10-2018

Validating professional standards for teachers: A practical guide for research design: Snapshot literature review

Jen Jackson

Yung Nietschke

SNAPSHOT LITERATURE REVIEW

Validating Professional Standards for Teachers: A Practical Guide for Research Design

October 2018

The Australian Council *for* Educational Research
19 Prospect Hill Road
Camberwell VIC 3124
Phone: (03) 9277 5555
ABN 19 004 398 145

www.acer.org

ISBN 978-1-74286-536-2

Recommended citation

Jackson, J. & Nietschke, Y. (2018). *Validating professional standards for teachers: A practical guide for research design*. Camberwell, Australia: Australian Council for Educational Research.

Acknowledgements

This report was prepared by Dr Jen Jackson, Research Fellow, and Yung Nietschke, Associate Partner, at the Australian Council for Educational Research. The authors gratefully acknowledge the advisory support of Dr Lawrence Ingvarson, Ray Adams, and Dr Hilary Hollingsworth in conducting the review.

The original research on which this report is based was conducted as part of the Myanmar Education Quality Improvement Program (My-EQIP). My-EQIP is a joint initiative of the Government of Myanmar and the Government of Australia. The research on teacher professional standards was funded by the Australian Government through the Department of Foreign Affairs and Trade.

The original research was undertaken to support the development of the Myanmar Teacher Competency Standards Framework (TCSF), part of the Strengthening Pre-service Teacher Education in Myanmar (STEM) project. STEM receives technical assistance from UNESCO and financial contributions from the Governments of Australia, Finland and the United Kingdom.

The views expressed in this publication are the authors' alone, and are not necessarily the views of the Australian Government, the Government of Myanmar, UNESCO, or any other development partners.

Contents

Acronyms

AERA	American Educational Research Association
APST	Australian Professional Standards for Teachers
ASEAN	Association of Southeast Asian Nations
CVI	content validity index
CVR	content validity ratio
ERIC	Education Resources Information Center
IRT	item response theory
HLM	hierarchical linear modelling
My-EQIP	Myanmar Education Quality Improvement Program
NBPTS	National Board for Professional Teaching Standards
NTES	National Teacher Evaluation System
PACT	Performance Assessment for California Teachers
PCK	pedagogical content knowledge
SDG	Sustainable Development Goals
SITE II	Self-Assessment Instrument for Teacher Evaluation
SMEs	subject matter experts
TCSF	Teaching Competency Standards Framework
TTI	Teacher Training Institution

Executive summary

This report presents findings from a rapid review of international literature on the validation of professional standards for teachers. The review was originally undertaken to inform the design of a validation study for the draft Teacher Competency Standards Framework (TCSF) in Myanmar, but has been adapted in this report for a general international audience. Studies reviewed in this report were sourced primarily from the Education Resources Information Center (ERIC) database.

”Validity”" may be defined in many different ways. This report uses Messick’s (1989) six components of *construct validity* as a conceptual framework, for organising the findings of the review. The review explored how each type of validity has been demonstrated in previous validation studies:

- **Consequential validity** concerns the benefits of using teaching standards, relative to the risks. It is often demonstrated with stakeholder surveys or consultations, or documentation of impact.
- **Content validity** and **substantive validity** concern whether teaching standards describe quality teaching practice, as it is demonstrated in the classroom, and articulated in theory and research. Content validity is often demonstrated by review of standards by subject-matter experts (SMEs).
- **Structural validity** concerns whether the components of the standards show patterns in empirical data that are consistent with expected patterns, based on theories of effective teaching practice. It is often demonstrated using psychometric methods, including item response theory (IRT).
- **External validity** concerns whether teaching standards have a relationship to other measures that may demonstrate teacher effectiveness. Several studies explored the relationship between teaching practice and student learning outcomes, showing that the relationship varies widely.
- **Generalisability** concerns whether standards are equally applicable to different types of teachers, regardless of their characteristics and contexts. Most studies address this through representative sampling, with two pursuing deeper analysis of the applicability of standards across contexts.

This analysis generated eight recommendations for validation studies of teaching standards:

1. Begin by setting out a clear definition of validity, informed by international research.
2. Examine the impact and benefits of the standards, relative to actual or potential costs and risks.
3. Include a systematic subject-matter expert (SME) review of the content of the draft standards.
4. Consider options for the use of psychometric methods, recognising that the robustness of any chosen psychometric method will depend upon the availability of data about teaching practice.
5. Do not rely on the relationship with student learning outcomes data to demonstrate validity, but consider other relevant variables through which external validity may be demonstrated.
6. Give particular attention to the generalisability of standards, including across school settings, geographic areas, ethnic and linguistic groups, and diverse socio-economic status communities.
7. Design the study to establish a basis for reliable methods of teacher assessment against the standards, including piloting methods for teacher assessment with potential for scaling up.
8. Take into account policy considerations: consultation; long-term planning; cost-effectiveness; appropriateness to context; scope for ongoing improvement; and a multi-method approach.

1 Introduction

This report aims to establish a strong evidence base for planning a validation study of professional standards for teachers. It presents findings from a rapid "snapshot" review of relevant research literature, to identify previous examples of validation studies, and extract lessons from these about worthwhile methods and considerations in research design. This review was originally conducted to inform the design of a validation study of the draft Myanmar Teacher Competency Standards Framework, but may also have wider relevance to other education systems pursuing similar standards-based reforms.

This report presents findings from the literature review. It is divided into four sections:

1. **What does "validity" mean?**
   How validity has been defined in previous research on teaching standards.
2. **How can validity be demonstrated?**
   Methods that are commonly used to demonstrate each type of validity.
3. **Validity and reliability**
   Methods for ensuring reliability (alongside validity) in assessments of teaching practice.
4. **Policy considerations**
   Policy issues to be considered alongside methodological issues, in the study design.

The report also includes two recent detailed case studies, from Vietnam and the Philippines.

A key finding is that validation of teaching standards is a long and complex process, which may involve different methods. Some researchers describe validation as a "long and winding road", and a "politically and methodologically complex journey" (Taut, Santelices, & Stecher, 2012, p. 163). It is therefore worthwhile giving careful consideration to the types of validity that may need to be demonstrated at different stages in this journey, and the best methods for doing so.

