Paper Code NGU06308

Evaluating Psychometric Properties of the Primary School Teacher Standards: Implications for Instrument Development and Assessment Practice.

Patrick Griffin
Nguyen Thi Kim Cuc
Shelley Gillis
Wally Fung

Assessment Research Centre
The University of Melbourne

Evaluating Psychometric Properties of the Primary School Teacher Standards: Implications for Instrument Development and Assessment Practice.

Abstract

This paper addresses the World Bank funded study in which the University of Melbourne was commissioned to develop and validate a set of competency profiles and assessment strategies for primary school teachers in Vietnam. Through extensive expert consultation, 64 criteria which belong to 14 Requirements were developed within three broad strands: 'Personality and Ideology', 'Knowledge' and 'Pedagogical Skills'. In 2003, the draft profiles and accompanying assessment procedures were trialed (Griffin, Nguyen, Gillis, and Mai, 2006). In 2004, the validated profiles and assessment procedures were used to assess a further 25,000 teachers in 10 provinces in Vietnam. This paper reports on the findings from the final stage of validating the teacher profiles with 25,000 teachers. The paper details the process of checking if the 14 Requirements and their quality levels which were developed after the trial with 2180 teachers were valid and reliable instruments to measure the three strands of the teacher profiles. Analysis was also conducted to explore if there is any Requirement that exhibits differential item functioning due to assessment practice of different provinces. The findings are to identify further refinement in the profiles and ways to improve assessment practice for future roll-out.

Introduction

This article presents an account of the development of competency standards and profiles for primary teachers in Vietnam. The project has taken more than four years and has used a combination of consultative, actuarial and item response modelling procedures to develop and validate a scale of teacher competence. In the overall project more than 27000 teachers have been assessed, over 1000 assessors trained, a set of teacher professional requirements has become available and a data management system has been trialled for the Vietnamese government. After reviewing the international literature on teacher standards and competencies in which this study of Vietnam teacher standards is grounded, this article reports on the findings from an initial study in which 2281 teachers were assessed in 10 provinces in Vietnam. The major aim of the study was to empirically validate and refine the standards for primary teachers in Vietnam as well as determine the most appropriate way in which evidence could be gathered and scored for future roll-out.

Background

According to Shaw (2004), economic development has created a demand for literate, trained populations and its advance has aroused a consciousness in parents that their children must be literate and skilled if they are to enjoy some of the benefits of the increased wealth being generated. Governments around the world have committed to a broader industrial base and are trying to address the issues arising from the resultant demand for a literate and highly trained population. In line with the declarations of the UNESCO/UNICEF conference in DAKAR 2000, there has emerged an imperative for education for all (EFA) and the implementation of universal education. The three goals of education established at the conference (i.e., equity, access and quality) have been difficult to implement as coexisting properties in developing systems. Access for all has tended to be linked to differential quality and equal opportunity and resourcing tends to be beyond developing economies.

As countries develop they have been able to give more attention to the precise nature of their schools' curriculum and to the quality of the teaching delivered in the realisation of that curriculum. Pre-service training programs have been progressively extended in duration. Inspection and reporting systems have been established for assessing the capability and performance of practising teachers, in part to identify areas where further in-service training has been required, but also to identify those teachers most able to take on supervisory or leadership responsibilities.

However, the sheer size of the required teaching 'force' and public costs associated with its provision have remained as important factors throughout this development. Increasingly, attention has focused on how the quality of both pre-service and in-service teacher training and teachers' in-school performance might be improved. From time to time, even in countries with mature economies and fully developed systems of universal schooling, moments of heightened concern have arisen over the overall costs of schooling. The systems have been challenged to do better with the resources they have. Ideas have been explored and strategies sought to provide a more clearly directed application of the resources and energies dedicated to teacher training and improvement. Governments, education administrators, school leaders and teachers looked for ways in which teacher development might be more explicitly 'tracked' so that those responsible for it could plan and map its progress and teachers could more readily demonstrate their attainment of knowledge, skill and other aspects of capability.

Increasingly, governments are moving from an input mode of financing education to emphasise throughput or process, output and outcomes approaches. However an outcome focus approach still tends to emphasise student achievement rather than the end result of schooling and lifelong learning. As part of the throughput or process, teacher qualifications and competencies are increasingly being examined and measured. Minimal threshold levels of standards are being established and teachers are increasingly being expected to demonstrate these levels. Professional development of teachers is central to the reforms in the UK, the USA and Australia, for instance; and governments are shifting their funding base from one of inputs required, to one based on the demonstration of improved performance and competencies demonstrated. This in turn shifts to the notion of improved performance of teachers being linked to improved performance of students. The implications are that student learning will become a central theme of funding models and this is itself linked to improved

teacher and teaching competencies. However, outcomes defined as student performances have been clearly shown to be flawed.

Most notably, this has been a first in the development of teacher standards. While the format of the standards is similar to those used in the United Kingdom, their content is quite different. Moreover, while the record system is similar to those reported in the Denver Public Schools (2005) system, this study has illustrated how it is feasible to develop the standards empirically.

The Knowledge Base and Competency-based Schemes for Teaching

Attempts to define, organise and adequately describe the knowledge base of teaching have been numerous. Shulman (1987) described a framework that has become something of a benchmark in the on-going quest for an appropriate set of categories. It can be summarised as follows:

- content knowledge;
- general pedagogical knowledge including principles and strategies for classroom management and organisation;
- curriculum knowledge including materials and programs used as the 'tools of trade';
- pedagogical content knowledge - an amalgamation of content and pedagogy that is a teacher's special form of professional understanding;
- knowledge of the learners and their characteristics;
- knowledge of educational contexts, including the characteristics of classrooms, schools, communities and cultures; and
- knowledge of educational ends, purposes and values, and their philosophical and historical grounds.

Delineation of categories within the knowledge base is seen as a starting point for building a broad and comprehensive competency-based scheme. It not only dis-aggregates the body of knowledge which teachers possess and build up in the progression from trainee to experienced practitioner, but it identifies the information and understandings that teachers draw upon when they engage in the many strategic thinking processes and actions which their practice requires.

In more recent applications of competency-based ideas to teaching, the construction of schemes for planning and assessing teacher development begins with comprehensive developmental maps of the knowledge, understandings and appreciations considered by a range of stakeholders to be necessary for successful teaching performance (Griffin, Poynter, Nguyen, Ry, Thiep and Nguyen, 2001). They identify the required capacities for action and skills that flow from the intellectual interpretation or 'reading' of teaching tasks and which transform aspects of knowledge into teaching action. In addition, schemes may identify values and commitments that a teacher must have or take up, and they may also include developing capabilities that a teacher is expected to build with experience.

Broad areas of qualities such as these (knowledge/understandings/appreciations; capacities and skills; values and commitments; developing capabilities) provide a more elaborate framework of strands or dimensions for a scheme. Within a strand (for example: pedagogical knowledge and skills) a number of descriptors or statements is used to detail the qualities or competencies that make up the strand (for example: capacity to develop positive attitudes towards learning; skill in providing opportunities for cooperative learning etc).

International Competency-based Schemes of Teaching Standards

Teaching standards are necessarily culturally-based. This can be seen by investigating developments in the United States, the United Kingdom and in Australia where the purpose and accountability links of teacher standards differ as presented in Figure 1.

	USA		Australia	United Kingdom	
Key Characteristic	Denver Public Schools (2005)	Danielson's (1996) Framework for Teaching	Australian Teaching Council (1996)	Teacher Training Agency (1996)	Scottish Office Education Dept (1993)
Instruction	Instruction	Instructional planning			
Assessment	Assessment		Monitoring and assessing student progress and learning outcomes	Monitoring, assessment, recording, reporting and accountability	
Planning	Curriculum and Planning		Planning and managing the teaching and learning processes		
Environment	Learning Environment				School related competencies
Professionalism	Professional responsibilities	Professional responsibilities (ideology and philosophy)	Using and developing professional knowledge and values		Attitudes and commitments
Pedagogy		Instructional interactions (pedagogy)			
Classroom Management		Classroom management		Planning, teaching and classroom management	Classroom (communication, methodology, classroom management and assessment)
Content knowledge				Subject knowledge and understanding	Subject and the content of teaching
Reflection			Reflecting, evaluating and planning for continuous improvement.		

Figure 1. A comparison of major standards implemented in the USA, UK and Australia according to key characteristics.

It can be seen in Figure 1 that whilst there are a number of common characteristics across a number of international standards, such as assessment and professionalism related competencies, there does not appear to be a single set of universal standards that are common across these three locations. It is no surprise therefore when developing standards for teachers in Vietnam that the culture and

government goals and directions influenced the development of standards and requirements of teachers.

In 1994 the OECD published its survey of teacher quality in its member states. It concentrated on the characteristics of teachers of high quality in relation to:

- knowledge of substantive curriculum areas and content;
- pedagogical skill including the acquisition of knowledge and ability to use a repertoire of teaching strategies;
- reflection and the ability to be self-critical;
- empathy and commitment to the acknowledgment of the dignity of others; and
- managerial competence in a range of responsibilities within and outside the classroom;

(Organisation for Economic Cooperation and Development, 1994)

This work was notable because of the characteristics it identified. The succinct statements illustrated the advantage of building up concepts from studies of highly successful practice. Observing that teacher commitment was the quality that made all other qualities possible, the report noted that high quality teachers:

- demonstrate commitment;
- have subject specific knowledge and know their craft;
- love children;
- set an example of moral conduct;
- manage groups effectively;
- incorporate new technology;
- master multiple models of teaching and learning;
- adjust and improvise their practice;
- know their students as individuals;
- exchange ideas with other teachers;
- reflect on their practice;
- collaborate with other teachers;
- advance the profession of teaching; and
- contribute to society at large.

