THE UNIVERSITY of ADELAIDE

Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines

Wilham M. Hailaya

This thesis is submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

School of Education
Faculty of the Professions
University of Adelaide

September 2014

Table of Contents

List of Tables

List of Figures

Abstract

This study examined teachers' assessment literacy and its probable impact on student achievement and aptitude (the outcome variables) through the intervening variables at the teacher and student levels. It likewise explored the effects of demographic variables on factors at the two levels and on the outcome variables. The study had 582 teacher samples and 2,077 student samples taken from Grade Six, Second Year and Third Year high school classes in the province of Tawi-Tawi, Philippines. It employed a mixed-methods design using quantitative method as a primary approach and qualitative method as a supporting approach. It utilised a number of statistical techniques, including Rasch modeling, structural equation modeling and hierarchical linear modeling, thematic analysis, and through the use of a number of software applications and include SPSS 16.0, LISREL 8.80, and HLM 6.08 to analyse the data.

The results revealed that the elementary and secondary school teachers in Tawi-Tawi, Philippines possessed relatively low assessment literacy. In terms of the specific assessment areas, the teachers performed highest on "choosing assessment methods appropriate for instructional decisions" and lowest on "developing assessment methods appropriate for instructional decisions". The qualitative finding concerning teachers' knowledge on validity and reliability supported the low assessment literacy results. Moreover, teachers generally indicated that they practised "assessment purpose", "assessment design", and "assessment communication" frequently, and "direct transmission method" and "alternative approach" of teaching in more than half of their lessons. Furthermore, the Grade Six, Second Year, and Fourth Year high school students generally exhibited positive "perceptions of assessment" and positive "attitude towards assessment". Besides, the Grade Six and Second Year high school students obtained below average "academic achievement", and Fourth Year high school students obtained below average "aptitude".

The results further revealed that teachers' assessment literacy negatively influenced their teaching practices while their assessment practices positively impacted on their teaching practices. No relationship was evident between their assessment literacy and assessment practices. However, analysis of

relevant sub-variables showed some degree of positive effect of assessment literacy on assessment practices. Additionally, the students' "perceptions of assessment" appeared to positively influence their "attitude towards assessment". The Grade Six and Second Year high school students' "perceptions of assessment" and "attitude towards assessment" likewise showed significant positive effects on their "academic achievement". The Fourth Year high school students' "perceptions of assessment" and "attitude towards assessment" exerted negative and positive effects, respectively, on their "aptitude".

Some demographic factors had moderating effects on the variables tested. Teachers' age range (60 years and above), school type, and gender appeared to moderate effects on "academic achievement" while teachers' age range (below 25 years), academic qualification, and years of teaching experience (16-20 years) had moderating effects on "aptitude".

The study's results generally serve as empirical evidence and additional information on in-service teachers' assessment literacy and its relations with other relevant variables. The results have implications for further research using other contextual variables and for the formulation of relevant policies, launching of assessment reform, development of assessment and research programs, and re-examination of assessment component of the Licensure Examination for Teachers. Furthermore, the findings in this study are relevant to pre-service teacher education programs and professional development of elementary and secondary school teachers, especially those from rural communities like Tawi-Tawi in the Philippines.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signature: ……Date: 15 September 2014

Acknowledgements

The contributions of a number of institutions and the many people have been so instrumental in the accomplishment of my PhD research and in the successful completion of my PhD program. For these, I'm deeply indebted that words of thanks are not sufficed to express my gratitude to them. Indeed, they greatly helped me achieve a milestone that should have been impossible without their invaluable support, advice, and guidance. In particular, I'm incredibly grateful to the following:

The Australian Government, through the Australian Agency for International Development (AusAID) and the Philippines-Australia Human Resource and Organisational Development Facility, for the Australian Leadership Awards (ALA) Scholarship, which made my PhD study in Australia possible. The scholarship afforded me a great opportunity to pursue the highest degree in my field, and to be more contributory to the future development of my community/country;

Associate Professor Sivakumar Alagumalai, my principal supervisor, for seeing me through in my entire PhD journey. The road was so bumpy that without his guidance the journey could have certainly been a failure. Moreover, his many pieces of advice beyond my PhD work served to enlighten me on the way I conduct myself towards my profession and life in general. These provided essential lessons that when combined with my PhD experiences constitute a truly meaningful journey. A well-principled educator, he is indeed a role model that deserves emulation;

Dr. I Gusti Darmawan, the former postgraduate research coordinator of the University of Adelaide's School of Education and my co-supervisor, for his valuable advice, especially on the methodology section of my research study. His insightful suggestions contributed a lot to the success of this research work;

Dr. Francisco Ben, also my co-supervisor, who had been so kind and patient in entertaining my concerns even beyond his official consultation time. His guidance in the statistical treatment of my data using specialised statistical software and in organising some chapters in my thesis had been enormously helpful. Moreover, his family had been so generous in extending help whenever I needed it during my stay

in Adelaide. They have been my family away from home for which I'll be forever grateful. Francis, Ivee, and Nikolai, you are all part of my success!

The Philippine Department of Education (DepEd) Secretary, Br. Armin B. Luistro, through his Undersecretary Rizalino D. Rivera of the DepEd National Office, former Regional Secretary Atty. Baratucal L. Caudang of the DepEd Regional Office in the Autonomous Region in Muslim Mindanao (ARMM), and Superintendent Dr. Kiram Irilis of the DepEd Tawi-Tawi Division, for the permission to administer my research study in the Division of Tawi-Tawi, Philippines and to access the National Achievement Test (NAT) and the National Career Assessment Examination (NCAE) results;

The DepEd National Educational Testing and Research Center under the directorship of Dr. Nelia V. Benito for the references/information in relation to NAT and NCAE;

The Mindanao State University Tawi-Tawi College of Technology and Oceanography (MSU-TCTO), my home university, for my study leave and the permission to make use of its Science High School, Preparatory High School, and Laboratory Elementary School as the pilot schools for my research study, and to administer my research questionnaires to its different community high schools;

The Tawi-Tawi State Agricultural College, Mahardika Institute of Technology, and Abubakar Computer Learning Foundation, Inc. for their permission to conduct my research study in their respective secondary and elementary schools;

Dr. Craig A. Mertler, currently the Dean of the Ross College of Education at Lynn University, Florida USA, for the permission to use the Assessment Literacy Inventory and for the needed literature on their instrument;

Prof. Eddie M. Alih, our former chancellor at the MSU-TCTO and under whose leadership I received the ALA Scholarship, for his unwavering support and permission to go on study leave, and for his constant encouragement to grow professionally; Sir Ed, I am also grateful to you for what I am now;

Atty. Lorenzo R. Reyes, our current chancellor at the MSU-TCTO, also for his support and encouragement to complete my PhD study;

Prof. Almuzrin B. Jubaira, my poet-friend whose life is poetry and whose poetry is life, for his appreciation, constant encouragement, and great ideas;

Prof. Felisa B. Halun, our vice-chancellor for academic affairs, and Dr. Elvinia Alivio, our Dean at the College of Education of the MSU-TCTO, for their constant advice to pursue a PhD;

Prof. Clarita A. Taup, Prof. Lucita R. Galarosa, and Prof. Manuel G. Pon of the MSU-TCTO for the validation of my instruments; Mr. Ibba Asakil of the MSU-TCTO Secondary Education Department for his help in coordinating the schools under their jurisdiction;

Mr. Noor Saada, the ARMM Undersecretary, Mr. Marjuni Maddi, the Assistant to the ARMM Regional Governor, and Dr. Abdurizal Aripin, Division Supervisor at the DepEd-Tawi-Tawi Division, for their help in facilitating and channeling the DepEd permissions at the regional and divisional levels;

Mr. Atari A. Idjiran, the DepEd District Supervisor of East South Ubian District, for his support during the conduct of the study in his district and in the municipalities of Bongao and Panglima Sugala;

Mr. Mohammad Nur Tidal, the principal of the Tawi-Tawi School of Arts and Trades, for the 2010-2011 NAT and NCAE results of the Tawi-Tawi Division;

Atty. Anwar Ito, Education Specialist III at the DepEd National Office, and Mr. Rajis Abdulwahid, the DepEd Tawi-Tawi Division Administrative Officer, for the DepEd demographic data;

Mr. Abdulwahid S. Dawang & Mrs. Nena Y. Dawang of DepEd-Tawi-Tawi division for their precious time and effort in arranging some meetings with school officials and in coordinating with some schools in Bongao and other municipalities;

Mr. Mohammad Jalam Eraham of the Tawi-Tawi School of Arts and Trades, Mr. Abdulnadi B. Hailaya, my uncle, Mr. Widin M. Hailaya, my brother, and Mr. Jansal Abdulpatta, my brother-in-law, for accompanying me to the different islands and for assisting me during my data collection;

Mr. Ricky Mohammadsali, principal of Simalak Elementary School; Mr. Elwan Matanio, principal of Likud Tabawan Elementary School; Mr. William Baird of the MSU-TCTO; Mr. Rio K. Hailaya, my uncle; Mr. Wilson M. Hailaya, my brother; Mr. Nursirim Kalim and Mr. Nur Perong of Tawi-Tawi School of Arts and

Tradesl; Ms. Jurifatol S. Huglay, Ms. Elenda Sahilaja, and Ms. Friselma Demsio of the Abubakar Computer Learning Center, Inc.; Mr. Alhajan Ellehero of the South Ubian National High School; Mr. Nashier Patani of the Ligayan National High School; Mr. Jaymar Gummoh, Mr. Bernasi Bernabi, and Mr. Alham Abdulhatam of the MSU-TCTO Tabawan Community High School; Mr. Nijal Kausad, Mr. Hamran Sairuna, and Mr. Alnajin Najalin of the Notre Dame of Tabawan High School; Mr. Herman Elemero of Lawm Tabawan Central Elementary School; Mr. Hahmin K. Beljium of West Tabawan Elementary School; Mr. Khalid G. Muyong of Talisay Elementary School; Ms. Rajima M. Sappayani of Datu Jaafar Central Elementary School; Mr. Mohammad Region Laison of Sipangkot National High School; Mr. Faiser Launion of Tandubas National High School; Ms. Kalsum Telso of the Sanga-Sanga National High School; Mr. Saupi Kalbi of the Tawi-Tawi West Coast Agricultural High School; Mr. Gabra Buhari, Mr. Kasmal Saraie, and Ms. Samsura Buhari of Mantabuan, Tawi-Tawi; Mr. Mark-Ben Francisco of Simunul, Tawi-Tawi; and Mr. Alpirin Julpati of Tabawan, Tawi-Tawi, for their help in the distribution and collection of the research questionnaires;

All district supervisors, directors/directress, and principals of the elementary and secondary schools in the province of Tawi-Tawi for having facilitated the administration of my survey questionnaires to their respective teachers and students, and my interview to their selected teachers;

All Grade 6, Second Year and Fourth Year high school teachers and students during the school year 2010-2011 in the province of Tawi-Tawi for serving as respondents to my study;

The concerned teachers and students of the MSU-TCTO Science High School, Preparatory High School, and Laboratory Elementary School for serving as pilot participants of my study;

My parents and my entire family for the moral support and for constantly praying for my success; and especially to my wife, Riddang, for the unwavering love, constant support and encouragement, and for standing beside me during the darkest moments of my PhD journey;

And finally, the Almighty ***Allah*** for the grace and blessings, and for the strength and determination to make this PhD thesis a reality.

Wilham M. Hailaya

Chapter 1: The Problem and Its Scope

1.1 Introduction

Countries have carried out measures to improve and maintain quality of education as an acknowledgement to its impact on society and individual development. Part of these efforts has been to undertake relevant studies to help examine and address important areas of concern. As a result, educational researchers have investigated a multitude of factors. Of these factors, teachers appear to be of utmost importance.

Teachers are regarded as one of the agents who can potentially contribute to the enhancement of the quality of education, and consequently to the attainment of the avowed educational goal – the desirable student outcomes. Their potential to influence the success of any educational effort stems from their role as the direct participants in the educational process (Ornstein, 1973). In fact, they are at the forefront in providing activities that could help enhance student learning, thus providing students the opportunities to develop skills that they need to thrive in the future. Hence, by virtue of this professional role, they are in a position to help make education more efficient and effective (Churchill, Ferguson, Godinho, Johnson, Keddie, Letts, Mackay, McGill, Moss, Nagel, Nicholson, and Vick, 2011; OECD, 2005).

Because of teachers' possible influence in the successful execution of the educational process and in the improvement of education quality, it has been stressed that they need to possess attributes that make them more capable in performing their teaching role and in bringing about student learning (Gonczi, Hager, and Oliver, 1990, as cited in Marsh, 2008; 2010). Two of these attributes are having sound knowledge of, and appropriate application of educational concepts that include assessment (Churchill, et al., 2011). Thus, it is vital for teachers to be assessment literate (Stiggins, 1991a; Schafer, 1993; Popham, 2009).

Experts have underscored the importance of assessment literacy due to its key role in the teaching-learning process and on teachers' professional competence. Specifically, the need for assessment literacy

arises from the role that assessment plays in the educational process (Popham, 2009). It has been acknowledged that assessment has the potential to support and improve teaching and learning (Brookhart, 1999; Pellegrino, Chudowsky, & Glaser, 2001). Hence, it needs to be properly executed to maximise and fully benefit from its potential. The success for the assessment to be appropriately applied requires relevant expertise from its users and implementers – the teachers. Moreover, as assessment is an important component of the educational process, teachers carry out assessment in the classroom. Stiggins and Conklin (1992) estimated that up to 50% of teachers' instructional time is spent in carrying out assessment-related activities. This significant time that teachers spend in doing assessment make it necessary for them to be familiar with the relevant concepts and skills so that they are in a position to integrate assessment with instruction (McMillan, 2000) and increase student achievement (Stiggins, 2002; Black & William, 1998a; 1998b). Furthermore, it is inevitable for teachers to employ assessment, as it is part of their professional responsibilities (Mertler, 2003; 2005; Popham, 2009). Teachers have the accountability to establish and improve student learning through assessment and it is incumbent upon them to use appropriate means and provide evidence of that learning (Phye, 1997). For teachers to be able to ascertain and enhance student learning, they ought to be competent in the area of student assessment [American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), and National Education Association (NEA), 1990].

However, despite the importance of assessment literacy in the education of students and the professional competence of teachers, research studies on teachers' assessment literacy have been less widespread than expected (Leighton, Gokiert, Cor, & Heffernan, 2010). In the Philippines, and specifically in the province of Tawi-Tawi, a study of this kind among in-service teachers has not been conducted. While there have been education reforms launched to improve the quality of education in the country, efforts to directly assess teachers' assessment knowledge and skills to provide the basis for supporting and improving teachers' assessment capabilities have not been given attention. It was for these reasons that this study was conceived.

This chapter focuses on the problem and its scope that this study intended to cover and examine. To provide the background and introduce the problem, the relevant topics are presented under the 'statement of the problem' section. In the succeeding sections, the general research questions are presented to provide an idea on the issues that this study attempted to investigate. The aims/objectives and significance of the study are likewise provided to justify the rationale and relevance of this research undertaking. To help delineate the scope, the coverage and limitations of the study are also described. A summary is presented at the end to highlight the key points discussed in the chapter.

1.2 Statement of the Problem

This section provides the overview of the education systems, the state of education, the relevant reforms, and the shortcomings of the educational assessment programs in the Philippines. It generally presents the context and highlights the problem that provided the rationale for this study.

1.2.1 Overview of the Philippine Education Systems

The 1987 Philippine Constitution provides a basis for the provision of education in the Philippines. It is stated in Section 2 of its Article XIV that "the State shall establish, maintain, and support a complete, adequate, and integrated system of education relevant to the needs of the people and society" (Philippine Constitution, 1987, p. 43). This mandate provided the grounds for the education systems in the country.

The Philippine education system generally includes the basic and higher education. These two systems are under the management of the Department of Education (DepEd) and the Commission on Higher Education (CHED), respectively. The technical and vocational education, which is under the higher education, is separately managed by the Technical Education and Skills Development Authority (TESDA). (SEAMEO RIHED, 2011; UNESCO-IBE, 2011; Lapus, 2008). The country's education setup is patterned after the education system of the United States [Senate Economic Planning Office (SEPO), 2011]. The structure of the Philippine education system, reflecting the newly implemented basic education program, is shown in Figure 1.1.

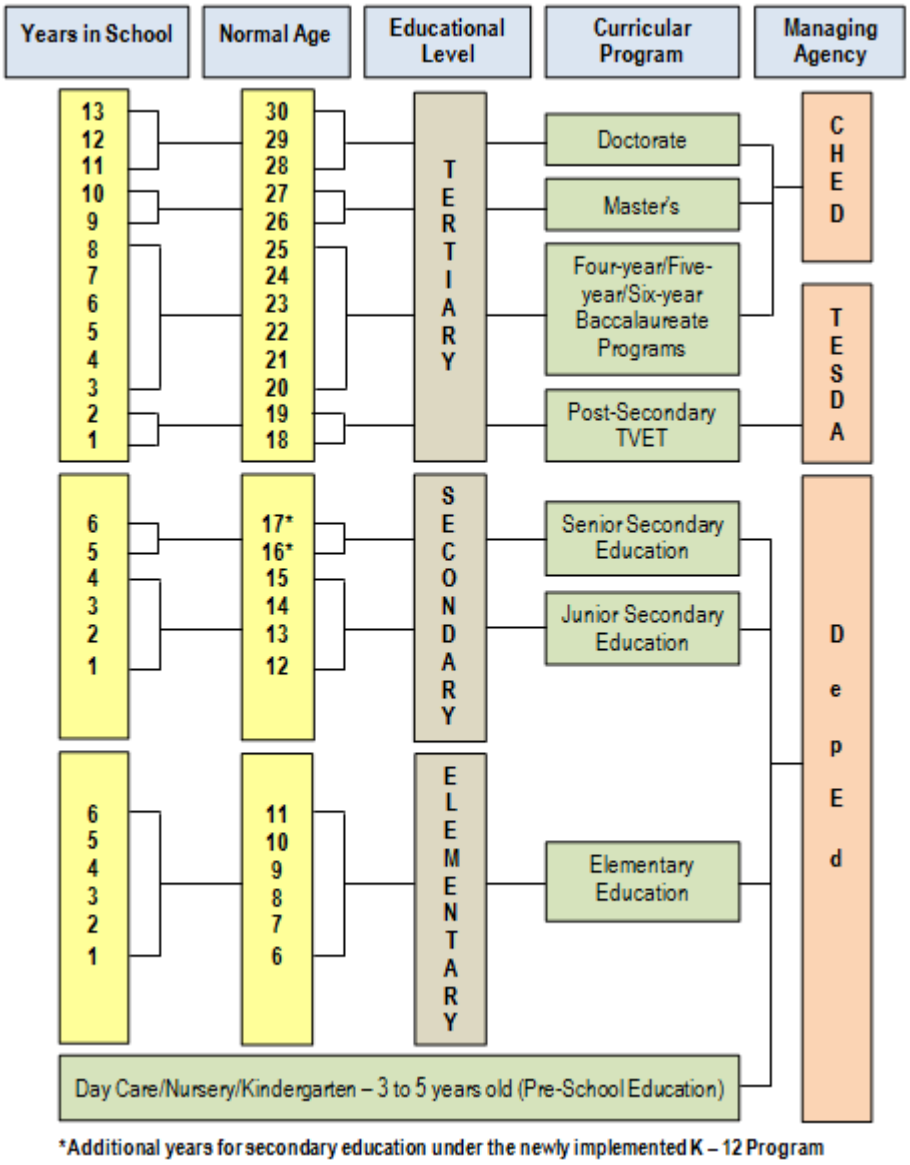


Figure 1.1. The Philippine Education System
(adapted from Ben, 2010; Syjuco, n.d.; & Luistro, 2012)

The basic education system comprises the elementary and secondary education levels. Currently, the country is in the transition of changing its basic education program from 10 - year to K + 12 - year education cycle. It has just implemented the enhanced K + 12 basic education program, which is expected to be fully operational by School Year (S.Y.) 2017-2018. The basic education under the newly adopted program has its goal defined by Republic Act (R.A.) No. 10533, otherwise known as the "Enhanced Basic Education Act of 2012". It serves to “develop productive and responsible citizens equipped with the

essential competencies, skills and values for both life-long learning and employment" (R. A. 10533, p. 1; DepEd, 2010, p. 7).

Under the old basic education program, the elementary and secondary levels lasted for six years and four years, respectively, covering a total of ten years. This number of school years has increased under the new program. Specifically, the K + 12 basic education program has officially included the kindergarten level and adds two more years in the secondary level. The newly adopted education model is K-6-4-2, which means that the program covers one year of kindergarten education, six years of elementary education (Grades 1-6), four years of junior high school (Grades 7-10), and two years of senior high school (Grades 11-12). The official school age is five years old for the kindergarten stage, 6 to 11 years old for the elementary grade, 12 to 15 years old for junior high school level, and 16 to 17 years old for senior high school level. The program is to be implemented in all public (government-funded) and private (privately-funded) schools, the existing major school type in the country. Moreover, the basic education is made compulsory and provided free in public schools (SEPO, 2011; DepEd, 2010; R. A. No. 10533).

As part of the system accountability to monitor and gauge learning outcomes, and to foster quality education in the country, the DepEd, through the National Educational Testing and Research Center (NETRC), administers the annual national tests nationwide (DepEd Order No. 5, s. 2005). The tests are called the National Achievement Test (NAT) and the National Career Assessment Examination (NCAE). The NAT is administered to determine the achievement level of the pupils/students while the NCAE is to assess the skills of high school students who intend to pursue post-secondary courses and other career options. Although annually conducted in all public and private basic education schools at the end of the school year, these examinations are not generally fixed to any particular grade. For instance, the elementary NAT was first administered to Grade 4 in S. Y. 2003-2004. However, it was conducted to Grade 6 beginning in S. Y. 2004-2005. Moreover, the NAT for the secondary level was conducted to Fourth Year in S. Y. 2003-2004 and 2005-2006. But from S. Y. 2006-2007 it has been conducted to Second Year (SEPO,

2011). The NCAE has been conducted to Fourth Year from S.Y. 2006-2007 to S.Y. 2010-2011. It has also been given to Third Year starting in S.Y. 2011-2012 (DepEd-NETRC, 2013).

Students can choose from two pathways after completing their basic education. They can either continue a post-secondary course under the technical-vocational education and training (TVET) program, which is overseen by TESDA, or continue their tertiary education, which is managed by CHED.

The TVET, which is also offered for secondary school leavers, college/university graduates and undergraduates who are interested to gain competencies in different vocational fields, and the unemployed clients who are looking for jobs, aims to equip students with technical/vocational skills that offer the chances for immediate employment. In this program, there are four basic ways of training. The first mode is school-based or formal delivery in which TVET programs are offered from one to three years in schools. The next mode is called Center-based in which short-term non-formal trainings are delivered in TESDA regional and provincial training centers. The third mode is community-based in which skills training programs that facilitate self-employment are done right in the community. The last mode involves enterprise-based programs that are carried out within the firms or industries (UNESCO-IBE, 2011; Syjuco, n.d.).

The higher education is the highest level of the education system. It generally aims to empower and make Filipinos globally competitive (SEAMEO RIHED, 2011). It offers opportunities for high school graduates and professionals to continue and enrich their academic studies through the offering of a wide range of academic programs and specialisations, including the pre-service teacher education programs. These degree programs are handled by public and private higher education institutions across the country (CHED, 2010).

1.2.2 The State of Basic Education in the Philippines

The state of Philippine basic education has been continuously assessed as part of the efforts to check whether the education system has been successful in its goal of providing quality education. This was also to provide the basis for any measures to be taken to address any relevant weaknesses and to

continue to adopt supportive programs and policies. To this end, a number of performance indicators have been examined and the results mostly revealed that education in the country is of poor quality (SEPO, 2011; Miralao, 2004; Lapus, 2008). One prominent gauge that can be cited is the students' academic performance, which is a direct indication of education quality. Tables 1.1 and 1.2 below provide specific data pertaining to the academic achievement of these pupils/students.

Table 1.1. The NAT achievement rates in MPS of Grade 6, Second Year and Fourth Year high school students in S.Y. 2006-2010

Grade/Year Level	NAT Achievement Rate in MPS					
	S.Y. 2005-2006	S.Y. 2006-2007	S.Y. 2007-2008	S.Y. 2008-2009	S.Y. 2009-2010	S.Y. 2010-2011
Grade 6	54.7	59.9	64.8	65.6	68.0	68.2
2nd Year		46.6	49.3	46.7	45.6	47.9
4th Year	44.3					

(Sources: Luistro, 2012; SEPO, 2011)

Table 1.1 shows the levels of academic achievement of the elementary school pupils and high school students in the six consecutive years. The achievement levels are in terms of the mean percentage scores (MPS), which indicate the percentages of correctly answered items in the NAT. As can be gleaned from the table, Grade 6 pupils obtained low achievement levels in the entire six-year period. It appears that there was a continuous increase in the elementary NAT student performances, which was a promising trend. However, the results were still below the acceptable level. In fact, their highest MPS, 68.2 in S. Y. 2010-2011, was well below the minimum competency level of 75% indicating that remedial measures need to be implemented. As for the high school students, the NAT results were even lower. Fourth year high school students obtained an MPS of 44.3 that was far below the acceptable minimum level. Moreover, the Second Year high school students exhibited consistently low and fluctuating results in the period of five years. Their highest MPS of 49.3 in S. Y. 2007-2008 was likewise far below the acceptable level of 75%, further suggesting that efforts to improve student performances are warranted (SEPO, 2011).

The academic standing of Filipino students can also be understood from the international tests such as the Trends in International Mathematics and Science Study (TIMSS). Table 1.2 provides the 2003 and 2008 TIMSS data pertaining to the performances of Filipino students in science and mathematics. As shown, similar picture of poor academic standing appeared. The Filipino students' scores on the tested subject areas were all below the international averages, indicating that they were outperformed by many students from the other participating countries. The Philippines ranked fifth and fourth from the bottom in the Second Year high school mathematics and science, respectively, and third from the bottom in both Grade Four mathematics and science in the 2003 TIMSS (SEPO, 2011; Gonzales, Guzman, Partelow, Pahlke, Jocelyn, Kastberg, & Williams, 2004; Maligalig & Albert, 2008). In the 2008 TIMSS results, the country ranked last (10th out of 10 participating countries) in Advanced Mathematics despite the participation of science high schools, the country's best and brightest (SEPO, 2011). These TIMSS results provide further indication that the quality of education in the country is poor and is far below the global or regional standard. Thus, relevant measures are needed.

Table 1.2. Science and Mathematics scores of Filipino students in the 2003 and 2008 TIMSS

Grade/Year Level/School	2003 Results					2008 Results
	Subject Area	Score	International Average	Rank	Number of Participating Countries	Score
Grade 4	Science	332	489	23rd	25	
	Mathematics	358	495	23rd	25	
2nd year high school	Science	377	473	43rd	46	
	Mathematics	378	466	34th	38	
Science High Schools	Advanced Mathematics		500	10th	10	355

(Adapted from SEPO, 2011; DepEd, 2010)

In the southernmost province of Tawi-Tawi, the state of education was even more alarming. The DepEd record showed that the province joined the rest of the areas in the Autonomous Region in Muslim Mindanao (ARMM) in having the most deteriorating basic education in the country. The 2007 Regional Assessment in Mathematics, Science and English (RAMSE) test, conducted under the Philippines-Australia

Basic Education Assistance for Mindanao (BEAM) Project, indicated that pupils or students in the ARMM did not only fail to reach the required minimum mastery level, but also had the difficulty in answering the items requiring higher order thinking skills (Philippine Human Development Report, 2008/2009). This is supported by the report that most schools in the ARMM areas were the least performers in the NAT. Tawi-Tawi, specifically, has been consistently cited as one of the country's least performing provinces in the primary and secondary NAT from 2003 to 2007 (Maligalig, Caoli-Rodriguez, Martinez, & Cuevas, 2010). Again, this reflects the poor education quality that needs to be addressed.

