assessment tools according to their effectiveness for classroom purposes
28. use digital assessment results to plan future teaching

Part III

Below are 46 statements about assessment. Please indicate your level of agreement with each statement (1=strongly disagree; 2=disagree; 3=somewhat disagree; 4=somewhat agree; 5=agree; 6=strongly agree).

1. I can develop assessment criteria that meet the assessment aims.
2. The assessment criteria I have developed can be consistent with the relevant assessment policies and requirements of the department, school, and country.
3. When the assessment criteria do not match the actual assessment situation, I have confidence in modifying it.
4. When the assessment criteria do not match the actual assessment situation, I still use the previous assessment criteria.
5. I can choose different assessment strategies to assess students' learning outcomes.

6. I am willing to use e-learning platforms, apps, and other digital tools to assess students' learning.
7. I can collect effective formative assessment results through digital tools such as e-learning platforms or apps.
8. I feel overwhelmed by using digital tools for assessment and refuse to use them.
9. I can accurately interpret the assessment results.
10. Whether in person or online environments, I can provide timely feedback to students based on the assessment results.
11. I will provide feedback based on each student's own situation.
12. I can organize and document all the mutual feedback.
13. When I find problems with the previous assessment methods, I can adjust the assessment strategies.
14. I can adjust teaching content or methods based on the assessment results.
15. When I find that the teaching content or methods that need to be adjusted are complex, I will give up.
16. When I find that the teaching content or methods that need to be adjusted are complex, I am confident in making corresponding adjustments.
17. I am willing to discuss assessment information and results with other teachers.
18. When necessary, I am willing to share digital assessment information with other teachers of the same course.
19. I can obtain useful information from the assessment experience of others to improve my own assessment work.
20. When there are problems with the assessment, I am willing to collaborate with others to solve problems.
21. I am willing to adopt student peer-assessment.
22. I am willing to adopt student self-assessment.
23. I can use digital tools to promote student self-assessment and reflection, such as computerized practice tests or writing reflection blogs.
24. I can use digital tools to promote students' understanding of peer assessment, such as online discussions or posting comments.
25. I am not very willing to use student self-assessment because I believe their assessments are not fair enough.
26. I am not very willing to use student peer-assessment because I believe their assessments are not fair enough.
27. I can collaborate with colleagues to improve the assessment plan.
28. I can collaborate with platform staff to improve online assessment.
29. I can collaborate with relevant institutions or organizations to improve assessment systems .
30. I don't think there is a need to collaborate with relevant institutions or organizations when improving assessment systems.
31. I believe that I have fulfilled my due responsibility in the assessment.
32. I always keep open-minded towards assessment.
33. I believe that I have applied critical thinking in making assessment decisions.
34. I view the assessment process as a valuable research process on teaching.
35. When using digital tools (e.g., online tests) for assessment, I can try to avoid students taking shortcuts due to their proficiency in software technology or platform familiarity.
36. I can handle the issue of online cheating properly (e.g., AI writing).

37. I try my best to avoid teaching to the test.
38. When some students' assessment performance is affected by emergent situations (e.g., sudden internet disconnection, illness), I will take appropriate measures to ensure fairness and equity in the assessment.
39. When designing assessments, I will provide students with multiple options to complete the assessment.
40. I can handle my relationship with students properly during the assessment process.
41. When designing assessments, I will consider the impact of assessment difficulty on student engagement in learning.
42. When the assessment destroys students' learning enthusiasm, I will try to take action to solve this problem.
43. I can take action to alleviate students' exam anxiety.
44. I can pay attention to students' excessive emphasis on GPA and provide appropriate emotional support.
45. When students feel frustrated due to unfamiliarity with digital tools to complete assessment, I will assist them and provide emotional support.
46. When providing assessment feedback to students, I can notice their negative emotions and take appropriate actions to improve them.

Part IV: Open-ended Question

Below is an open-ended question about your assessment practices. We highly appreciate that you could share your experience and thoughts.

Question: Please briefly introduce all your assessment work and the use of digital tools in assessment.

Appendix E

The Final Instrument (Chinese version)

高校教师测评素养

第一部分 基本信息

1. 我的性别是
   - 男
   - 女
   - 其他
   - 不回答

2. 我的年龄是

3. 我所属的院系是

4. 我在高校工作了多少年？

5. 我工作的院校属于是
   - 985 工程
   - 211 工程
   - 普通一本
   - 普通二本
   - 高职高专
   - 不回答

6. 我曾经学习了
   - 两个学期及以上的测评课程
   - 一个学期的测评课程
   - 一门短期测评课程
   - 一个有关测评的学习模块（例如一次研讨会，一次在线会议等）
   - 教师培训中穿插的有关测评的知识
   - 无测评相关培训
   - 我不记得了

第二部分

这一部分包含 28 条有关测评知识的项目。对于每一条项目，请选择您对该测评知识的掌握程度（1 = 根本不懂；2 = 略懂；3 = 懂一些；4 = 懂的比较多；5 = 懂的非常多）。