More than any other analysis, this set of expectations has influenced the work in Vietnam through the World Bank education sector report.

Moreover, the establishment of standards and their implementation must be based on a number of principles articulated by Brock (2000):

- the identification of any professional standards must involve full discussion with and ultimately ownership of such standards by the teaching profession;
- accomplished teachers make a difference [in pupil performance];
- any attempt to establish professional teacher standards must be firmly grounded in accurate and comprehensive understanding of both the timeless and evolving nature of the work of teachers, principals and other school leaders;
- any construction of professional standards should facilitate the concept of career-long continuum from probationary teacher to retirement – with possibility of moving within as well as outside of and returning to the professional and be applicable to all ranks across the spectrum from beginning or newly appointed to experienced teachers, principals and school leaders; and
- the articulation and commitment to professional standards must be flexible enough to enable, indeed celebrate, the quality of individuality which is a hallmark of being a professional.

As such, a standards framework needs to acknowledge that an accomplished teacher likes children, likes working with them and to have high expectations. Teachers need to have appropriate

intellectual mastery of the subjects and be able to keep abreast of evolving knowledge and teaching methods. They need to be reflective learners themselves and continually attempt to increase their knowledge and practice expertise. The standards must also acknowledge that knowledge, understanding and practices are interdependent and that individual competencies interact.

Glaser (1987) and Berliner (1999) provided insights into who can be considered as expert teachers. Expert teachers excel mainly in their own domain and in particular contexts. They develop automaticity for repetitive operations that are needed to accomplish their goals. Expert teachers are more opportunistic and flexible in their teaching than are novices. They are more sensitive to the task demands and social situations surrounding them when solving problems. Expert teachers can represent problems in qualitatively different ways than do novices, have faster and more accurate pattern recognition capabilities, perceive more meaningful patterns in the domain in which they are experienced and begin to solve problems slower, but bring a richer and more personal resources of information to bear on the problems they are trying to solve. They make better use of knowledge, have extensive pedagogical knowledge, including deep representations of subject matter knowledge, better problem-solving strategies, better adaptation and modification of goals for diverse learners and have better skills for improvisation. They are better at decision-making, deal with more challenging objectives, establish a better classroom climate, have better perception of classroom events and a better ability to read the cues from students. Expert teachers have a greater sensitivity to context. They are better at monitoring and providing feedback to students. They more frequently test hypotheses about teaching and learning, give greater respect to students and display more passion for teaching. Their students have higher self efficacy and motivation to learn, pursue deep learning activities and have higher levels of achievement. Expert teachers have a better understanding of how to translate expertise in discipline to a form that is understood by pupils and have greater knowledge of discipline and of pedagogy interact.

Teacher qualities and competencies change and grow through experience and teachers adapt to the circumstances in which they find themselves at varying stages of their career. School authorities seek to recognise this or allocate additional responsibilities to selected experienced teachers and schemes are often structured according to levels or stages. The capacity to adapt and demonstrate increasingly sophisticated competencies is expected through successive levels.

Developing Primary School Teacher Standards in Vietnam

Indicators that describe ways in which teachers could demonstrate evidence of those qualities in their work are often needed. Indicators assist teachers to monitor their own development and provide an idea of what is expected at particular levels. They also assist those who are responsible for supporting or assessing teachers in their development. Monitoring or assessment of a teacher's development also needs to take account of the context within which the teacher works and the quality with which the teacher demonstrates or adapts performance to the demands of the context. Ideally, an assessment would occur across the range of competencies and would be qualified according to how well the teacher performed specific duties and adapted to the context. Stages of development of a teacher's competence could then be identified and a profile drawn up to assist the teacher and those responsible for her/his development to plan for improvement. This is not the same as adjusting an assessment for the effect of context.

In developing the primary teacher standards for Vietnam, these background studies were taken into account in the development of the prototype standards developed in the year 2000. It was decided that it should be a standards or competency-based approach in which the focus was on what teachers were required to know or do in the school rather than on time served. This represented a radical shift in thinking and needed a long gestation period for the government to publicise and gain the acceptance of the teaching profession and the community. A national program through the media was launched to gain this acceptance. A period of two years elapsed after the initial feasibility study (Griffin et al., 2001) before the competency approach was further explored. After reviewing the international scene in standards and teacher evaluation, a committee established by the Ministry of Education and Training (MoET) set the parameters for the development of standards and for profiling

teacher development. For example, the number of levels was set by the government working party after a series of consultations and functional analyses of teachers' duties according to the Government regulations. The number of levels was set to accommodate the government regulation defining the ranks of teachers as 'Teacher', 'Senior Teacher' and 'Leading Teacher'. The study reported here sought to develop a set of professional standards for defining the skills and knowledge required of teaching at each of these levels in Vietnam. There were three main purposes of the procedure developed for this study. They were:

- to empirically validate and refine the standards;
- to identify efficient and standardised scoring procedures for making professional judgements of the competence level of the teacher; and
- to determine the most appropriate way in which to gather evidence of teacher competence in school settings.

Background Development Work on Defining the Standards

The construction of the standards was based on a combination of both theoretical and psychometric approaches to scale development. A set of prototype standards were initially developed by the MoET, in which three 'strands' or areas of competence were drafted, with each strand having three levels. The prototype standards contained no procedural advice; they were simply broad statements and description of levels of development among teachers. A series of forums with key stakeholder groups (including academics, government officials, teacher education providers) were used to review the standards and to make recommendations about procedures to ensure that the assessment process matched the existing procedure as closely as possible but allowed for change in expectations to be introduced.

At the end of the drafting process, three strands were agreed upon (*Ideology and Personality, Pedagogy and Discipline Knowledge*). Specific requirements (competencies) were agreed upon for each strand. These were defined as the professional expectations of teachers. There were four requirements in the *ideology* strand and five requirements for each of the *pedagogy* and *knowledge* strands. Each requirement was defined by a series of indicative behaviors, knowledge or skills that the teacher was expected to be able to exhibit. These were called performance indicators (PI). Each indicative behaviour (PI) was then further refined according to the quality of the behaviour, knowledge or skill exhibited. These were called quality criteria (QC) and they essentially answered the question of 'how well' was the indicative behaviour demonstrated such that it was possible to differentiate between teachers based on evidence produced. As such, the structure of the standards addressed four issues:

- What is expected of teachers? (requirement)
- What evidence would a teacher have to demonstrate to indicate that this was present? (performance indicator)
- How well did the teacher demonstrate this? (quality criterion)
- How do the quality criteria differentiate between teachers?

Methodology

The first three questions listed above addressed the overall definitions of teacher requirements. The fourth question was treated as an empirical question, and was subject to a survey of teachers and an investigation of the efficacy of the assessment procedures developed in parallel to the standards. The content and substance of the requirements and the assessment procedures were subjected to a series of reviews and examinations including a series of expert review panels and a pilot study to examine the proposed assessment procedures and the potential impact on the teachers. The feedback from the panel and pilot studies was used for a final revision before trials begin.

Teacher Training Institute (TTI) staff, district officers and leading teachers filled the role of assessors. They were selected by the Ministry of Education and Training and hence were assumed to have high levels of teacher competence as well as honourable status in the community. Eleven assessors were selected from each of the ten (of a total of 61) provinces that were selected by the

Government to participate in the study reported here. They were also trained to become 'assessor trainers' for later scaling up of the procedure. This would enable continuous training of assessors to occur for future roll-out in which over 380,000 teachers are expected eventually to be assessed.

Assessors were trained in the procedures and interpretation of evidence obtained using portfolio, interview, third party reports; and direct observation. Data were forwarded to the central project office. A data checking exercise was performed to ensure that there were no incorrect or inappropriate codes in the data and to check the data for accuracy and reasonableness. The data were then analysed using item response modelling procedures involving *ConQuest* (Wu, Adams and Wilson, 1998).

Recording Instruments

The assessors were required to complete a questionnaire on both the teacher's performance level as well as the sources of evidence (i.e., portfolio, interview, classroom observation and third party). The assessor recorded the numerical code for the quality criterion that best described the teacher's performance. The requirements, performance indicators and quality criteria were presented in a rating scale format. A sample item is shown in Figure 2.

Requirement 3.1 Knows how to design lessons plans which reflects by identifying the right objectives, contents of the lessons, intended teaching methods and aids, and appropriate allocation of time according to lessons procedures

Criterion 3.1.2: *Lesson plans must present sufficiently objectives of the lessons.*	3.1.2.1 Lesson plans must be developed in accordance with objectives of the lessons	1
	3.1.2.2 Lesson plans must present sufficiently objectives of the lesson on the knowledge, skill and attitude	2
	3.1.2.3 Lesson plans must present sufficiently objectives of the lesson in the detailed manner for observation and evaluation	3
	Not enough Information to make a decision	0

Figure 2. Candidate questionnaire: A behavioural rating scale.

As shown in Figure 2, an hierarchical rating scale was used to record the teacher's performance. The number of levels varied, depending on the nature of the indicator. A zero was used if the assessor could not identify any evidence of the criterion.