1.2.3 Relevant Education Reforms

The 1987 Philippine Constitution does not only mandate the provision of education but ensures the right of every Filipino to have access to quality education. Section 1 of its Article XIV provides that "the State shall protect and promote the right of all citizens to quality education at all levels, and shall take appropriate steps to make such education accessible to all" (Philippine Constitution, 1987, p. 43). Consistent with this mandate, a number of actions and/or reforms have been launched. For instance, the country in 1990 joined the global movement on 'Education For All (EFA)' that envisions to improve the quality of basic education and make every Filipino functionally literate by 2015 (Lapus, 2008). Also, the management of education system was *trifocalised* in 1994 to allow for improved governance. The control, regulation, and supervision of the basic, technical-vocational, and higher education that were used to be under DepEd were now handled by three agencies: DepEd (Republic Act No. 9155 – Governance of Basic Education Act of 2001; DepEd Order No. 1, s. 2003) for basic education, TESDA (Republic Act No. 7796 - Technical Education and Skills Development Act of 1994) for technical-vocational education, and CHED (Republic Act No. 7722 - Higher Education Act of 1994) for higher education (SEAMEO RIHED, 2011; UNESCO-IBE, 2011; Lapus, 2008; de Guzman, 2003).

To help realise the EFA, the Basic Education Sector Reform Agenda (BESRA) was likewise formulated. The BESRA provides consistent policy actions developed under the five key reform thrusts namely (Lapus, 2008, p. 12):

> KRT1- Continuous school improvement facilitated by active involvement of stakeholders through School-Based Management (SBM); KRT2 - Better learning outcomes improved by teacher standards; KRT3 - Desired learning outcomes enhanced by national learning strategies, multi-sectoral coordination, and quality assurance; KRT4 - Improved impact on outcomes resulting from complementary services for early childhood education, alternative learning services, and private sector participation; and KRT5 - institutional culture change within DepEd to facilitate attainment of the first four thrusts.

Following the goals/reform thrusts as embodied in the EFA and BESRA and to address the poor state of education in the country, the DepEd has implemented a number of policies and programs on key relevant aspects, such as curriculum, teaching, and assessment. In 2002, it adopted the 'Basic Education Curriculum (BEC)' (DepEd Order No. 43, s. 2002) as the new national curriculum for the elementary and secondary levels. The BEC sets the minimum learning competencies that the students are expected to achieve, and the teaching standards that basic education teachers are expected to adhere. It embraces constructivism approach in teaching and learning (Lapus, 2008). Included in the guidelines on the implementation of BEC learning areas was the grading system and reporting of pupil performance, which was amended, upon the enforcement of performance-based grading system in S. Y. 2004-2005. The amendments included the setting of performance standard at 75%; the use of a table of equivalence (direct percentage weights) in converting scores into grades; test construction/design to include 60% easy items for basic content and skills, 30% medium-level items for higher level skills, and 10% difficult items for desirable content or skills that aim to determine the fast learners; the use of the averaging method in determining the final marks/grades; and the use of rubrics and non-traditional assessments (open-ended questions, performance and portfolio assessments) to complement the long-standing traditional

assessments (tests of different types) (DepEd Order No. 79, s. 2003; DepEd Order No. 04, s. 2004; DepEd Order No. 33, s. 2004). Moreover, the National Competency-Based Teacher Standards (NCBTS) were adopted (DepEd Order No. 32, s. 2009). The NCBTS is an integrated framework that defines standards for effective teaching. It was developed to help guide and improve teachers in their professional activities. The NCBTS has seven domains each of which has specific strands. One of these domains pertains to planning, assessing and reporting. This domain is defined as follows (DepEd, 2006, p. 32):

> The domain of Planning, Assessing and Reporting refers to the aligned use of assessment and planning activities to ensure that the teaching-learning activities are maximally appropriate to the students' current knowledge and learning levels. In particular, the domain focuses on the use of assessment data to plan and revise teaching-learning plans, as well as the integration of formative assessment procedures in the plan and implementation of teaching-learning activities.

In addition, this domain has three specific strands. The first strand requires teachers to communicate promptly and clearly to learners, parents, and superiors about the progress of the learners; the second strand expects teachers to develop and use a variety of appropriate assessment strategies to monitor and evaluate learning; and the last strand obliges teachers to regularly monitor and provide feedback on the learners' understanding of content (Ballada, 2013; DepEd, 2006).

Just recently, the DepEd has implemented the K + 12 program as described earlier in this chapter. This program aims to help increase the quality of basic education in the country. It offers a decongested curriculum, prepares students for higher education and labor market, and conforms to the global standards of at least 12-year basic education cycle (SEPO, 2011; DepEd, 2010). It adopts its own version of BEC containing the restructured/renamed core learning areas and a new set of assessment guidelines. On the assessment and rating of student learning, it requires a holistic approach. It prescribes that students are assessed on four levels which shall be weighted as follows: knowledge – 15%; process or skills – 25%; understanding – 30%; and products/performances – 30% with a total of 100% (DepEd Order No. 31, s. 2012).

To complement the reforms on basic education and to help ensure qualified teaching force, the CHED has improved and strengthened the pre-service teacher education programs. Beginning in S. Y. 2005-2006, the undergraduate teacher education curriculum prescribes the offering of the Bachelor of Elementary Education (BEEd) and the Bachelor of Secondary Education (BSEd) courses. The BEEd program aims to develop students to become elementary school teachers who are either generalists who can teach different learning areas in grade level, special education teachers who can teach children with special needs, or pre-school teachers who can handle children at the nursery or kindergarten level. It has been "structured to meet the needs of professional teachers for elementary schools and special education programs in the Philippines". Moreover, the BSEd program seeks to develop students to become high school teachers who can teach one of the prescribed specialisation subjects. It has been structured to meet the needs of professional teachers for secondary schools in the country (CMO No. 30, s. 2004, p. 2).

Some of the subject areas prescribed under the BEEd and BSEd programs have been expanded. One of these is the assessment subject. Under the old curriculum, only one assessment subject that focused on testing was offered (CMO No. 11, s. 1999). In the current curricular offerings, this subject has been increased to three in response to the assessment requirements prescribed by DepEd in its basic education curricula and to enhance teachers' understanding and application of both traditional and non-traditional assessments in the classroom. These assessment subject areas, which fall under the Methods and Strategies Courses (Professional Education Courses), are called "Assessment of Student Learning 1", which focuses on traditional assessment (pen-and-paper assessment), and "Assessment of Student Learning 2", which focuses on alternative assessment. In addition, there is a one-unit field of study course that also focuses on learning assessment (CMO No. 30, s. 2004; Balagtas et. al., 2010).

Under the recent major reforms, teacher development has been considered as one of the imperatives. It was recognised that teachers can significantly contribute to the enhancement of student learning and the attainment of the desired education standard. As such and because of the nature of their role, teachers have become the essential focus of the development efforts (Human Development Network,

2008; Lapus, 2008; Caoli-Rodriguez, 2007; Maligalig, et al., 2010). Thus, programs on in-service training for teachers have been strengthened and more professional development activities have been conducted. In fact, for every new curriculum that the DepEd intends to adopt, teachers, who are the direct personnel involved in the implementation, have always been subjected to undergo relevant training to make them ready for the curriculum rollout. Moreover, classroom assessment has also been articulated as one of the competency areas that teachers need to possess. Assessment, specifically student assessment, has been recognised as an essential part of the curriculum that needs to be given more attention to help improve the learning quality of the students. With the implementation of the new basic education and undergraduate teacher education curricula, there has been an expansion in the assessment focus to include now the use of the alternative assessment methods. As a result, the DepEd has conducted a massive training on classroom assessment and other curriculum areas to all basic education teachers in the country to prepare and make them all set for the reform implementation.

However, despite all these efforts, there are a number of relevant shortcomings in the assessment programs and trainings. First, it has been observed that most teachers completed their degrees during the era when only one assessment course that focused on testing was offered. They are more familiar with it and have long been used to the traditional pencil-paper tests and one-shot training covering new alternative assessment methods does not seem to guarantee assessment competency. Second, most DepEd orders pertaining to classroom/student assessment lack the necessary details to guide teachers in the assessment implementation (Ballada, 2013). Because of this, teachers tend to implement assessment according to their own interpretations. Third, the NCBTS assessment domain has been found as not comprehensive. It does not provide specific assessment competencies that teachers need to possess in order to successfully carry out assessment-related activities (Ballada, 2013). As Magno (2013, p. 46) stressed, the specific assessment strands of NCBTS “do not capture the ideals of assessment literacy”. Fourth, many teacher education institutions still offer one assessment subject that focuses only on testing despite the CHED 2004 curricular revision (Balagtas, et al., 2010). Fifth, while student assessment has been articulated as one of

the important key areas to be addressed, a standalone assessment reform that includes comprehensive assessment standards for teachers does not currently exist. Sixth, empirical evidence on teachers' assessment literacy to base assessment trainings/interventions has been unavailable. Finally, the absence of research on teacher assessment literacy has left teachers' assessment knowledge and practices unchecked and unprioritised. In view of these shortcomings and in view of the need to make teachers competent in the area of classroom or student assessment, a relevant undertaking is highly warranted.

In the province of Tawi-Tawi, where the quality of basic education has been in dismal condition, effective capability building on student assessment is vital, and any effort that helps pinpoint relevant teachers' needs is equally important. Hence, this study has been conceived in response to these needs and shortcomings.

1.3 Research Questions

This study generally attempted to investigate teacher assessment literacy and its influence on student outcomes in the province of Tawi-Tawi, Philippines. In the main, it sought to answer the following questions:

1. What is the assessment literacy of the elementary and secondary school teachers?
2. How does teacher assessment literacy impact on student outcomes through the intervening variables at the teacher and student levels?
3. How do the demographic variables affect factors at the teacher and student levels and ultimately on the outcome variables?

1.4 Aims/Objectives of the Study

As initially reflected in Section 1.3, the study aimed to examine teachers' assessment literacy and its possible influence on student outcomes through mediating variables at the teacher and student levels. Specifically, it intended to investigate the impact of teacher assessment literacy on student academic achievement and aptitude through assessment practices, teaching practices, assessment perceptions, and

assessment attitude. Additionally, it sought to explore the effects of demographic variables such as gender, age range, academic qualification, years of teaching experience, and school type on factors at the two levels and ultimately on academic achievement and aptitude. This research undertaking was likewise conducted to provide empirical evidence and/or relevant information on in-service teachers' assessment literacy and its relations with related variables to support the development of related programs and the formulation of pertinent policies at the local, regional, and national levels of education system. Furthermore, the study was administered to contribute to any assessment reform, to develop and increase the capabilities of local teachers especially on classroom/student assessment, and to help improve the quality of education particularly in the province of Tawi-Tawi and generally in the Philippines.

1.5 Significance of the Study

It was described in Sub-section 1.2.2 that the state of basic education in the Philippines has been deteriorating. This is especially true to the public elementary and secondary schools in the rural areas such as those in the Autonomous Region in Muslim Mindanao in which Tawi-Tawi is a part. As a result, relevant development programs and projects have been undertaken to help address the declining quality of basic education. This research study was conducted to help and support those efforts.

Specifically, the study is expected to be of help to the Tawi-Tawi local teachers, as the results will provide them with feedback on aspects that are important in the improvement of their capabilities in student assessment and instruction, and in bringing about effective learning in the classroom. In addition, the research findings will serve as an important basis for the Philippine Department of Education at the provincial, regional and national levels to regulate relevant policies and develop assessment programs for the elementary and secondary school teachers. This will pave the way for the formulation of professional development programs on classroom assessment and launching of comprehensive assessment reforms. The study's results will likewise serve as basis for the Philippine Professional Regulation Commission to revisit the assessment component of the Licensure Examination for Teachers. Moreover, the outcomes of

this research will also be a valuable reference for teacher education institutions in the region and the country in developing standard pre-service teacher education assessment courses and in designing in-service teacher assessment training, especially for teachers who come from the rural areas. Furthermore, this study will assist curriculum and textbook writers in the country in designing assessment techniques and strategies and in coming up with learning assessment tools that will help teachers authentically assess their student learning in the classroom. Finally, the findings of this study will be a vital reference to those who will do similar research studies in other parts of the country and abroad.

In terms of its contribution to the assessment discipline, the study will provide additional information on teachers' assessment literacy and its influence on student outcomes. At a time when research on teacher assessment literacy is less widespread, the study should be useful in providing better understanding of this educational issue. Particularly, the study will add value to and help fill in the gaps as identified in the literature. Firstly, it will help address the insufficient research on teacher assessment literacy (Stiggins, 1991a; 2002; Schafer, 1993) and practices (Mertler, 2003; Kennedy, Chan, Fok, & Yu, 2008) by providing more information using a different context; secondly, the study will contribute to the insufficient research linking teachers' assessment literacy with other teacher attributes and student outcomes as it attempted to establish the directional influences among these factors (Stiggins, 1991b; Stiggins, Conklin & Bridgeford, 1986); and thirdly, this project will help highlight classroom/student assessment research and teacher assessment literacy issues as a study of this kind has not been conducted in the Philippines and particularly in the province of Tawi-Tawi.

1.6 Scope and Limitations of the Study

The study's main focus is on the assessment literacy of elementary and secondary schoolteachers in the province of Tawi-Tawi, Philippines. The assessment literacy investigated in this study pertains to the knowledge and skills on basic assessment areas or principles as defined by the 'Standards for Teacher Competence in Educational Assessment of Students' (AFT, NCME, & NEA, 1990). Moreover, the study

covers the examination of any directional influences of teacher assessment literacy on student outcomes through the intervening variables at the teacher and student levels. Intervening factors at the teacher level include assessment practices that comprise assessment purpose, assessment design, and assessment communication, and teaching practices that consist of structuring activities, student orientation, and enhanced activities. Those at the student level include assessment perceptions that comprise perception towards assignment and perception towards test, and assessment attitude as a one-construct variable. While the student-level factors can be generally considered as outcomes, this study operationalises student outcomes as the tested outcome variables that include student academic achievement and aptitude. The study also involves examination of the effects of demographic variables like gender, age range, academic qualification, years of teaching experience, and school type on assessment practices, teaching practices, assessment perceptions, and assessment attitude, and ultimately on student academic achievement and aptitude. Furthermore, it involves both teacher and student participants who represented Grade Six, Second Year and Fourth Year classes of the public and private elementary and secondary schools in the province of Tawi-Tawi and who were present during S. Y. 2010-2011.

By defining the scope, it is acknowledged that this study is far from being complete and perfect. In the educational context, there are complex webs of factors and relationships that can hardly be covered in a single study. As Mullens and Kasprzyk (1999, p. 678) stated, “teaching and learning in any context is a complex human endeavor that cannot yet be adequately represented by responses to a single survey”. Needless to say, this study has a number of limitations. Firstly, the variables/constructs examined in this study were limited to assessment literacy and related factors as mentioned above. It is believed that there are other factors that can interact with teacher assessment literacy and student outcomes. However, due to time and financial constraints, the study only included the intended constructs/variables/factors. Secondly, participants were taken only from the researcher’s home province and were selected using purposive sampling. It is also acknowledged that more areas and more samples could have been covered to make this study more generalisable. However, the decision to collect data from only one area was made to ensure

the success of the research undertaking as the researcher has the familiarity of the place, and as the archipelagic nature and peace and order conditions of other areas did not warrant safe travel. And thirdly, this study has limitation due to its nature as a cross-sectional type. A study of this kind has a lesser generalisability compared to other types such as the longitudinal studies. However, it has also its own advantages of being less expensive, able to offer quick findings, and could include more samples than longitudinal designs (Creswell, 2008; Ben, 2010).

1.7 Summary

This chapter provided an introduction and set up the context of this study. It presented the issues that the study attempted to investigate. To highlight the source and the context of the issues, this chapter provided the background by introducing the education system, the state of basic education, and the education reforms in the Philippines. To bring to light the issues, the chapter briefly discussed the teacher assessment literacy, including its relevance to the teaching-learning process and its position in the context of education reforms in the country. To define the scope of the problem, the research questions, and the study's coverage and limitations were presented. The chapter also provided the relevance of this study to the research venue and assessment literature by discussing the research aims/objectives and significance. The issues presented in this chapter are described in more detail in Chapter 2 (Review of Related Literature and Studies).

Chapter 2: Review of Related Literature and Studies

2.1 Introduction

Assessment is considered one of the essential aspects of education. Its importance stems from its potential to generally influence educational success and better education quality (Black, 2011). According to Chatterji (2003), assessment is employed for numerous purposes in different levels of education system. For instance, at the systemic level, assessment is used for accountability purposes. It is utilised to ascertain the effectiveness of educational endeavors (Popham, 2009). Educational plan and activities that include reforms, programs, school performance, and the general progress of student achievement are evaluated and monitored through assessment. At the classroom level, assessment is more associated with the teaching-learning process (Hargreaves, 1997; Cizek, 1997; Shepard, 2000). It is used for purposes that are directly related to instruction and learning. Specifically, it is employed to support teaching, and to ascertain and enhance student learning, which is the overarching purpose of assessment (Shepard, 2000; Phye, 1997). However, while both system and classroom assessments are important, those at the classroom level need to be given more attention as it is where students are being developed. Stiggins (1999a; 2012) asserted that 99% of the assessments that happen in the student's life occur in the classroom, and if assessment does not work effectively in this level, it is possible that assessment at the other levels (national, regional, provincial, district) becomes irrelevant.

Stiggins' (1999a; 2012) assertion concerning classroom assessment calls for teachers to be developed and be competent in this area. As he also stressed, teachers are directly involved in the development and execution of assessments in the classroom. In fact, it has been estimated that they use 30% to 50% of their professional time to carry out assessment-related activities (Stiggins & Conklin, 1992). Thus, they need relevant competence for them to be able to do high-quality classroom assessment and to

effectively carry out instruction and maximise student learning. This requires teachers to be assessment literate (Stiggins, 1991a; 1999a;1999b; 2012; Schafer, 1993; Popham, 2009).

From what experts have stressed, assessment literacy can be viewed as one of the prerequisite attributes that make teachers in a better position to carry out their professional activities, which, in turn, affect student outcomes. As such, assessment literacy is related to other teacher and student variables such as assessment practices, teaching practices, student perceptions of assessment, student attitude towards assessment, academic achievement, and aptitude. This chapter presents the relevant literature and studies concerning teacher assessment literacy and these relevant variables.

The chapter initially discusses assessment and its role in the teaching-learning process to provide background information and the rationale for considering assessment literacy. After which, assessment literacy and relevant studies are considered in detail. Moreover, literature and studies concerning assessment practices, teaching practices, assessment perceptions, assessment attitude, academic achievement, and aptitude are discussed. The theoretical framework that supports the link among these tested variables is likewise described to provide the basis for the model adopted in this study. A summary is provided at the end to reiterate the key points addressed in this chapter.

2.2 Assessment and Its Role in the Teaching-Learning Process

The meaning of the term “assessment” is generally not quite straightforward. According to Cizek (1997), assessment has no standard usage as it is used in so many different ways and contexts and for so many different purposes. However, based on the usages of the term in the current assessment literature, he proposed four facets that should be incorporated in the assessment definition. Firstly, an assessment definition should be universally applicable to any condition, format, and context. By this, the definition should be encompassing enough to include more relevant dimensions. Secondly, a preferable definition is one that appeals to educators and enhances the position of assessment in instruction. This suggests that assessment definition should be readily acceptable and useful to teachers and/or educators. Thirdly, a

definition is preferable if it recognises that assessment serves or supports instruction. This suggests that the meaning of assessment should specify the functions of assessment in relation to teaching. And fourthly, an assessment definition should have a strong relevance to students' needs and development, suggesting that assessment definition should embody the overarching purpose of improving student learning and relevant outcomes. Incorporating these facets, he proposed the following definition for educational assessment:

> ***Assessment*** *\uh ses' ment\ (1) v.t.: the planned process of gathering and synthesising information relevant to the purposes of (a) discovering and documenting students' strengths and weaknesses, (b) planning and enhancing instruction, or (c) evaluating progress and making decisions about students. (2) n.: the process, instrument, or method used to gather the information (p. 10).*

Hence, assessment generally needs to be viewed as a planned process designed to accomplish a specific educational purpose and intended to benefit students (Cizek, 1997).

There are a number of assessment definitions or concepts that satisfy one or more facets proposed by Cizek (1997). For instance, Airasian (1994) suggested that assessment should include complete information that helps teachers understand their pupils, monitor their instruction, and establish a viable classroom culture. Marsh (2008, p. 261; 2010, p. 310) also defined assessment as the term that is "typically used to describe the activities undertaken by a teacher to obtain information about the knowledge, skills, and attitudes of students". Another definition is given by the American Federation of Teachers (AFT), National Council on Measurement & Evaluation (NCME), and National Education Association (NEA) (1990, p. 1) that describes assessment as the "process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his/her progress, strengths & weaknesses, to judge instructional effectiveness & curricular adequacy and to inform policy".

Assessment can be categorised in terms of purpose as employed in a particular level/unit or in different levels/units of education system. Generally, it can be classified into *system* and *classroom* assessments (Asaad & Hailaya, 2004). These are what other experts described as external and internal

purposes with reference to classroom setting. The system assessment embodies a set of assessments made by policy-makers in the national, regional, provincial, district, and school levels to appraise and improve curriculum, school, and education system. This is akin to what Popham (2009) called as *accountability assessments* or what are referred to as *external assessments* (Kennedy, 2007; Alagumalai & Ben, 2006). These assessments take the form of standardised devices, like the high-stake or standardised examinations, which government entities employ to ascertain the effectiveness of the educational endeavors (Popham, 2009). On the other hand, classroom assessment is what is sometimes referred to as *internal assessment.* This type of assessment encompasses those that support and enhance teaching and learning. Popham (2006, p. 6) described these assessments as "those formal and informal procedures that teachers employ in an effort to make accurate inferences about what their students know and can do." These two general types of assessment need to be complementary to each other to make assessment more contributory to the educational process and more beneficial for students (Stiggins, 2002; 2012; Alagumalai & Ben, 2006).

Classroom assessment can further be classified into many types corresponding to the different purposes of assessment in the classroom. One classification includes *assessment of learning*, *assessment for learning*, and *assessment as learning* (Earl, 2003). The *assessment of learning* encompasses the concept of summative assessment and evaluation or the use of mostly examinations to assess learning at the end of instruction. The general purpose of using this type of assessment is to gather and interpret learning evidence and to grade and report learning (Dunn & Mulvenon, 2009; Spencer, 2005; Stiggins, 2002). As this type of assessment has been used with traditional methods and tools such as tests of different types (e.g. multiple choice, true-false, matching, completion), it has sometimes been labeled as traditional or conventional assessment (Kennedy, et al., 2008). On the other hand, *assessment for learning* encompasses formative assessment and evaluation, which are used to support teaching and improve learning (Black & Wiliam, 1998a; 1998b; Assessment Reform Group, 2002). It is a recently introduced and recommended type that seeks to integrate assessment into teaching; it emphasises the use of feedback

and alternative methods/tools like observation, projects, presentations, performance activities, self-assessment and peer assessment. As this model of assessment recommends the use of new alternative strategies and tools, it has sometimes been called as non-traditional, alternative or authentic assessment. The *assessment as learning* has emerged as an offshoot of assessment for learning. This type emphasises that assessment may support and even serve as learning for students. It provides immediate learning by allowing students to discover strategies for further learning. In a sense, it makes them "learn how to learn". This is associated with self-assessment, wherein students are allowed to assess themselves for them to draw feedback about their own achievement and devise strategies as a result of that feedback for them to gain more learning (Kennedy, et al., 2008, p. 201).

Other concepts/terminologies include formal and informal assessments, diagnostic assessments, and norm-referenced and criterion-referenced assessment frameworks. Formal assessments are those that are planned and structured and that are intended to elicit information about student performance or achievement. These include tests of various types that are either high-stake or teacher-made. On the other hand, informal assessments are those that are done by teachers while in the course of delivering instruction. Assessments of this type are usually intended to gather immediate information to check student learning and behaviour and to offer immediate feedback. This type includes oral recitation/classroom interaction and observation. Moreover, diagnostic assessments are those that are employed to find the strengths and weaknesses of students in relation to classroom lessons in order to provide the necessary remediation. Furthermore, the criterion-referenced and the norm-referenced assessments are employed when comparison of student performance is desired. The criterion-referenced assessment/criterion-based assessment is the framework that compares student's work to a set of defined, explicit and objective criteria (benchmarks/standards) while the norm-referenced assessment/norm-based assessment is the framework that compares student's work to a norm that is usually a set of outcomes from similar students (Churchill, et. al., 2011; Asaad & Hailaya, 2004).

There have been views about which of the classroom assessment or internal types should be practiced in the classroom. Black & Wiliam (1998a; 1998b) and the Assessment Reform Group (2002) have promoted assessment for learning because it provides greater learning gains. However, Dunn and Mulvenon (2009) and Phye (1997) have argued that each of the types can be used interchangeably depending on the purpose for which assessment is to be employed. This view is supported by other experts who see the applicability of all the models in the classroom. For instance, Taylor (1994) advanced the idea that the success of any assessment model depends on the context, saying that it even makes performance-based assessment a failure if it doesn't serve the purpose or goal of that context. Additionally, Stiggins (2002) has pointed out that for assessment to be useful in the classroom there should be a balance between assessment for learning and assessment of learning. This is further in agreement with Kennedy (2007; et al., 2008) who stressed that in a context where examination culture is prevalent, assessment of learning can be employed and integrated by using it formatively.

However, as pointed out earlier, the success of classroom assessment depends on how it contributes to teaching and learning. This entails understanding of the link of classroom assessment to the teaching-learning process. Hargreaves (1997) and Black (2004) have stressed the inextricable interplay between pedagogy, classroom assessment and learning. They argue that this link allows effective educational process to operate and occur in the classroom. Assessment plays a significant role in the relationship because it facilitates instruction and student learning (Brookhart, 1999; Pellegrino, Chudowsky, & Glaser, 2001; Shepard, 2000). As Klenowski (2008, p. 1) pointed out, "assessment is inseparable from curriculum and has become a powerful driver for change; it is at the heart of the teaching-learning dynamic."

Specifically, classroom assessment validates and enhances teachers' decision-making process (klenowski, 2008; Popham, 2009; Schafer, 1993; Stiggins, 2012). Teachers make decisions in relation to teaching and learning (Clark & Peterson, 1984). For instance, they need to decide on the learning goals consistent with the curriculum, on the subject contents and activities that need to be given emphasis, on methods and strategies to effectively deliver both the targets and the contents, on what and how to assess

student learning, on judgment whether the learning aims have been achieved, and on improvement that needs to be done in the case of the emerging flaws. The decisions involving key instructional tasks are crucial to the instructional success and therefore teachers need information or evidence to aid them in deciding soundly. These vital information or evidence can be effectively provided by sound classroom assessment (Stiggins, Conklin & Bridgeford, 1986; Black & William, 1998a; 1998b; Guskey, 2003). Moreover, quality assessment helps evaluate and improve teachers' instructional process (Stiggins, 1991a; 1991b; 2002; Plake, Impara & Fager, 1993; Black & William, 1998a; 1998b; Assessment Reform Group, 1999; Shepard, 2000). Teachers also need to find out and reflect on the effectiveness of their instructional tasks. They need to be constantly informed of the status of their teaching activities and to learn about their students and student learning while on the job to be able to improve their practices (Campbell, et al., 2004). The information about teaching activities and student learning can be reliably provided by the results of effective assessment. With the assessment results, teachers will be able to check if their instructional approaches, methods, strategies and techniques meet students' learning needs. Using this as a basis, teachers can adopt better alternative ways that can improve their teaching activities and effect learning (Stiggins, Arter, Chappuis, & Chappuis, 2006). For instance, teachers can devise instructional and assessment activities that serve students' capability and interests; they can also do remedial activities like re-teaching, review, extensive exercises and drills that can boost student learning (Plake, Impara & Fager, 1993; Kellaghan & Greany, 2001). Additionally, through effective assessment, teachers will be able to reliably measure and establish students' performance, diagnose and identify learning problems, determine students' readiness, judge the quality of students' learning, and predict students' academic success (Asaad & Hailaya, 2004). These activities allow teachers to decide what to do next with their instructional activities and strategies to motivate and develop the students (Plake, Impara & Fager 1993; Kellaghan & Greany, 2001; Guskey, 2003).