1.1 Method

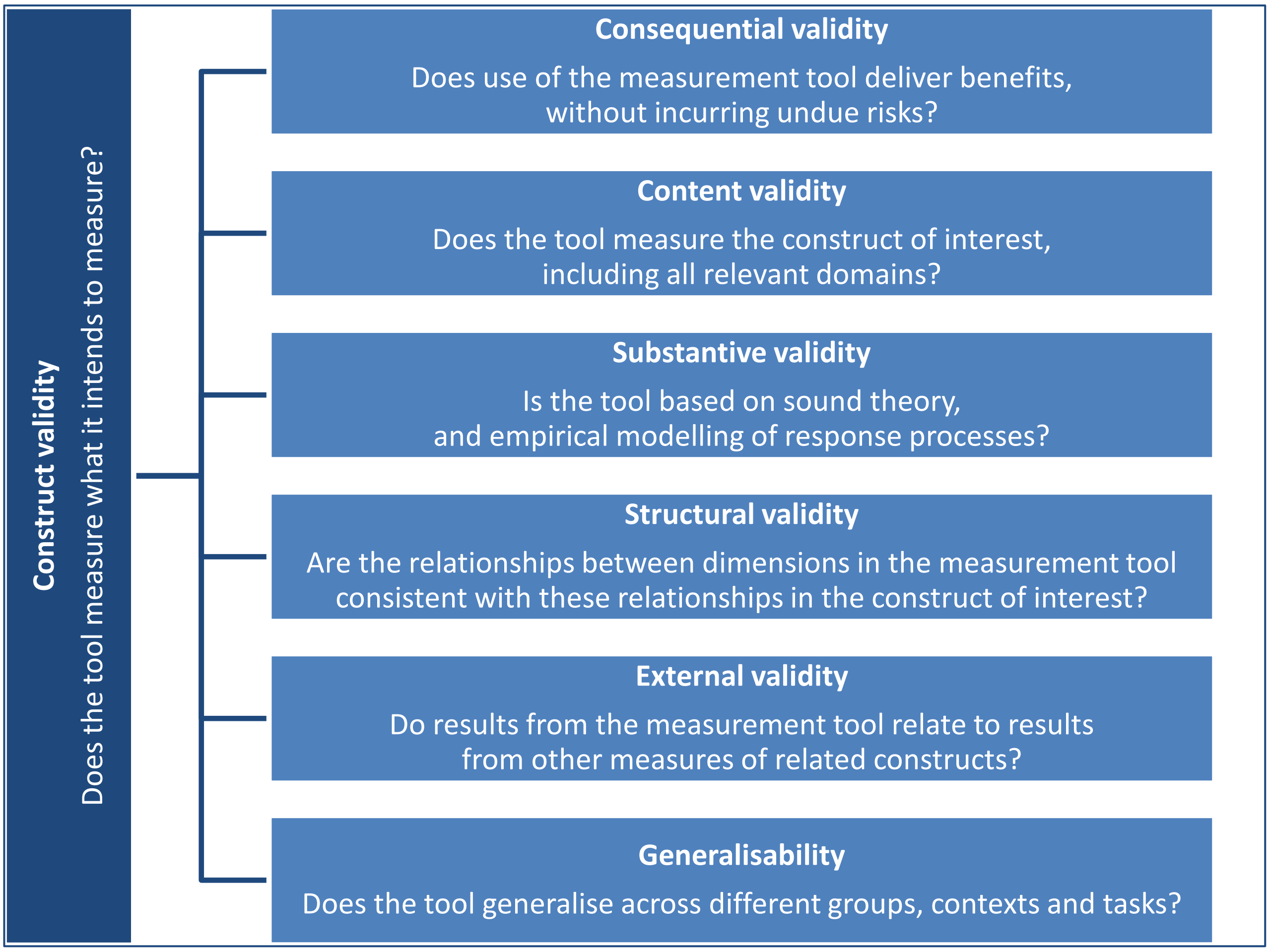
The literature search was conducted in the Education Resources Information Center (ERIC), a leading global database of education research, using the terms valid*, teach*, and standard*. Two further searches were undertaken in a leading academic library, and on the internet, to check for studies in other databases, and grey literature. A total of 27 studies were identified as relevant to the review. Some general literature on teaching standards was also reviewed, where it directly elaborated on issues raised in these studies; although this review did not involve a comprehensive search of general literature.

The 27 studies were analysed to determine: how validity was defined in each study; the methods used to demonstrate validity; how reliability was supported in any assessments of teaching practice; and any policy considerations identified. Findings of this analysis are presented in this report, under the four sections above. In each section, major headings indicate key concepts, followed by a brief explanation; sub-headings present key findings relevant to the design of validation studies for teacher professional standards; and dot points provide examples of studies supporting each finding. Each section concludes with a recommendation, to guide researchers in developing an effective validation study design.

2 What does "validity" mean?

Validity in research means that a measurement tool provides a true representation of what it claims to measure. This is often called *construct validity*, which is a broad term encompassing all other forms of validity (Newton, 2012). Construct validity concerns the relationship between a measure's content and what it is intended to measure. For example, if a tool is designed to measure "effective teaching", then evidence of its construct validity would demonstrate that it is measuring this construct comprehensively and fully, and not measuring something else.

This report adopts Messick's (1989, 1995) definition of construct validity, which has been widely used in research. This definition recognises that the validity of a measurement tool depends not only on its intrinsic attributes, but on its usefulness. Messick identified six components of construct validity that may be demonstrated for any measurement tool, illustrated in Figure 1.



Construct validity

Does the tool measure what it intends to measure?

Consequential validity

Does use of the measurement tool deliver benefits, without incurring undue risks?

Content validity

Does the tool measure the construct of interest, including all relevant domains?

Substantive validity

Is the tool based on sound theory, and empirical modelling of response processes?

Structural validity

Are the relationships between dimensions in the measurement tool consistent with these relationships in the construct of interest?

External validity

Do results from the measurement tool relate to results from other measures of related constructs?

Generalisability

Does the tool generalise across different groups, contexts and tasks?

Figure 1: Six components of construct validity (Messick, 1989)

These six components of validity provide a useful framework for thinking about how validity might be demonstrated in relation to teaching standards. They are also helpful for organising the findings of the literature review, to show how each type of validity has been demonstrated in research. The next section of this report is therefore organised according to these themes.

2.1 Other definitions of validity

Some studies in the literature review explored other general definitions of validity, and how they have been applied in validation studies of teaching standards. Key findings are shown below.

2.1.1 Few studies of teaching standards have used explicit frameworks to define validity

- Ingvarson (2002) noted that frameworks for defining the validity of personnel standards have seldom been used in the development of teaching standards. He cited the *Personnel Evaluation Standards* (PES) (Stufflebeam, 1988), as a way to define "reliable and valid measurement of educational personnel" (Ingvarson, 2002, p. 15). The PES involve four principles:

    - Utility (making evaluations more useful and more often used)
    - Feasibility (feasibly conducting evaluations, including in complex contexts)
    - Propriety (ensuring propriety in all aspects of the evaluation)
    - Accuracy (promoting accurate and dependable evaluation).

  The PES have been used in a range of international contexts; including developing contexts, although not in school education (for example, in university faculties: Ahmady et al., 2008).