There were many assessors (almost 1000) involved in this study and in the final roll out of the project there will be thousands of assessors. Differences in judgements of assessors and the effects of differential stringency were to be expected. For this study, the assumption was made that the variability of stringency was randomly distributed among assessors and hence the effects of other factors such as location, teacher gender and so forth needed to be checked. This was especially the case when sharp differences in competence were identified across provinces. The question as to whether the assessors in some provinces were more stringent than assessors in other provinces needed to be addressed. If this were found to be the case, there was an unfair system being implemented in that competence depended on where a teacher was employed rather than actual ability.

Wherever judgement is involved there is a possibility of different judgement levels of stringency being applied and different scores being assigned to candidates of equal ability on the same item. The question in this study was whether the difference in stringency was systematically related to the provincial location of the assessor and the teacher. When this occurs systematically over groups of individuals, the effect is called differential item functioning (dif). Usually the effect (dif) refers to differences in item performance across comparable groups of examinees (Dorans and Holland, 1992). Test items, for example, are said to exhibit differential item functioning if the item scores of equally able examinees from different groups (e.g., of different race, sex, or age) are systematically different (Kelderman and Macready, 1990, p307). It refers to a psychometric difference in how an item functions across the groups. It also refers to a difference in item performance between comparable

groups of examinees, that is, groups that are matched with respect to the construct being measured by the test. The two groups have been typically referred to as the reference (R) group and the focal (F) group. However, these definitions apply to the case where only two groups are compared. With recent developments in software such as *ConQuest*, (Wu, Adams and Wilson, 1998) and RUMM (Andrich, Lyne, Sheridan, and Luo, 2003) these labels are redundant and the definition of diff can be broadened. The reference group can be thought of as a 'benchmark' for item behaviour and the focus group can be any of a number of groups compared to that benchmark. In the case of item response modelling (IRM) and its use in the examination of dif the modelled item characteristic curve can be used as the benchmark and the focus groups are the groups of interest that are compared to this benchmark. Alternatively, it is possible to use a known or accepted group behaviour with the test item as a benchmark. This is the case in the present study.

Data Analysis Method

The assigned scores for the assessment were derived from 64 separate criterion items and 14 requirements. The requirements were the main focus of the study as these were scored over all 10 trial provinces. The criterion items were scored in only two provinces as a cross check of the earlier validation of the requirements (Griffin, Nguyen, Gillis and Mai, 2006). Each was scored using a partial credit approach. Scoring each item in this manner treated them as 14 independent polytomous items, in which each teacher, n_v, had a competency θ_v and each item had a set of difficulty parameters δ_{i1}, δ_{i2}, δ_{i3} ... δ_{ik} representing the difficulty of attaining each of the scores from 1 to k for item *i*. Each of these parameters governed the likelihood of a teacher with ability, θ_v, being given a score of *k* rather than *k*-1. The analysis modelled the relationship between teacher competence and the difficulty parameters on each of the 14 requirements. The Rasch model was used to estimate teacher competence independent of which particular items were used for the estimation. The natural logarithm of the odds of achieving a specific score of *k* rather than *k*-1 is obtained from the simple relationship

$$\ln \frac{p_k}{p_{k-1}} = \ln \frac{n_k}{n_{k-1}} \qquad (1)$$

where p_k and p_{k-1} are the proportions of teachers scoring k and k-1 respectively.

Most items were scaled using IRT (Item Response Theory) scaling methodology. With the Simple Logistic Rasch model (Rasch 1960/1980) for dichotomous items, the probability of assigning a score of 1 instead of 0 is modelled as

$$P_i(\theta) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \qquad (2)$$

where $P_i(\theta)$, the probability of person *n* scoring 1 on item *i*. θ_n, is the estimated latent teacher competence trait of person *n* and δ_i the estimated location of item *i* on this dimension. For each item, item responses are modelled as a function of the latent trait θ_n.

In the case of items with more than two (*k*) scoring categories (as for example with a maximum *x*=score greater than 1) this model can be generalised to the *Partial Credit Model* (Masters and Wright, 1997)[1]. The Partial Credit Model developed by Masters (1982) is an extension of the Simple Logistic Model, and overcame the restrictions to dichotomous scoring or fixed rating scale categories. The model was developed by estimating parameters for the difficulties associated with a series of performance levels within each item. Masters (1982) argued that the difficulty of the k^{th} level in an

[1] An alternative is the Rating Scale Model (RSM) which has the same step parameters for all items in a scale (see Andrich, 1978).

item governs the probability of responding in category *k* rather than in category *k* - 1. The probability of person *n* completing the k^{th} level is specified by Masters (1982: 158) as:

$$P(X_{ni} = x) = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_{ik})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h} (\theta_n - \delta_{ik})} \quad (1)$$

The model estimates the probability of a person *n* scoring *x* on the m_i performance level of item *i* as a function of the person ability on the variable being measured and the difficulties of the m_i levels in item *i.* The observation *x* is a *count* of the successfully *completed* item levels, while only the difficulties of these completed levels appear in the numerator of the model. The model provides estimates of person ability θ_n and item step level difficulty δ_{ik} and $P_{x_i}(\theta)$ denotes the probability of person *n* scoring *x* on item *i*. θ_n denotes the person's position on the latent trait, the item parameter δ_{ik} gives the location of the item step, k, on the latent continuum and denotes an additional step parameter.

The IRT approach also enabled an assessment of how closely the obtained data was predicted by the mathematical model. The predicted data were called the modelled data and a comparison of observed and modelled data allowed an examination of fit. Fit was important because if the data did not conform to the model, it was assumed that the relationship between ability and the chances of success had broken down. If this were the case, then it was possible that the item may have not been measuring the same relationship between ability and chance of success and from this it was also possible that the ability being measured by the item may not have been the same as that measured by other items. Such an item might be omitted from the assessment instrument. The same applies to the way a person's data conform to the pattern predicted by the model across all items. When the person's response pattern differs markedly from the modelled data pattern, it is possible that this person may be exhibiting a different type of ability to that being demonstrated by other candidates. If this is the case further investigation of the person's response pattern is warranted and a decision is required as to whether this person should be assessed at all using this particular assessment instrument.

Item fit was assessed using the information weighted mean-square fit statistic (infit), which is a residual-based fit statistic (Wright and Mok, 2000). Weighted infit statistics were reviewed both for item and step parameters. The *ConQuest* software (Wu, Adams and Wilson, 1998) was used for the estimation of item parameters and the analysis of item fit. The software package *Quest* (Adams and Khoo, 1996) was used to estimate the person fit.

Given that the 64 criterion items and the 14 requirement items had variable maximum scores, the partial credit model (Wright and Masters, 1982) using the computer program *ConQuest* (Wu, Adams and Wilson, 1998) was used to derive the estimates of item difficulty and teacher competence.

Under an IRM framework, an item was showing dif if the Item Characteristic Curve (ICC) was not the same for the groups being assessed and if these were different in important ways to the benchmark curve; that is candidates who are equal in terms of the latent trait or ability do not have the same probability of being assigned the same score on the item (Embretson and Reise, 2000). Dif is therefore the effect when candidates who are equal in terms of the ability being measured by an assessment instrument come from different subgroups and in general membership of the subgroup systematically affects the probability of being assigned a specified score (Camilli and Shepard, 1994). Membership of a group is then a determining factor in terms of scores on the assessment instrument.

Typically, classical methods use an internal criterion such as total test score or 'other items in the test' as the criterion for matching examinees to see if 'comparable examinees from different groups' performed the same on individual test items.

In this model the probability of an assigned score is determined by the ability of the candidate and the difficulty of each score point for each item. It is an extension of the simple model. The relationship between probability and the score assigned is described in a plot called the characteristic curve. For a dichotomous item, these are as in Figure 3, where only the curve for a score of one is shown (the curve for a zero score provides redundant information but can be described as the symmetric opposite of the curve for a score of one).

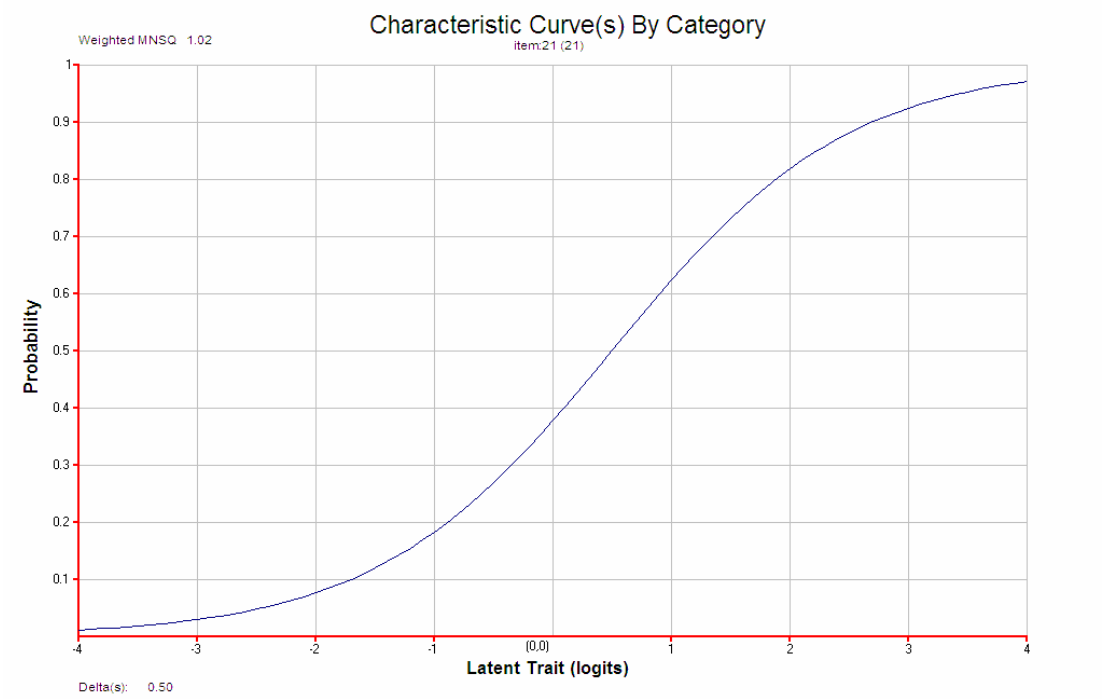

Figure 3. Characteristic curve for a dichotomously scored item.