Classroom assessment also contributes to students by increasing their learning (Black & William, 1998a; 1998b; Stiggins, 2002; Shepard, 2000; Guskey, 2003). It helps clarify the learning targets, helps

identify the learning gap, and engages and motivates students to learn (Black & Wiliam, 1998a; 1998b). Clear learning targets essentially provide the direction and the standards for students to follow and achieve. In the process of attaining the targets, students would discover the gap between what they possess and the expected learning outcomes, prompting them to put more extra efforts to boost their learning. In a sense, it engages and motivates students to learn more and eventually achieve the standard (Stiggins, 2002; Black & Wiliam, 1998a; 1998b).

Assessment of different types at the classroom level has been stressed to support instruction and to ascertain and maximise student learning. As discussed in the preceding sections, these assessments have the potential to make instruction effective and to make learning more meaningful. However, while assessment promises effective instruction and meaningful learning, there is no guarantee that this is so until classroom teachers are adept in using assessment. In other words, the successful application of assessment in the classroom is contingent upon teachers. As Churchill, et. al. (2011, p.398) put it,

> "teachers as educational professionals need to be able to correctly use the language and concepts of the profession. Teachers are not only expected to know the language of the content areas they are teaching, but also the language and key concepts of education itself. One aspect of this professional knowledge is the terminology and concepts associated with assessment and reporting".

Hence, teachers need to be knowledgeable and skillful in the area of classroom or student assessment (Stiggins, 1991a; Schafer, 1993; Popham, 2009; AFT, NCME, NEA, 1990). The next subsection discusses the teacher assessment literacy.

2.3 Teacher Assessment Literacy

Literacy has a broad meaning in education. In fact, it is one that sometimes renders vagueness. J. C. Cairns (as cited in de Castell, Luke & MacLennan, 1981, p.12) quoted the definition of literacy (pertaining to a literate person) from the 1971 UNESCO Committee for the Standardization of Educational Statistics as follows: "A person is literate when he has acquired the essential knowledge and skills which enable him to

engage in all those activities in which literacy is required for effective functioning in his group or community". Moreover, Walter (1999) described literacy as a broad knowledge about different subjects enabling a person to be critical and analytical of what he/she is doing.

In line with the literacy concept as presented above, there are assessment literacy definitions that can be cited and adopted. The first is the straightforward meaning by the Center for Assessment and Evaluation of Student Learning (2004, p.1), which defines assessment literacy as "having an adequate amount of information and understanding about how student learning is assessed and tested". The second meaning is one given by Paterno (2001, as cited by Mertler, 2003, p. 9) who defined assessment literacy as the "possession of knowledge about the basic principles of sound assessment practice, including terminology, the development and use of assessment methodologies and techniques, familiarity of standards of quality in assessment, and familiarity with alternative to traditional measurements of learning". Stiggins (1991a) further described assessment literate persons as follows:

> *"Assessment literates are those who have a basic understanding of high-and low-quality assessment and are able to apply that knowledge to various measures of student achievement; assessment literates ask two key questions about all assessments of student achievement: What does this assessment tell students about the achievement outcomes we value? And what is likely to be the effect of this assessment on students? Assessment literates seek and use assessments that communicate clear, specific, and rich definitions of the achievement that is valued; assessment literates know what constitutes high-quality assessment; they know the importance of using an assessment method that reflects a precisely defined achievement target; they are aware of extraneous factors that can interfere with assessment results; and they know when the results are in a form that they understand and can use" (p. 535).*

Related to the definitions above, assessment literacy frameworks have been developed and proposed. For instance, Stiggins (1999b) put forward seven assessment competencies that teachers need to possess. These include: 1) connecting assessment to clear purposes; 2) clarifying achievement

expectations; 3) applying proper assessment methods; 4) developing quality assessment exercises and scoring criteria and sampling appropriately; 5) avoiding bias in assessment; 6) communicating effectively about student achievement; and 7) using assessment as an instructional intervention. This framework underscores that when executing assessment it is important to consider the purpose of doing it. It should initially be clear what is the purpose of doing assessment and for whom the results are intended. Having achievement or learning targets is likewise essential in carrying out assessment. Clear and good targets are those that teachers want to assess. These should be well defined to help provide better direction for doing assessment. Once the purpose and the targets have been set, the next consideration is the methods to assess the targets. For accurate assessment, it is important that targets and assessment methods should be matched. It is likewise vital that the chosen methods should be developed properly. They should sample knowledge and skills appropriately and should be developed with proper scoring procedures to avoid bias and to ensure dependable results. These results need to be communicated to stakeholders in a timely and understandable manner and should be utilised to improve teaching and learning (Stiggins, 1999b; Stiggins, et. al., 2006).

Additionally, and relevant to what Stiggins (1999b) had proposed, Rowntree (1987) stressed on five assessment dimensions, which teachers need to know in assessing students. These dimensions include purpose or expected outcomes (Why assess?), content and/or skill (What to assess?), methods or means (How to assess?), interpretation, explanation, and/or appreciation (How to interpret?), and response, communication, and/or intervention (How to respond?). This framework stresses the importance of why assessment needs to be done, what to look for in doing it, what appropriate strategies and how assessment is being carried out, how the meaning from the assessment information can be derived and explained, what appropriate response or intervention is needed and how it can be communicated to stakeholders. The framework views these dimensions as strongly contributory to the effectiveness of assessing students in the classroom.

Subscribing to the view that assessment is an important part of teaching and that good teaching doesn't happen without sound student assessment, the AFT, NCME, & NEA (1990, pp. 2-5) also developed the seven assessment principles called "standards" that teachers need to possess to be able to practice sound assessment. These standards, which overlap with those formulated by Stiggins (1999b) and Rowntree (1987), state as follows:

1) Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.

> *"Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to use the concepts of assessment error and validity when developing or selecting their approaches to classroom assessment of students. They will understand how valid assessment data can support instructional activities such as providing appropriate feedback to students, diagnosing group and individual learning needs, planning for individualized educational programs, motivating students, and evaluating instructional procedures. They will understand how invalid information can affect instructional decisions about students. They will also be able to use and evaluate assessment options available to them, considering among other things, the cultural, social, economic, and language backgrounds of students. They will be aware that different assessment approaches can be incompatible with certain instructional goals and may impact quite differently on their teaching.*
>
> *Teachers will know, for each assessment approach they use, its appropriateness for making decisions about their pupils. Moreover, teachers will know of where to find information about and/or reviews of various assessment methods. Assessment options are diverse and include text- and curriculum-embedded questions and tests, standardized criterion-referenced and norm-referenced tests, oral questioning, spontaneous and structured performance assessments, portfolios, exhibitions, demonstrations, rating scales, writing samples, paper-and-pencil tests, seatwork and homework, peer- and self-assessments, student records, observations, questionnaires, interviews, projects, products, and others' opinions" (p. 3).*

2) Teachers should be skilled in developing assessment methods appropriate for instructional decisions.

> *"Teachers who meet this standard will have the conceptual and application skills that follow. Teachers will be skilled in planning the collection of information that facilitates the decisions they will make. They will know and follow appropriate principles for developing and using assessment methods in their teaching, avoiding common pitfalls in student assessment. Such techniques may include several of the options listed at the end of the first standard. The teacher will select the techniques, which are appropriate to the intent of the teacher's instruction.*
>
> *Teachers meeting this standard will also be skilled in using student data to analyze the quality of each assessment technique they use. Since most teachers do not have access to assessment specialists, they must be prepared to do these analyses themselves" (p. 3).*

3) Teachers should be skilled in administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods.

> *"Teachers who meet this standard will have the conceptual and application skills that follow. They will be skilled in interpreting informal and formal teacher-produced assessment results, including pupils' performances in class and on homework assignments. Teachers will be able to use guides for scoring essay questions and projects, stencils for scoring response-choice questions, and scales for rating performance assessments. They will be able to use these in ways that produce consistent results.*
>
> *Teachers will be able to administer standardized achievement tests and be able to interpret the commonly reported scores: percentile ranks, percentile band scores, standard scores, and grade equivalents. They will have a conceptual understanding of the summary indexes commonly reported with assessment results: measures of central tendency, dispersion, relationships, reliability, and errors of measurement.*
>
> *Teachers will be able to apply these concepts of score and summary indices in ways that enhance their use of the assessments that they develop. They will be able to analyze assessment results to*

identify pupils' strengths and errors. If they get inconsistent results, they will seek other explanations for the discrepancy or other data to attempt to resolve the uncertainty before arriving at a decision. They will be able to use assessment methods in ways that encourage students' educational development and that do not inappropriately increase students' anxiety levels" (p. 3).

4) Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.

"Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to use accumulated assessment information to organize a sound instructional plan for facilitating students' educational development. When using assessment results to plan and/or evaluate instruction and curriculum, teachers will interpret the results correctly and avoid common misinterpretations, such as basing decisions on scores that lack curriculum validity. They will be informed about the results of local, regional, state, and national assessments and about their appropriate use for pupil, classroom, school, district, state, and national educational improvement" (p. 4).

5) Teachers should be skilled in developing valid pupil grading procedures, which use pupil assessments.

"Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to devise, implement, and explain a procedure for developing grades composed of marks from various assignments, projects, inclass activities, quizzes, tests, and/or other assessments that they may use. Teachers will understand and be able to articulate why the grades they assign are rational, justified, and fair, acknowledging that such grades reflect their preferences and judgments. Teachers will be able to recognize and to avoid faulty grading procedures such as using grades as punishment. They will be able to evaluate and to modify their grading procedures in order to improve the validity of the interpretations made from them about students' attainments" (p. 4).

6) Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.

> *"Teachers who meet this standard will have the conceptual and application skills that follow. Teachers will understand and be able to give appropriate explanations of how the interpretation of student assessments must be moderated by the student's socio-economic, cultural, language, and other background factors. Teachers will be able to explain that assessment results do not imply that such background factors limit a student's ultimate educational development. They will be able to communicate to students and to their parents or guardians how they may assess the student's educational progress. Teachers will understand and be able to explain the importance of taking measurement errors into account when using assessments to make decisions about individual students. Teachers will be able to explain the limitations of different informal and formal assessment methods. They will be able to explain printed reports of the results of pupil assessments at the classroom, school district, state, and national levels" (p.4).*

7) Teachers should be skilled in recognising unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

> *"Teachers who meet this standard will have the conceptual and application skills that follow. They will know those laws and case decisions which affect their classroom, school district, and state assessment practices. Teachers will be aware that various assessment procedures can be misused or overused resulting in harmful consequences such as embarrassing students, violating a student's right to confidentiality, and inappropriately using students' standardized achievement test scores to measure teaching effectiveness" (p. 5).*

As presented above, the framework proposed by AFT, NCME, & NEA (1990) includes seven assessment areas on which teachers need to be literate. It generally stresses on the need for teachers to be knowledgeable on assessment and related concepts and be skillful in applying those concepts when assessing the students in the class. In the main, it underlines that assessment literacy should cover the

seven principles or standards, which encompass familiarity with and proper execution of assessment, development and appropriate application of assessment approaches/methods/techniques, and tools, and proper utilisation of assessment data. The standards likewise underscore the importance of relating assessment to the context, and specifically to the cultural, social, economic, and language background of students; the need to consider the validity, reliability, errors, and limitations of assessment procedures as well as linking assessment methods with instructional goals/purposes and decisions; the need to appropriately interpret and use assessment results for vital educational decisions and activities such as to plan the lessons, to diagnose the learning needs, to provide feedback, and to evaluate the effectiveness of teaching; the need to properly communicate assessment results to stakeholders, especially the students; and awareness of pertinent laws and assessment misuses.

A number of studies related to one or more of the competencies/standards as presented in Section 2.3 have been conducted. However, studies that directly address the actual knowledge of teachers on the proposed competency areas/standards are not widespread. Hence, assessment literacy studies are limited (Leighton, et al., 2010; Plake & Impara, 1997; Abell & Siegel, 2011). The assessment literacy studies available in the literature are presented below.

Using the standards developed by AFT, NCME, & NEA (1990), Plake, Impara and Fager (1993) conducted a national survey on assessment competencies of teachers in the US. The study involved samples of 555 teachers and 286 administrators. In this study, the level of assessment literacy was described in terms of raw scores that teachers obtained from the instrument. It was revealed that the overall teachers' mean raw score was 23.20 (near 66% correct), which was below the set minimum passing score of 70%. In terms of performance on the specific standards, teachers were best in the areas of administering, scoring and interpreting test results (Standard 3) but poorest in communicating assessment results (Standard 6). In addition, teachers who had measurement training and who expressed comfort with interpreting standardised tests scored significantly higher than those who had not and who expressed discomfort with standardised tests. Campbell et al. (2002, as cited in Mertler, 2003) extended this study to

220 undergraduate pre-service teachers who were taking test and measurement courses. The study disclosed that the undergraduate pre-service teachers' mean raw score was 21 out of 35 (60% correct), which was also below the set minimum passing score of 70%. In terms of the standards, these pre-service teachers were strongest in choosing appropriate assessment methods (Standard 1) but also weakest in communicating assessment results (Standard 6). In an attempt to compare pre-service and in-service teachers, Mertler (2005) conducted a parallel study involving 67 undergraduate students and 101 teachers. This study revealed that, on average, pre-service teachers answered slightly lower than 19 out of 35 items correctly. For the in-service teachers, on average, they answered slightly less than 22 out of 35 items correctly, which was about the same with what Plake, et. al. (1993) study found. Mertler's (2005) study confirmed the finding of Campbell et al. (2002, cited in Mertler, 2003) that pre-service teachers were strongest in choosing appropriate assessment methods (Standard 1). However, the group was found to be weakest in developing valid grading procedures (Standard 5). As for the in-service teachers, Mertler's (2005) study confirmed the finding of Plake, et al. (1993) that the group was best in administering, scoring and interpreting test results (Standard 3). However, in-service teachers were poorest in developing valid grading procedures (Standard 5). In addition, the assessment literacy scores of both groups of teachers were compared. The results indicated that in cases where significant differences existed, the in-service teachers significantly scored higher (i.e. more assessment literate) than their pre-service counterparts.

Volante & Fazio (2007) conducted a relevant study in Canada. Involving 69 teacher candidates, the study examined the self-described assessment literacy level, purposes of assessment, utilisation of different assessment methods and the need for further training, and suggested methods for promoting assessment literacy in university and practice teaching settings. The study likewise attempted to compare teacher candidates' responses in terms of the tested demographic variables. This study revealed that the teacher candidates exhibited relatively low assessment literacy (self-efficacy) ratings across each of the year levels of the tested program. Besides, the candidates' assessment literacy ratings significantly differed in terms of previous teaching experience. Additionally, majority of teacher candidates suggested the traditional

summative than formative purposes of assessment. Moreover, the candidates tended to utilise observational techniques and personal communication but need further training on performance and portfolio assessments. Furthermore, the teacher candidates overwhelmingly endorsed the development of specific courses on classroom assessment and evaluation.

Balagtas, et al. (2010) administered a similar study in the Philippines. The study employed the standards developed by AFT, NCME, and NEA (1990) as its assessment literacy framework. It examined the assessment literacy levels of 90 Bachelor of Elementary Education (BEED) students, 100 Bachelor of Secondary Education (BSE) students, 125 Certificate of Teaching Program (CTP) students, and 142 graduate students. The study disclosed that the BEED, BSE, CTP, and graduate students obtained overall mean scores of 14.98 (about 42% of the items answered correctly), 15.08 (about 43% correct), 15.14 (about 43% correct), and 15.48 (about 44% correct), respectively. These scores are below 75%, the minimum passing score in Philippine schools. On this basis, the researchers concluded that the groups of respondents covered in the study exhibited relatively low assessment literacy and thus need to be re-trained in assessment. The researchers likewise recommended that the assessment component of the teacher education program needs to be enhanced. In terms of specific standards, the BEED, BSE, and graduate students performed highest in the area of developing assessment methods appropriate for instructional decisions (Standard 2) while the CTP respondents performed highest in the area of recognizing inappropriate assessment methods and uses of assessment information (Standard 7). However, these groups of respondents all performed lowest in the area of communicating assessment results (Standard 6). Besides, the assessment literacy level of graduate students did not differ significantly when compared with those of their undergraduate counterparts.

The studies cited above appear to consistently point out that teachers possess low assessment literacy level and thus a relevant improvement is needed. In terms of particular assessment areas, teachers' strength appear to be mostly on the first three AFT, NCME, and NEA's (1990) standards (Choosing assessment methods appropriate for instructional decisions – Standard 1; Developing assessment methods

appropriate for instructional decisions – Standard 2; and Administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods – Standard 3) and their weakest point appears to be on developing valid pupil grading procedures (Standard 5) and on communicating assessment results (Standard 6). However, more studies are warranted to provide more relevant evidence and to confirm these findings.

It was pointed earlier in this chapter that assessment literacy possibly interplays with other education variables. The literature provides that assessment is associated with teaching and learning. As such, the relationship of assessment literacy with other teacher-level and student-level variables possibly exists. The succeeding sections present and discuss other tested variables and related studies.

2.4 Assessment Practices

Closely related to assessment literacy are the assessment practices. In fact, the need for assessment competence is spurred by the demand for high-quality classroom assessment practices (Stiggings, 2012). These practices are called for as they help examine and improve instructional practices, and monitor and promote students' learning (Shepard, 2000). Stiggins (2002) and Stiggins and Concklin (1992) have stressed that sound assessment activities help students gain deeper and meaningful learning. Thus, assessment practices have been part of the foci of educational research. Studies on assessment practices have particularly been conducted to draw feedback on teachers' needs and activities in assessment and to implement any intervention/program to improve their capability to carry out assessment in the classroom.

Zhang and Burrow-Stock (2003, p. 324) described the concept of classroom assessment as one that "embraces a broad spectrum of activities from constructing paper-pencil tests and performance measures, to grading, interpreting standardized scores, communicating test results, and using test results in decision-making." This description provides a framework for research on assessment practices. More encompassing frameworks are provided by experts such as those proposed by Stiggins (1999b), Rowntree

(1987), and the AFT, NCME, and NEA (1990) as described in Section 2.3. The framework developed by Stiggins (1999b) appeared in the book of Stiggins, Arter, Chappuis, & Chappuis (2007) and is described as the keys to quality classroom assessment. These keys emphasise four core areas that are essential in making classroom assessment accurate and effective. These include 'assessment purpose', 'learning target', 'assessment design', and 'assessment communication'.

A number of relevant research studies that captured some of the activities described in the frameworks above and in Section 2.3 have been conducted. For instance, in the review of literature on classroom assessment by McMillan & Workman (1998), Airasian (1984, as cited in McMillan & Workman, 1998) was reported to have suggested that teachers focus classroom assessment on two areas: academic achievement and social behaviour; these factors were found to vary with grade level, with elementary teachers placing greater importance on social behavior. Airasian (1984, as cited in McMillan & Workman, 1998) also found that teachers' informal assessments are used most of the time and this has influenced student self-perceptions of ability. Another reviewed study was that of Fleming and Chambers (1983, as cited in McMillan & Workman, 1998) who, in their analysis of nearly 400 teacher-developed classroom tests, concluded that short-answer questions are used most frequently while essay questions are being avoided. The study also revealed that matching items are used more than multiple choice or true-false items and that most of the items are at the knowledge level (of Bloom's taxonomy). As a follow up to the study of Fleming and Chambers (1983, as cited in McMillan & Workman, 1998), Carter (1984, as cited in McMillan & Workman, 1998) was cited to have measured test development skills of high school teachers and found that teachers had difficulty in recognising or writing items that measure higher order thinking skills. Carter's (1984, as cited in McMillan & Workman, 1998) study is supported by the findings of Stiggins and Conklin (1992), which revealed that recall knowledge items are used approximately 50% of the time. As a result of their study, Stiggins and Conklin (1992) proposed classroom assessment profiles that can be used to characterise diverse assessment practices and environments. These profiles include assessment purposes, assessment methods, criteria used in selecting assessment methods, quality of assessment, feedback to

students, teacher as assessor (background, preparation), teacher perception of the students, and assessment-policy environment. Another related study cited by McMillan & Workman (1998) was that of Cross and Weber (1993) who surveyed self-report practices and beliefs of 536 high school teachers on classroom assessment. Cross and Weber's (1993, as cited in McMillan & Workman, 1998) study found that short answer (56%), multiple choice (52%), essay (38%), performance (37%) and true false (19%) were practised by teachers. In another cited study, Cizek (1988, as cited in McMillan & Workman, 1998) found that assessment practices of 143 elementary and secondary teachers vary widely and unpredictably. Cizek (1988, as cited in McMillan & Workman, 1998) disclosed, too, that teachers, especially the more experienced ones, employed commercial sources as their primary source for major tests and quizzes, and majority of them are not aware of their district's assessment policies and colleagues' practices. McMillan, Myran & Workman (2002) conducted a study involving 900 Grades 3-5 teachers and found that they used three types of assessments: constructed-response such as projects, essays & presentations; objective assessment such as multiple choice, matching & short answer; and teacher-made major examinations. The said authors also confirmed the findings of other studies that teachers use a "hodgepodge" of factors when assessing and grading students. In the study of Stiggins and Bridgeford (1985), which involved 228 teachers, the findings revealed that classroom teachers used their own tests rather than published and performance tests in classrooms. Specifically, and as also found in other studies, teachers utilised multiple choice, true-false, completion, and matching types of test. Besides, science and mathematics teachers were found to use traditional assessment more than the performance tests while writing and speaking teachers tended to use performance type than traditional mode. However, in a related study by Mertler (1998) which examined the assessment practices of 625 K-12 Ohio teachers, it was found that teachers at the elementary level used alternative assessment techniques like observations and questions, as well as portfolios, more frequently than middle and high school teachers. Moreover, teachers with fewer years of experience tended to use alternative assessments more frequently than teachers with 30 or more years of experience in the classroom. Furthermore, the author found that classroom assessment practices differ by

school. In the study of Mcnair, Bhargava, Adams, Edgerton, and Kypros (2003), paper-and-pencil tests appeared as the regular assessment tools used by teachers in grades 3 and 4, and rarely or occasionally used by teachers below these levels; the early grade teachers and those in grades 3 and 4 also appeared to use observation for summative rather than formative analysis, checklists primarily for summative purposes, and portfolios primarily for summative rather than formative purposes. Zhang & Burry-Stock (2003) examined in their study additional teachers' assessment practices and self-perceived assessment skills across teaching levels and content areas. In this study, it was found out that as grade increases, teachers rely more on objective tests in classroom assessment and show an increased concern for assessment quality. In terms of content areas, teachers' involvement in assessment activities reflects the nature and importance of the subjects they teach. Besides, regardless of teaching experience, teachers with measurement training appeared to have higher level of self-perceived assessment skills in using performance measures, standardised testing, test revision, and instructional improvement, as well as in communicating assessment results than those without measurement training.

While a number of research studies concerning assessment practices have already been conducted such as those cited above, these studies are far from being complete. In fact, the focus of most of the studies have been on assessment methods/tools, standardised testing, and grading practices, leaving other classroom assessment areas understudied (Stiggins & Conklin, 1992). Hence, studies that cover more aspects of assessment practices, such as those related to the keys to quality classroom assessment (Stiggins, et al., 2007), are still warranted.

2.5 Teaching Practices

A teacher is widely held as the primary factor that directly affects student learning. Because of teacher's influence on student achievement, teacher attributes have become the subjects of educational research (Churchill, et. al., 2011; OECD, 2009a). The Teaching and Learning International Survey (TALIS) acknowledges that factors such as teachers' professional competence and professional activities impact on

student learning. Thus, teaching practices can be covered and studied as they affect the way students learn.

To examine instructional practices of teachers, two general approaches can be considered. These are the *direct instruction* and the *alternative* or *constructivist* approaches (Rowe, 2006; OECD, 2009a; 2010). The direct instruction or what is sometimes called as *explicit teaching* is a teaching strategy that is employed "for presenting material in small steps, pausing to check for student understanding, and eliciting active and successful participation from all students" (Rosenshine, 1986, p. 60). Proposed by Engelmann and his colleagues in the 1960s (Magliaro, Lockee, & Burton, 2005), this method has its roots in the behavioral theory as put forward by behavioral psychologists. As such, it views knowledge as independent of the learner that needs to be transferred from external reality to the internal reality of the learner (Applefield, Huber, & Moallem, n.d.). This teaching method requires teachers to engage students in learning basic skills and knowledge through the design of effective lessons, corrective feedback, and opportunities for practice. It has three important components namely, introduction, main presentation of the lesson, and practice. Other description of this approach includes "using reinforcement and mastery learning principles, assessing regularly and directly, breaking tasks into smaller components via tasks analysis, and teaching of prerequisite skills". One of the known characteristics of this strategy is that teachers make all instructional decisions (Kinder & Carnine, 1991, p. 193). According to Rowe (2006, p. 103), direct instruction is based on three principles as follows:

- All children can learn, regardless of their intrinsic and context characteristics;
- The teaching of basic skills and their application in higher-order skills is essential to intelligent behaviour and should be the main focus of any instructional program, and certainly prior to student-directed learning activities; and
- Instruction with student experiencing learning difficulties must be highly structured and permit large amount of practice.

In other words, under the direct instruction method, teachers transmit knowledge directly to students in a controlled or structured way. Conversely, the alternative or constructivist approach sprang from the views of educational theorists like Dewey, Piaget, Vygotsky, and Bruner (Harris & Graham, 1994). It views knowledge as dependent within the learner and as such it promotes knowledge construction rather than knowledge transmission (Applefield, et. al., n.d.). This method emphasises the importance of active construction of knowledge among learners (Harris & Graham, 1994). Thus, in contrast with direct instruction method, this strategy requires teachers to provide students with activities and allow these students to work independently or with minimal guidance/supervision. Rowe (2006, p. 101) stated that "student-centred" methods of teaching tend to be aligned with this approach, which has the following underlying rationale:

- Students should be intrinsically motivated and actively involved in the learning process; and
- Subject matter studied should, as far as possible, be 'authentic', 'interesting', and 'relevant'.

The constructivism approach views the learner as "an active contributor to the learning process, and that teaching methods should focus on what the student can bring to the learning situation as much as on what is received from the environment" (Rowe, 2006, pp. 101-102).

There have been studies conducted on the two general teaching methods, particularly on the effects of these approaches on student learning. For example, Din (2000) investigated whether direct instruction strategy helps improve students' mathematics skills. Nineteen students aged 7 to 16 years were subjected to individualised treatment for three weeks. The results indicated that direct instruction, when used appropriately, can help students improve their mathematics skills. This result is in agreement with the findings from Project Follow Through, the world's largest educational experiment (Grossen, 1995). In this experiment, which involved 70,000 students in more than 180 schools, the effect of both teacher-directed and student-directed models of teaching on student learning was examined (Rowe, 2006). The results revealed that the teacher-directed or direct instruction model provided consistent positive results, bringing

children close to 50th percentile in all subject areas tested (Grossen, 1995). Kim (2005) conducted a study on constructivist teaching. In this study, the effects of a constructivist approach on academic achievement, self-concept and learning strategies, and student preference were examined. Seventy-six (76) graders were subjected to experiment between constructivist approach (experimental group) and traditional approach (control group) for a total of 40 hours over nine weeks. The results showed that constructivist approach was more effective than traditional approach in terms of academic achievement but ineffective in relation to self-concept and learning strategy. Besides, Constructivist approach had some effect on motivation, anxiety towards learning and self-monitoring, and constructivist environment was preferred to a traditional classroom. However, while studies on the effects of direct instruction and constructivist models of teaching on student-level variables have been carried out, only few studies concerning teachers' use of these approaches in the classroom are available. One of the few studies that can be cited is the OECD's (2009a; 2010) Teaching and Learning International Survey (TALIS), a large-scale study that involved 23 member countries of the Organisation for Economic Cooperation and Development (OECD). In this survey, teaching practices of teachers covering the two general teaching approaches - direct transmission and alternative or constructivist approaches – were investigated. The general findings showed that teachers in most participating countries employed traditional mode of instruction that is characterised by structured setting and mere transmission of knowledge than the alternative student-oriented and constructivist models of teaching that is characterised by student-oriented and enhanced activities.