- Goe, Holdheide, and Miller (2014) also reported that early efforts to establish teacher evaluation systems in the US were disparate and often "perfunctory" (p. 2). Their comprehensive guide to designing teacher evaluation system includes specific questions to consider, in determining a clear understanding of effective teaching, and establishing standards for teaching practice.

2.1.2 Most teaching standards have not demonstrated all possible kinds of validity

- Rothenberg and Hessling (1990) used the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999) to critique validation studies of teacher performance assessment systems in three US states. These standards are widely used across a range of educational and psychological research. They demand that validity is demonstrated through five sources of evidence (which approximately correspond to the six types of validity above): content, response processes, internal structure, relations to other variables, and testing consequences. The study found that teaching standards in all three states needed further evidence of validity, when the assessment instruments were used in different contexts.

- Milanowski, Heneman, and Kimball (2011) compared eight systems for standards-based teacher assessment, also in multiple US states. This included analysis of validity. They found that *content validity* studies were available for five of the eight systems, and *criterion (external) validity* studies were available for four, with some external validity studies still underway.

Recommendation 1 – That validation studies of teacher professional standards begin by setting out a clear definition of validity, informed by frameworks and definitions in international research.

3 How can validity be demonstrated?

This section examines methodological options for demonstrating the six kinds of validity identified above. It presents key findings in relation to each type of validity, supported by research examples.

3.1 *Consequential validity:* Do the standards achieve their desired purpose?

The first type of validity relates to the consequences of measurement or assessment, and relative risks and benefits involved. **Teaching standards with *consequential validity* achieve the desired effects on teaching and learning, without incurring undue risks or costs.**

Several studies in this review examined whether the introduction of teaching standards had achieved the desired effects on teaching and learning. The methods used in these studies typically included surveys, or contextual analysis of related policy documents. Key findings are presented below.

3.1.1 Demonstrating impact takes time, and involves complex interactions across the system

- **National Research Council (2008)**, in describing the development and validation of the National Board for Professional Teaching Standards (NBPTS) in the US, noted the time it takes for such reforms to achieve the desired impact on practice. They also noted the complex ways in which teaching standards interact with other dynamics within an education system; for example, some teachers concealed their achievement of the standards, to "not be seen as showing off" (p. 257). This demonstrates that pathways to impact may not be as direct as anticipated in policy.

3.1.2 Teaching standards may achieve impact simply by raising awareness of good teaching

- **Montecinos, Rittershaussen, Solis, Contreras, and Contreras (2010)** focused on the consequential validity of the Chilean Samples of Teaching Performance (STP) instrument. The methods used involved surveys and focus groups with teaching students assessed using the STP, and teacher educators. Almost all participants reported that the STP resulted in improvements to students' understanding of teaching practice, and expanded conversations about practice between students and teacher educators.

- **Australian Institute for Teaching and School Leadership (2016)**, in a survey of Australian teachers, found that not all teachers were using the Australian Professional Standards for Teachers (APST). Effective leadership, and sufficient time to engage with the standards, were critical factors in uptake. Pre-service teachers were most likely to be using the APST.

3.1.3 Teaching standards may fail to achieve impact due to issues in implementation

- **Tandon and Fukao (2015)**, in their overview of teacher quality in Cambodia, demonstrated the introduction of teaching standards had not achieved desired effects. Fewer than 10 per cent of lower secondary teacher education students, and fewer than one-quarter of primary teacher education students, were aware of the standards. Furthermore, only around 40 per cent of teacher trainers were aware of the standards. The study concluded that standards were not playing a central role in teacher preparation, as they were intended to do.

- **SEAMEO INNOTECH (2010)**, in Vietnam, found that even legislation of the national teaching standards has not resulted in their full adoption. Wide disparities in implementation of the standards were reported, particularly between urban and rural areas.
- **Maharaj (2014)** investigated the views of school administrators on the use of classroom observations in the Teacher Performance Appraisal process in Ontario, Canada. The study concluded that administrators generally did not see the assessment process as an adequate measure of teaching practice, and did not feel that it was effective in driving improvement.

3.1.4 High-stakes assessment can increase impact, but also increases the need for accuracy

- **Taut and Sun (2014)** report that the Chilean National Teacher Evaluation System (NTES) was made legally binding in 2005. Uptake of the standards has been high, because refusal to participate in the NTES assessments entailed negative consequences for the teachers.
- **Kimball and Milanowski (2009)** cautioned that higher stakes attached to teacher competency assessments (such as for promotion) increase the need for validity and reliability.
- **Maharaj (2014)** found that many Canadian school administrators felt that the reason teaching standards had not led to improvements in teacher practice, was because they were not used for high-stakes purposes, such as employment decisions.

3.1.5 The cost incurred by a teacher evaluation system must be appropriate to its purpose

- **Berliner (2018)** noted that tests of teacher competence with high levels of psychometric validity can be expensive to develop and implement, but acknowledges that this cost may be necessary if they are used for high-stakes decision-making (such as hiring decisions).
- **Taut and Sun (2014)** reported that the Chilean NTES has been criticised as being too expensive, although it is still less costly than assessment systems in the US.

Recommendation 2 – That validation studies of teacher professional standards consider consequential validity, by examining actual and potential benefits of the standards, relative to costs and risks.

3.2 *Content and substantive validity:* Do the standards measure the right things?

Content and substantive validity are grouped together in this discussion, because they both relate to whether a set of standards measure the right thing. **Teaching standards with *content* and *substantive validity* are an accurate representation of quality teaching practice, as it is demonstrated in the classroom, and articulated in theory and research.**

Content validity can be applied to a set of standards as a whole; to its composite domains; or the items within each domain (Sireci & Faulkner-Bond, 2014). Determining content validity in teacher performance measures can be complicated, due to potential overlap between domains; for example, between "effective teaching" and "classroom management" (Rothenberg & Hessling, 1990, p. 12).

Judgements about content and substantive validity are often based on a combination of theoretical notions about the constructs to be assessed, empirical research into teachers' everyday practices, and experts' judgements (Ingvarson & Hattie, 2008; Messick, 1995). In the studies identified for this review, analysis by subject matter experts (SMEs) was a common method used to demonstrate content validity. This method involves testing whether the domain definition underlying a measurement tool is aligned with the notion of the domain held by experts in the field. Experts are usually drawn from academia with experience in teaching or training teachers in a relevant field, or may be experienced practitioners or education administrators. The greater the level of consensus between the SMEs, the higher the content validity. Key findings and examples are presented below.