For the partial credit item the characteristic curve is quite often of the form shown in Figure 4.

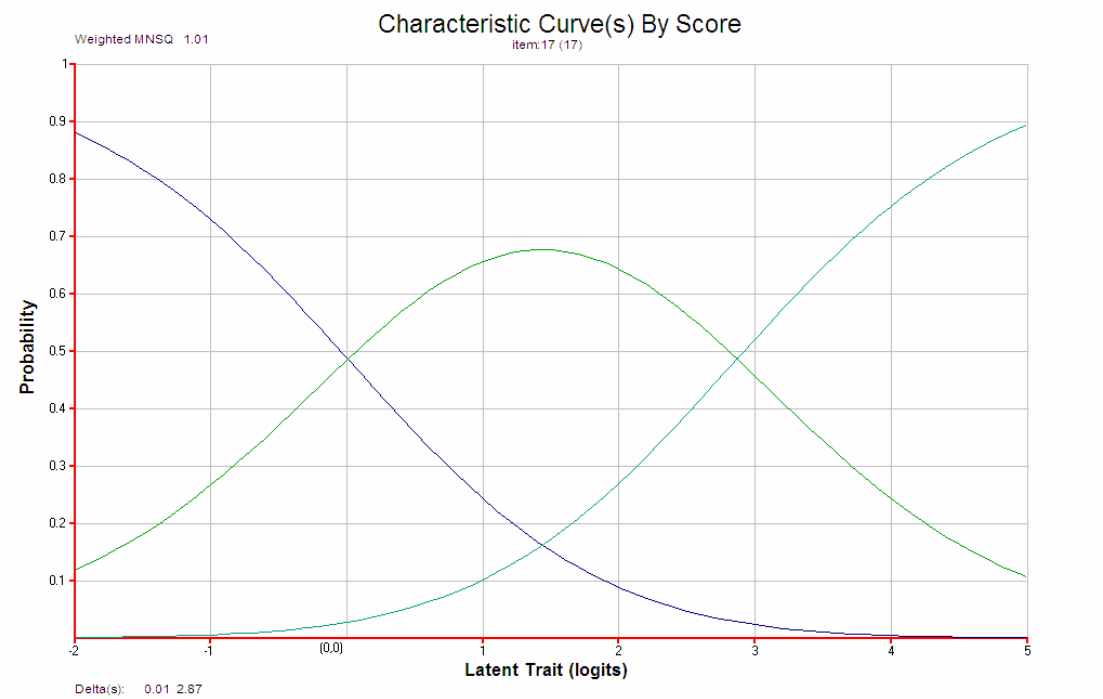

Figure 4. Characteristic curve for a polytomous item with three score categories.

When the data obtained for the same item over different groups of candidates are analyzed separately, the ICCs should be coincident. Small variations are tolerable because of measurement error, but in general the ICCs for different groups of teacher groups should not be separated. This is because of the requirement that the probability of a given score being assigned should be identical for all teachers of the same competence. When the ICCs for separate groups of teachers differ on a systematic basis, then a secondary effect is operating and in this case we have assumed that it is the assessor location and that systematic bias is operating. Such an effect is unacceptable and would have to be remedied either through retraining or a change in the approach to teacher assessment. It would mean that a harsher assessor required a higher competence of a teacher for a specific score and a more lenient assessor required a lower competence of a teacher. In this case the harsher assessor group would generate a set of scores that would be represented by an ICC that moves to the right of the modeled curve in Figure 4 and a lenient assessor group would generate scores that would lead to an ICC to the left of the modeled curve. If the movement were such that the chances of promotion, salary increment or assignment to responsibility or other decision was affected by the systematic difference in assessor stringency, then there would have been an important difference in the ICCs and some remedial action would be needed.

Sample

In 2004, 24886 teachers in 10 provinces of Vietnam were assessed by 990 assessors. The number of teachers and assessors are presented in Table 1

Table 1

Number of Assessors and Teachers involved in the Assessments.

Province	Number of Assessors	Teachers	Districts	Commune/Sub-districts
Son La	128	3202	11	110
Vinh Phuc	84	2100	9	98
Hai Phong	112	2795	13	95
Ninh Binh	66	1800	8	67
Quang Binh	72	1788	7	85
Phu Yen	83	2100	8	70
Kon Tum	48	1200	8	47
Binh Phuoc	72	1811	8	55
HCM	229	5800	25	182
Ben Tre	96	2400	8	89
Total	990	24996	105	898

Teachers were assessed against each of the 64 criteria and the assessment results were first recoded on the Criterion form. The scores for each Requirement and strand were then calculated using score conversion rules and were recorded on the Requirement form. While all the Requirement forms were collected for analysis, only 3 995 Criterion forms of teachers from Hai Phong and Kom Tum provinces were collected for the purpose of validation.

Results

One of the first impressions when the data were examined was that there were systematic and profound differences across provinces. The starkest were the differences between the provinces Ho Chi Minh City and Hai Phong. It appeared that teachers in Hai Phong were least competent in terms of pedagogical skills and the teachers in Ho Chi Minh City were mostly outstanding. Bearing in mind that this study aimed at validating an instrument rather than assessing teacher competence, there was a real possibility that the instrument had been misused and provided different results depending on the location of the assessor and the teacher. Hence it was essential that the data were interrogated to determine if location had introduced a systematic source of bias.

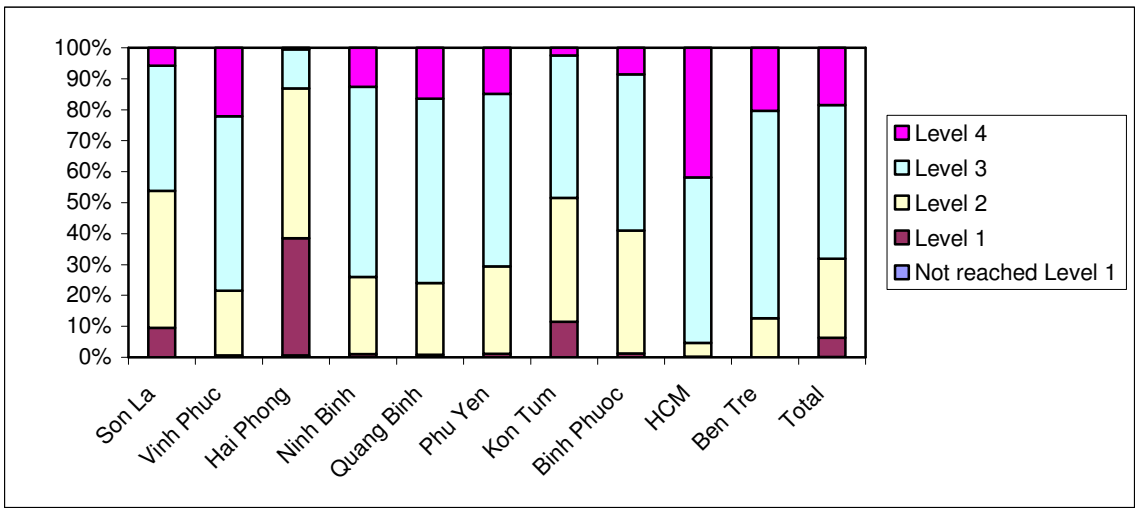

Figure 5. Distribution of competence for pedagogical skills across provinces.

The 14 requirement data patterns were explored using *ConQuest* to determine if the differences in assessor group (represented by province identification) stringency were responsible for the

differences in teacher competence assessments. The analysis result of one strand is detailed in this section as an example and is presented in Tables 2 and 3.

Table 2

Item Parameter Estimates, Measurement Error and INFIT of Requirements of Strand 3

Item	δ_1	SE_1	δ_2	SE_2	δ_3	SE_3	δ_4	SE_4	INFIT
3.1	-11.10	.36	-3.64	0.07	4.14	0.05	7.88	0.03	0.87
3.2	-12.18	.69	-3.75	0.06	0.34	0.06	5.58	0.05	0.84
3.3	-9.72	.18	-0.62	0.05	3.46	0.03	8.05	0.05	0.91
3.4			-2.50	0.06	1.93	0.04	5.30	0.04	0.9
3.5	-10.70	.30	-0.33	0.02	6.11	0.04	10.20	0.12	0.93

It can be noted from Table 2 that the INFIT values are all within the acceptable range of 0.77 to 1.3 and hence there is evidence of a single underlying continuum in Requirement 3.1 which is being measured, which are acceptable (Adam and Khoo, 1996; Wright and Masters, 1982).