The limited studies on the teaching practices of teachers using both direct instruction and alternative or constructivist strategies necessitate more studies to help provide more information on teachers' professional practices. The use of these approaches as mediating variables to link teachers' personal and professional attributes with student learning/achievement and aptitude are likewise warranted to obtain a picture of how teachers can potentially contribute to the success of the teaching-learning process.

2.6 The Student Outcomes

The term 'student outcomes' generally refers to student attributes that can be affected by teacher characteristics. In this study, it is operationally taken to refer to student academic achievement and general aptitude as the outcome variables. There are a number of student-level variables that can affect academic achievement and aptitude. However, this study only considered those that are directly related to assessment and based on its objectives.

The literature on assessment-related variables at the student level is insufficient. Nevertheless, some of these variables can be identified from the available literature to support the questions and the proposed model tested in this study. The first variable that is directly related to classroom assessment is the *student learning and/or achievement*, which is the ultimate aim of education. Campbell, et al. (2004) stressed that assessment directly or indirectly has the potential to impact on learning. This is supported by Black and Wiliam (1998a; 1998b) who, in their meta-analytic study of 250 articles, found the significant learning gains from using formative assessments. In a related study, which involved nearly 300 Hong Kong teachers, Brown, et al. (2009) found that Hong Kong teachers believed that learning outcomes were improved by using assessments to make students accountable and by preparing them for the examination. The common indication of learning that has been employed as an outcome variable in many research studies is the *academic achievement*. In most cases, the results of the large-scale examinations such as those conducted by the education department are used as the secondary data for the academic achievement. In the Philippines, the results of the standardised tests such as the National Achievement Test (NAT) have been examined leading to the documentation of student performance as cited in Chapter 1. *Aptitude* is likewise a student-level attribute that is related to assessment and that can be affected by teacher-level factors such as assessment literacy. Aptitude can be viewed "as a general attribute of which a person had a particular amount or capacity". It is considered as a necessary part upon which learning also depends (Nichols & Mittelholtz, 1993, p. 129). Studies on aptitude are not widespread, too. In the Philippines, the measurement of aptitude is done through the National Career Assessment Examination

(NCAE), which determines the abilities of students and the career paths that students may take when going to the university (NETRC-DepEd, 2013). This examination contains core learning areas that constitute the general scholastic aptitude, which can be measured as an outcome variable. Another assessment-related variable at the student level that can be affected by teacher characteristics is the *attitude*. Attitude can be defined as "a mental or neutral state of readiness, organized through experience, exerting a directive or dynamic influence on the individual's response to all objects and situations to which it is related" (Pickens, 2005, p. 44). Mickelson (1990) explained about attitude and its possible link with achievement. The author categorised attitude into abstract and concrete attitudes and asserted that each of these attitude types may influence achievement. Moreover, a review by Rohaan, Taconis and Jochems (2010) pointed out the existence of the relationship between teachers' knowledge and beliefs and pupils' attitude. Hence, attitude is a student-level variable that can possibly be related to teachers' assessment literacy (Stiggins, 1991a; 2002). A number of studies pertaining to attitude can be found in the literature. However, studies on assessment attitude, especially those that concern test and assignment, are scarce. Finally, students' *perception of assessment* is also a student-level variable that is related to teachers' assessment literacy and practices. Perception is generally defined as "a process which involves the recognition and interpretation of the stimuli which register on our senses" (Rookes & Willson, 2000, p. 1). Waldrip, et al. (2009) justified the link between students' perceptions of assessment process and teachers' effectiveness and success as both teachers and students should be involved in framing that process. In a sense, perceptions of the assessment process reflect teachers' competence. Research studies on assessment perceptions are also limited (Waldrip, et al., 2009).

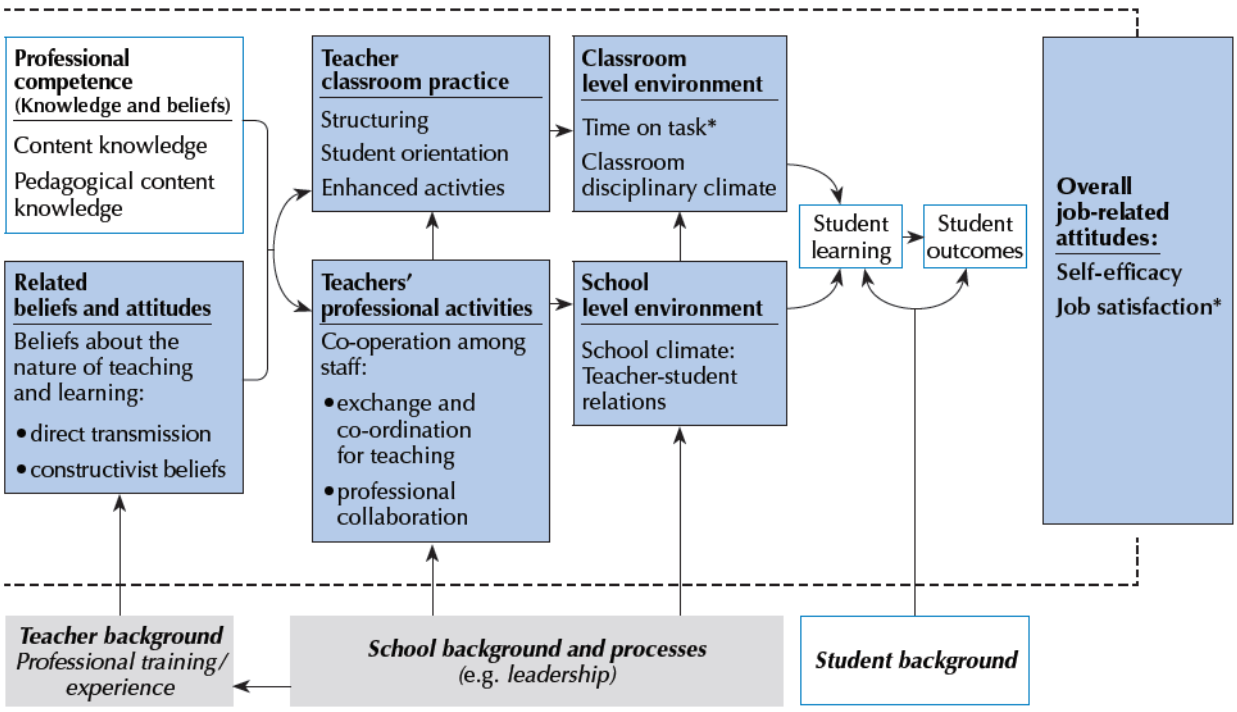
Teachers' attributes are expected to influence student outcomes in any educational context. It is generally assumed that teacher-related factors affect students in the classroom (Maligalig, et al., 2008; Campbell, et al., 2004). Hence, teachers' assessment literacy through the intervening variables at the teacher and student levels can possibly impact on academic achievement and aptitude. However, no

research concerning the direct relationship between teachers' assessment literacy and student outcomes has been found in the literature. This gap is what this study attempted to fill in.

To justify the relationships among the factors or major constructs covered in this study, a model was proposed. The proposed model is presented and discussed in the succeeding section (Section 2.7).

2.7 Proposed Model

The proposed model that was tested in this study has been patterned from the theoretical framework of the Teaching and Learning International Survey (TALIS) (OECD, 2009a; 2010). It has also been based on Bigg's 3-P (Presage-Process-Product) Model (Biggs, 1989). The factors/constructs covered in the tested research framework have been taken from the literature as discussed above to ensure that they are based on established and authoritative knowledge.



Note: Constructs that are covered by the survey are highlighted in blue; single item measures are indicated by an asterisk (*).
Source: *Creating Effective Teaching and Learning Environments: First Results from TALIS* (OECD, 2009).

Figure 2.1. TALIS Theoretical Framework (OECD, 2010, p. 32)

The TALIS' (OECD, 2009a; 2010) framework (Figure 2.1) views that the quality of learning environment is the most significant causal factor of student learning and other outcomes. This learning

environment includes both school and classroom levels. Thus, supportive conditions need to be present in these two levels. However, it is also stressed under the TALIS framework that the learning environment provided at the classroom level has a stronger impact on student learning. This makes conditions and factors at the classroom level even more important.

The quality of learning environment at the classroom level is said to be dependent on the kind of practices present within that classroom. This makes teachers as one of the most important factors because they are the ones exercising these practices. In other words, teachers' professional activities are significant as these activities can positively influence better learning in the classroom. In turn, this calls for adequate teachers' professional competence for the classroom activities to be successfully carried out and to yield positive effect on the student outcomes. Hence, the framework considers teacher attributes and practices as important factors that influence student learning (OECD, 2009a; 2010).

Based on the view presented above, the framework stresses that professional competence, which includes knowledge and related attributes like beliefs and attitudes, impacts on classroom-level environment variables like teaching practices, which, in turn, affect student learning and other outcomes (OECD, 2009a; 2010). Stated in other way, this framework views that teacher competence and other characteristics influence teacher professional activities (OECD, 2010), which, in turn, affect student attributes (Wilkins, 2008).

The *Bigg's 3-P (Presage-Process-Product) Model* (Biggs, 1993) shown in Figure 2.2 also provides similar view of factors and their relations. This model covers three groups of factors, each of which positions to affect each other. As the name of the model indicates, these groups of factors are the 'presage factors', 'process factors', and the 'product factors'. The model considers both teaching and learning contexts. The presage factors for these two contexts include student characteristics like prior knowledge, abilities, preferred learning ways, value, and expectations, and teaching contexts like curriculum, teaching method, classroom climate, and assessment; the process factors include task processing and approaches to

learning and teaching; the product factors are the outcomes (Biggs, 1993; Freeth & Reeves, 2004; Watkins & Hattie, 1990).

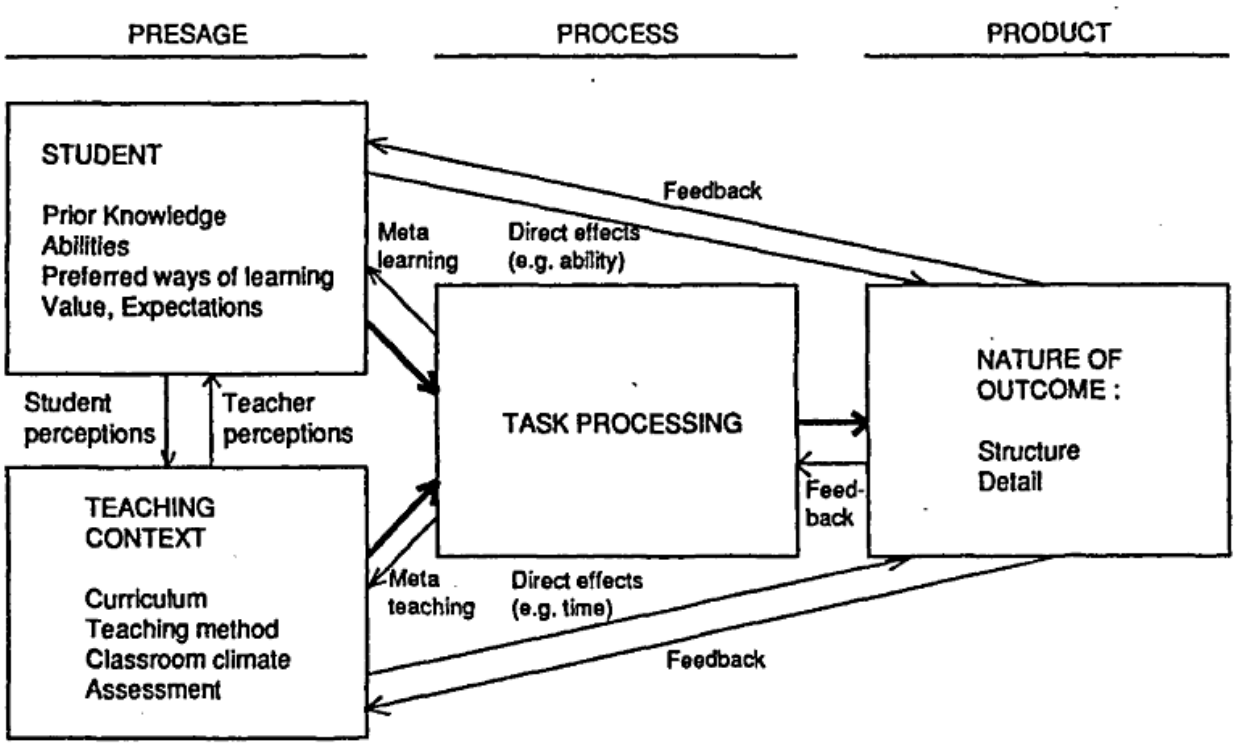


Figure 2.2. Bigg's 3P Model of classroom learning (adopted from Biggs, 1993, p. 75)

The presage factors provide the contexts under which the outcome variables depend. It is possible for these factors to influence each other right at the presage level. For the process factors, they serve to facilitate the effect on the outcome variables. The resulting variables that may be influenced by the presage and process factors constitute the product factors (Freeth & Reeves, 2004). This model specifies a flow from the presage stage, through process level, to the product stage. As explained by Biggs (1989), the presage factors that include the teacher expertise in assessment and instruction will influence the classroom activities (process). These activities can result to either surface or deep learning. These factors will, in turn, affect the cognitive and affective characteristics of students (product). However, the directions of influence are not solely a one-way type. Direct and indirect relationships exist within and among the three stages. Considering the teaching context of Bigg's 3P Model, teacher expertise and characteristics can be

assigned as presage factors influencing classroom-based activities that, in turn, influence student outcomes like perceptions, attitudes, knowledge, skills, behavior, and practice (Freeth & Reeves, 2004).

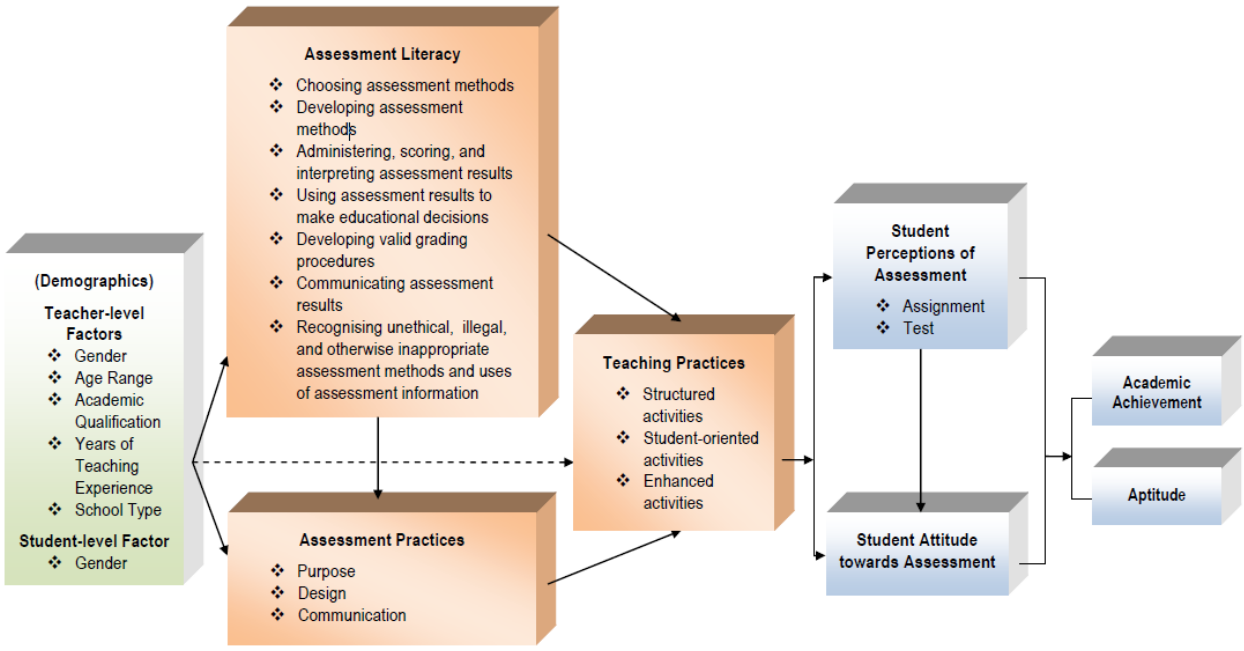


Figure 2.3. Proposed Theoretical Model

As stated earlier in this section, the proposed theoretical model as shown in Figure 2.3 operates within the views/concepts of the TALIS theoretical framework and Bigg's 3-P model. Specifically, the proposed model contains teacher-level and student-level factors. The teacher-level factors include gender, age range, academic qualification, years of teaching experience, school type, assessment literacy, assessment practices, and teaching practices. Moreover, the student-level factors include gender, perceptions of assessment, attitude towards assessment, academic achievement, and aptitude. The model depicts the study's general objective of examining the teachers' assessment literacy and its probable effect on student achievement and aptitude through the intervening variables at the teacher and student levels, and of exploring the effects of demographic variables on factors at the two levels and eventually on the outcome variables. The one-way arrows indicate the relationships among the tested factors/constructs. The one-way arrow implies that one factor/construct directly impacts on another. The research model, together

with the research design and methods as described in the next chapter (Chapter 3) guided the entire process of conducting this study.

2.8 Summary

In this chapter, the literature and studies covering the variables tested in this study have been reviewed. The chapter started with the discussion of the concepts and the types of assessment and the connection between assessment and the teaching-learning process to provide the conceptual background. Specifically, the uses and purposes of assessment at different levels of education system have been discussed to flag the importance of assessment and the need to possess assessment competence. From the review, it appears that assessment has no standard usage but can be employed to capture certain aspects of the teaching-learning process. Besides, assessment types such as systemic/accountability/external, classroom/internal, formal/informal, normative-referenced, and criterion-referenced assessments, as well as assessment of learning, assessment as learning, and assessment for learning have been discussed. Following the conceptual background and the rationale, the concepts and studies associated with teacher assessment literacy, assessment practices, teaching practices, assessment perceptions, assessment attitude, academic achievement, and aptitude were presented.

The theoretical framework adopted in this study and its rationale have been discussed. The model proposed that teacher assessment literacy affects other teacher-level variables namely, assessment practices and teaching practices, and that assessment practices affect teaching practices. These teacher-level variables in turn affect the student-level variables namely, perceptions of assessment and attitude towards assessment. The factors at the teacher and student levels were likewise posited to eventually affect the outcome variables - academic achievement and aptitude. The demographic factors that include gender, age range, academic qualification, years of teaching experience, and school type at the teacher level and gender at the student level were likewise posited to influence the variables at the two levels and ultimately on the outcome variables.

Chapter 3: Research Design and Methods

3.1 Introduction

This study was the first of its kind to be administered in the province of Tawi-Tawi, Philippines. It involved factors for which data were gathered in a new context employing new samples/population. As such, suitable design and methods were needed to secure reliable data for analysis and for appropriate findings, conclusions, and recommendations.

This chapter discusses the research design and methods that were used to gather, analyse, and interpret the data. It begins with the planning stage to provide the background for the selection of the research topic, research design and methods. The chapter proceeds with the discussion of the sampling/sampling method, the participants, the research locale, and data collection. After which, the chapter deals with a section that describes the different scales employed in the study. In this particular section, the instrument development and validation are generally discussed. It continues with the discussion of data analysis to provide the procedure/steps on how the gathered data were treated and interpreted. The chapter ends with a summary that highlights the key points.

3.2 Planning Stage

This section describes the planning process through which this study had been conceived, developed, and conducted. It specifically includes the identification of the research topic and the selection of design and methods, which are considered in detail in the following subsections.

3.2.1 Focus of the study

In any research process, identification of the topic comes at the initial stage. It is important to clearly identify the research topic right at the beginning as it provides structure and direction for the remaining

steps of the study. It encompasses important problems and/or issues that need to be justified and addressed by the researcher (Creswell, 2008; Gay & Airasian, 2003).

Gay & Airasian (2003) mentioned four possible sources for research topic. This includes theories, personal experiences, replications of other similar studies, and library searches. Moreover, in justifying the research topic, Creswell (2008) also pointed out the need to cite the recommendations of other researchers and experts as found in the literature, the issues that arise in the workplace, and/or the researcher's personal experiences. The focus of this study emanated from these sources.

The main topic of this study had been conceived out of the researcher's observations in his more than a decade of teaching and in his involvement in many teacher trainings in his home province in the Philippines. His readings on the topic had also contributed to the pursuit of this research undertaking. In his university, which caters to the education needs of pre-service teachers in his province, the researcher observed that only one assessment course was offered in the entire education program. His experience in teaching the assessment course for years allowed him to take note of the inadequate pre-service teacher preparation on assessment. Moreover, teacher training on classroom assessment, especially for teachers from the remote areas, was scarce. Oftentimes, teachers were only provided with a transmutation table for converting scores into grades without a proper training on the development and applications of assessment methods, strategies, and tools, and proper interpretation and uses of assessment data. There were seminars/seminar-workshops conducted by the Philippine Department of Education and other institutions/organisations but these were believed to be fragmented and insufficient. The lack of follow up on teachers' application of knowledge gained from the assessment training and the absence of relevant research led to the question of whether basic education teachers in the province have the required competence in the area of educational assessment. Furthermore, the readings on the topic helped provide the idea of examining in-service teachers' assessment competence in response to experts' recommendations and gap in the assessment literature. All these sources have provided the basis and the impetus to undertake this study on teachers' assessment literacy and student outcomes.

3.2.2 Design and Methods

This study employed a mixed-methods design. Generally, this design is referred to as "a procedure for collecting, analysing, and 'mixing' both quantitative and qualitative research and methods in a single study to understand a research problem" (Creswell & Plano Clark, 2007, as cited in Creswell, 2008, p. 552). The 'embedded type' of this design was specifically used as the study considered one form of research method as the main source of data and the other form as a support.

The embedded mixed-methods design is a procedure in which collection of quantitative and qualitative data is done at one time and one form of data plays a supportive role to the other form of data. The rationale for using this design is to tap the strengths of both quantitative and qualitative data in understanding the problem that the study attempted to address (Creswell, 2008). This design is believed to provide rich interpretations of the data through the use of both quantitative and qualitative methods. As Creswell (2008, p. 552) stressed:

> Quantitative data, such as scores on the instruments, yield specific numbers that can be statistically analyzed, can provide results to assess the frequency and magnitude of trends, and can provide useful information to describe trends about a large number of people. However, qualitative data, such as open-ended interviews that provide actual words of people in the study, offer different perspectives on the study topic and provide a complex picture of the situation.

Moreover, while the quantitative method offers objectivity in treating the data, the qualitative method provides deeper insights into the participants' perspectives, making the integration of both methods powerful in understanding the situations surrounding the participants (Gay & Airasian, 2003). These methods are considered complementary to each other and can help improve the quality of research as their strengths are being maximised and their weaknesses are being minimised (Johnson & Christensen, 2004). Krathwohl (1998, as cited in Gay & Airasian, 2003, p.184) in support of this design noted that:

> Research, however, is a creative act; don't confine your thinking to specific approaches. Researchers creatively combine the elements of methods in any way that makes the best sense for the study they want to do. Their own limits are their own imagination and the necessity of presenting their findings convincingly. The research question to be answered really determines the method.

Another support in using the design is given by Teddlie and Tashakkori (2009, p. 33) who emphasised that mixed methods research is superior to single approach designs as it can "simultaneously address a range of confirmatory and exploratory questions with both the qualitative and quantitative approaches", can "provide better or stronger inferences", and as it can "provide the opportunity for a greater assortment of divergent views". Thus, this design was considered for this research.

This study used quantitative research as the main source of data and qualitative research as the source of the supporting data. It employed quantitative approach to draw objective, valid, and reliable information and qualitative approach to collect data that could enrich the interpretation of the quantitative data. Through these methods, it was believed that authentic findings and conclusions could be deduced and appropriate recommendations could be offered.

3.2.3 Ethics Clearance/Approval

Prior to the administration of the study and to satisfy the research ethics requirement, ethics clearance was sought from the University of Adelaide Human Research Ethics Committee. The study labeled with Project No. H-159-2010 (Appendix A) was granted approval on 15 September 2010. The committee's approval was with the conditions that: a) information sheet about the study was to be provided to every participant; b) consent form to participate in the study was to be secured from every respondent; c) for participants who were below 18 years old, consent from the parents/guardians were to be obtained; d) in conducting the survey questionnaires and interviews, the identity of every participant was to be kept confidential; e) instructional and learning time were not to be infringed; f) teachers' employment and students' academic standing were not to be affected in any way; and g) participation was to be made voluntary and that every respondent was free to discontinue at any time during the conduct of the study.

These conditions were observed in the administration of the study. Where appropriate, some of these conditions were even made clear to the participants just before the administration of survey to answer doubts and to ensure that the research was living up to the ethical requirement.

As the study was to be administered in the Philippines, a similar approval was likewise sought from the Philippine Department of Education (DepEd) at the national, regional, and provincial levels. The permission (Appendix A) to conduct the study in the province of Tawi-Tawi was granted on 16 August 2010 by DepEd National Office, 28 August 2010 by DepEd-Autonomous Region in Muslim Mindanao Regional Office, and on 9 August 2010 by the DepEd-Tawi-Tawi Division/Provincial Office. The study was undertaken with all the necessary permissions.

3.3 Sampling and Data Collection

The schools and the population were first defined prior to the administration of the survey. All elementary and secondary schools in the research venue were considered for this study. Specifically, all Grade 6 (elementary level), Second Year, and Fourth Year (secondary level) classes were targeted for data collection. These classes were identified, as they were involved in the national examinations. The national examinations namely, National Achievement Test (NAT) and National Career Assessment Examination (NCAE), became the bases for the selection of classes, as the results of these tests were part of the study. In addition, assessment in these classes is believed to be more or less varied and similar, and provides the setting where teachers engage in formal and informal assessment activities and in assessment-related decision making. Moreover, teachers and students were identified as the two groups of participants in this research. Teacher participants included all those who handled subjects in the three-targeted classes during the school year 2010-2011 and student respondents from the same classes and school year were selected purposively on the basis of the available NAT and NCAE results. Hence, the samples for this study were selected through purposive sampling. Table 3.1 shows the participants (shaded) by level and class.

Table 3.1. The study participants

Teacher Participants		Student Participants	
Elementary Level	Secondary Level	Elementary Level	Secondary Level
Grade 6	4th Year	Grade 6	4th Year
Grade 5	3rd Year	Grade5	3rd Year
Grade 4	2nd Year	Grade 4	2nd Year
Grade 3	1st Year	Grade 3	1st Year
Grade 2		Grade 2	
Grade 1		Grade 1	

3.3.1 Identification of Schools and Participants

The schools involved in the study were taken from the list provided by the Department of Education (DepEd)-Tawi-Tawi Division Office in Bongao, Tawi-Tawi, Philippines. All public and private elementary and secondary schools were initially identified. However, as the schools are located in the different islands throughout the province, only those that could be reached and accessed, and that posed no danger to the researcher were finally selected. The schools that were too remote and that were located in chaotic areas at the time of data collection were excluded from the study. Tables 3.2, 3.3, and 3.4 present the number of schools that were considered in the study.