3.2.1 SMEs may include academics and practitioners, as well as end-users of the standards

- **Walkowiak, Berry, Meyer, Rimm-Kaufman, and Ottmar (2014)** used multiple methods to validate a set of standards for mathematics teachers. Content validity was demonstrated through SME review, with one university expert and one expert practitioner providing feedback on each dimension. The study also analysed response processes for coders of teaching practice, to confirm that the patterns of their coding were consistent, and matched to the coding guide.
- **Banerjee, Chopra, and DiPalma (2017)** used an expert panel to validate the content of standards for paraprofessionals in early intervention. The draft standards had been created by an expert panel of 11 sector leaders, who identified a further 49 experts to participate in the validation process, through a "snowball sampling" process.
- **Montecinos et al. (2010)** confirmed content validity for a set of standards for beginning teachers by using three expert panels. The panels included students who had undergone assessment.
- **Polit, Beck, and Owen (2007)**, in an informative study of content validity for nursing standards, noted that careful selection of SMEs at the beginning of the validation process is critical to rule out any bias, erratic behaviour or proficiency issues. They suggested using well-defined criteria, as proposed by Grant and Davis (1997). They also suggested that the first iteration of content validation requires a large panel of eight to twelve experts. This is consistent with a more recent review of the literature by Sireci and Faulkner-Bond (2014) on validity evidence, based on test content, which recommended the use of at least 10 SMEs in content validity studies.

3.2.2 Various statistical methods can be used to confirm the agreement between SMEs

- **Dally and Dempsey (2015)** used SMEs to develop a set of statements describing the skills and knowledge of special education teachers. The seven SMEs were either practitioners or academics with extensive experience in special education, or knowledge about the role of special education teachers. SMEs began by providing qualitative feedback on the draft standards, leading to initial revisions. SMEs then rated the relevance of each statement in the revised standards, based on a four-point Likert scale. Data was analysed using a content validity index (CVI) – a common statistical method for demonstrating alignment among SME ratings. The results showed that the statements met the criteria for content validity, although SME had various suggestions for further improvements to the standards outside of the rating process.
- **Akcamete, Kayhan, and Yildirim (2017)** examined the content validity of a scale of professional ethics for special education practitioners. In the study, 285 SMEs in special education rated the items on the scale as "appropriate", "not appropriate" or "should be changed". A content validity ratio (CVR)

was calculated based on the percentages of approval by experts on each scale item. Based on the CVR and SMEs comments on their recommendations, the scale was revised from 27 items to 33 items. The authors also used factor analysis to identify dimensions on the scale.

- **Banerjee, Chopra, and DiPalma (2017)** asked SMEs to rate each teaching standard against four criteria: (a) essential for safe and effective practice; (b) desirable, but not essential for safe and effective practice; (c) unnecessary for safe and effective practice; or (d) other. The analysis used simple percentages of how respondents rated each standard to determine their validity.

3.2.3 SME review may be most effective when it involves multiple rounds of rating/review

- **van der Schaff and Stokking (2011)** collected SME judgements about the content validity of teaching standards using a Delphi method. A group of 21 SMEs (including education academics, administrators, and lead practitioners) rated the standards on four criteria:
    - content relevance
    - thoroughness of formulation
    - clarity of formulation
    - correspondence with everyday teaching practice.

  After each rating, the standards were revised and rated again, resulting in three rounds of rating over three months. This method achieved a high level of consensus about the standards' validity. The criterion with least agreement was "correspondence with everyday teaching practice", perhaps reflecting the fact that experience of teaching practice varied across the SMEs.

3.2.4 SME review may include qualitative data, to provide deeper understanding

- **van der Schaff and Stokking (2011)** also asked SMEs to explain the reasons for their rating, to understand their underlying assumptions and preferences. While there was a high degree of statistical consensus across the ratings, the SMEs' underlying assumptions differed. The authors concluded that SMEs from different organisational contexts might bring different assumptions and preferences to the task of assessing content validity.

3.2.5 SME review must consider the relevance of standards to their cultural context

- **Vesamavibool, Urwongse, Hanpanich, Thongnoum, and Watcharin (2015)** compared the content of Thai teaching standards to standards for teachers in the ASEAN region. This included comparison of standards, and interviews with selected experts. They concluded that the Thai teaching standards required revision.

Recommendation 3 – That validation studies of teacher professional standards include a systematic subject-matter expert (SME) review, including a broad range of expertise in teaching practice.

3.3 *Structural validity:* Do components of the standards inter-relate as expected?

Structural validity concerns the consistency between the structure of a measurement tool as demonstrated in data, and the structure of the construct as anticipated by theory. **Teaching standards with *structural validity* show patterns in empirical data that are consistent with expected patterns,**

based on theories of effective teaching practice. For example, if one standard is expected to be harder to meet than another, this pattern should be visible in the data. Similarly, if a group of standards is expected to cluster together (for example, within a domain), then this pattern should also be evident in the relationships within the data.

Unlike content validity, this type of validity requires empirical data to be collected – it cannot be demonstrated by examining the standard alone. Demonstrating structural validity is also likely to require psychometric analysis, to identify the patterns in the data. Common psychometric methods include factor analysis, to check which items are clustered together; or item response theory (IRT) techniques, which is an advanced type of factor analysis commonly used in validation of psychometric tests. Key findings and relevant examples of studies are presented below.

3.3.1 Robust psychometric analysis of teaching practice requires time and resources

- **Griffin, Nguyen, and Gillis (2006)** validated the Primary School Teacher Standards in Vietnam using detailed psychometric analysis. The project took more than four years, using data from observational assessments of 2281 teachers. The analysis tested the whether the standards adequately discriminated between teachers based on the different levels of competency, and was used to identify descriptors of four levels of teaching practice, based on how the items fitted together.

3.3.2 Less costly (and less robust) psychometric analysis may use teacher surveys or tests

- **SiMERR (n.d.)** used psychometric methods in the Draft Design Clarification Study for the draft professional standards for teachers in the Philippines. The study analysed data from a survey of primary and secondary teachers, about the clarity of meaning and the perceived level of difficulty of the indicators. A Rasch (IRT) analysis showed that the structure of the standards was a good fit for the model; that is, that there was good alignment between increases in teachers' apparent ability, and their likelihood of achieving the standards that were rated more difficult.
- **Joscon and McPhan's (2015)** Philippines study used a content knowledge test to assess teachers' pedagogical content knowledge (PCK). The test assessed four dimensions of PCK: knowledge of specific content, aptitude for teaching the subject, and knowledge of tasks relating to the subject matter. When the test was piloted, psychometric analysis using IRT was used to identify items that did not fit the Rasch model, and which were therefore not good indicators of PCK. This study also formed the foundation for the subsequent validation study of the Philippine Developmental National Competency Based Teacher Standards.