Table 3

Mean Ability and Requirement Discrimination Index of all Requirements of Strand 3

Requirement		Quality indicator					Discrimination
		0	1	2	3	4	Index
3.1	Mean ability	-5.66	-5.88	1.34	5.44	8.32	0.86
3.2	Mean ability		-6.22	-0.51	3.00	6.74	0.89
3.3	Mean ability	-7.28	-3.00	1.62	5.19	8.46	0.88
3.4	Mean ability	0.00	-5.21	0.41	3.71	6.46	0.90
3.5	Mean ability	-7.16	-2.78	2.98	6.92	8.76	0.83

It can be seen from Table 3 that the Requirement levels were ordered correctly, as when the levels increase the mean ability of the groups corresponding to these levels also increases. The discrimination indexes of all Requirements were higher than 0.80 which showed that the Requirements discriminated well between teachers in regard to their pedagogical skills

The Dif analysis for 14 requirements are presented in Figures 6 to 19.

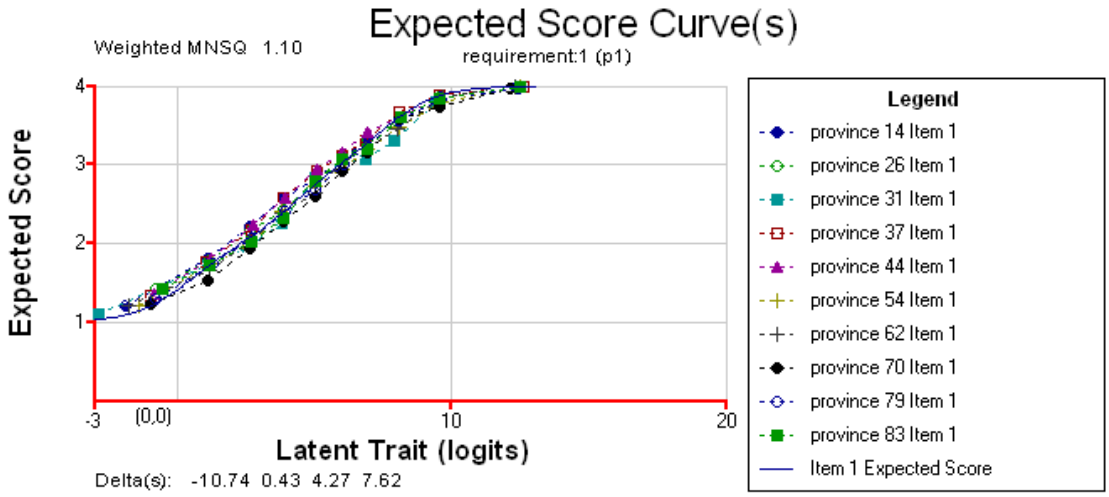

Figure 6. Dif analysis for requirement 1.1.

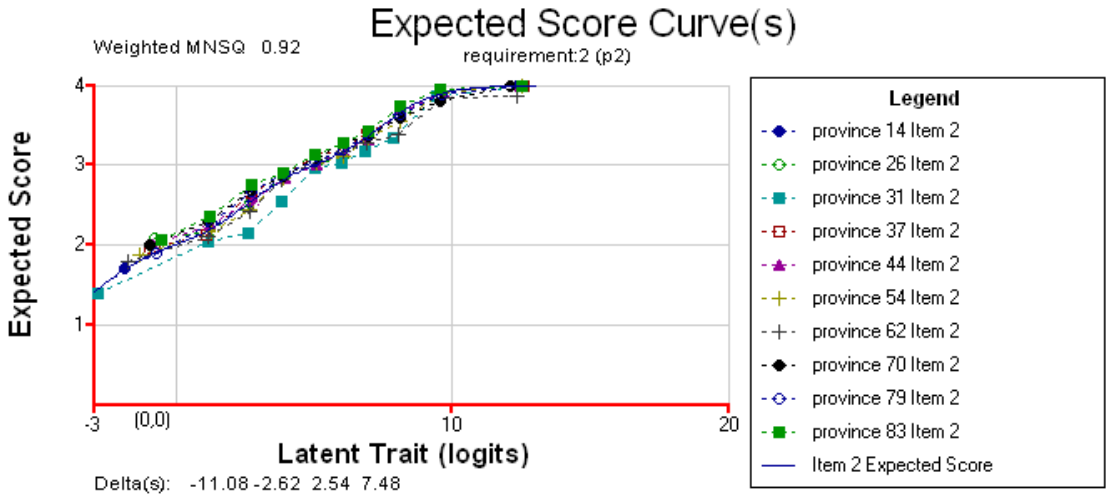

Figure 7. Dif analysis for the requirement 1.2.

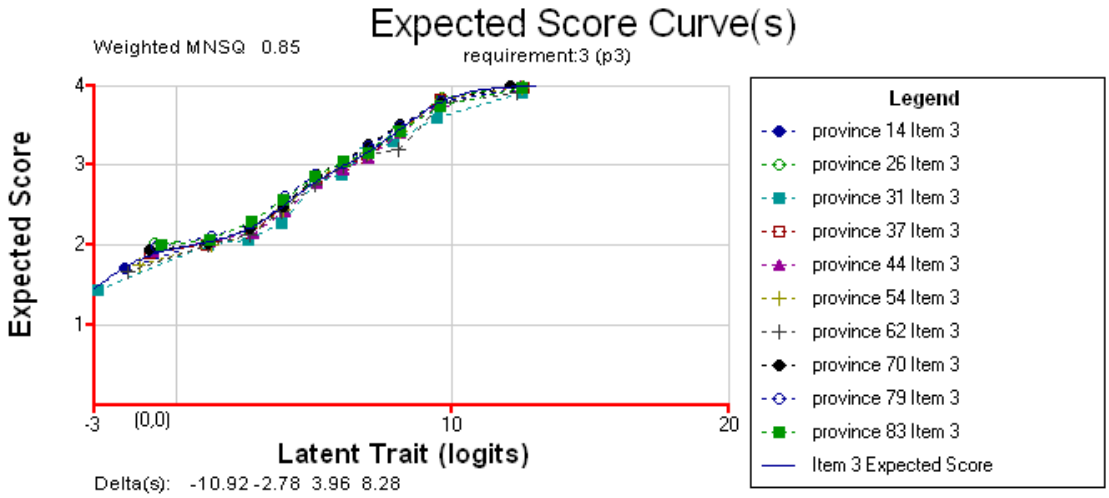

Figure 8. Dif analysis for the requirement 1.3.

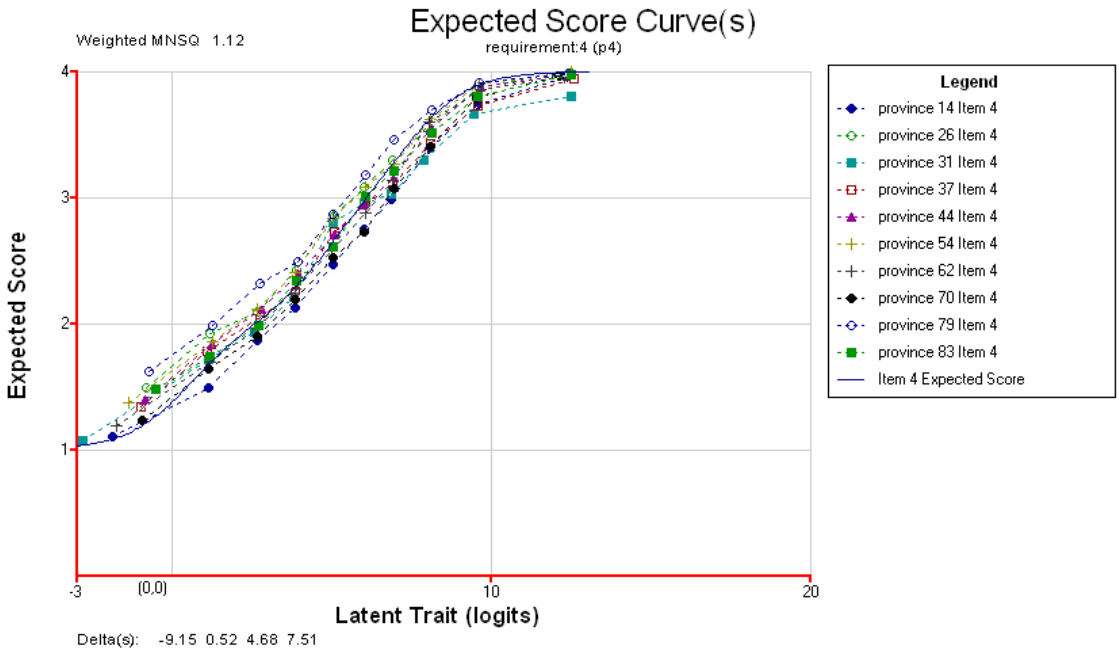

Figure 9. Dif analysis for the requirement 1.4.