Table 3.2. Number of participating elementary schools by type

Type of School	Number
Public Elementary School	89
Private Elementary School	2
Total	91

Table 3.3. Number of participating secondary schools by type

Type of School	Number
Public High School	31
Private High School	6
Total	37

Table 3.4. Distribution of Schools by municipality and school level

Municipality	School Level		Total
	Elementary	Secondary	
Bongao	21	8	29
Languyan	5	3	7
Mapun	0	2	2
Panglima Sugala	7	1	8
Sapa-Sapa	12	3	15
Sibutu	10	3	15
Simunul	15	7	23
Sitangkai	8	3	9
South Ubian	5	3	8
Taganak/Tutrtle Island	0	1	1
Tandubas	8	3	11
Total	91	37	128

After the final selection of the schools, the researcher coordinated with the appropriate offices in the DepEd-Tawi-Tawi Division and with all the district supervisors and school principals/head teachers to obtain the actual lists/number of teachers and students in the targeted classes. From the lists, teacher respondents and student respondents were identified. Tables 3.5 and 3.6 show the number of teacher and student respondents.

Table 3.5. Number of teacher participants by level and school type

Class	School Type		Total
	Public School	Private School	
Grade 6	310	11	321
Second Year	114	21	135
Fourth Year	104	22	126
Total	528	54	582

Table 3.6. Number of student participants by level and school type

Class	School Type		Total
	Public School	Private School	
Grade 6	893	22	915
Second Year	408	107	515
Fourth Year	521	126	647
Total	1822	255	2077

3.3.2 Research Locale

Tawi-Tawi is an archipelagic province that comprises 307 islands and islets. Located in the southwestern tip of the Philippines, it has a combined area of about 462 square miles. Politically, Tawi-Tawi is one of the 79 provinces and is part of the Autonomous Region in Muslim Mindanao (ARMM). It has 11 island municipalities and 203 *barangays* (villages or smallest political units). The province is headed by a governor, each municipality by a mayor, and each *barangay* by a *barangay* captain (Country Reports on Local Government Systems, Philippines, 2002; Tawi-Tawi Geography, 2010). Figure 3.1 shows the location/map of Tawi-Tawi.

Figure 3.1. Map of Tawi-Tawi, Philippines
(Source: GraphicMaps.com)

Educationally, Tawi-Tawi is one of the country's 182 divisions. Its 18 school districts cover 168 public elementary schools, with the addition of one (1) laboratory school under the supervision of the Mindanao State University at Tawi-Tawi (MSU-Tawi-Tawi) and three (3) private schools. It also has 46 secondary schools, of which 22 are public under the Department of Education (DepEd), 20 community high schools under the MSU-Tawi-Tawi, and four (4) private institutions. As of school year 2007-2008, the primary and secondary schools have total enrolments of 62,937 pupils and 15,618 students, respectively. These schools, which are distributed across the 11 municipalities, are served by 1,455 elementary school teachers and 477 secondary school teachers (DepEd-Tawi-Tawi Division Report, 2008; MSU-Tawi-Tawi Secondary Education Department Report, 2009).

Similar to the rest of the areas in the ARMM, Tawi-Tawi is having the most deteriorating quality of education. The 2007 Regional Assessment in Mathematics, Science and English (RAMSE) test, conducted under the Philippines-Australia Basic Education Assistance for Mindanao (BEAM) Project, indicated that

pupils or students in the ARMM did not only fail to reach the required minimum mastery level, but also have the difficulty in answering the items requiring higher order thinking skills (Philippine Human Development Report, 2008/2009). This information is supported by the report that most schools in the ARMM areas are the worst performers in the recent NAT. Tawi-Tawi, specifically, has been consistently cited as one of the country's least performing provinces in the primary and secondary NAT from 2003 to 2007 (Maligalig, et al., 2010).

As a result of the declining trend in education, the DepEd, in partnership with international agencies, has been pouring in more education projects/initiatives such as BEAM (http://www.beam.org.ph), an AusAID-funded project and the Education Quality and Access for Learning and Livelihood Skills (EQuALLS-http://www.equalls.org), a USAID-funded project, in the ARMM, including Tawi-Tawi. These projects have been targeting instructional capacity of teachers as area for development. Trainings (the researcher had been involved in some of these trainings as trainer) that focus on pedagogy and classroom assessment have been conducted to increase teachers' competence and, in turn, student learning. However, as reported in Chapter 1, despite these trainings, teachers' literacy and practices on the competency-focused areas and the possible impacts on student outcomes have not been assessed or studied. This helped provide the rationale for this research study.

The selection of Tawi-Tawi as a research site had been spurred by three reasons. Firstly, being one of the least performing provinces as mentioned above, Tawi-Tawi is in dire need of educational improvement and relevant research such as this study could help address this need. Secondly, Tawi-Tawi, a rural province, could be a good pilot study site for the ARMM region and for the country as a whole, as most of the least performing schools are of rural type. The study can be replicated in the entire ARMM and in any part of the country to help devise teacher education development programs and policies, especially on the area of classroom assessment. The second reason was believed to be a strong justification as this study was the first of its kind to be conducted in the region/country. And thirdly, Tawi-Tawi is the home place

of the researcher where he has the familiarity of the community and thus making the research a successful undertaking.

3.3.3. Data Collection Methods

There were three ways of collecting the data for this study. These were through the survey questionnaires, open-ended interviews, and records of secondary data. The surveys used instruments that had been applied in previous studies and developed by the researcher using existing scales as guide. Besides, the open-ended interviews for teacher respondents employed guide questions, which the researcher developed in accordance with the purpose of the study. Moreover, the records of secondary data contained student results in the National Achievement Test (NAT) and the National Career Assessment Examination (NCAE) for school year 2010-2011.

Separate questionnaires were developed and used for teacher and student respondents. For the teacher questionnaire, two existing scales (with minor modifications) were utilised and one instrument was developed to capture teachers' attributes that included assessment literacy, assessment practices, and teaching practices. For the student questionnaire, two instruments that were modified from the existing scales were used to capture students' attributes that included perceptions of assessment and attitude towards assessment. The secondary data were likewise employed to indicate students' academic achievement (NAT results) and students' general aptitude (NCAE results). In addition, the open-ended interviews were developed and used to elicit teachers' qualitative responses. The interview guide questions pertained to assessment tools and their qualities. The development and use of survey questionnaires and interviews, and the employment of secondary data were decided on the bases of the factors investigated in this study and on the objectives and questions that this study attempted to attain/address.

The use of survey questionnaires for data collection is advantageous as it can accommodate large amount of data, which are needed to represent the entire samples/population, to "reflect attitudes, beliefs, practices, and trends of the population", and to reduce sampling error (Creswell, 2008, p. 394). It is also

considered as “much efficient as it requires less time, and is less expensive” (Gay & Airasian, 2003, p. 307). Moreover, as expounded by Creswell (2008), the use of open-ended interviews serves to support or augment the quantitative data from the survey questionnaires permitting in-depth interpretations and better understanding of the research problems. Furthermore, the use of secondary data from the standardised examinations helps ensure validity and reliability of the data, allows comparison of responses/performances, and helps reduce error.

3.4 Survey Instruments

The survey instruments employed in this study were either adopted with modifications or developed by the researcher using some existing scales as guide and in consultation with his supervisors. In adopting or developing the instruments, the objectives/research questions and factors answered/examined in the study were used as the primary basis. Also, applicability of the items in the research context was also considered to make the instruments suitable and useful. Hence, items that were found irrelevant were discarded in the final instruments. The following subsections discuss the survey questionnaires used in the study.

3.4.1. Adoption/Modification/Development of Instruments

The scales for the assessment literacy, teaching practices, assessment practices, perceptions of assessment, and attitude towards assessment were either adopted/modified from the existing instruments or developed using the existing scales or literature as guide. In adopting/modifying and developing the scales, some steps were taken to ensure that the respondents answered the items without or with less difficulty and with less time as possible. Specifically, clarity of language, brevity, clear format or structure, single cognitive load per item, clear directions, and applicability of all items to teacher and student respondents were observed.

The scale employed for the teachers’ assessment literacy was the *Assessment Literacy Inventory* (ALI) developed by Mertler and Campbell (2005). The ALI uses the AFT, NCME, and NEA’s (1990)

standards as its assessment literacy framework (see Chapter 2 for the discussion of these standards). It consists of 35 multiple-choice items that are given under the five classroom-based scenarios. Each scenario has seven items that are aligned to the used standards. Each item has four options containing one correct answer and three distracters.

As the ALI was applied to the new group of samples and in a different context, it was necessary to modify its scenarios and items to make it appropriate and useful. However, in modifying the scenarios and the items, only some names and irrelevant situations were changed/rephrased to contextualise the ALI. Moreover, in rephrasing the inappropriate situations, a care was taken to ensure that the rephrased situations were parallel to the original scenarios to preserve the integrity of the instrument. Due to the needed modification, the ALI was further validated/recalibrated in this study (see Chapter 4).

To capture the teachers' assessment practices, a relevant instrument called the *Assessment Practices Inventory* (API) was developed by the researcher. The API was developed initially to parallel the ALI. As such, the ALI's framework/standards and structure were adopted in the API. The only differences with the ALI were that the API questions were of Likert type (5-point scale) and more items were developed for every standard to ensure that sufficient items were retained after the validation process. Additionally, in developing the API, the teacher questionnaire of the Pan-Canadian Assessment Program (PCAP) (Canadian Council of Ministers of Education-CCME, 2010), the Practices of Assessment Inventory (Brown, Kennedy, Fok, Chan, & Yu, 2009), and the Third International Mathematics and Science Study (TIMSS) science teacher questionnaire (International Association for the Evaluation of Educational Achievement – IEA, 1999) were consulted. All items that pertained to teacher assessment practices and that were related to the purpose of the API/study were adopted, modified, and/or used as guide in constructing new items. To ensure that the items were relevant to the Philippine or Tawi-Tawi context, the National Competency-Based Teacher Standards (NCBTS) (DepEd, 2009) and the department orders concerning the practices of assessment in Philippine schools (DepEd Order Nos. 4, 33, and 92, s. 2004) were also referred to. However, the conceptual paradigm for the development of the API was changed and some items were

revised, regrouped, and deleted based on the results of the expert/judgment/face validation. Hence, the API finally adopted the 'Keys to Quality Classroom Assessment' by Stiggins, et al. (2007) as its development framework. As a newly developed instrument, the API underwent a rigorous validation process. The API and its validation are described in more details in Chapter 5.

The *Teaching Practices Scale* (TPS) was used to capture the teachers' instructional practices. The TPS was adapted from the TALIS teacher questionnaire (OECD, 2009a; 2010). As the TPS was applied to a different context (Philippine/Tawi-Tawi context in this study), some items were modified to make them relevant. Modification of TPS was likewise undertaken to suit the purpose of the scale and the study. In modifying the TPS, the NCBTS (DepEd, 2009) was also consulted to align the items to the Philippine teaching standards. Because of the modification, the TPS was re-validated (see Chapter 6).

The *Student Perceptions of Assessment Scale* (SPAS) was employed to capture the students' perceptions of assessment. The SPAS was a modified version of the Students' Perceptions of Assessment Questionnaire developed by Waldrip, Fisher, & Dorman (2008). The items in this questionnaire were adopted, modified, and/or used as basis to construct similar but suitable items with respect to the context. Specifically, the questions were changed to generic items instead of subject-specific and were made to represent assignment and test as the major constructs, as the purpose of this research instrument was to capture the general perceptions of students on assignments and tests in all learning areas. Some items were reworded to suit the participants and those that were believed to be irrelevant were excluded in the final form. As a modified scale, the SPAS was likewise recalibrated (see Chapter 7).

To elicit assessment attitude of students, the *Student Attitude towards Assessment Scale* (SATAS) was used. This instrument aimed to measure the general attitude of students towards assessment in the classroom or school. The SATAS questions were mostly based on the items of the 'Attitude Scale' developed by Mickelson (1990). Mickelson's (1990) scale was the main basis, as its items appeared to reflect the prevailing beliefs about education and the strong association among education, assessment, and effort in the context where the study was conducted. However, as the SATAS was meant to measure

assessment attitude among students, the adopted questions were rephrased to make SATAS appropriate for the study and for the intended context. Hence, in using/adopting the items from the Mickelson's (1990) scale, only those that were relevant to the context were considered.

This study had also the aim of exploring the effects of demographic variables on factors at the teacher and student levels and on the outcome variables. Thus, items covering demographic variables were developed by the researcher and were included in the teacher and student questionnaires.

The scales/instruments employed in the study are summarised and presented in Figure 3.2.

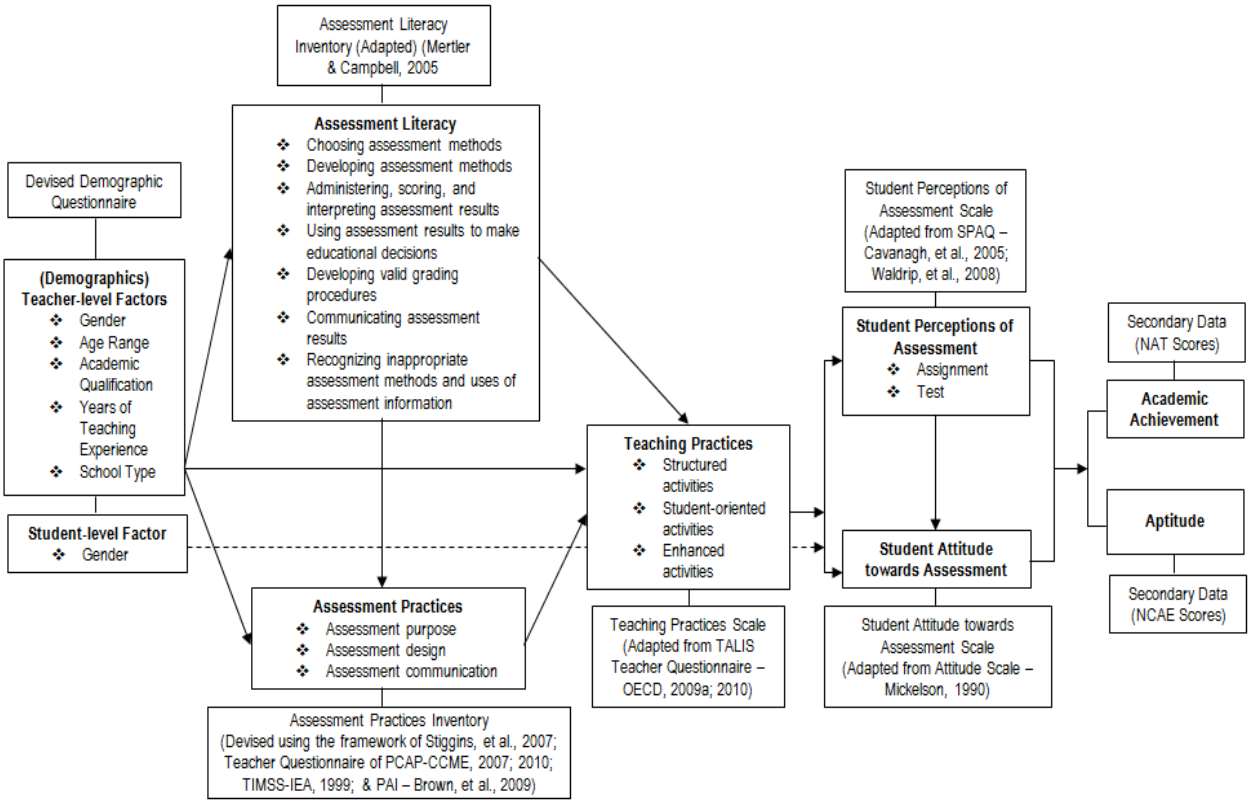


Figure 3.2. Scales/instruments employed in the study

3.4.2. Development of Interview Questions

The semi-structured, open-ended interview was employed in this study. As such, a written protocol or guide that outlined the kind and order of the questions asked, and the manner of the interview was developed. Similar to the construction of survey instruments, the interview questions were developed

following some basic guidelines. Each question was constructed with reference to the topic and the purpose of the study. It was made as brief and as clear as possible. Words that were believed to carry some bias and negative implications to teachers were likewise avoided. The interview questions for teacher participants specifically pertained to the assessment methods/tools and their qualities such as validity and reliability. The questions were developed to provide further details and to obtain in-depth interpretation of the quantitative data obtained from the questionnaires.

3.4.3. The Pilot Study

As mentioned earlier, some items were modified and/or developed, and the instruments were applied to a different context. As such, it was necessary to trial the entire survey. This was to ensure that the instruments were adequate for the targeted context and the data collection procedure was feasible. Hence, a pilot test was conducted to Grade 6 teachers and students of the MSU-Tawi-Tawi Laboratory Elementary School, and 2nd year and 4th year teachers and students of the MSU-Tawi-Tawi Preparatory High School and Science High School to determine the suitability of the items, the reliability of the questionnaires, and the amount of time needed to complete the survey. The pilot study participants were excluded from the final research study.

Initially, the research instruments and the interview questions were checked by the researcher and his supervisors and by the three experts from MSU-Tawi-Tawi who passed judgment on the relevance, appropriateness, and acceptability of the items. After incorporating the suggestions and modifying some items on the basis of the comments from the face validation, the instruments were pilot tested to 45 teachers and 30 students of the above-mentioned schools. Reliability of the teacher and student questionnaires was computed using SPSS software (v.16). The reliability coefficients indicated that the questionnaires had internal consistency, confirming early report on the reliability of the existing scales employed in this study and/or supporting the reliability of the newly developed instruments. These reliability coefficients are reported in the succeeding validation chapters. Moreover, comments and/or suggestions from the pilot study participants were noted and used for further modification/revision of the instruments.

The time to finish the survey was also recorded for implementation in the actual conduct of the study. Specifically, for all the items including the interview questions, suggestions/comments were on the need to improve the structure, to contextualise some items and sample teacher names, and especially for student respondents, to simplify the terms/language used in the instruments. The results of the pilot study were used in improving and finalising the questionnaires.

3.4.4. Instruments for the Main Study

The instruments were finalised after all the relevant modifications/revisions. The final instruments were called "Teacher Questionnaire" and "Student Questionnaire". The Teacher Questionnaire was of two parts. The first part was the modified Assessment Literacy Inventory. This scale has two sections: Section A that contained seven general information items and Section B which consisted of 35 multiple-choice questions with four options in every item. The second part of the Teacher Questionnaire pertained to other assessment and teaching practices and has two main sections: Section A was on assessment practices and Section B was on teaching practices. The Student Questionnaire covered three sections as follows: Section A was on General Information, Section B on assessment perceptions, and Section C on assessment attitude.

The teacher and student questionnaires were separately prepared. Attached to each of the questionnaires was a cover letter, which explained the rationale and the importance of the study to the participants, community, and the teaching-learning process/education quality, and the relevant permissions from the concerned agencies/institutions to ensure ethics and proper administration of the instruments.

3.4.5. Validity and Reliability of the Instruments

It has been stressed that for any measuring instrument to be effectively adopted/developed, its good qualities need to be ascertained. This step is necessary, especially in an empirical research and/or quantitative study (White, 2011; Ben, 2010), to ensure that the instrument has the capacity to gather quality

data (Kline, 2011). To establish the qualities of a good measuring instrument, a number of criteria/properties can be used. Foremost of these are validity and reliability (Asaad & Hailaya, 2004; Creswell, 2008; Field, 2009; Shute & Becker, 2010).

The term *validity* is commonly defined as the property or capacity of the instrument to measure what it intends to measure (Gipps, 1994; Mueller, 1996). In terms of the broader concept of research, validity refers to whether the method has the capacity to examine a phenomenon that it intends to examine (Kvale, 1995, as cited in Ben, 2010). However, this notion of validity is considered limited. Validity is an encompassing concept that involves issues related to research purposes, questions, methods, and results (White, 2011). While the concept of validity may be regarded as broad and complex, its two important aspects – meaningfulness and usefulness – can be considered when viewing it (Keeves & Masters, 1999). Hence, in the context of this study, validity means that the respondents' individual scores gathered from the research instruments make sense, are meaningful, and enable the researcher to draw good conclusions from the sample being studied to the population (Creswell, 2008).

The concept of *reliability* generally relates to consistency and accuracy of test results. However, some experts have stressed on the relations of reliability not only to the results but also to the purpose, measurement process, and the conditions under which the instrument is conducted. According to Field (2009), reliability is the property of the instrument to produce the same results under the same conditions. Mueller (1996) and White (2011) also stressed that reliability relates to the degree to which the instrument or research is consistent in what it measures. Additionally, Gipps (1994) views reliability as not only related to the consistency of performance but also to the consistency in assessing that performance. In quantitative studies, the concept of reliability presupposes the consistency and stability of an instrument to measure what it sets to measure with only minimal errors in the scores (Ben, 2010).

The relationship between validity and reliability generally stems from the view that sound empirical research should be both valid and reliable (White, 2011). It may be viewed as complementary (Ben, 2010) as each supports the other and as the existence of one per se does not ensure the good quality of

instrument/research. However, between these two criteria/properties, validity is more crucial since it permits sound inference from the results and as it indicates the extent to which the research has achieved its aims (White, 2011; Shute & Becker, 2010). Nevertheless, reliability is required. As Thompson (2004) described, validity relates to whether the scores measure the 'correct something'; for scores to measure something, they should first be reliable. But, as Thompson stated, reliability is only a necessary and not a sufficient condition for validity.

This study adopts the notion that "the more reliable the scores from an instrument, the more valid the scores may be" (Creswell, 2008, p. 169). However, reliability of scores within the framework of Classical Test Theory (CTT) is affected by a number of factors such as measurement precision, group heterogeneity, length of the instrument, and time limit given to the respondents (Alagumalai & Curtis, 2005). Hence, these factors were taken into consideration in developing/adopting/modifying the instruments for this study. Figure 3.3 on the next page shows a graphical representation of how the qualities of the scales used in this study were established.

3.4.6. Validation of the Scales

As described in Section 3.4.1, the instruments employed in this study were adopted, modified, and/or developed from the existing scales/available literature. It is implicit that the adopted/modified instruments have already been validated by their respective authors/developers while the newly devised instrument needs rigorous validation. It was important to re-validate/validate the employed instruments to ensure that they measure the constructs/variables/factors covered in this study, to ensure their suitability to the Tawi-Tawi/Philippine context, and as new validation techniques were to be used. The new validation methods were employed to test each item within a scale and the structure of each instrument for their consistency and coherence by calculating different model fit indices and comparing them with the established values for model fit. The re-calibration/calibration was carried out using the Rasch Model and the confirmatory factor analysis (CFA)/structural equation modeling (SEM), which are discussed in Section

3.6.1. Detailed discussions/descriptions of the employed scales and their validation are given in Chapters 4 – 8.

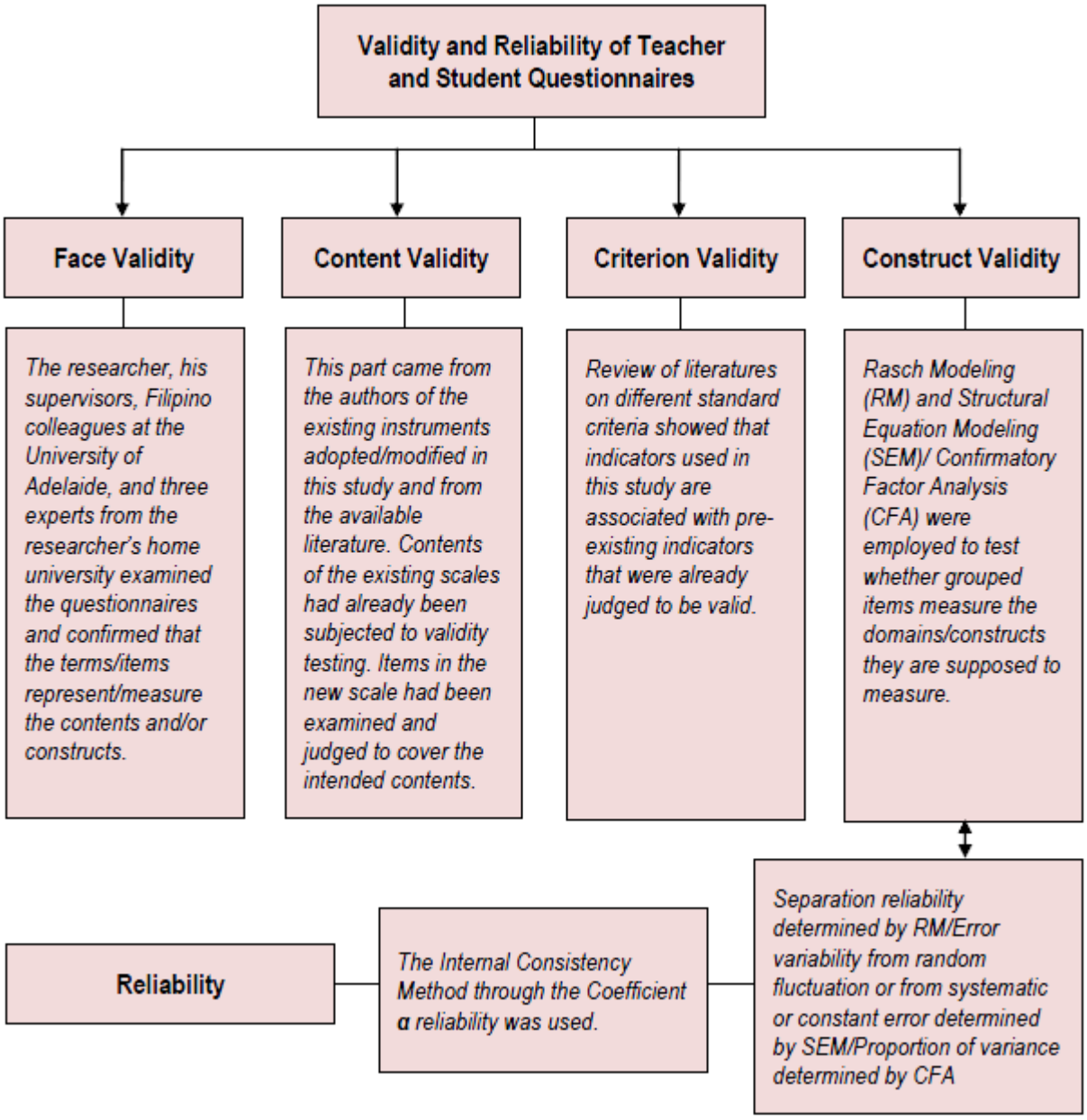


Figure 3.3. Validity and reliability of the employed scales
(Adapted from Ben, 2010)

3.5 The Survey

Once the necessary approval/permissions from the involved agencies/institutions and the survey questionnaires were ready, the survey was conducted to the participants in the province of Tawi-Tawi,

Philippines. The survey started on 17 December 2010 and ended on 31 July 2011. The following subsections describe the administration of the survey instruments and the collection of other needed data.

3.5.1. Administration of the Instruments

The process of administering the questionnaires and interviews commenced with the coordination of all involved agencies and/or institutions in the province of Tawi-Tawi, Philippines. Initially, the researcher coordinated with DepEd-Tawi-Tawi division office, MSU-Tawi-Tawi Secondary Department, Tawi-Tawi Reginal Agricultural College High School Department, and all private institutions. The researcher likewise met with all key people in the division/province in the Municipality of Bongao, Tawi-Tawi for information and the scheduled visit to all involved schools for the survey, and/or for the logistics. Once everything was ready, the administration of the questionnaires proceeded as planned.

The researcher, with the help of his relatives and colleagues, conducted the questionnaires to teacher and student respondents in the schools province-wide. In administering the questionnaires, the research team had to travel via *pumpboat* or *motorised banca* (small boat) from one island to another. Also, in reaching the respondents, the team had to walk from one school to another and traveled via a motorcycle or local *banca* from one *barangay* (village) to another.