1. 理解信度和效度的概念。
2. 区分不同的测评策略（例如形成性考核，终结性考核等）。
3. 设计测评方法（例如项目，作业，报告，汇报等）。
4. 设计既适合差生又适合优等生的考试。
5. 制定评分量表和准则。
6. 评估学生的知识和技能的发展变化。
7. 使用学生自评。
8. 使用学生互评。
9. 根据学生的各项表现确定学生的成绩。
10. 在评分时避免个人偏好和刻板印象。
11. 确定评分时要考虑的不同因素（例如参与度、学习行为等）。
12. 根据主要学习目标进行教学、考核和评分。
13. 分析定性的测评数据。
14. 分析定量的测评数据。
15. 解读测量误差。
16. 解读学生分数所对应的个人能力。
17. 确定评估是否符合专业领域的要求。
18. 确定评估是否符合学校的要求。
19. 用数字工具交流测评结果。
20. 给学生提供书面反馈。
21. 给学生提供口头反馈。
22. 将测评结果汇报给相关负责人（例如同事，院系负责人等）。
23. 对每位学生的评估结果保密。
24. 防止学生在测评中使用作弊手段。
25. 避免应试教育。
26. 用数字工具测评学生（例如计算机化练习测试、应用程序、讨论板、博客等）。
27. 根据数字工具在课堂上的有效性来选择不同的工具完成测评。
28. 利用数字测评结果规划未来的教学。

第三部分

这一部分包含 46 条项目。请选择您对每条项目的同意程度（1=强烈不同意；2=不同意；3=比较不同意；4=比较同意；5=同意；6=强烈同意）。

1. 我能够制定出符合测评目的的测评标准。
2. 我制定的测评标准能够与系部，学校，国家的相关考核政策和要求相一致。
3. 当测评标准与实际测评情况不符时，我有信心修改好它。
4. 当测评标准与实际测评情况不符时，我选择继续使用之前的测评标准。
5. 我能够选择不同的测评策略去考核学生的学习成果。

6. 我愿意使用在线学习平台或 APP 等数字工具来评估学生的学习。
7. 我可以通过在线学习平台或 APP 等数字工具收集到有效的形成性考核结果。
8. 我对使用数字工具进行测评感到不知所措，并拒绝使用它们。
9. 我能够准确解读测评结果。
10. 无论是面对面还是网络环境，我都能够根据测评结果及时地给予学生反馈。
11. 我会根据学生自身情况给予相应的反馈。
12. 我能够整理并记录所有的师生反馈。
13. 当发现之前的测评方式有问题时，我能够调整测评策略。
14. 我能够根据测评结果调整教学内容或教学方法。
15. 当发现需要调整的教学内容或教学方法复杂时，我会放弃调整。
16. 当发现需要调整的教学内容或教学方法复杂时，我有信心去做出相应的调整。
17. 我愿意与其他教师共同探讨测评信息和测评结果。
18. 必要时，我愿意与负责同一课程的其他教师共享数字评估信息。
19. 我能够从他人的测评经验中获取有用的信息，用来完善自己的测评工作。
20. 当测评出现问题时，我愿意与其他教师合作解决问题。
21. 我愿意采用学生自评。
22. 我愿意采用学生互评。
23. 我能够使用数字工具去促进学生自评与反思（例如计算机化练习测试、反思博客等）。
24. 我能够使用数字工具去促进学生对互评的理解（例如在线讨论、发帖留言等）。
25. 我不太愿意使用学生自评，因为我认为他们的评价不够公正。
26. 我不太愿意使用学生互评，因为我认为他们的评价不够公正。
27. 我能够与同事合作完善测评方案。
28. 我能够与平台工作人员合作完善在线测评。
29. 我能够与相关单位或组织合作改进考核体系。
30. 我觉得改进考核体系时不需要与相关单位或组织合作。
31. 我认为我在测评工作中尽了应尽的责任。
32. 我一直以开放的心态对待测评。
33. 我认为我在测评工作中做决策时具有批判性思维。
34. 我将评价过程视为一个有价值的教学研究过程。
35. 在运用数字工具进行测评时，我能够尽量避免学生因熟练的软件技术或平台熟悉程度而投机取巧。
36. 我能够妥善处理网络作弊问题 （例如 AI 代写等）。
37. 我尽量避免应试教学。
38. 当某些学生的测评表现受紧急情况影响时（例如突然断网，疾病等），我会酌情处理以保证考试公平公正。
39. 在设计测评时，我会给学生提供多个选择去完成考核。
40. 我能够在测评过程中妥善处理好与学生之间的关系。
41. 在设计测评时，我会考虑测评难度对学生学习参与度的影响。
42. 当测评打消了学生的学习积极性时，我会尝试采取措施解决此问题。
43. 我会采取措施去缓解学生的考试焦虑。

44. 我会关注到学生过于重视 GPA 的情绪并适当安抚。
45. 当学生因为不知道如何用数字工具完成测评而感觉沮丧时，我会及时安抚并提供帮助。
46. 在向学生提供评估反馈时，我可以注意到他们的负面情绪，并采取适当的行动来改善。

第四部分

这一部分是一道关于您的评估经验的问题。我们非常感谢您能分享您的经验和想法。

1. 请详细描述您的测评工作以及在测评中的数字工具使用情况。