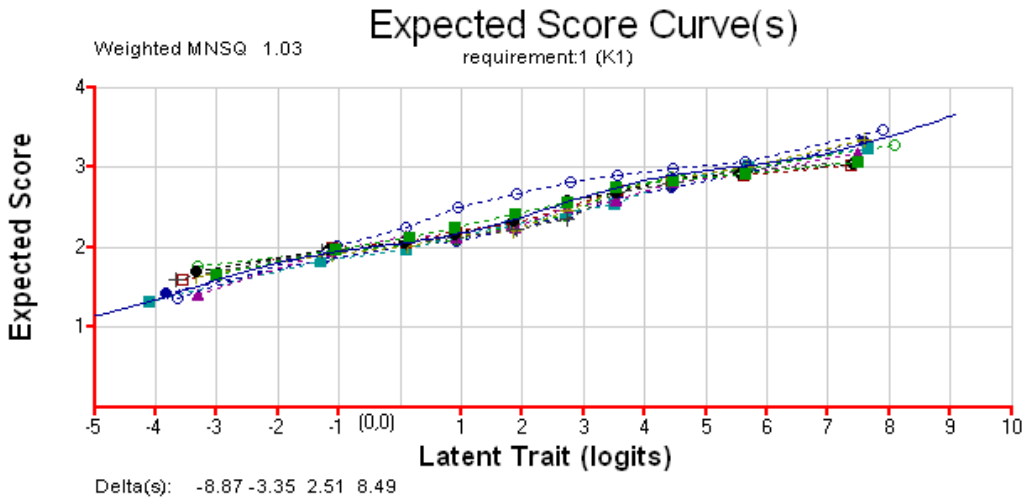

Figure 10. Dif analysis for the requirement 2.1.

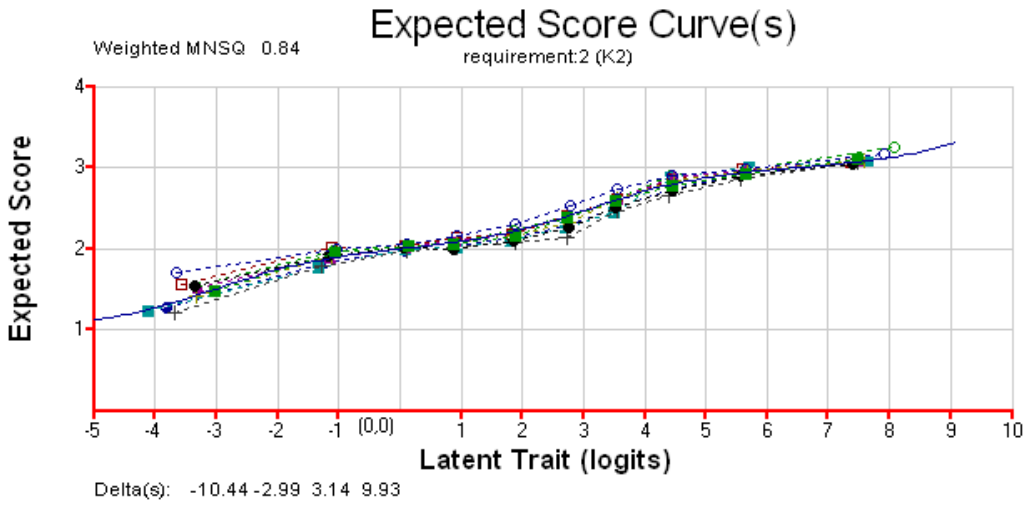

Figure 11. Dif analysis for the requirement 2.2.

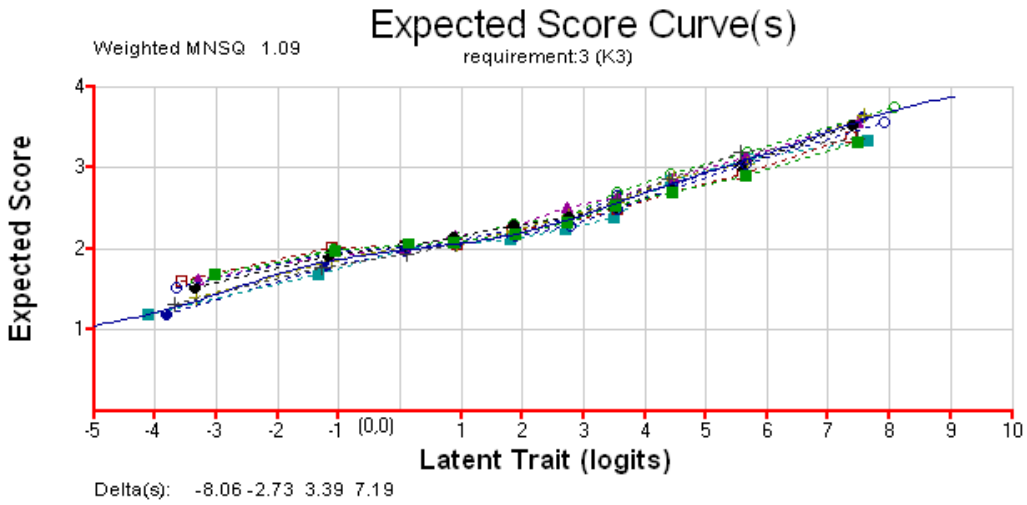

Figure 12. Dif analysis for the requirement 2.3.

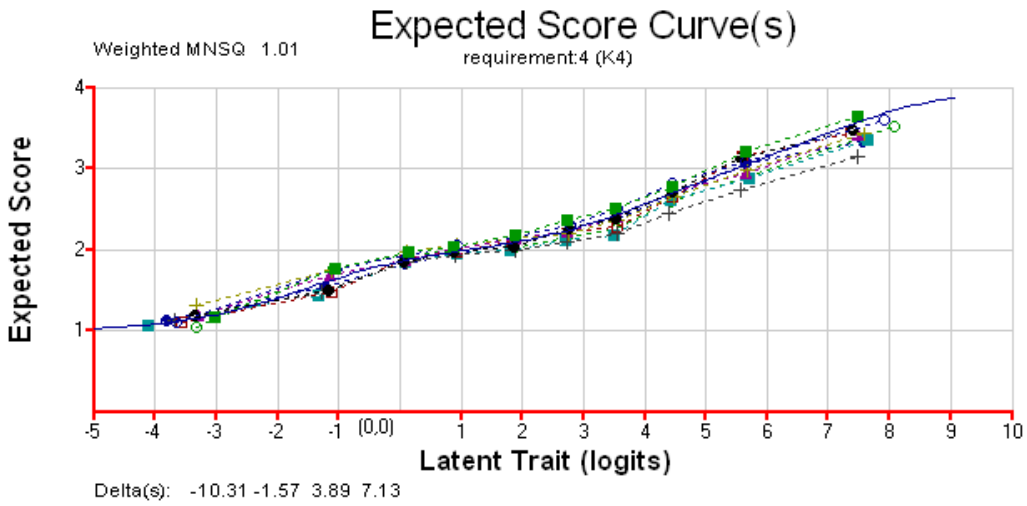

Figure 13. Dif analysis for the requirement 2.4.

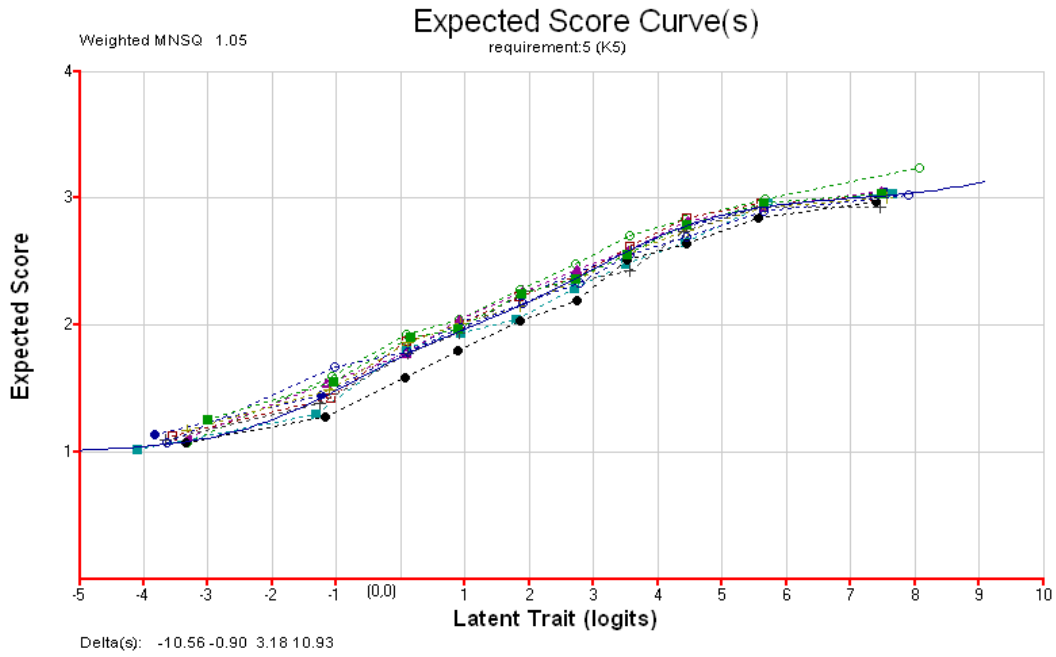

Figure 14. Dif analysis for the requirement 2.5.