While in school, teacher respondents were handed the two parts of the questionnaire: the first part containing the general information and the Assessment Literacy Inventory (ALI) and the second part containing the assessment practices and the teaching practices (see Appendix B). For the first part, teachers were requested to respond at a convenient spot during their free time in school and return it at the end of school day. For the second part, teachers were allowed to bring it home to avoid disruption of their teaching duty. Teachers were also requested to return the second part of the questionnaire when completed but not beyond the stay of the research team. It was necessary for the research team to be around during the survey to answer any query and/or doubts anent the instruments and to ensure efficient collection of the questionnaires. However, due to unavoidable circumstances, there were times that the researcher was not

present, requesting the colleagues and relatives to conduct the survey after proper briefing. As for the student participants, class advisers in the three-targeted classes (Grade 6, 2nd Year, and 4th Year) were requested to conduct the survey to ensure efficient and standardised administration and collection of the Student Questionnaires. However, prior to the administration of the instrument, class advisers were briefed on the pertinent information. Instructions were explained and teachers were allowed to translate any word and gave examples to students when necessary. This was to facilitate better understanding and administration of the instrument items/questionnaire. The Student Questionnaires were collected at the end of the class or school day depending on the situation in schools. For schools that happened to be busy with their major activities, it was requested that the completed questionnaires be sent to either DepEd-Tawi-Tawi division office or directly to the researcher at the soonest time possible.

Due to the archipelagic geography of Tawi-Tawi (see Figure 3.1 on page 58), the weather factor, and peace and order condition, the data collection took about seven (7) months to complete.

3.5.2. Administration of the Interviews

While the conduct of the survey questionnaires was going on, teacher interviews were arranged and/or scheduled. At first, identification of the respondents was done through purposive sampling. As the schools were spread throughout the province, the interview participants were selected only from teachers of the municipalities of Sibutu, Bongao, and South Ubian. These areas were specifically identified as there was easy access and as the schools appeared to represent both rural and town schools.

Teachers were selected from the three-targeted classes, from high and low age ranges, and generally on the basis of their potential to provide responses needed in the study. Once the schedule was set and the corresponding appointments with the participants were made, interview questions were handed to let them prepare and provide answers as honestly and as detailed as possible. After which, interviews were conducted in schools in the above-mentioned municipalities.

There were 34 teacher participants in the interview. In conducting the interview, the researcher at the outset explained the purpose, the importance, and the confidentiality of the interview, and informed the participants that the whole interview process was audio-recorded. After which, the researcher started to pose questions using the interview protocol. Questioning strategies such as probing and redirection were carefully observed to elicit the needed answers. During the interview, rephrasing of questions and providing examples and/or related scenarios were done as needed to ensure good grasp of the questions. At the end of the interviews, the researcher expressed due thanks and again assured the participants of the confidentiality of their responses.

3.5.3. Collection of Secondary Data

This study employed secondary data to indicate student academic achievement and general scholastic aptitude. These were the National Achievement Test (NAT) and the National Career Assessment Examination (NCAE) results. Specifically, the NAT results were the overall standard scores representing achievements in five key curricular subject areas: Filipino (Philippine national language), Mathematics, English, Science, and HEKASI (Heograpiya, Kasaysayan, and Sibika or Geography, History, and Civics) for Grade 6 and Filipino, Mathematics, English, Science, and Araling Panlipunan (Social Studies) for Second Year. The NCAE results were the standard scores representing the general scholastic aptitude and involving scientific ability, reading comprehension, verbal ability, and mathematical ability (for Fourth Year level). To collect these data, the researcher coordinated with DepEd-Tawi-Tawi division office and was referred to Mr. Mohammad Nur Tidal, the principal of the Tawi-Tawi School of Arts and Trades, for the test results. Observing the conditions and the ethics in handling secondary data, the researcher secured the available NAT and NCAE results of all concerned schools in electronic form in July 2011.

3.6 Statistical Procedures Employed in the Study

As stated earlier, this study involved a number of instruments and factors to be tested and analysed. Hence, appropriate statistical procedures were needed to ascertain the validity and reliability of

the questionnaires and the resulting data, to ensure that the findings somehow reflect the true situation about the samples/population in the research context in relation to the factors examined in this study, and to offer appropriate recommendations for action or implementation by the concerned agencies, institutions, and individuals. These statistical procedures are discussed in the following subsections.

3.6.1 Statistical Procedures Employed in Validating the Instruments

It was necessary to calibrate/recalibrate the instruments in order to obtain data that could be useful and meaningful for analysis and measurement of factors involved in the study. Hence, certain statistical procedures were employed to validate the instruments.

There are two broad statistical frameworks that can be utilised to validate the instrument. These are the Classical Test Theory (CTT) and the Item Response Theory (IRT) (Hambleton & Jones, 1993). A "CTT is a psychometric theory that allows the prediction of outcomes of testing, such as the ability of the test-takers and the difficulty of items" (Alagumalai & Curtis, 2005, p. 5). It introduces three basic concepts: test score (observed score), true score, and error scores (latent scores) (Hambleton & Jones, 1993). The test score (X) is defined as the sum of the true (T) and error (E) scores and is designated by the formula, X = T + E (Alagumalai & Curtis, 2005; Hambleton & Jones, 1993). The CTT has three basic assumptions: a) true scores and error scores are uncorrelated, b) the average error score in the population of all examinees is zero, and c) error scores on parallel tests are uncorrelated (Hambleton & Jones, 1993, p. 255). Although it is primarily concerned with the test-level scores, it also establishes item-level statistics such as item discrimination and facility (Alagumalai & Curtis, 2005; Hambleton & Jones, 1993). However, the CTT has been critisised for its serious limitations. Foremost of these include: a) dependence of item discrimination and difficulty on the sample of examinees; b) dependence of observed and true scores on the entire test; and c) assumption of equal errors of measurement for all examinees (Alagumalai & Curtis, 2005; Hambleton & Jones, 1993). These limitations have weakened the utility of CTT. As Alagumalai and Curtis (2005, p. 10) noted:

> CTT has limited effectiveness in educational measurement. When different tests that seek to measure the same content are administered to different cohorts of students, comparisons of test items and examinees are not sound. Various equating processes, which make assumptions about ability distributions, have been implemented, but there is little theoretical justification for them. Raw scores add further ambiguity to measurement, as student abilities, which are based on the total score obtained on a test, cannot be compared. Although z-score is used as standardisation criteria to overcome the problem, it is assumed that the examinees are from the same population.

These shortcomings are addressed by and make IRT a better statistical framework. The IRT "is a complex body of methods used in the analysis of test and attitude data" (Alagumalai & Curtis, 2005, p. 2). It includes a one-, two-, and three-parameter item response models (Alagumalai & Curtis, 2005; Ben, 2010; Humbleton & Jones, 1993). It is item based and probabilistic in nature and is thought to be more flexible than the deterministic CTT (Hambleton & Jones, 1993). The strengths of IRT are highlighted by Hambleton and Jones (1993, p. 259) as follows: 1) Item statistics are independent of the groups from which they were estimated; 2) Scores describing examinee proficiency are not dependent on test difficulty; 3) Test models provide a basis for matching test items to ability levels; and 4) Test models do not require strict parallel tests for assessing reliability. As this study involved test and attitude items/data and different groups of respondents, and while the procedures under CTT were still carried out due to its strength in some analyses, IRT was used as the main validation analytic approach, especially for the item-level validation.

Generally, the instruments were calibrated at the item and structural levels to ascertain the true function of every item and the connection of all the items to the identified constructs. Validation of instrument items (micro-level) and structure (macro-level) was done using the Rasch Model and the confirmatory factor analysis (CFA)/structural equation modeling (SEM), respectively.

3.6.1.1 Rasch Model

The *Rasch Model* is the popular one-parameter item response model (Ben, Alagumalai, & Hailaya, 2012) that can be utilised to judge items at the pilot or validation stage (Wu & Adams, 2007). It is a

technique that can be used to establish the psychometric properties of a newly constructed scale, to review the psychometric properties of existing ordinal scales, to examine the hypothesised dimensional structure of ordinal scales, to construct items banks, and to calculate the change scores from ordinal scales (Tennant & Conaghan, 2007). This model was developed by and bears the name of Georg Rasch, a Danish mathematician, in the 1960s (Baker, 2001). It stresses on the probability of a specified response as contingent upon the test takers' ability and the item difficulty. The probability of success in getting the item right is modeled as a logistic function of the difference between the person ability and the item difficulty (Van Alphen, Halfens, Hasman, & Imbos, 1994). The advantage of the model lies in its objectivity. It puts person and item parameters on the same scale and both parameters are sample independent (Tinsley & Dawis, 1975; Hambleton & Jones, 1993; Van Alphen, Halfens, Hasman, & Imbos, 1994).

There are a number of Rasch measurement models that can be used for item analysis, depending on the kind of instrument and/or the nature of responses/data. As this study involved questionnaires that were in the form of multiple choice and Likert-type questions, Rasch models that were appropriate for dichotomous and polytomous data were employed. Particularly, the Dichotomous Rasch Model and the Rating Scale Model were used in analysing the items.

The Dichotomous Rasch Model is the simplest of the family of Rasch models. Mathematically, it is denoted by:

$$P_{ni} = \frac{\exp(B_n - D_i)}{[1+\exp(B_n - D_i)}$$

where P_{ni} is the probability of a person *n* with ability B_n succeeding on item i which has difficulty level D_i (Wright & Mok, 2004, p. 11). It stresses that the probability of success to get the item right is dependent upon the person's ability and the item difficulty. As the name indicates, this model analyses items that requires binary responses. Instrument that has a two-category treatment of responses as in the case of a multiple-choice test (1,0 for "right", "wrong" responses) can be subjected to analysis under this model.

When the instruments are of Likert type and adopt three or more response categories, two Rasch models can be used. These are the Rating Scale Model (RSM) (Andrich, 1978) and the Partial Credit Model (PCM) (Masters, 1982). The RSM works well for response categories that include ordered ratings such as those used in the 4-point Likert-scale of "Strongly Disagree", "Disagree", "Agree", and "Strongly Agree", which can be coded as 1, 2, 3, and 4. The scales and the corresponding codes are recognised as hierarchical categories only, in which one category is higher than the previous category by an unspecified amount. This model transforms ordinal data to interval scale to make the data useful and assumes that the thresholds ("the point at which the probability of opting for the next category is equal to that of the previous one") are equal for all ratings. The model provides analysis of an item estimate for each Likert stem and a set of estimates for the thresholds between Likert categories (Wright & Mok, 2004, p. 19; Bond & Fox, 2007). However, it is sometimes not possible for all the respondents to regard all categories as equidistant. Some participants tend to be harder in endorsing some items while others tend to be more likely to endorse the items, making the category thresholds unequal. When this occurs, the PCM is the appropriate model for the analysis. The PCM is similar to RSM in treating the polytomous data except that each item has its own threshold that is independent of other items' thresholds. It works well for items where "credits are given for partially correct answers", where "there is hierarchy of cognitive demand on participants in each item", where "each item requires a sequence of tasks to be completed", and/or where "there is a batch of ordered response items with individual thresholds for each item" (Wright & Mok, 2004, pp. 22-23). In this study, the RSM was used instead of PCM. The justification for the use of RSM can be taken from the assertion of Masters (1999, p. 103, as cited in Ben, 2010) who pointed out that:

> … the fact that response alternatives are defined in the same way for all items introduce the possibility of simplifying the partial credit model by assuming that, in questionnaire of this type, the pattern … will be the same for all items in the questionnaire and that the only difference between items will be a difference in location on the measurement variable (e.g. difficulty of endorsement).

As mentioned in Subsection 3.4.5, validity and reliability are the two criteria that can be used in establishing the good quality of the instruments. The Rasch model helps establish these important criteria through construct validity and person/item reliability. It requires that the unidimensionality as ascertained by the item fit be established, thus guaranteeing that all items measure a single attribute or construct. It likewise "provides indices that help determine whether there are enough items spread along the continuum" (item reliability index) and "enough spread of ability among persons" (person reliability index), ensuring consistency of inferences about the items and the takers (Bond & Fox, 2007, p. 40).

In this study, the Rasch method was employed to estimate measures of individuals and item characteristics on a particular scale. It determined whether the responses conform to the requirements of a measurement model. In judging the responses/items, fit indicators, which the model provided, were used. Items that conformed to the measurement requirements were retained while those that failed to satisfy the requirements were removed (Curtis & Boman, 2004; Ben, Hailaya, & Alagumalai, 2012).

To judge the acceptability of the items, the residual-based fit statistics were used. The infit weighted mean square (IWMS)/unweighted mean square (UMS) and the t-statistic (t) were particularly employed to indicate whether or not an item conforms to the Rasch model. For the ALI, which was basically of the test form, a range of 0.80 to 1.20 was used for the IWMS; for other scales, a range of 0.70 to 1.30 was employed for IWMS (Linacre, 2002). Moreover, a range of -2 to +2 for t (Wu & Adams, 2007) was used to indicate acceptable item fit for the ALI and other scales. Items that fell outside the adopted ranges for IWMS and t were removed one at a time as they violated the measurement requirements. To understand the items/response pattern, guides in interpreting the resulting IWMS/UMS and t values were used. For the IWMS/UMS, a value of more than 1 indicates noise while value less than 1 indicates lack of stochastic fit to the Rasch model. For the t, a value greater than 2.0 "would indicate an unexpected or irregular response pattern across items, i.e., noise or lack of unidimensionality. On the other hand, a t value of less than – 2.0 "would indicate possible redundancy in item responses, i.e., a lack of expected stochastic fit or violation of local item independence" (Schumacker, 2004, pp. 235-236).

3.6.1.2 Confirmatory Factor Analysis (CFA)/Structural Equation Modeling (SEM)

The CFA is a confirmatory technique that is used to verify the factor structures of any scale (Schreiber, Stage, Barlow, & King, 2006). It is employed to provide evidence of construct validity (Probst, 2003). This technique assumes that a theory underpins the structural relationships of the factors or constructs. As such, it is described as a theory-driven technique in which analysis is governed by the theoretical relationships among the observed and latent variables. The theoretical relationships are empirically tested and confirmed by a set of data (Schreiber, et al., 2006).

The Structural Equation Modeling (SEM) was used to carry out CFA. "SEM is a statistical methodology that takes a confirmatory (i.e., hypothesis-testing) approach to the multivariate analysis of a structural theory bearing on some phenomenon". This modeling technique has two important aspects of the procedure: "a) that the causal processes under study are represented by a series of structural (i.e., regression) equations, and b) that these structural relations can be modeled pictorially to enable a clearer conceptualisation of the theory under study (Byrne, 1998, p. 3). Under this technique, theoretical models that hypothesised how set of items define factors/constructs and how these factors/constructs are related to each other can be tested (Schumacker & Lomax, 2010). Further discussion about SEM is given in Chapter 10.

As CFA is done through SEM, it has also five basic steps. The first step is the model specification that is concerned with the proposition of the theoretical model. In here, the researcher is tasked to specify the relationships and parameters in the model. The second step is the model identification in which information from the data is used to determine whether parameter estimation is possible. Once the model has been identified, model estimation, as the third step, ensues. In this step, parameters in the model are estimated using a number of methods. The fourth step is the model testing in which the researcher determines how well the data fit the model using a number of fit indices. The final step is the model modification that aims to improve the fit and the model. The last step should be done with great care as

improvement of model fit should not defeat the purpose of CFA and the theoretical model (Schumacker & Lomax, 2010).

CFA analyses also involve the measurement and the structural parts of the model. The measurement model needs to be assessed to ensure that significant relationships between manifest variables and the latent variables exist. This can be indicated by the magnitude of the factor loading (Diamantopoulos & Siguaw, 2000). The minimum acceptable value has been set at +0.4 (-0.4), although values greater than +0.5 (-0.5) are needed for practical significance (Hair, Jr., Black, Babin, & Anderson, 2010). However, for the instrument that is in the form of a test such as a multiple-choice type (right or wrong response or dichotomous data), 0.3 and even lower can be accepted as the nature of the data leads to factors of small variance (Kline, 1994). The measurement model needs to be established to ensure construct validity and composite reliability. The structural model that determines the significant relationships among latent variables can be assessed once the measurement model has been found to be satisfactory (Diamantopoulos & Siguaw, 2000).

How well the proposed structure/model fits the data was assessed using several fit indices. These indices assess how well the sample covariances were reproduced by the covariances predicted from the parameter estimates. The indices used in the analysis included the chi-square (χ^2) statistic, ratio of chi-square to its degrees of freedom, root mean square error of approximation (RMSEA), standardised root mean square residual (SRMR), goodness-of–fit index (GFI), adjusted goodness-of-fit index (AGFI), comparative fit index (CFI), and parsimony goodness-of-fit index (PGFI).

The χ^2 is described as an index of 'exact fit' as it evaluates the perfect fit of a model to empirical data (Matsunaga, 2010). However, although often used, it has been considered to be sensitive to sample size and is almost always indicative of bad model fit. Thus, there is a need to divide the χ^2 by the number of degrees of freedom (df) to further assess the model (Probst, 2003). The RMSEA is an index of 'approximate fit' (Schermelleh-Engel, et al., 2003) and it determines how close the model fits to the data (Matsunaga, 2010). Considered as one of the most informative fit indices and that represents error due to approximation

(Diamantopoulos & Siguaw, 2000), it shows "how well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available?" (Byrne, 2001, p. 82; 2010, p. 80). The SRMR is a residual-based index that shows the average value of the standardised residuals between observed and predicted covariances (Matsunaga, 2010). It is a summary measure of standardised residuals (Diamantopoulos & Siguaw, 2000). The GFI and AGFI are absolute fit indices that estimate the extent to which the sample variances and covariances are reproduced by the hypothesised model (Bollen & Long, 1993). The AGFI's defining characteristic that differs from GFI is that it adjusts for the number of degrees of freedom in the specified model. However, caution was taken with the use of these fit indices as they can be overly influenced by sample size (Fan, Thompson & Wang, 1999, as cited in Byrne, 2001). The CFI is one of the major incremental fit indices that "measure the proportionate improvement in fit by comparing a target model with a more restricted, nested baseline model" (Diamantopoulos & Siguaw, 2000, p. 87). The PGFI "indicates model complexity (the number of estimated parameters)" (Ben, 2010, p. 100; Byrne, 2010).

The nonsignificant result of χ^2 indicates good fit (Matsunaga, 2010). For χ^2/df, $0 \leq \chi^2/df \leq 2$ and $2 < \chi^2/df \leq 3$ indicate good and acceptable fit, respectively (Schermelleh-Engel, et al., 2003). For the RMSEA, values less than the critical value of 0.05 indicate good fit (Schermelleh-Engel et al., 2003). However, some researchers such as Schulz (2004) indicate that values around 0.08 indicate reasonable error of approximation, and for some, (e.g., Hu and Bentler, 1999) 0.06 is considered as the critical value for the RMSEA. Values more than 0.10 for RMSEA indicate poor fit (Diamantopoulos & Siguaw, 2000). SRMR values of less than 0.05 indicate a good fit while values between 0.05 and 0.10 indicate acceptable fit (Schermelleh-Engel et al., 2003). Threshold values for GFI and AGFI are 0.90 and 0.85, respectively; for GFI, values between 0.95 and 1.00 indicate good fit while values between 0.95 and 0.90 indicate acceptable fit; for AGFI, values that fall between 0.90 and 1.00 indicate good fit while values between 0.85 and 0.90 indicate acceptable fit (Schermelleh-Engel et al., 2003). For CFI, the conventional threshold of 0.90 can be used (Matsunaga, 2010). And for the PGFI, a threshold of 0.90 can be adopted (Ben, 2010),

although "nonsignificant χ^2 statistics and goodness-of-fit indices in the .90s accompanied by parsimonious-fit indices in the 50s, are not unexpected" (Byrne, 2010, p. 78). Table 3.7 shows a summary of the different fit indices and their corresponding threshold values, which were used to judge and accept the model fit in this study.

Table 3.7. Summary of model fit indices and their corresponding permissible values

Fit Index	Permissible Values/Ranges
χ^2	Result not statistically significant
χ^2/df	$0 \leq \chi^2/df \leq 3$
RMSEA	≤ 0.10
SRMR	≤ 0.10
GFI	≥ 0.90
AGFI	≥ 0.85
CFI	≥ 0.90
PGFI	≥ 0.90

Note: χ^2 = chi square; df = degrees of freedom; RMSEA = root mean square error of approximation; SRMR = standardised root mean square residual; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CFI = comparative fit index; and PGFI = parsimony goodness-of-fit index.

3.6.2 Statistical Procedures Employed in Data Analysis

In treating the data from the questionnaires, appropriate statistical procedures were also employed. These procedures involved statistics that is associated with descriptive and inferential analyses, including single-level and multi-level analyses. These procedures are described in the following section and in the relevant chapters (Chapters 9, 10, and 11).

3.7 Data Analysis

After the administration of the survey, the gathered data were made ready for analysis. To ensure data utility, some steps were taken. The following subsections describe the steps.

3.7.1 Preparation of Data

After the conduct of the study, questionnaires were collated and/or organised by the type of respondents and class level. Once the questionnaires were ready, the data were directly encoded using Microsoft Excel. Teacher data were encoded first, followed by student data, and then the interview data. The quantitative data were cleaned using the Statistical Package for Social Science (SPSS) software (v.16) (SPSS Inc., 2007a). Also, to prepare the data for analysis, the files were converted to SPSS format. Codes for the demographic and other data were assigned. The missing responses were coded as "-99". The qualitative (interview) data were manually transcribed and were entered as text data in Microsoft Word.

3.7.2 Data Analysis Techniques

For the instrument calibration, the item-level analysis was done using the Rasch model. In running the analysis, ConQuest Version 2.0 software (Wu, Adams, Wilson, & Haldane, 2007) was employed. For the structural-level analysis, CFA/SEM was used to evaluate the measurement model. The CFA/SEM was carried out using LISREL 8.80 (Jöreskog, & Sörbom, 1993; 2006). LISREL 8.80 is one of the most widely used CFA/SEM software programs (Matsunaga, 2010). The use of this highly specialized software for CFA has been recommended as it "works most effectively in a confirmatory context" (Diamantopoulos & Siguaw, 2000, as cited in Ben, 2010, p. 99). It is considered superior due to its robustness in standard error calculation and parameter estimation (Byrne, 1998 & von Eye & Fuller, 2003, as cited in Matsunaga, 2010). Moreover, AMOS software (v.18) (Arbuckle, 2007) was utilised for the path diagrams.

Quantitative and qualitative data analyses were done in accordance with the study's objectives and questions and employing the corresponding statistical techniques and procedures. For the descriptive analysis of the quantitative data, the frequency, percentage, mean (the measure of central tendency and is the average value/score), and the standard deviation (the measure of dispersion/variability and is simply the square root of the variance) were computed using SPSS 16.0 (SPSS Inc., 2007a). Inferential analysis that involved t-test of independent samples and one-way analysis of variance (ANOVA) was carried out through

the same software. To examine the possible relationships among variables at a particular level, the structural equation modeling (SEM) employing LISREL 8.80 (Jöreskog & Sörbom, 2006) was used. Furthermore, to investigate the possible effects of teacher-level factors on student-level factors and the influence of factors from the two levels on the outcome variables, and to explore the effects of the demographic variables on factors at the two levels and ultimately on the outcome variables, the hierarchical linear modeling (HLM) was also carried out using HLM 6.08 software (Raudenbush, Bryk & Congdon, 2009). The qualitative data were examined using thematic analysis and employing SPSS text analysis (SPSS Inc., 2007b).

3.8 Summary

This chapter generally describes how the study was conceived and highlights the steps that were taken to gather, analyse, and interpret the data. Specifically, planning and focus of the study were decided based on the researcher's observations and experiences in handling assessment course/training, and readings on the topic. The embedded mixed-methods design employing quantitative method as the primary approach and qualitative method as the supporting approach was used to collect, examine, and interpret the data. The data were obtained from the teacher and student participants who were identified and selected through purposive sampling from the elementary and secondary schools in the province of Tawi-Tawi, Philippines. Data collection methods included surveys, interviews, and the use of secondary sources. The Teacher and Student Questionnaires, which comprised adopted/modified scales such as the Assessment Literacy Inventory (ALI), Teaching Practices Scale (TPS), Student Perceptions of Assessment Scale (SPAS), and Student Attitude towards Assessment Scale (SATAS), and the newly devised instrument, the Assessment Practices Inventory (API), were devised and employed to collect the quantitative data. The interview questions were also developed and used to gather the qualitative data. The employed instruments were subjected to rigorous validation using Rasch Model and confirmatory factor analysis (CFA)/structural equation modeling (SEM) and employing ConQuest 2.0 and LISREL 8.80

software, respectively. These instruments, as well as the interview questions, were conducted with the permissions of the University of Adelaide's ethics committee and involved agencies/institutions in the research venue. The secondary data were drawn from the results of the National Achievement Test (NAT) and the National Career Assessment Examination (NCAE). The gathered data were analysed and interpreted through the use of descriptive and inferential statistics, including structural equation modeling (SEM) and hierarchical linear modeling (HLM), and thematic analysis. The statistical and thematic analyses were carried out using SPSS (v.16), LISREL 8.80, and HLM 6.08 software.

Chapter 4: The Assessment Literacy Inventory

4.1 Introduction

The main focus of this study was to investigate the teacher's assessment literacy and its influence on academic achievement and aptitude through the intervening variables at the teacher and student levels. Figure 4.1 below shows a graphical representation of the proposed effects of teacher assessment literacy on factors at the two levels and ultimately on the outcome variables. The relationships depicted in the figure had been drawn from the literature.

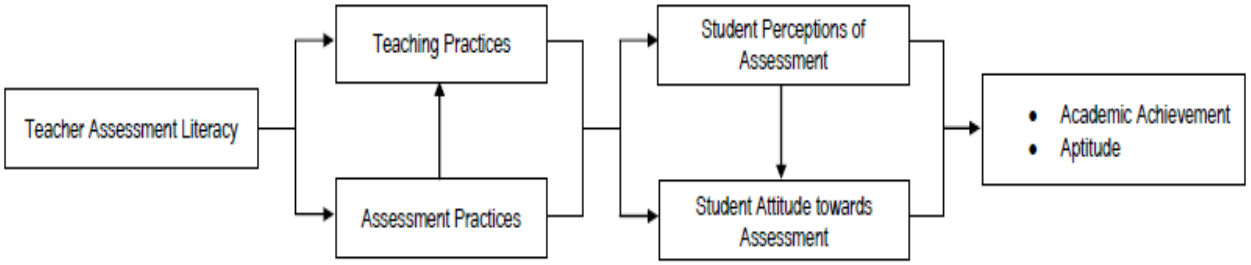


Figure 4.1. Effects of teacher assessment literacy on academic achievement and aptitude through the intervening factors at the teacher and student levels

To measure the assessment literacy and to answer the research questions involving the effects of teacher assessment literacy on the outcome variables through the intervening factors at the teacher and student levels, the Assessment Literacy Inventory (ALI) developed by Mertler and Campbell (2005) was employed. In the light of different context to which the instrument was applied and to ensure that the data gathered through the ALI were reliable for subsequent analysis, the instrument was validated. This chapter presents the validation of the ALI instrument.

The chapter begins with the discussion of the ALI and its development to provide background information about the scale. After which, the previous analytic study of the ALI by its authors is reported to

provide information on its psychometric properties. To give an idea on how the scale was made applicable to the intended context, modification and pilot testing of the ALI are also described. The ALI's validation that includes both the micro-level (items) and the macro-level (structure) analyses is then discussed. The chapter ends with a summary, which highlights the essential points.