3.3.3 Psychometric analysis can be applied to systemic data on teaching practice, if available

- **Duckor, Castellano, Téllez, Wihardini, and Wilson (2014)** provide an example of complex psychometric analysis being applied to the validation of the Performance Assessment for California Teachers (PACT). The study applied IRT modelling techniques to analyse a large body of data, which had been collected from regular system-wide assessments of teaching students in two Californian university systems. The IRT analysis found that the whole PACT instrument was a relatively good fit for the psychometric model, but that there were limitations and inconsistencies in the fit for the five domains. They concluded that further work is needed to validate PACT, using a wider variety of methods.

Recommendation 4 – That validation studies of teacher professional standards consider options for the use of psychometric methods to demonstrate structural validity, recognising that the robustness of any chosen psychometric method will depend upon the availability of data about teaching practice.

3.4 *External validity:* Do the standards relate to other relevant indicators?

External validity concerns the relationship between a measure, and other measures that are expected to be related. The relationship may occur either concurrently (at the same time), or predictively (over time). **Teaching standards with *external validity* have a proven relationship to other measures that may demonstrate teacher effectiveness.**

Measures of teacher competency may be expected to relate to measures of student learning. Several studies in this review used this method to demonstrate the validity of teaching standards, as shown in the examples below. However, this measure is controversial: Berliner (2018) reports that only a small amount of variance in student achievement on standardised tests can be attributed to the teacher, and they are therefore unreliable to use as evidence of good teaching. There are also differing views about how to measure student learning outcomes so they can be used reliably and validly to detect differences in teaching performance (Santelices & Taut, 2011). Milanowski et al. (2011) argue that student learning outcomes provide useful data for some purposes, but are not sufficient to validate teacher effectiveness by themselves.

3.4.1 The relationship between teacher assessments and student learning outcomes varies

- **Kimball and Milanowski (2009)** examined the relationship between school leaders' ratings of teacher performance, and measures of student learning. School leaders assessed teacher performance using a range of evidence, including teachers' self-assessment, pre-observation data (including a lesson plan), classroom and non-classroom observations, a reflection form, and instructional artefacts (such as records of student learning, contact with parents, or professional activities). The study found substantial variation in the relationship between school leaders' assessments of teachers, and teachers' impact on student learning.
- **Xu, Grant, and Ward (2016)** examined the relationship between teacher assessments on a state-wide teacher evaluation system in Virginia (US), compared to student learning outcomes. Teacher assessments used multiple forms of data, including classroom observations, documentation logs, and student surveys. They found that teacher assessments by external evaluators were only modestly related to student academic progress, especially assessments on standards related to planning, assessment, and professionalism. Assessments of teachers by principals were not related to student progress, suggesting a need for more principal training.
- **Wilson, Hallam, Pecheone, and Moss (2014)**, in Connecticut (US), investigated the relationship between teachers' portfolio assessment scores, a test of teacher knowledge, and changes in student reading achievement. Hierarchical linear modelling (HLM) analysis showed a statistically significant but moderate relationship between portfolio assessment and student learning outcomes, with

portfolio assessment being a better predictor of student learning outcomes than the test of teacher knowledge. They concluded that this supported the validity of the portfolio measure.

- **Waggoner and Carroll (2014)** compared student learning outcomes against standards-based teacher assessments in Oregon (US). They found the correlations differed for different types of assessments. They concluded that this confirms the need for multiple methods to assess the competency of beginning teachers, rather than a single, high-stakes assessment.
- **Akram and Zepeda (2015)** collected self-assessment surveys from 279 mathematics and English teachers based on five competency standards, in a study validating the Self-Assessment Instrument for Teacher Evaluation (SITE II) in Pakistan. The authors found a positive relationship between the teachers' self-assessment scores and their students' achievement.

3.4.2 Binary teacher assessments may relate most clearly to student learning

- **Bond, Smith and Baker (2000)** described a study design that planned to use classroom observations to determine whether teachers who had been certified against the standards demonstrated better practice than those who had not. The study also planned to analyse whether certification was related to differences in student learning, using samples of student work (recognising that standardised tests do not reliably demonstrate the impact of good teaching).
- **Santelices and Taut (2011)** examined the evidence of external validity for the Chilean NTES. The study collected in-depth teaching performance data on 58 teachers who were evaluated by NTES as either "outstanding" or "unsatisfactory", including: gains in student achievement scores, observation log data, expert ratings of teaching materials, and teachers' scores on a subject and pedagogical knowledge test. The study found that the NTES ratings for these two extreme groups did differ significantly on half the performance indicators. The other indicators also showed predicted (but non-significant) differences between the "outstanding" and "unsatisfactory" groups.

3.4.3 Use of student learning to measure teacher effectiveness may be politically sensitive

- **Taut and Sun (2014)** note that the Chilean NTES deliberately did not use student achievement data as an indicator of teaching performance, for political reasons. Despite this, the authors still found that teacher performance on the NTES was a significant predictor of student learning.

3.4.4 The relationship between teaching and learning outcomes may differ across subjects

- **Sorola (2014)** compared standards-based teacher portfolio assessments with student learning outcomes in Texas (US), and found differences in results between reading and mathematics.

3.4.5 Relationships between other teacher and student indicators may also be explored

- **Choi, Benson and Shudak (2016)** used observations of student behaviour to test the validity of an instrument measuring teacher dispositions – that is, to test the assumption that better teacher dispositions would relate to better student engagement. The study found that teachers' ability to engage students did not appear to be related to their dispositions, as assessed by the instrument.

Recommendation 5 – That validation studies of teacher professional standards do not rely on the relationship between teacher competency and student learning outcomes data to demonstrate the validity of the standards, but consider other variables through which external validity may be demonstrated.

3.5 *Generalisability:* Do the standards apply to all teachers, in all contexts?

Generalisability concerns the applicability of measurement tools across different groups and contexts. **Teaching standards with *generalisability* are equally applicable to different types of teachers, regardless of their characteristics and contexts.**

Generalisability may be seen as an important component of the fairness of standards. If standards are not generalisable, they may disadvantage particular groups of teachers, or teachers in certain types of schools. This is especially important where there is great diversity across the school system.

Most studies in this review addressed generalisability through representative sampling methods. Two studies also paid particular attention to generalisability in their data analysis, shown below.

3.5.1 Generalisability in teacher assessments can be analysed using psychometric methods

- **Griffin et al. (2006)**, in Vietnam, used psychometric methods to check whether any systematic bias was evident in the ratings (for example, whether teachers in a particular geographic location were assessed more harshly). This was because their results showed major differences in how teachers' competence had been rated, across Vietnamese provinces. The analysis confirmed that the differences were not due to systematic geographical bias in the assessments.