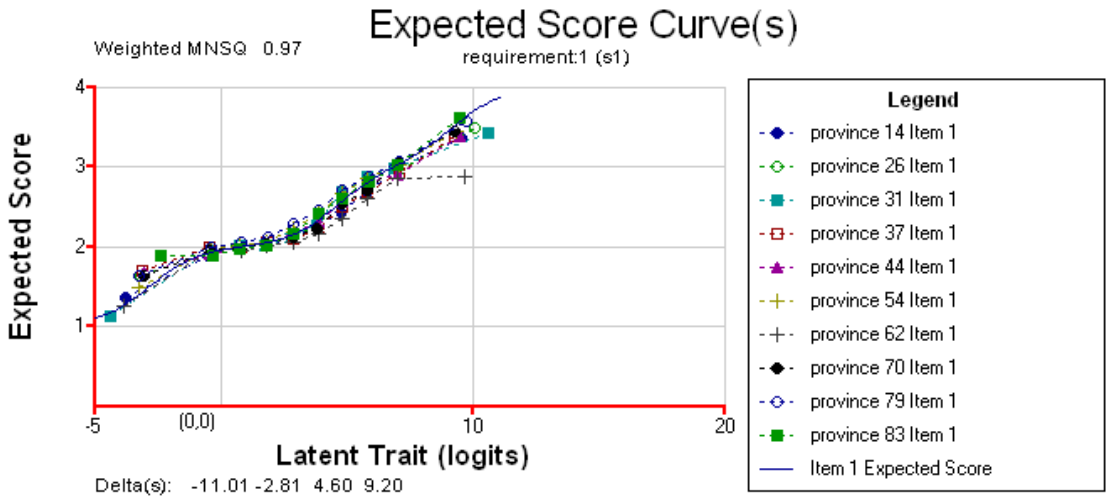

Figure 15. Dif analysis for the requirement 3.1.

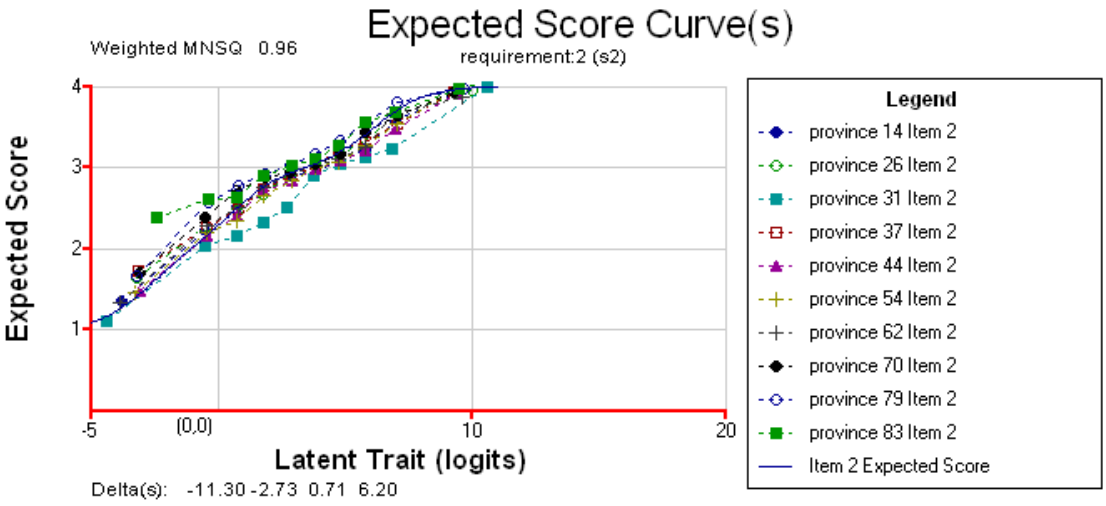

Figure 16. Dif analysis for the requirement 3.2.

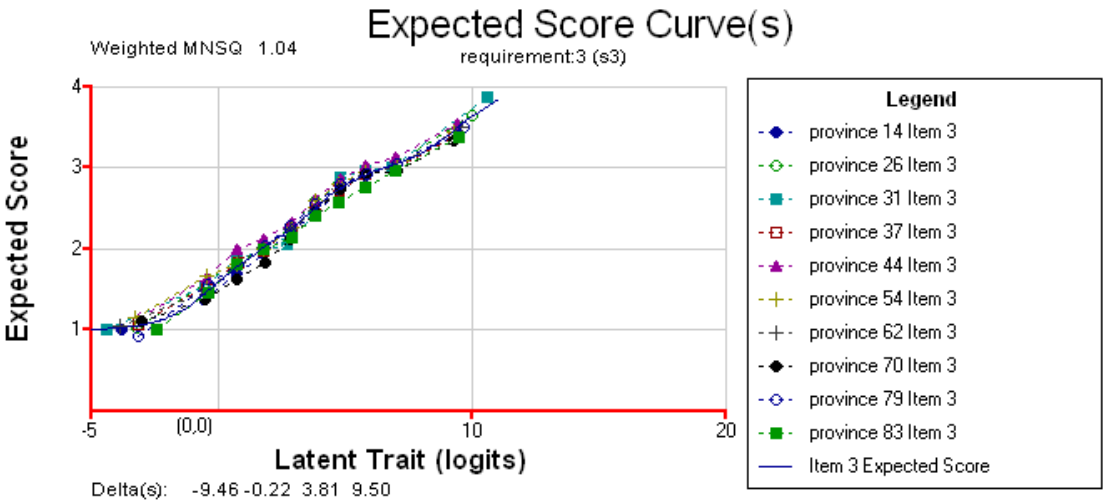

Figure 17. Dif analysis for the requirement 3.3.

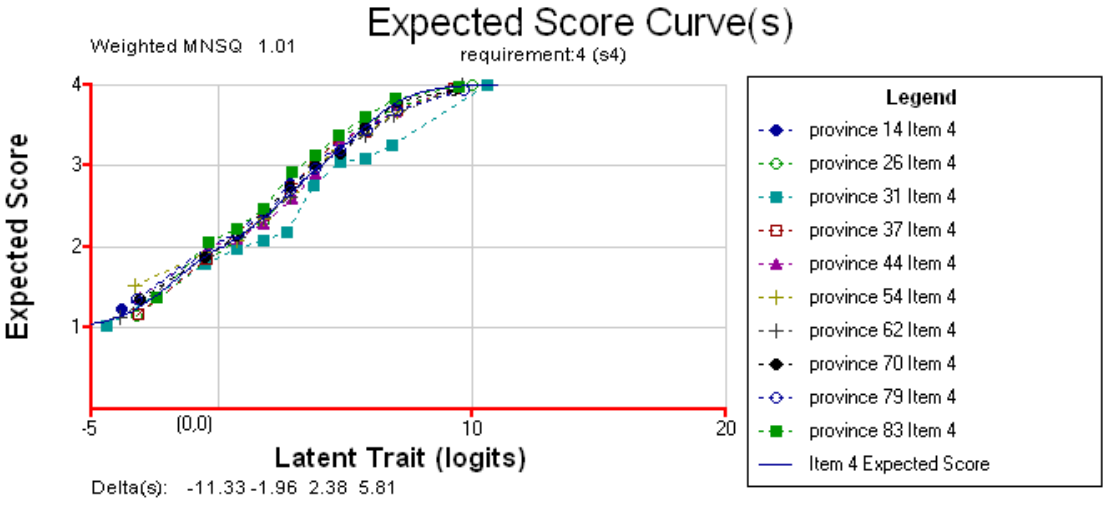

Figure 18. Dif analysis for the requirement 3.4.

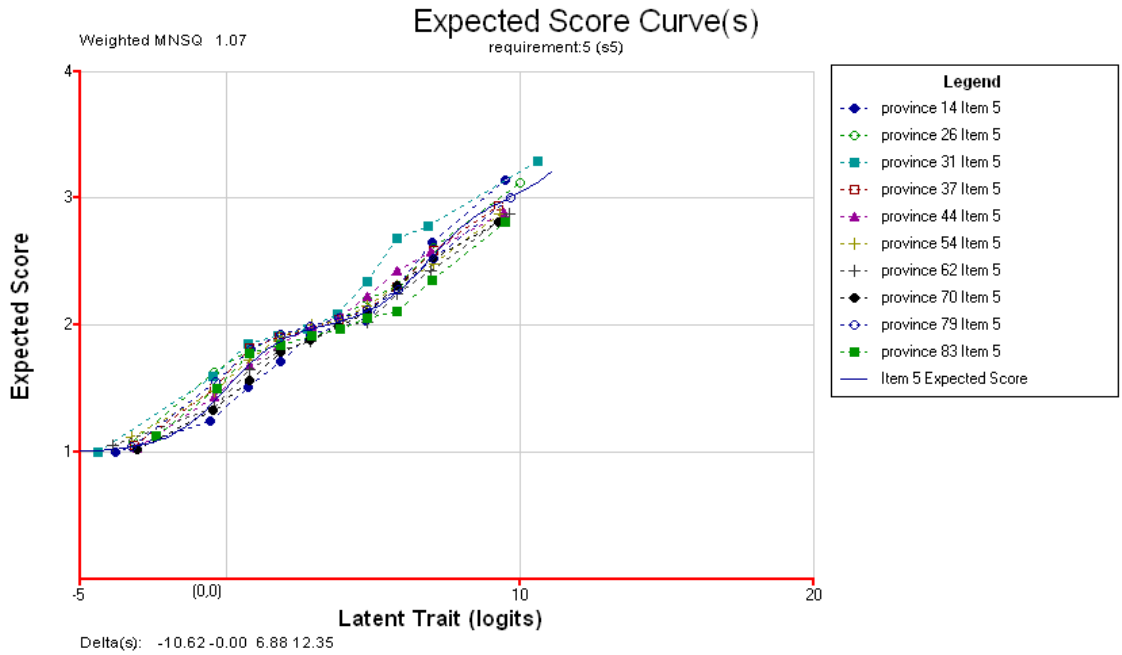

Figure 19. Dif analysis for the requirement 3.5.