4.2 The Assessment Literacy Inventory (ALI)

The development of the ALI was spurred by the poor validation results of the earlier scales on assessment literacy. In 1991, the first scale, the "Teacher Assessment Literacy Questionnaire (TALQ)", developed by Plake (1993) was employed in a national survey both to establish its psychometric qualities and to measure the teacher assessment literacy. Using a sample of 555 in-service teachers from across the U.S., the reliability result for the whole test employing KR_{20} was 0.54 (Plake, Impara, and Fager, 1993). The survey found that out of 35 items, the teacher respondents obtained a score of 23 (66%), which led the researchers to conclude that the teachers were not adequately prepared to assess student learning (Campbell & Mertler, 2005). In 2002, Campbell et al. conducted a similar study employing the identical scale called the "Assessment Literacy Inventory (ALI)" to the 220 undergraduate students (Campbell & Mertler, 2005). The data from this study yielded a reliability result of 0.74 using the same statistical technique; as revealed, the reliability value was higher compared to the study of Plake, et al., (1993). Campbell et al. (2002, cited in Mertler & Campbell, 2005) study also found that the pre-service teachers obtained an average score of 21 out of 35 items (60%), two questions fewer that their in-service counterparts from the study of Plake et al. (1993). In 2003, Mertler tried to combine the two groups in his study. He examined and compared the assessment literacy of both in-service and pre-service teachers. Like Campbell et al. (2002), he used a slightly modified version of TALQ (Plake, 1993) and called the instrument, the "Classroom Assessment Literacy Inventory (CALI)". Mertler (2003) noted that the results of his study yielded similar results with those of Plake et al. (1993) and Campbell et al. (2002). Using KR_{20}, Mertler (2003) obtained reliability results of 0.57 for the in-service teachers (Plake et al. study, KR_{20}=0.54) and 0.74 for the pre-

service teachers (Campbell et al. study, KR20=0.74). On the levels of assessment literacy, Mertler (2003) found that the in-service teachers' mean score was 22, quite similar with the results obtained by Plake et al. (1993), and the pre-service teachers' average score was 19, also about the same with the finding of Campbell et al. (2002) study (Campbell & Mertler, 2005).

Having employed identical instruments as TALQ and having obtained consistently low reliability results, both studies of Campbell et al. (2002) and Mertler (2003) concluded that the original instrument (TALQ, to include the identical scales of ALI and CALI) possessed poor psychometric qualities. Their criticisms of the original scale as "difficult to read, extremely lengthy, and contained items that were presented in a decontextualized way" led them to recommend for a complete revision or development of a new assessment literacy scale. Hence, the new ALI, which contains different items and structure from the earlier instruments, was developed by Mertler and Campbell in 2003 (Mertler & Campbell, 2005, pp. 8-9). This new assessment literacy scale was intended to be a context-based instrument to appropriately capture the teacher assessment literacy (Campbell & Mertler, 2005).

The ALI consists of 35 multiple-choice items that are embedded in five classroom-based scenarios. Each scenario reflects a classroom situation that features a teacher doing assessment-related activities and making assessment-related decisions. The situation in each scenario is followed by seven items that are aligned to the Standards for Teacher Competence in the Educational Assessment of Students (STCEAS) developed by the American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), and the National Education Association (NEA), 1990. Each stem has four options, with a distribution of one correct answer and three distracters (Campbell & Mertler, 2005).

4.3 Previous Analytic Practices

As a new scale, the ALI was initially validated (face validation) by its authors being the experts in classroom assessment themselves. After the development of the ALI, the authors reviewed the items to ensure their alignment with the standards and to check for the item clarity, readability, and the accuracy of

the keyed answers. Items that had issues with any of these qualities were revised. They continued with their judgmental review until consensus was reached regarding the item appropriateness and quality (Mertler & Campbell, 2005).

After the face validation, the ALI was trialed twice to establish its psychometric properties. In the first trial, which was done in 2003, the ALI was administered to 152 undergraduate pre-service students who took the introductory assessment courses that were aligned with the STCEAS (AFT, NCME, & NEA, 1990). The authors analysed the resulting data using the Test Analysis Program (TAP) of Brooks and Johanson (2003, as cited in Mertler & Campbell, 2005). After undertaking the item analysis, they made appropriate revisions on the ALI problematic items. In the second trial, which was done in 2004, the revised ALI was conducted to the 250 undergraduate students after completing their tests and measurement course. The authors analysed the data using the SPSS software (v.11) and TAP (v.5.2.7). The results of the two trials were used by the authors to judge the acceptability of the ALI scale.

In the initial pilot test, the results revealed that the ALI had an overall KR_{20} of 0.75, mean item difficulty of 0.64, and the mean item discrimination of 0.32. The authors reported that these values already indicated acceptability of the ALI from a psychometric perspective. Further analysis also disclosed that when four of the 35 items were removed, there was an improvement on the overall reliability. As a result, the ALI was slightly revised. Moreover, the results of the second phase of pilot testing appeared to further indicate the utility of the ALI as an assessment literacy scale. The overall KR_{20} of 0.74, mean item difficulty of 0.68, and the mean item discrimination of 0.31 confirmed both the results of the first phase of the pilot test and the acceptability of the ALI as an assessment literacy scale. The authors cited research studies as a support to their claim that ALI is an acceptable instrument. For instance, they reported that Kehoe (1995) recommended a reliability value as low as 0.50 for a short test (10-15 items), though tests with over 50 items should yield KR_{20} values of 0.80 or higher. They also cited Chase (1999, cited in Mertler & Campbell, 2005) who suggested that a test of this type should have a reliability coefficient not lower than 0.65, but preferably higher. Similar suggestion from Nitko (2001, as cited in Mertler & Campbell, 2005) who

advocated the acceptable range of reliability coefficient as between 0.70 and 1.00 was also reported. Looking at the results, the ALI authors reported that the ALI reliability fell within the acceptable values. As to the item difficulty results, the authors presented that 25 of the 35 ALI items were answered correctly by a percent of examinees that fell within 30% and 80%, a range that is acceptable according to Kehoe (1995). The other support cited was from Chase (1999, as cited in Mertler & Campbell, 2005) who said that the range for effective item difficulties was from 0.20 to 0.85. Again, as 28 of ALI's items fell within these values, acceptability of the ALI was justified by this difficulty index. Finally, the authors reported that by item discrimination index, the ALI was also found to be acceptable. As cited, Chase (1999) stated that discrimination values of 0.30 and higher indicate fairly good item quality. Using this range as a basis, 20 of the ALI's items were acceptable. The authors justified the remaining items by saying that it is mathematically impossible to obtain high discrimination value on items that have high difficulty value. The authors concluded that as ALI had acceptable psychometric qualities by the indices they used, it is an appropriate assessment literacy instrument. As a further justification, the authors did note that although when used with pre-service teachers the reliability result of the ALI was the same with Campbell et al. (2002) study, the "user-friendly format of the ALI" which served to reduce the cognitive overload relevant to the 35 unrelated items as found in the early scales and the unique classroom-based scenarios featured in the instrument made the ALI more relevant in terms of measuring the teacher's assessment literacy.

From the validation of the ALI, the authors provided relevant recommendations. Specifically, Campbell and Mertler (2005) encouraged the employment of ALI with pre-service teachers in future studies to further improve and validate the instrument. They likewise recommended that the ALI be used with in-service teachers to further establish its utility and to ascertain its status as a valid assessment literacy scale. The recommendations to capture the teaching experience and to use the scale with in-service teachers had been considered and thus helped provide the rationale for the use of the ALI in this research study. In addition, the ALI was transported to the Tawi -Tawi context based on the objectives of this study and as the scale of this kind is not yet available in the Philippines. Moreover, the ALI is believed to be applicable to the

Tawi –Tawi context, as the Philippines and the US, where the ALI was developed and employed, have similar education systems.

The use of techniques under the Classical Test Theory (CTT) in validating any instrument has shortcomings (see Chapter 3). In view of this criticism, there is a need to recalibrate the ALI. The validation of the ALI in this study involved the use of newer psychometric methods that include the Rasch Model for the item-level analysis and the confirmatory factor analysis (CFA)/structural equation modeling (SEM) for the structure-level analysis.

4.4 ALI Modification to Suit the Tawi-Tawi Context

To make the ALI usable to the Tawi-Tawi/Philippine context where the study was conducted, it was slightly modified by the researcher. Modifications were done mainly on teacher names and topics in the scenarios and the corresponding items. The original teacher names were changed to local names to help contextualize the scale. Where the topic/s in any of the situations/items was/were found to be irrelevant to the context, parallel topics were used. However, in changing any of the topics, the original structure of the ALI scenarios and items were preserved to maintain the integrity of the scale. Sample original and modified ALI items are presented in Table 4.1 (also see Appendix B) to show the modifications in the instrument. After modification, the ALI was validated by the researcher's supervisors and three experts from the researcher's home university, the Mindanao State University in Tawi-Tawi (MSU Tawi-Tawi), for the appropriateness and suitability of the items. From the expert validation, the ALI's items were judged as acceptable for the Tawi-Tawi context and thus the instrument could be administered. After the expert validation, the ALI was pilot tested with the 45 elementary and secondary school teachers of MSU Tawi-Tawi to check again for its reliability and for reasons as mentioned in Chapter 3 (Subsection 3.4.3). The reliability was determined using the SPSS software (v.16). A Cronbach Alpha of 0.75, which indicated acceptable reliability, was obtained. Hence, the adapted ALI was made part of the Teacher Questionnaire, which was employed to collect data from teacher respondents in this study.

Table 4.1. Sample original and modified ALI items

Original ALI Items	Modified ALI Items
Scenario #1	Scenario #1
Ms. O'cannor, a math teacher, questions how well her 10th grade students are able to apply what they have learned in class to situations encountered in their everyday lives. Although the teacher's manual contains numerous items to test understanding of mathematical concepts, she is not convinced that giving a paper-and-pencil test is the best method for determining what she wants to know.	Mr. Kalim, a math teacher, questions how well his fourth year high school students are able to apply what they have learned in class to situations encountered in their everyday lives. Although the teacher's manual contains numerous items to test understanding of mathematical concepts, he is not convinced that giving a paper-and-pencil test is the best method for determining what he wants to know.
1. Based on the above scenario, the type of assessment that would ***best*** answer Ms. O'cannor's question is called a/an _____. a) performance assessment b) extended response assessment c) authentic assessment d) standardized test	1. Based on the above scenario, the type of assessment that would ***best*** answer Mr. Kalim's question is called a/an _____. a) performance assessment b) extended response assessment c) authentic assessment d) standardized test
2. In order to grade her students' knowledge accurately and consistently, Ms. O'cannor would be well advised to ___. a) identify criteria from the unit objectives and create a scoring rubric b) develop a scoring rubric after getting a feel for what students can do c) consider student performance on similar types of assignments d) consult with experienced colleagues about criteria that has been used in the past	2. In order to grade his students' knowledge accurately and consistently, Mr. Kalim would be well advised to _____. a) identify criteria from the unit objectives and create a scoring rubric b) develop a scoring rubric after getting a feel for what students can do c) consider student performance on similar types of assignments d) consult with experienced colleagues about criteria that has been used in the past

4.5 Current Validation of the ALI

As the ALI had been modified and applied to a different context, and to ensure that the data obtained from the instrument were reliable for further analysis and valid inferences, it had been subjected to

further validation by the researcher. However, in validating the ALI, the researcher adopted a different approach from what the authors of the scale employed in their analytic study. The ALI authors employed a deterministic approach often referred to as the Classical Test Theory or CTT. In view of the shortcomings of the CTT (Hambleton & Jones, 1993; Alagumalai & Curtis, 2005), the researcher employed a probabilistic approach. As previously mentioned, the Rasch model and the confirmatory factor analysis (CFA)/structural equation modeling (SEM) were used to validate the ALI at the item-level (micro-level analysis) and the structure-level (Macro-level analysis), respectively (see Chapter 3 for details).

4.6 Item Analysis of the ALI using the Rasch Model

The item-level analysis was carried out to examine the ALI at the 'micro level'. Its main purpose was to find out how each of the items fits the model and to examine the unidimensionality of the ALI scale. The Item Response Theory using the Rasch Model was employed for the item-level analysis. To run the analysis, the ConQuest 2.0 software (Wu, Adams, Wilson, & Haldane, 2007) was used.

The analysis of the ALI items was carried out using the responses from the 582 elementary (Grade 6) and secondary (Fourth and second Year) school teachers. All teachers in the targeted levels were considered for this study. To judge the acceptability of the items, the residual-based fit statistics were used. The Infit Weighted Mean Square (IWMS) and the t-statistic (t) were particularly employed to indicate whether or not an item conforms to the Rasch Model. As ALI is in the form of a test, a range of 0.80 to 1.20 (Linacre, 2002) for IWMS, and -2 to +2 for t (Wu & Adams, 2007) were used to indicate acceptable item fit. Items that fell beyond these ranges were removed one at a time as they violated the measurement requirements.

The first analysis involved the seven-factor model corresponding to the seven standards or principles of assessment as developed by AFT, NCME, and NEA (1990). Each of these standards was analysed separately using the Rasch Model.

The initial and final analysis for Standard 1 included the 5 items and the responses from all the participants. The fit statistics for each item were obtained. The results are presented in Table 4.2. As can be seen, the first run of the data provided results in which all items possessed acceptable fit statistic values. All the IWMS values were within the acceptable range of 0.80 – 1.20 and all t values were within the allowed range of -2 to +2. The results indicate that the items fit the Rasch model. Moreover, the value of separation reliability (0.99) indicates that measurement error was small and there was a high discriminating power (Alagumalai & Curtis, 2005; Ben, 2010). This further indicates that the items have more precise measurement and reliability (Wright & Stone, 1999). Hence, the five items (1, 8, 15, 22, & 29) can be finally taken to measure Standard 1.

Table 4.2. Results of the initial and final item analysis of the ALI items under Standard 1

Item #	Standard*	Estimate (Difficulty)	Error	IWMS	t
1	1	1.03	0.07	0.97	- 0.4
8	1	- 0.13	0.06	0.98	- 0.8
15	1	1.09	0.07	0.98	- 0.3
22	1	- 0.65	0.06	1.00	- 0.2
29	1	- 1.334	0.13	1.02	0.6

Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 562.67; df=4; Sig level=0.000;
**1 - Choosing Assessment Methods*

Table 4.3 on the next page presents the Rasch analysis results of the five items under Standard 2. As can be spotted, all the items were within the acceptable ranges of IWMS and t. These results indicate that the items fit the Rasch Model. Besides, the separation reliability value of 0.99 implies less error and high discriminating power (Alagumalai & Curtis, 2005; Ben, 2010). It also indicates that the items have high measurement capacity (Wright & Stone, 1999). Thus, items 2, 9, 16, 23, and 30 can be employed to represent Standard 2.

Table 4.3. Results of the initial and final item analysis of the ALI items under Standard 2

Item #	Standard*	Estimate (Difficulty)	Error	IWMS	t
2	2	- 1.12	0.07	1.00	0.2
9	2	- 0.59	0.07	1.00	- 0.2
16	2	0.71	0.08	1.01	0.2
23	2	0.40	0.08	1.03	0.4
30	2	0.602	0.14	1.03	0.4

Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 462.98; df=4; Sig level=0.000;
**2 - Developing Assessment Methods*

Similar results appear for all the items under Standard 3. As shown in Table 4.4, IWMS and t values were within the acceptable range, which imply that items 3, 10, 17, 24, and 31 fit the Rasch Model. In other words, these items are acceptable. In addition, the value of separation reliability (0.99) indicates less error, high discriminating power ((Alagumalai & Curtis, 2005; Ben, 2010), and high measurement capacity (Wright & Stone, 1999). Thus, the items reflect the assessment principle as delineated in Standard 3.

Table 4.4. Results of the initial and final item analysis of the ALI items under Standard 3

Item #	Standard*	Estimate (Difficulty)	Error	IWMS	t
3	3	0.73	0.07	0.98	- 0.2
10	3	0.01	0.07	1.01	0.4
17	3	- 1.25	0.06	0.97	- 1.2
24	3	0.61	0.07	1.02	0.3
31	3	- 0.113	0.14	1.01	0.2

Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 573.17; df=4; Sig level=0.000;
**3 - Administrating, Scoring, and Interpreting Assessment Results*

Table 4.5 shows the results of the initial and final item analysis of the five items under Standard 4. It can be spotted from the table that similar pattern appears for items 4, 11, 18, 25, and 32. The acceptable IWMS and t values indicate that these items follow the Rasch model and can be accepted. Moreover, the separation reliability value of 0.97 provides that the items are capable of measuring the assessment principle as represented in Standard 4. Hence, the items can be taken to measure assessment literacy in Standard 4.

Table 4.5. Results of the initial and final item analysis of the ALI items under Standard 4

Item #	Standard*	Estimate (Difficulty)	Error	IWMS	t
4	4	- 0.38	0.06	1.01	0.6
11	4	0.23	0.06	1.03	0.6
18	4	0.28	0.06	1.01	0.3
25	4	0.46	0.06	1.01	0.2
32	4	- 0.685	0.13	1.00	0.2

Separation Reliability = 0.97; Chi-Square Test of Parameter Equality = 135.67; df=4; Sig level=0.000;
**4 - Using Assessment Results*

The results of the initial and final analysis for the five items under Standard 5 are shown in Table 4.6. As revealed, items 5, 12, 19, 26, and 33 had acceptable IWMS and t values that indicate acceptability of these items according to the Rasch model. A value of 0.87, though lower than those in Standards 1, 2, 3, and 4, for separation reliability provides support that the five items possessed acceptable psychometric property. Thus, the items can measure assessment principle/literacy as represented in Standard 5.

Table 4.6. Results of the initial and final item analysis of the ALI items under Standard 5

Item #	Standard*	Estimate (Difficulty)	Error	IWMS	t
5	5	0.05	0.06	0.99	- 0.3
12	5	0.43	0.06	1.00	- 0.0
19	5	0.08	0.06	0.99	- 0.3
26	5	0.30	0.06	0.99	- 0.1
33	5	- 0.865	0.13	0.99	- 0.6

Separation Reliability = 0.87; Chi-Square Test of Parameter Equality = 67.13; df=4; Sig level=0.000;
**5 - Developing Valid Grading Procedures*

The results in Table 4.7 provide similar picture for the five items under Standard 6. As can be gleaned, all IWMS values are within the acceptable range. Similarly, all t values are within the allowed range. These results suggest that the items fit the Rasch Model. In addition, the separation reliability value of 0.99 indicates desirable psychometric property. Thus, items 6, 13, 20, 27, and 34 can be utilised to measure the literacy on assessment principle as delineated in Standard 6.

Table 4.7. Results of the initial and final item analysis of the ALI items under Standard 6

Item #	Standard*	Estimate (Difficulty)	Error	IWMS	t
6	6	0.39	0.07	1.00	0.0
13	6	0.70	0.07	0.99	- 0.1
20	6	- 0.03	0.06	1.02	0.5
27	6	- 0.76	0.06	1.01	0.4
34	6	- 0.296	0.13	0.95	- 2.0

Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 292.88; df=4; Sig level=0.000;
**6 - Communicating Assessment Results*

Table 4.8 above shows the results of the initial and final item analysis of the five items under Standard 7. As can be spotted, similar results appear for items 7, 14, 21, 28, and 35. With the acceptable IWMS and t values and high separation reliability, these items appear to fit the Rasch model and to possess good psychometric property. Thus, the items reflect the assessment principle described in Standard 7 and can be used to measure the relevant literacy.

Table 4.8. Results of the initial and final item analysis of the ALI items under Standard 7

Item #	Standard*	Estimate (Difficulty)	Error	IWMS	t
7	7	- 0.90	0.06	1.00	0.0
14	7	0.00	0.07	1.00	- 0.1
21	7	- 0.12	0.07	1.01	0.3
28	7	0.98	0.07	1.00	0.1
35	7	0.030	0.14	1.04	1.1

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 376.23; df=4; Sig level=0.000; *7 – Recognizing unethical, illegal, and otherwise inappropriate Assessment Methods and Uses of Assessment Results*

The 35 ALI items were also subjected to Rasch analysis as a one-factor structure. This was to determine if all the items reflect a single or a dominant dimension called 'assessment literacy'. This model was tested as the assessment principles represented by the standards all pertain to teachers' knowledge and skills in the area of student assessment. The analysis and results are presented below.

The initial analysis included all the items and the responses from all the participants. The fit statistics for each item were obtained. The results are presented in Table 4.9. As can be seen, the first run of the data provided results in which all items possessed acceptable fit statistic values except Item 22 that was found to be misfitting due to t-value of below the acceptable minimum range (-2.0). This item was

removed, as it indicated possible redundancy in the participants' responses (lack of expected stochastic fit or violation of local item independence) (Schumacker, 2004).

Table 4.9. Results of the initial analysis of the ALI items

Item #	Standard*	Estimate (Difficulty)	Error	IWMS	t
1	1	0.75	0.07	0.95	- 0.7
2	2	- 0.99	0.06	1.00	0.1
3	3	0.80	0.07	1.00	0.1
4	4	- 0.42	0.06	0.99	- 0.4
5	5	0.14	0.07	0.95	- 1.1
6	6	0.39	0.07	0.97	- 0.5
7	7	- 0.89	0.06	1.01	0.6
8	1	- 0.40	0.06	1.04	1.4
9	2	- 0.43	0.06	1.04	1.6
10	3	0.08	0.07	1.04	0.9
11	4	0.31	0.07	1.04	0.8
12	5	0.52	0.07	1.02	0.3
13	6	0.70	0.07	0.95	- 0.7
14	7	- 0.03	0.06	1.03	0.7
15	1	0.81	0.07	0.96	- 0.6
16	2	0.91	0.07	0.98	- 0.3
17	3	- 1.18	0.06	1.00	0.1
18	4	0.27	0.07	1.03	0.7
19	5	0.17	0.07	0.95	- 1.1
20	6	- 0.03	0.06	1.02	0.6
21	7	- 0.14	0.06	1.02	0.6
22	1	- 0.92	0.06	0.94	-2.7**
23	2	0.60	0.07	1.02	0.3
24	3	0.68	0.07	1.06	0.9
25	4	0.45	0.07	1.01	0.3
26	5	0.39	0.07	1.01	0.2
27	6	- 0.77	0.06	1.02	1.0
28	7	0.92	0.07	0.95	- 0.6
29	1	- 1.60	0.06	1.05	1.4
30	2	0.80	0.07	0.99	- 0.1
31	3	- 0.05	0.06	0.99	- 0.3
32	4	- 0.73	0.06	1.00	- 0.1
33	5	- 0.80	0.06	1.02	0.9
34	6	- 0.30	0.06	0.99	- 0.4
35	7	0.001	0.39	0.97	- 0.7

1 - Choosing Assessment Methods; 2 – Developing Assessment Methods; 3 – Administrating, Scoring, and Interpreting Assessment Results; 4 – Using Assessment Results; 5 – Developing Valid Grading Procedures; 6 – Communicating Assessment Results; 7 – Recognizing unethical, illegal, and otherwise inappropriate Assessment Methods and Uses of Assessment Results;Misfitting*

After the removal of item 22, the whole data set was recalibrated. The results of the second run are shown in Table 4.10 on the next page. As presented, all items were found to fit the Rasch Model in the second and final run as indicated by their corresponding fit statistic values. The value of separation reliability (0.99) was indicative of the small measurement error, high discriminating power (Alagumalai & Curtis, 2005), and high measurement capacity (Wright & Stone, 1999). The acceptable fit statistic and separation reliability values provide evidence that the ALI items are of acceptable quality. The Rasch item analysis results indicated that the 34 remaining items of the ALI conformed to the Rasch Model and satisfied the unidimensionality requirement. Hence, the ALI items were deemed appropriate in measuring the teacher assessment literacy as a single or dominant dimension.

Item 22 asks about the appropriate method of evaluating the pupils' writing skills. It is believed that this item is familiar to many Tawi-Tawi's elementary and secondary school teachers as it asks about the common evaluation method and skills. As such, it was unexpected that this item did not fit the Rasch Model. Perhaps, the over-fitted t value for this item is influenced by some identical or similar pattern of responses from the participants. However, further analysis is warranted.

Table 4.10. Results of the final item analysis of the ALI items

Item #	Standard*	Estimate (Difficulty)	Error	IWMS	t
1	1	0.72	0.07	0.92	- 1.2
2	2	- 1.01	0.06	1.01	0.3
3	3	0.77	0.07	0.99	- 0.1
4	4	- 0.44	0.06	1.01	0.2
5	5	0.11	0.07	0.96	- 1.0
6	6	0.36	0.07	0.98	- 0.5
7	7	- 0.92	0.06	1.03	1.3
8	1	- 0.43	0.06	1.02	0.6
9	2	- 0.45	0.06	1.05	1.8
10	3	0.05	0.07	1.03	0.8
11	4	0.28	0.07	1.01	0.1
12	5	0.49	0.07	0.99	- 0.1
13	6	0.67	0.07	0.95	- 0.9
14	7	- 0.05	0.06	1.00	0.0
15	1	0.78	0.07	0.95	- 0.7
16	2	0.88	0.07	0.97	- 0.3
17	3	- 1.20	0.06	1.01	0.2
18	4	0.24	0.07	1.02	0.4
19	5	0.14	0.07	0.91	- 2.0
20	6	- 0.06	0.06	1.00	0.0
21	7	- 0.17	0.06	1.02	0.7
22	2	0.57	0.07	1.00	0.0
23	3	0.65	0.07	1.03	0.4
24	4	0.42	0.07	0.99	- 0.3
25	5	0.36	0.07	1.01	0.2
26	6	- 0.80	0.06	1.02	1.0
27	7	0.90	0.07	0.95	- 0.7
28	1	- 1.62	0.06	1.06	1.9
29	2	0.78	0.07	0.98	- 0.3
30	3	- 0.08	0.06	1.00	- 0.0
31	4	- 0.76	0.06	0.98	- 0.8
32	5	- 0.83	0.06	1.03	1.3
33	6	- 0.33	0.06	0.99	- 0.5
34	7	- 0.03	0.38	0.98	- 0.4

Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 3306.81; df=33; Sig level=0.000

4.7 Analysis of the ALI Structure using Confirmatory Factor Analysis (CFA)

To determine the ALI scale at the macro-level, the structure-level analysis was carried out. The main purpose of doing this analysis was to examine the structure of the ALI scale and if the proposed model for the ALI fits the data. As the ALI was devised using the AFT, NCME, and NEA's (1990) standards, the

CFA was employed. To run the analysis, LISREL 8.80 software (Jöreskog & Sörbom, 2006) was used. The succeeding subsections present and discuss CFA results.

4.7.1 Structural Analysis using CFA

The 34 ALI items that fit the Rasch Model were subjected to CFA. The items were tested in terms of the seven-factor model as hypothesised by the authors of the ALI. The seven-factor structure of the ALI corresponded to the Standards (Standard 1 to Standard 7) for Teacher Competence in the Educational Assessment of Students developed by AFT, NCME, & NEA (1990). Under this model, the standard represents the latent factor (the unobserved factor) and the items serve as the manifest variables (the observed factors). Standard 1 has four items (items 1, 8, 15, and 29) while the rest of the standards have five items each: Standard 2 - items 2, 9, 16, 23, and 30; Standard 3 - items 3, 10, 17, 24, and 31; Standard 4 - items 4, 11, 18, 25, and 32; Standard 5 - items 5, 12, 19, 26, and 33; Standard 6 - items 6, 13, 20, 27, and 34; and Standard 7 - items 7, 14, 21, 28, and 35. The structure of the seven-factor model is presented in Figure 4.2.

4.7.1.1. Model Fit

In evaluating the ALI's hypothesised model, the overall model fit to the data was first examined using the results of chi-square (χ^2) statistic, ratio of chi-square to its degrees of freedom, root mean square error of approximation (RMSEA), standardised root mean square residual (SRMR), goodness-of–fit index (GFI), adjusted goodness-of-fit index (AGFI), comparative fit index (CFI), and parsimony goodness-of-fit index (PGFI). These indices and their corresponding permissible values/ranges were discussed in Chapter 3.