3.5.2 Generalisability may be informed by theories of diversity in teaching and learning

- **Ladson-Billings and Darling-Hammond (2000)** designed a study to examine whether assessments against the US NBPTS disadvantaged teachers in urban, minority communities, compared to teachers in other contexts. The study design was motivated by analysis of previous results from NBPTS assessments, which suggests that the assessments – comprising portfolio assessments tools and essay-writing exercises – systemically disadvantaged minority teachers. The study was also motivated by theories of specific pedagogical practices for minority groups, which the authors argued were not adequately reflected in the content of the NBPTS.

Recommendation 6 – That validation studies of teacher professional standards give particular attention to the generalisability of the standards, including across school settings, geographic areas, diverse ethnic and linguistic groups (of both teachers and students), and diverse socio-economic status communities.

4 Validity and reliability

The above discussion focused on the validity of teaching standards, and the various ways in which validity may be demonstrated. Any validation study that involves evaluation or assessment of teachers against standards must also consider the reliability of the evaluation.

Validity and reliability are similar but distinct concepts in research. Validity relates to measuring the right things in the right way, while reliability relates to the consistency and accuracy of measurements. Both reliability and validity are necessary to have consistent, accurate measures of the target construct. Reliability was an important consideration in all of the studies in this review that involved standards-based assessments or evaluations of teacher performance. The studies used a variety of methods for evaluating teachers against teaching standards, listed below:

- **Classroom observation** was the most common method of teacher assessment used in the studies. Observations of practice are often used to evaluate teachers, as they are assumed to be direct measures of teacher competency (Berliner, 2018). However, observations may be unreliable, as teacher performance in the classroom may not be stable over time (Berliner, 2018), and may be influenced by various contextual factors, such as the time of day observations are made, the unit being taught, student behaviour, or the teacher's personal circumstances. Classroom observations also may not capture all dimensions of teacher competency, as not all aspects of competency are directly observable in classroom practice (Milanowski et al., 2011).

- **Teaching portfolios** can also be used as a source of evidence to assess teacher competency. A typical portfolio includes goals and lesson plans for a predefined set of lessons, instructional artefacts, samples of student work, and a personal reflection on teaching practice. As such, portfolios are representative of teachers' best practice, rather than their typical performance. Taut and Sun (2014) found that teacher portfolios were the most "technically robust" of the four teacher assessment instruments used in the Chilean NTES.

- **Self-assessment** has been identified as a powerful tool for measuring teacher performance, particularly when used for the formative purpose of professional development (Akram & Zepeda, 2015; Taut & Sun, 2014). This is because the self-assessment process allows teachers to make judgements about the effectiveness of their knowledge, skills and performance to inform self-improvement. Self-assessments may encourage teachers to identify strengths and weaknesses, promote collegial interactions, and assist in school improvement (Akram & Zepeda, 2015).

- **Integration of multiple sources of evidence** was identified in several studies as the most reliable method for assessing teacher competency (Kane, 2006; Taut, Santelices, & Stecher, 2012; Walkowiak et al., 2013). For example, Griffin et al. (2006), in their major validation of teaching standards in Vietnam, used a combination of methods for measuring teacher performance. Assessors completed a questionnaire on the teacher's performance, and used other sources of evidence, including classroom observation, portfolio, and third-party opinions.

Each of these methods for assessing teachers requires reliable judgements to be made about the quality of teaching practice that is demonstrated. These judgements may be made by a variety of people, including external assessors, school leaders, or teachers themselves. Judgements are inherently value-laden, and it therefore essential to minimise the influence of assessors' personal assumptions and prejudices, to create a reliable assessment process (Messick, 1995; van der Schaff & Stokking, 2011). Key findings in relation to strategies for improving reliability are provided below.

4.1 Repeated or unannounced assessments can help capture "typical" teaching practice

- **Walkowiak et al. (2013)**, conducted three to five video-taped observations, which not only captured what teachers were doing but also focused on student's responses, group work and writing on the board. They then used statistical analysis to determine a "reliability coefficient" for the set of multiple observations for each teacher.

- **Maharaj (2014)** suggested that classroom observations should be unannounced, which would give assessors a more accurate picture of typical teacher practice (rather than peak performance if they have prepared for the observation), therefore allowing more meaningful assessment.

4.2 Evaluation by multiple assessors reduces the effects of assessors' subjectivity

- **Wilson et al. (2014)** used a process where two trained assessors evaluated teacher portfolios, and if significant differences were found, a third assessor would be required to reconcile the scores. The assessors were required to decide on one of four performance levels based on a scoring rubric, and provide evidence for each guiding question to arrive at a score. Scores were audited by an assessor trainer who provided further training for those who were deemed to drift off calibration. The level of inter-rater reliability was determined by the per cent of exact and adjacent scores, with scores plus or minus two points triggering a third independent evaluation.

- **Taut and Sun (2014)** report that, in the Chilean NTES, 20 per cent of teacher portfolios for each subject and grade level are selected randomly for double assessment. If the two rater scores differ significantly, the supervisor acts as a third rater to reconcile the differences. The study reported that rater errors were generally small (between 3 and 10 per cent), possibly due to revisions to improve the internal structure of the portfolio and the double scoring process.

4.3 Comprehensive training of assessors is essential to achieve reliability

- **Griffin et al. (2006)**, in Vietnam, described how all assessors were trained and then tested against the requirements they were testing, to demonstrate the knowledge and skills required to collect evidence from various sources. Training was an iterative process, with provincial officers regularly reviewing assessors' decision patterns, to identify those who needed further training.

- **Milanowski et al. (2011)**, in a review of eight teacher assessment systems currently used in the US, found that all systems included rigorous protocols for assessor training. These included demonstration of master level (Cincinnati TES, CLASS, PRAXIS III and TAP), refresher training (TAP) and re-rating by a second assessor (NBPTS and PACT).

- **Walkowiak et al. (2013)** provides a particularly strong example of assessor training, in validating a set of mathematics teaching standards. Coders participated in an extensive four phase training program. The preparation phase involved reading the literature, studying the coding guides and sample coding practices. The training and mastery phase involved coder meetings to determine codes for at least three videotaped classroom lessons. The calibration phase required coders to code a subset of ten lessons. All coders participated in a monthly drift test, in which they coded lessons independently then came together to discuss the codes, and verify convergence among coders and with master coders. The results showed that alignment was above 86 per cent for most dimensions of teaching, demonstrating strong evidence of reliability.

4.4 Choosing knowledgeable assessors can also improve reliability

- **Milanowski et al. (2011)** found that various teacher assessment systems in the US (NBPTS, PACT and Cincinnati TESS) required teacher evaluators to have relevant subject expertise, and to be matched to the subject and grade level of the teacher being assessed.

- **Taut, Sandelices, and Stecher (2012)**, in their review of various validation studies of the Chilean NTES, found that characteristics such as age, title, institution and job experience did not predict assessor quality. However, there was evidence that assessor's teaching experience and number of hours they work at schools were important factors.