For each of the requirements, the Chi-square test showed that there was a difference in teacher performance in the 10 provinces.There was a small and significant dif in the set of Requirement score patterns of teacher performance across the 10 provinces. However for each of the 14 requirements it was clear that the expected score (vertical axis) of teachers was unaffected by the diff effect. Teachers of the same ability did not vary in terms of expected score across provinces. In other words, assessors' provincial practices were not the source of dif. That is, irrespective of where the teachers were assessed, the expected scores for each Requirement did not change. Their location was not the source of the large differences in competency scores. So the 14 Requirements were successful in defining teacher performance on the three competency strands.

Discussion

The differential effects of assessors were not systematically associated with the provincial identification. Differences in the distribution of teacher competence were more likely to be related to as sampling bias than a bias in the measurement instrument. That is, there was a sampling problem rather than an instrument problem. The instruments and the assessment procedure were not linked to location. Sampling was more than likely the major source of differential performance across provinces, since no control could be exerted over the sample design other than to indicate to the provincial teams that a uniform rectangular distribution was needed for the calibration.

The process of development and validation of the Vietnam Teacher Profiles thus supported the principles recommended by Brock (2000) and illustrated that they could be successfully implemented. The combination of extensive expert consultation and psychometric validation (Griffin et al., 2006) has been proved to be an effective way of developing and validating competency assessment instruments.

While the Profiles with 14 Requirements were linked to the Profiles with 64 Criteria through the score conversion rules, as for the 25 000 assessments conducted in 2004, the Profiles with 14 Requirements can be used independently. At some stage, when assessors are familiar with the methods of competency assessment trained by the University of Melbourne, the Vietnam Government may consider the use of the Teacher Profiles with 14 Requirements independently.

The documented process is useful for the Vietnam Government in future revision of the Teacher Profiles. Furthermore the process can be used in developing the teacher standards for secondary and higher education.

Reform in primary education in Vietnam has been an ambitious program. Reforms of curriculum, teaching and learning, resource and infrastructure were targeted in the World Bank strategy developed in conjunction with the Vietnamese government. Developing teacher standards had been identified as an important central aspect of their reform of the education system. This article has discussed the development of only one component of the reform of teaching and teaching standards. The overall reform was intended to include changes to teacher appraisal, their terms of service, opportunities for pre- and in-service teacher training and to a personnel management system. The assessment procedures developed and reported in this study were meant to be central to the overall reform. Links between the assessment outcomes and professional development opportunities were also meant to have been established coincidently. A three-tier progression for advancement in teaching was expected to be established as a framework for teacher promotion. Teachers would and could advance to the top of the first tier (beginning teacher) based on time served, but if a teacher sought promotion to 'advanced teacher' an assessment of competence would be required indicating that the teacher has at least met the standards for that second level. The teachers could then progress to the top of this second tier and when ready for promotion to the level of 'expert teacher' another assessment would be required. At the time of this article being written, the other components of the primary teacher reform had not been realised to a point where the assessment could point to professional development programs to raise performance from one tier to the next in terms of teacher professional progress. Nor has the management system been finalised, nor the terms of service that would define the regulations for such a system to be implemented. The fifth component of the system, capacity building, had been successful in part, at least for the assessment component, in that the local team had been trained in the methods of developing standards, and in the assessment strategies. More than 1000 assessors had been trained, technical teams had been trained in the three regions surrounding Hanoi, DaNang and Ho Chi Minh City, so an infrastructure had been put in place for the system to be rolled out as the remaining components emerged.

In this component of the reform, item response modelling was used to develop a simple-to-use questionnaire format for recording teacher competence against a range of standard requirements. The results showed that assessors could be trained, and that the requirements and the criteria discriminated between teachers on the basis of their professional competence. Assessors found the system usable and the training program was readily adapted to local Vietnamese conditions and educational culture. It was clear, however, that teaching and classroom practices and cultures were not amenable to western cultural competencies. What was regarded as superior teaching and classroom management was not the equivalent of western approaches, but it was not the purpose of the study to impose such systems. So, despite the similarity in structure to standards developed elsewhere, the content and orientation of the Vietnamese standard are more closely oriented to the culture of the existing system. The structure of the standards and the methodology was clearly transportable from a western system to the Vietnamese Confucian context.

While the nomenclature varied to suit the language and expectations of the Vietnamese government, the structure remained stable. Strands (domains), were broken down into requirements (competencies), which in turn required a checklist of evidence (performance indicators) and these in turn were qualified according to the quality of the performance embodied in the evidence (quality criteria). It was clear that professional standards could be developed for Vietnamese primary schools as much as they could be used in other fields such as emergency management, senior secondary schools and even for school principals.

The Vietnamese education system had several requirements of its own. It was clear that assessors had to be trained and credentialed to collate evidence from a range of sources before completing the assessment record forms. It was also necessary to train the assessors to prepare the assessment materials and procedure in advance of the visit to the school so that the time spent on any

individual teacher assessment in the school was minimised. The expense, in terms of teacher and assessor time, needed to be minimised. A time limit was placed on the assessment and a single form used to record all assessment data. Assessors will calculate a score for each requirement and also record this on a yet to be designed *Requirement Record Form*.

All assessors also had to be competent against the requirements. This meant that they were all expected to undergo a training program and be assessed against the knowledge and skills involved in conducting assessments and providing advice to teachers about career enhancement and professional development. Both the assessor and teacher signed the completed record and recommendation sheet at the end of the assessment debriefing session. In the event of a dispute over the assessment, an appeals process was established by MoET so that all appeals could be heard at the district office. Procedures for this were being developed and documented in the Terms of Service. District and provincial officers were also able to review decision patterns of assessors on a regular basis and identify assessors who required further training.

Most notably, this has been a first in the development of teacher standards. While the format of the standards is similar to those used in the United Kingdom, their content is quite different. Moreover, while the record system is similar to those reported in the Denver Public Schools (2005) system, this study has illustrated how it is feasible to develop the standards empirically with compatibility to cultural systems.

References

Adams R.J., & Khoo, S.T. (1996). *Quest : The Interactive Test Analysis System*. Camberwell, Vic: Australian Council for Educational Research.

Andrich, D. (1978). *A rating formulation for ordered response categories*. Psychometrika, 43, 561-573.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2003). *RUMM2020,* v. 4.0, Rumm Laboratory.

Berliner, D. C. (1999). Developing a commitment to social justice in teacher education. In R. Stevens (Ed.), *Teaching in American schools: A festschrift to honor Barak Rosenshine*. Upper Saddle River, NJ: Prentice-Hall.

Brock, P. (2000) *Standards of professional practice for accomplished teaching in Australian classrooms: A national discussion paper*. Australian College of Education, Australian Curriculum Studies Association and Australian Association for Research in Education

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage.

Denver Public Schools (2005). *2005-2006 performance evaluation handbook for teachers, student services professionals, student services professionals – itinerant, curriculum specialists, and evaluators*. Denver Public Schools, Colarado.

Dorans, Neil J., & Holland, P. W ((1992). *DIF detections & description : Mantel-Haenszel and standardization*. Educational Testing Service, Priceton, NJ.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Glaser, R. (Ed.) (1987). *Advances in instructional psychology.* Hillsdale: Lawrence Erlbaum.

Griffin, P. (2004). What were the levels of achievement of grade 5 pupils in reading and mathematics. *Vietnam reading and mathematics assessment study. 2*: 25-58.

Griffin, P., Poynter. G., Nguyen, P. N., Ry, V.T., Thiep, B.D. & Nguyen, T.K.C.. (2001). Report to the Project Management Unit of the *World Bank Primary Education Project*. Hanoi: World Bank Office, Hanoi, SR Vietnam.

Griffin, P., Nguyen, T.K.C., Gillis, S., & Mai, T.T. (2006). An empirical analysis of primary teacher standards in Vietnam. *Planning and changing on reforms in curriculum, teaching and learning in Southeast Asia*. (in press).

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (p 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kelderman, H., & Macready, G.B.. (1990). The use of loglinear models for sssessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27*(4), 307-327.

Masters, G. N. & Wright, B. D. (1997). The partial credit model. In W.J.van der Linden & R. K.Hambleton. *Handbook of modern item response theory*. Springer.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-173.

Organisation for Economic Cooperation and Development (1994) *Quality in teaching*. Centre for Educational Research and Innovation, Paris, OECD.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Shaw, C. (2004) *Vietnam reading and mathematics achievement study*. Hanoi: World Bank, retrieved from http://siteresources.worldbank.org/INTEAPREGTOPEDUCATION/Resources/0-Preface_v1.pdf.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, Volume 57, No 1.

Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Wright, B. D., & Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement, 1*, 83-106.

Wu, M., Adams, R. J., & Wilson, M. (1998). *ConQuest: Generalised item response modelling software*. Melbourne: ACER