- **Taut and Sun (2014)** note that Chilean law dictates that assessors of teacher portfolios should be in-service teachers with at least five years teaching experience, and knowledgeable in the subject area and grade level they are assessing. In any NTES assessment year, 450 raters are recruited to score around 15,000 portfolios over a four-week period.

4.4.1 Assessments that yield implausible results are unlikely to be reliable

- **Taut and Sun (2014)**, in their analysis of different methods used to assess teachers against the Chilean NTES, concluded that the current form of the NTES self-evaluation is not a reliable method for assessment, given that the mean score is very high (close to "outstanding"). They suggested that self-assessment is only useful for formative decision-making processes.

- **Ingvarson and Kleinhenz (2006)** report that implausibly high pass rates contributed to a lack of trust in the Performance Threshold teaching standards developed in England and Wales. Around 98 per cent of teachers who were assessed against the threshold were successful.

Recommendation 7 – That validation studies of teacher professional standards are designed to establish a basis for reliable methods of teacher assessment against the standards, drawing on international best practice. This may include piloting teacher assessment methods with potential for scaling up over time.

5 Policy considerations

The validation of teaching standards is an inherently political process. Questions of who sets the validation research agenda, how research results are communicated, and how any modifications to the standards will be implemented, are central to the discussion of validation research. Beyond the methodological considerations discussed in this report, there are a number of policy issues to take into account in designing a validation study for teacher professional standards:

- **Adopt a consultative approach.** The most successful examples of teaching standards in the literature have involved broad consultation throughout the development and validation process (Ingvarson & Kleinhenz, 2006). Two studies identified engagement with teachers' unions as especially critical to reaching consensus on standards. This also gave the assessment system greater legitimacy, and reduced resistance from teachers to participate (Taut & Sun, 2014; Maharaj, 2014). The example from the Philippines (see Case study 2), which involved a broad group of education stakeholders, also increased acceptance of the standards, and provided useful insights into how the standards could be implemented from the perspective of teachers. Consultation is also important in disseminating the results of validation, and in gaining agreement on any changes to the standards that may occur as the result of the validation process.

- **Successful validation studies of teaching standards require a long-term approach.** The Philippines and the Vietnam case studies (see above) describe processes occurring over a period of five years. Both involved extensive consultations in the initial phases of standards development. The validation phases employed psychometric methods to gather validity evidence, using large samples of informants. This long-term approach has provided evidence for the validity of the standards, with broad support from teachers and education stakeholders.

- **Successful validation studies of teaching standards require investment.** The literature notes that methods used in rigorous validation studies can be costly. Therefore, consideration should be given to appropriate resourcing. In choosing a method for validating standards, cost-effectiveness is an important consideration. The expense involved in a detailed psychometric study may not always be justified, if the top priority is to demonstrate the impact of teaching standards on practice. Conversely, a low-cost study is unlikely to be worthwhile, if the goal is to demonstrate validity to the highest possible standards of research. Choice of method therefore depends both upon the resources available, and the goals of the validation process.

- **Context should be considered in all phases of standards development and validation**, from the beginning of the design process to the implementation phase. The literature notes that the process of validation and evaluation is necessarily value-laden, because informants are asked to make judgements about the meaning of teaching constructs. While international experience informed the development of the teaching standards in Vietnam, the process of validation resulted in important refinements to the initial set of standards to reflect the current regulation concerning teacher rankings and local understanding of the requirements of teaching practice (Griffin et al., 2006). The methods used in validation studies, such as classroom observations, are also subject to contextual issues, which should be considered in validation study design.

- **Maintain scope for ongoing improvement.** Teaching is a complex practice, and no set of standards can describe it perfectly. Expectations for teachers may also change over time, as pedagogical practices shift, and new methods arise through research and innovation. Standards must remain consistent enough to provide a trustworthy foundation for teachers' practice, but flexible enough for further revision. This requires a balance between enshrining standards in policy or legislation, and remaining open to ongoing improvement (see Taut & Sun, 2014).
- **Adopt a multi-method approach to demonstrating validity.** As shown in this report, the validity of teaching standards may be demonstrated in many ways. The use of multiple methods for validation not only strengthens understanding of teaching practice, but provides opportunities to approach the issue of validation in different ways. This may be especially valuable for engaging diverse stakeholders with diverse expectations for the validation process, and for maintaining dialogue about how teaching standards can be strengthened.

Recommendation 8 – That validation studies of teacher professional standards take into account policy considerations identified in this report: consultation; long-term planning; cost-effectiveness; appropriateness to context; scope for ongoing improvement; and a multi-method approach.

6 Conclusion

This report summarises the international evidence base for the design of validation studies of teacher professional standards. While the literature does not identify any single best method for validating teaching standards, it does provide a rich array of possibilities for validation. This report has also aimed to clarify the key concepts involved in the validation of standards, and in the development of reliable measures for teacher assessment or appraisal. It is hoped that this will provide a strong research foundation, for policymakers and researchers to work with education stakeholders to develop rigorous and practical validation study designs.

A significant theme to emerge in the literature is that validation is both a political and methodological process. While the level of precision in teaching standards is important, it is equally (if not more) important that validation confirms that there is a high level of ownership of standards by the teaching profession (Call, 2018). This must include consideration of the different contexts in which the standards will apply, and the need for all teachers – regardless of the schools in which they teach, and the communities that they serve – to feel that the standards are a fair, accurate and useful representation of their work. These considerations will be priorities for the any validation study of professional standards for the teaching workforce.

References

Ahmady, S., Changiz, T., Brommels, M., Thor, J., Gaffney, F. A., & Masiello, I. (2009). Contextual adaptation of the Personnel Evaluation Standards for assessing faculty evaluation systems in developing countries: The case of Iran. *BMC Medical Education, 9*(1).

Akram, M., & Zepeda, S. J. (2015). Development and validation of a teacher self-assessment instrument. *Journal of Research & Reflections in Education (JRRE), 9*(2), 134–148.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Australian Institute for Teaching and School Leadership (2016). Final report – evaluation of the Australian professional standards for teachers. https://www.aitsl.edu.au/docs/default-source/default-document-library/final-report-of-the-evaluation-of-the-apst.pdf

Banerjee, R., Chopra, R. V., & DiPalma, G. (2017). Early intervention paraprofessional standards: development and field validation. *Journal of Early Intervention, 39*(4), 359–370. https://doi.org/10.1177/1053815117727114

Berliner, D. C. (2018). Between Scylla and Charybdis: Reflections on and problems associated with the evaluation of teachers in an era of metrification. *Education Policy Analysis Archives, 26*(54).

Bond, L., Baker, W. K., & Smith, T. (2000). *Preliminary analysis report [microform]: construct validity study of the National Board for Professional Teaching Standards.* Washington, DC: National Partnership for Excellence and Accountability in Teaching.

Call, K. (2018). Professional teaching standards: A comparative analysis of their history, implementation and efficacy. *Australian Journal of Teacher Education, 43*(3), 93–108.

Choi, H., Benson, N. F., & Shudak, N. J. (2016). Assessment of teacher candidate dispositions: Evidence of reliability and validity. *Teacher Education Quarterly, 43*(3), 71–89.

Dally, K. A., & Dempsey, I. (2015). Content validation of statements describing the essential work of Australian special education teachers. *Australian Journal of Teacher Education, 40*(2).

Duckor, B., Castellano, K. E., Téllez, K., Wihardini, D., & Wilson, M. (2014). Examining the internal structure evidence for the performance assessment for California teachers: A validation study of the elementary literacy teaching event for Tier I teacher licensure. *Journal of Teacher Education, 65*(5), 402–420.

Goe, L., Holdheide, L., Miller, T. (2014). *Practical guide to designing comprehensive teacher evaluation systems: A tool to assist in the development of teacher evaluation systems.* Revised edition. https://files.eric.ed.gov/fulltext/ED555655.pdf

Griffin, P., Nguyen, T. K. C., & Gillis, S. (2006). *Developing and validating primary school teacher standards in Vietnam.* Paper presented at Australian Association for Research in Education Conference, Melbourne. www.aare.edu.au/data/publications/2004/ngu04603.pdf

Ingvarson, L. (2002). *Development of a national standards framework for the teaching profession*. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1007&context=teaching_standards

Ingvarson, L., & Hattie, J. (2008). *Assessing teachers for professional certification, Vol. 9: The first decade of the National Board for Professional Teaching Standards.* Bingley: Emerald Group Publishing.

Ingvarson, L. & Kleinhenz, E. (2006). *Standards for advanced teaching: A review of national and international developments.* https://research.acer.edu.au/teaching_standards/2/

Jocson, J. & McPhan, G. (2015). Contexts and processes for the development of content tests to assess teachers' pedagogical content knowledge. The Online Journal of Quality in Higher Education, 2(4).Akcamete, G., Kayhan, N., & Yildirim, A. E. S. (2017). Scale of professional ethics for individuals working in the field of special education: Validity and Reliability Study. *Cypriot Journal of Educational Sciences, 12*(4), 202–217.

Kane, M. T. (2006). Educational measurement. In R. L. Brennan (Ed.), Validation (4th ed., pp. 17–64). Westport: American Council on Education/Praeger.

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45*(1), 34–70.

Ladson-Billings, G., & Darling-Hammond, L. (2000). The validity of National Board for Professional Teaching Standards (NBPTS)/Interstate New Teacher Assessment and Support Consortium (INTASC) assessments for effective urban teachers [microform] : findings and implications for assessments. Washington, DC]: NPEAT, University of Maryland, College of Education.

Maharaj, S. (2014). Administrators' views on teacher evaluation: Examining Ontario's teacher performance appraisal. *Canadian Journal of Educational Administration and Policy, 152.*

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. *American Psychologist, 50*, 741- 749.

Milanowski, A. T., Heneman, H. G., III, Kimball, S. M. (2011). *Teaching assessment for teacher human capital management: Learning from the current state of the art.* WCER Working Paper No. 2011-2. Wisconsin Center for Education Research.

Montecinos, C., Rittershaussen, S., Solis, M. C., Contreras, I., & Contreras, C. (2010). Standards-based performance assessment for the evaluation of student teachers: A consequential validity study. *Asia-Pacific Journal of Teacher Education, 38*(4), 285–300.

National Research Council (2008). *Assessing accomplished teaching: Advanced-level certification programs.* Washington, D.C.: The National Academies Press.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research & Perspectives, 10*(1-2), 1-29. DOI: 10.1080/15366367.2012.669666.

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

Rothenberg, L., & Hessling, P. A. (1990). *Applying the APA/AERA/NCME "Standards": Evidence for the validity and reliability of three statewide teaching assessment instruments.* https://files.eric.ed.gov/fulltext/ED318778.pdf

Santelices, M. V., & Taut, S. (2011). Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system. *Assessment in Education: Principles, Policy & Practice, 18*(1), 73–93.

SEAMEO INNOTECH (2010). *Teaching competency standards in Southeast Asian countries: Eleven country audit.* www.seameo-innotech.org/wp-content/uploads/2014/01/SIREP1%20-%20Teaching%20Standard%20FINAL.pdf

SiMERR (n.d.). *The validation of the Philippine Professional Standards for Teachers in the Philippines.* Manila: Research Center for Teacher Quality.

Sireci, S. & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*, 100–107.

Sorola, A. J. (2014). Validity of a standards-based teacher evaluation system. https://repositories.lib.utexas.edu/handle/2152/28492

Stufflebeam, D.L. (1988). *The Personnel Evaluation Standards: How to assess systems for evaluating educators.* Newbury Park, CA: Corwin Press.

Tandon, P., & Fukao, T. (2015). *Educating the next generation: Improving teacher quality in Cambodia.* Directions in Development. Retrieved from https://eric.ed.gov/?id=ED555622

Taut, S., Santelices, M. V., & Stecher, B. (2012). Validation of a national teacher assessment and improvement system. *Educational Assessment, 17*(4), 163–199.

Taut, S., & Sun, Y. (2014). The Development and Implementation of a national, standards-based, multi-method teacher performance assessment system in Chile. *Education Policy Analysis Archives, 22*(71).

van der Schaaf, M. F., & Stokking, K. M. (2011). Construct validation of content standards for teaching. Scandinavian Journal of Educational Research, 55(3), 273–289.

Vesamavibool, S., Urwongse, S., Hanpanich, B., Thongnoum, D., & Watcharin, K. (2015). The comparative study of professional standards for Thai Teachers and for ASEAN teachers. *Procedia - Social and Behavioral Sciences, 191*, 2280–2284.

Waggoner, J. & Carroll. J. B. (2014). Concurrent validity of standards-based assessments of teacher candidate readiness for licensure. *SAGE Open,* *4*(4). https://doi.org/10.1177/2158244014560545

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E. & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics, 85*(1), 109.

Wilson, M., Hallam, P. J., Pecheone, R., & Moss, P. A. (2014). Evaluating the validity of portfolio assessments for licensure decisions. *Education Policy Analysis Archives, 22*(6).

Xu, X., Grant, L. W., & Ward, T. J. (2016). Validation of a statewide teacher evaluation system: Relationship between scores from evaluation and student academic progress. *NASSP Bulletin, 100*(4), 203–222.