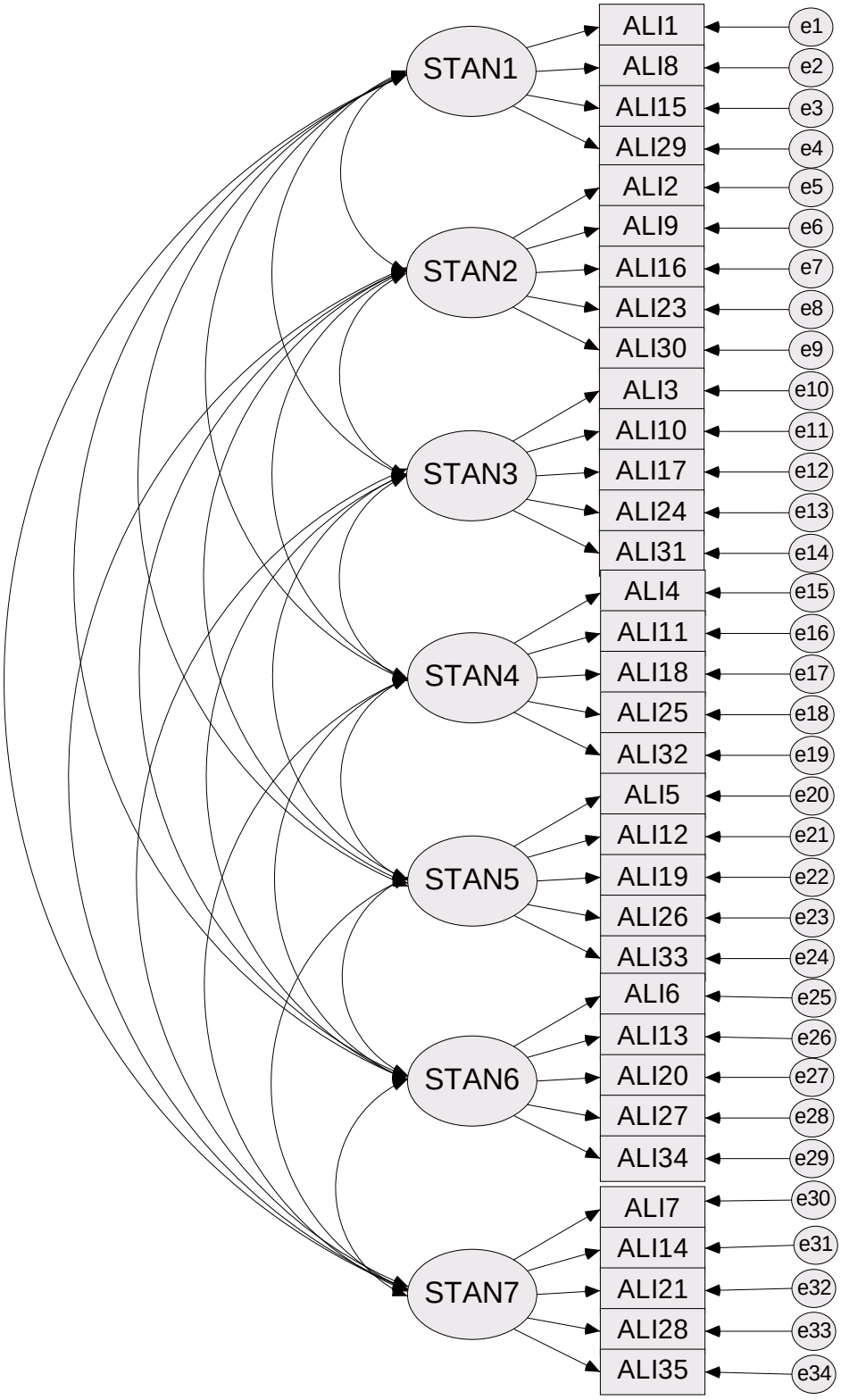


Figure 4.2. Structure of the Seven-Factor Model for the ALI

Table 4.11 presents the summary results of the used fit indices for the ALI's seven-factor structure. Of the eight fit indices reported, only two (RMSEA and SRMR) showed acceptable fit while the other five (X^2, X^2/df, GFI, AGFI, and CFI) revealed poor fit of the model to the data. Moreover, it was noted that although RMSEA and SRMR exhibited acceptable fit, their values were closer to the adopted thresholds. Furthermore, the PGFI indicated some complexity in the ALI's hypothesised model. These results generally indicate that the ALI's seven-factor structure did not fit the data well and therefore the current factorial structure can be challenged.

Table 4.11. Summary results of fit indices for the seven-factor ALI structure

Fit Index	Obtained Value	Remark
X^2	3433.70 (P = 0.0) (significant)	Poor fit
X^2/df	3433.70/506 = 6.79	Poor fit
RMSEA	0.09	Acceptable fit
SRMR	0.08	Acceptable fit
GFI	0.77	Poor fit
AGFI	0.73	Poor fit
CFI	0.53	Poor fit
PGFI	0.66	Some model complexity

4.7.1.2. CFA of the ALI Hypothesised Measurement Model

In addition to checking the overall model fit, the ALI items were examined using the factor loadings to gauge whether or not the items reflected the factors that they were presumed to represent. The scoring of the ALI was dichotomous with responses being scored as either correct (1) or incorrect (0). As such, a factor loading of 0.30 (Kline, 1994) was employed. The factor loadings are presented in Table 4.12.

As can be seen from Table 4.12, the majority of the ALI items have factor loadings below the adopted threshold of 0.30. The number of ALI's items (by standard) that exhibited acceptable factor loadings are as follows: Standard 1 – two items (ALI1 & ALI15); Standard 2 – none; Standard 3 – three items (ALI3, ALI17, & ALI31); Standard 4 – two items (ALI11 & ALI 32); Standard 5 – two items (ALI12 &

ALI19); Standard 6 – two items (ALI6 & ALI13); and Standard 7 – four items (ALI7, ALI21, ALI28, & ALI35). In total, only 15 items exhibited acceptable factor loadings. These results reveal that the items do not uniquely represent the factors under the seven-factor structure of the ALI and are consistent with the finding regarding the poor model fit of the ALI's seven-factor structure. Hence, it seems that the hypothesised seven-factor structure may not be appropriate for the ALI.

Table 4.12. Factor loadings of ALI items under the seven-factor model

Factor	Item	Loading(se)*
1 (Standard 1)	ALI1	0.55(0.05)
	ALI8	0.28(0.05)
	ALI15	0.59(0.05)
	ALI29	0.11(0.05)
2 (Standard 2)	ALI2	0.17(0.04)
	ALI9	0.13(0.04)
	ALI16	0.26(0.05)
	ALI23	0.18(0.04)
	ALI30	0.21(0.05)
3 (Standard 3)	ALI3	0.58(0.05)
	ALI10	0.19(0.05)
	ALI17	0.49(0.05)
	ALI24	0.26(0.05)
	ALI31	0.36(0.05)
4 (Standard 4)	ALI4	0.15(0.04)
	ALI11	0.31(0.05)
	ALI18	0.17(0.04)
	ALI25	0.29(0.05)
	ALI32	0.30(0.05)
5 (Standard 5)	ALI5	0.29(0.04)
	ALI12	0.31(0.04)
	ALI19	0.42(0.04)
	ALI26	0.23(0.04)
	ALI33	0.11(0.03)
6 (Standard 6)	ALI6	0.34(0.04)
	ALI13	0.43(0.04)
	ALI20	0.28(0.04)
	ALI27	0.21(0.04)
	ALI34	0.27(0.04)
7 (Standard 7)	ALI7	0.31(0.05)
	ALI14	0.28(0.05)
	ALI21	0.33(0.05)
	ALI28	0.60(0.05)
	ALI35	0.38(0.05)

**n = 582*

4.8 Confirmatory Factor Analysis of the Alternative Model

To provide alternative model to what the ALI proponents had hypothesised, a one-factor model was further tested for the ALI structure. This was carried out as the 34 ALI items could possibly represent a single or unitary dimension that can be labeled 'assessment literacy'. Although this has already been established through the Rasch Model, the analysis could be done to provide comparison and a better understanding of the ALI structure. The following subsections present and discuss the CFA results of the alternative model.

4.8.1. Structural Analysis of the Alternative Model using CFA

The 34 ALI items were further subjected to CFA. The items were examined in terms of one-factor structure. All items were loaded to one latent construct called the "assessment literacy (ASLIT)". This model was evaluated using the same technique and fit indices as discussed in Chapter 3. The structure of the model is presented in Figure 4.3.

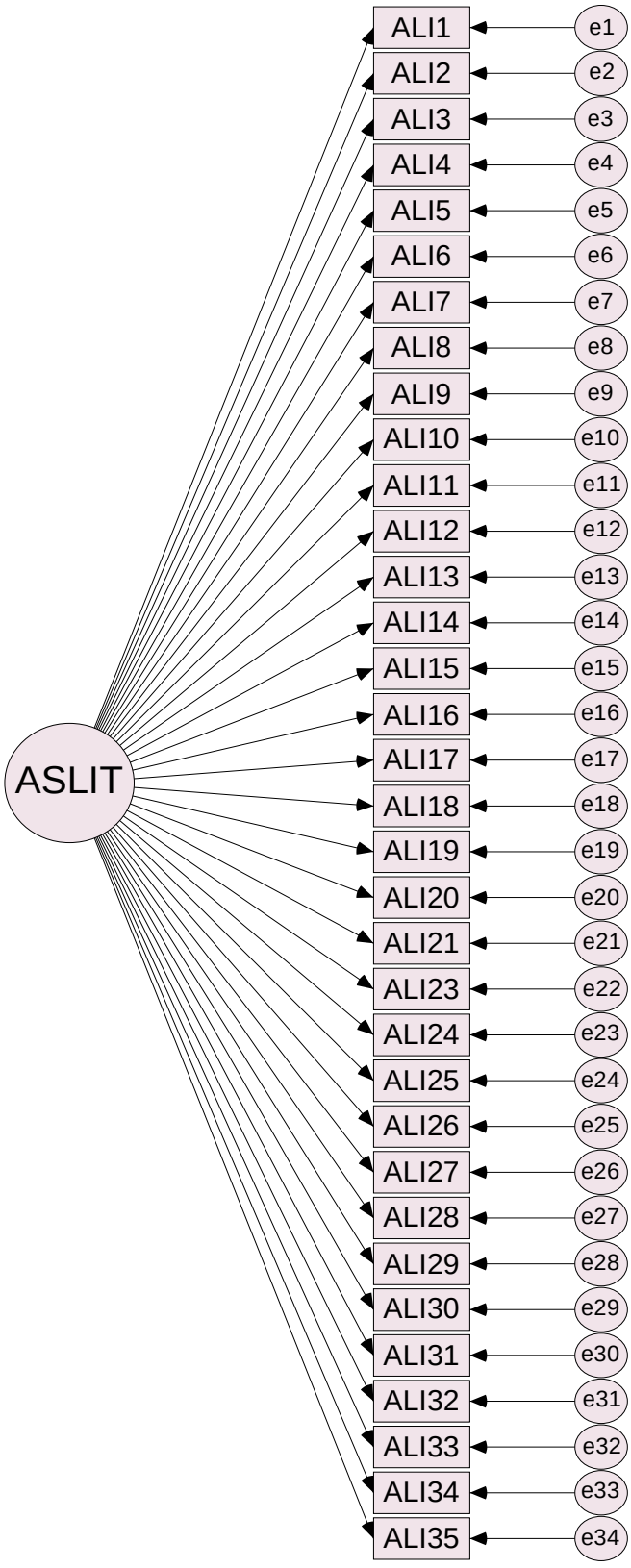


Figure 4.3. Structure of one-factor model for ALI

4.8.1.1. Model Fit

Similar to the previous analyses, the overall model fit of the ALI's hypothesised alternative model to the data was initially examined using the results of chi-square (χ^2) statistic, ratio of chi-square to its degrees of freedom, root mean square error of approximation (RMSEA), standardised root mean square residual (SRMR), goodness-of–fit index (GFI), adjusted goodness-of-fit index (AGFI), comparative fit index (CFI), and parsimony goodness-of-fit index (PGFI) (see Chapter 3 for details about these indices). The results are shown in Table 4.13.

It can be gleaned from Table 4.13 that the CFA results of one-factor model are similar with what were obtained under the seven-factor model. Although the resulting values of X^2, X^2/df, GFI, CFI, and PGFI increased/decreased, they likewise indicate poor fit and some model complexity. Hence, it can be deduced from CFA results that a one-factor structure still appears inappropriate for the ALI.

Table 4.13. Summary results of fit indices for the one-factor ALI structure

Fit Index	Obtained Value	Remark
X^2	3640.97 (P = 0.0) (significant)	Poor fit
X^2 /df	3640.97/527 = 6.91	Poor fit
RMSEA	0.09	Acceptable fit
SRMR	0.08	Acceptable fit
GFI	0.76	Poor fit
AGFI	0.73	Poor fit
CFI	0.50	Poor fit
PGFI	0.68	Some model complexity

4.8.1.2. CFA of the ALI Hypothesised Alternative Measurement Model

The ALI items were also examined under the one-factor structure. The results (factor loadings) are presented in Table 4.14. As can be spotted from the table, majority of the items (19 items) exhibited factor loadings of at least 0.30, the threshold adopted for the ALI scale. These items are ALI1, ALI3, ALI5, ALI6,

ALI11, ALI12, ALI13, ALI15, ALI16, ALI19, ALI20, ALI23, ALI26, ALI27, ALI29, ALI31, ALI32, ALI33, and ALI35. It appeared that more items could be accepted under this tested model. However, with still 15 items remaining unacceptable due to low factor loadings, the one-factor structure for the ALI can also be challenged. These results are consistent with the poor overall model fit of one-factor structure to the data. These suggest that ALI may have other factorial structure and thus further CFA of ALI's structure using other models is warranted in future studies.

Table 4.14. Factor loadings of ALI items under the one-factor model

Factor	Item	Loading(se)*
Assessment Literacy (ASLIT)	ALI1	0.56(0.04)
	ALI2	0.26(0.04)
	ALI3	0.37(0.04)
	ALI4	0.24(0.04)
	ALI5	0.42(0.04)
	ALI6	0.46(0.04)
	ALI7	0.23(0.04)
	ALI8	0.26(0.04)
	ALI9	0.22(0.04)
	ALI10	0.21(0.04)
	ALI11	0.33(0.04)
	ALI12	0.40(0.04)
	ALI13	0.46(0.04)
	ALI14	0.27(0.04)
	ALI15	0.49(0.04)
	ALI16	0.39(0.04)
	ALI17	0.28(0.04)
	ALI18	0.21(0.04)
	ALI19	0.53(0.04)
	ALI20	0.30(0.04)
	ALI21	0.28(0.04)
	ALI23	0.43(0.04)
	ALI24	0.27(0.04)
	ALI25	0.21(0.04)
	ALI26	0.34(0.04)
	ALI27	0.32(0.04)
	ALI28	0.14(0.05)
	ALI29	0.53(0.04)
	ALI30	0.12(0.05)
	ALI31	0.30(0.04)
	ALI32	0.37(0.04)
	ALI33	0.38(0.04)
	ALI34	0.18(0.05)
	ALI35	0.35(0.04)

**n = 582*

4.9 Model Used in the Study

The seven-factor and one-factor models for the ALI were tested in this study using the Rasch Model and CFA/SEM. The Rasch analysis results indicated that both models/factorial structures are appropriate for the tested instrument. However, CFA results revealed otherwise. While the results of Rasch Model and CFA appeared contradictory, the acceptable structures revealed by the former can be adopted. The factorial structure that fits the Rasch Model is believed to be sound due to the strength of Rasch Model as a validation technique. According to Cavanagh & Romanoski (2006), Rasch technique is powerful due to its strict adherence to measurement requirements; thus, it can be utilised to establish the measurement capacity of any instrument. Moreover, Rasch Model seems to be superior to those of Classical Test Theory (CTT), which includes CFA/SEM. Its theoretical and mathematical foundations are more grounded than its classical counterpart (Ewing, Salzberger & Sinkovics, 2005). Hence, both seven-factor model and one-factor model were used in this study.

4.10 Summary

In this chapter, the ALI scale was validated using the data collected from the 582 elementary and secondary school teachers. The ALI was validated at the "micro" and "macro" levels. For the micro-level analysis, items were analysed using the Rasch model. Rasch analysis involved items under the seven-factor structure and the one-factor structure. For the macro-level analysis, the ALI's seven-factor and one-factor models were determined using the CFA technique. The results of Rasch analysis indicated that both seven-factor and one-factor models are appropriate for the ALI. However, the CFA provided contradictory results. Based on Rasch analysis results, this study used both models for the ALI. The seven-factor model can be used to gauge the teacher assessment literacy on the seven specific assessment areas/principles covered in the study. Moreover, the one-factor model can be utilised to examine the general assessment literacy of teacher respondents.

Chapter 5: The Assessment Practices Inventory

5.1 Introduction

This study attempted to investigate the influence of teacher assessment literacy on other relevant attributes. The attempt was spurred by the view that assessment literacy can possibly affect other teachers' characteristics, which further impact on their performance in carrying out the teaching-learning process. One of the characteristics deemed affected by assessment literacy is the 'assessment practices'.

The consideration of the teacher assessment practices stems from the view that knowledge is related to practice. The close link between these attributes has long been assumed in education and allied disciplines. However, it does not always follow that one's knowledge results to one's practice. Thus, more evidence is needed to establish the direct effect of knowledge on practice. In this study, the influence of teachers' assessment literacy on their assessment practices was examined to substantiate the assumed link and to provide a context for professional development. Figure 5.1 depicts the general relationship concerning assessment practices and other tested variables. To carry out the investigation, an instrument was needed. Hence, the 'Assessment Practices Inventory (API)' was developed. In this chapter, the development and calibration of the API are described.

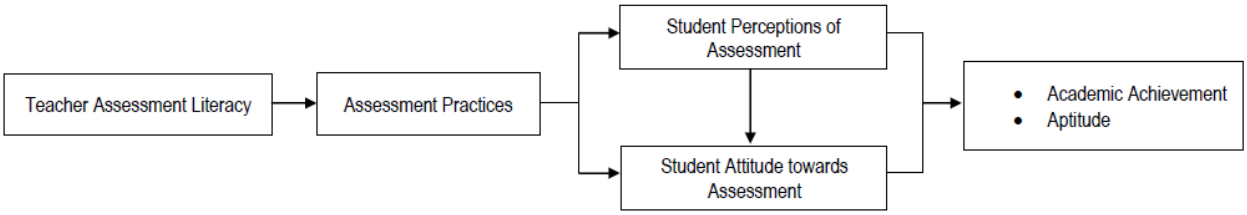


Figure 5.1. The relationship among teacher assessment literacy, assessment practices, and student outcomes

The chapter begins with the information on the instrument development, in which the conceptual framework and the process of constructing and revising the items are described. The next section deals

with the pilot test of the API. In this part, the purpose, process, and results of the pilot test are presented. The ensuing section focuses on the calibration of the API, first at the item level and finally at the structural level. The chapter concludes by reiterating the key points.

5.2 Development of the Assessment Practices Inventory (API)

The development of the API was generally based on the goals of the study. It was conceived that this study examined the influence of teacher assessment literacy on teacher assessment practices using parallel instruments to ascertain the causal relationship and to allow coherent interpretation of the results. As the ALI was employed in examining the teacher assessment literacy, it was then necessary that the API resembled the ALI in terms of development framework. Thus, it was intended that the API items and structure should also be aligned with the Standards for Teacher Competence in the Educational Assessment of Students (AFT, NCME, & NEA, 1990). In other words, the seven standards as used in the ALI served as the conceptual framework for the development of the API.

Prior to the development of the API, the search for any established instrument that matched the study's plan was initially conducted. A number of assessment practices questionnaires were found from the available literature. However, these questionnaires measured different contents and adopted different structures. None of the questionnaires was explicitly following the assessment standards as used in the ALI. Thus, the researcher developed the API using some existing scales and documents as guide.

The researcher specifically used the ALI, the assessment practices section of the teacher questionnaire of the Pan-Canadian Assessment Program (CCME, 2007; 2010), Practices of Assessment Inventory (Brown, Kennedy, Fok, Chan, & Yu, 2009), and the Third International Mathematics and Science Study (TIMSS) science teacher questionnaire (IEA, 1999) as guide for item construction and response category structure. These existing scales had been identified as guide due to their relevance to the purpose and context of the study, and to the intended scope and structure of the instrument. To ensure that the items were also consistent with the Tawi-Tawi or Philippine context, the National Competency-Based

Teacher Standards (NCBTS) (DepEd, 2009) and the department orders concerning the practices of assessment in Philippine schools (DepEd Order Nos. 4, 33, and 92, s. 2004) were likewise consulted. Using these scales and documents, the items for each of the seven standards were constructed.

During the item construction phase, effort was made to ensure that all statements in the API reflect assessment practices. Moreover, the principles of item construction were observed. Each item depicted a single scenario or a single concept and was phrased as clearly as possible in relation to the standard. The language and the scenarios used were made as applicable as possible to all participants. Each item was also made brief to elicit easy and quick response, and to draw the interest of the participants. All items adopted a five-point response category. The initial development of the API resulted to it having 66 items under a five-point Likert scale of "never", "seldom", "occasionally", "frequently", and "all the time" that were coded as 1, 2, 3, 4, and 5, respectively. These included items with subsections/extensions and the "open-ended" type. The distribution of the items with respect to the standards was as follows: 1st Standard – 21 items; 2nd Standard – 6 items; 3rd Standard – 6 items; 4th Standard – 5 items; 5th Standard – 17 items; 6th Standard – 6 items; and 7th Standard – 5 items. Items in the first six standards were all made positive while those in the 7th standard carried negative implications.

As a new scale, it was necessary for the API to be subjected to a rigorous calibration process to ascertain its validity and reliability and to ensure that the data derived from it are useful for further analysis. Thus, the instrument underwent validation phase. The initial stage was the face validation. To check the contents, relevance, and appropriateness of the statements for the Philippine/Tawi-Tawi context, the researcher, in consultation with his supervisors, reviewed the API items. After the review, the items were further face-validated by three colleagues at the University of Adelaide in Australia and by two colleagues at the Mindanao State University – Tawi-Tawi College of Technology and Oceanography in the Philippines. Based on their comments and suggestions, the conceptual paradigm for the development of the API was changed and some items were revised, regrouped, and deleted. Hence, the API finally followed the 'Keys to Quality Classroom Assessment' by Stiggins, et al. (2007) in its final development.

The keys to quality classroom assessment as proposed by Stiggins, et al. (2007) emphasise four components that are essential in making classroom assessment accurate and effective. These components include 'assessment purpose', 'learning target', 'assessment design', and 'assessment communication'.

The 'assessment purpose' emphasises that classroom assessment should begin with a clear purpose. Stiggins, et al. (2007) asserted that the initial consideration when doing assessment should be the information needs of assessment users that include both teachers and students. As expounded, teachers use results or evidence from assessment to support their instructional decisions that have bearing on the improvement of their teaching and student learning. Moreover, students use assessment results to decide which to focus in their studies and to devise other ways that will help them achieve better learning. The way teachers and students use assessment results should be part of the assessment purpose. The 'learning target' is the specific objective or achievement expectation that the teachers wish their students to attain. This encompasses any important subject matter or knowledge and skills that the students need to learn. In doing classroom assessment, it is essential that at the outset, teachers should have a clear sense of what they want their students to achieve, as this will dictate the assessment methods and procedures to be employed. Having a clear target is the foundation of good teaching and sound assessment (Stiggins, et al., 2007). The 'assessment design' stresses the use of appropriate methods and procedures to obtain accurate assessment results. This part is strongly tied to assessment purpose and learning target. If teachers begin with assessment purpose with respect to the information needs of the assessment users, and define clear target to assess or measure, then they should be able to employ assessment methods and procedures that are appropriate for the said purpose and target. This step should be able to help teachers in arriving at accurate results from which sound decision can be made. Furthermore, when utilising any assessment method and procedure, teachers should consider a particular context, an appropriate sampling of student achievement, and the quality of assessment methods and scoring procedures "to avoid all potential sources of bias" and to ensure dependable results. Having accurate assessment results would strengthen decisions, improve instruction, and accurately reflect and enhance student learning (Stiggins, et al., 2007, p. 16).

Finally, the 'assessment communication' focuses on the need and the way to effectively communicate assessment results to the intended users "in a timely and understandable manner". It has been pointed out that "assessment fails to achieve its learning ends" if the results are not properly conveyed to the assessment users. Thus, it is necessary that the assessment results should correctly and appropriately reach those intended to use them. However, it has also been pointed out that ineffective communication can result to poor-quality decisions, which are detrimental to student learning. To prevent ineffective communication, it is important that "everyone must understand the meaning of the achievement target, the information underpinning the communication must be accurate, everyone must understand the symbols being used to convey information, and the communication must be tailored to the intended audience, e.g., level of detail, timing, and format" (Stiggins, et al., 2007, p. 17).

The API was revised following the aforementioned 'keys to quality classroom assessment'. However, only 'assessment purpose', 'assessment design', and 'assessment communication' were used as bases, as the scale focused on general assessment practices. Moreover, the API items on 'assessment design' were separated into two groups. Items that involved assessment methods were grouped together to elicit response on whether or not the methods were used by the respondents. As these items reflected just the frequency and/or percentage of use, they were excluded from the construct validation. Hence, only items that pertained to assessment procedures were retained to reflect the 'assessment design' component. The API carried 21 items in its final form and the items were distributed as follows: 1st factor (assessment purpose) – 9 items; 2nd factor (assessment design) – 7 items; and 3rd factor (assessment communication) – 5 items. It was believed that this conceptual framework was also covered in the seven standards as employed in the ALI. Thus, the examination of the link and any effect of assessment literacy on assessment practices were believed to be still appropriate. The 21 API items (see Table 5.1/Appendix B) were initially pilot tested and were subjected to the calibration process at the item and structural levels.

Table 5.1. The API items

Item Number	Item Statement
1	I use assessment to check the attainment of lesson objectives.
2	I use assessment to establish student learning.
3	I use assessment to increase student learning.
4	I use assessment to develop students' higher order thinking skills.
5	I prepare table of specifications as my guide in constructing test.
6	I construct test that measures attribute/behavior as stated in my teaching objectives.
7	I use clear directions when giving assessment like tests and projects.
8	I use answer key when marking objective tests like multiple choice, true-false, and matching types.
9	I use rubrics when marking other assessment types such as essay tests, projects, and student demonstration.
10	I use reference table or standard procedure in transmuting scores into grades.
11	I use established procedure in deriving grades from different assessment methods.
12	I interpret assessment results according to the established scale.
13	I use assessment results to plan my instruction.
14	I use assessment results to determine the pace of my instruction.
15	I use assessment results to determine the strategies that suit my student learning needs.
16	I use assessment results to provide feedback to my students.
17	I explain to my students and their parents how grades are derived.
18	I explain to my students and their parents the meaning of assessment results.
19	I explain to my students and their parents the meaning of the national/regional examination results (e.g., average score, percentile rank, etc.).
20	I write comments on student test papers.
21	I write comments on student report card.

As the API instrument was developed by the researcher using the existing questionnaires and the assessment practices framework from the literature, it was necessary to subject it to expert validation to ensure content relevance.

5.3 Pilot Test of the API

The 21 API items were organised into one section and formed part of the teacher survey. The items were pilot tested, together with the other items in the study's teacher questionnaire, to 45 MSU Tawi-Tawi

elementary and secondary school teachers to obtain the initial reliability, elicit suggestions to improve the instrument, and to determine the time for questionnaire completion. During the pilot test, the participants were requested to write their suggestions to improve the instrument and to write the time to start and the time to complete the questionnaire. The suggestions of the pilot test participants and the amount of time to accomplish the questionnaire were noted and used in finalising the instrument. Specifically, the suggestions to make the items as brief and familiar as possible were followed in finally framing the API items. A Chronbach alpha of 0.89, which indicated acceptable reliability, was obtained for this scale.

5.4 Calibration of the API

The calibration of the API was done at the item and structural levels. At the item level, the Rasch Model, particularly the Rating Scale Model (Andrich, 1978), was employed. The item-level analysis was carried out using ConQuest 2.0 software (Wu, Adams, Wilson, & Haldane, 2007). At the structural level, CFA was used as the API had been developed using a priori. The CFA was performed using LISREL 8.80 software (Jöreskog & Sörbom, 2006). The Rasch Model and CFA have been described/discussed in Chapter 3. However, as the validation of this scale specifically employs the Rating Scale Model, this technique is further described in this chapter to provide additional relevant information. Discussion of the calibration results follows after the description of this validation technique.

5.4.1 The Rating Scale Model

The rating scale model is an extension of Rasch model. It is an item and category analysis technique for polytomous or Likert-type data. It describes a probabilistic relationship between item's inherent intensity or affective value and person's attitude towards the item. Under this technique, the latent trait is also assumed to be unidimensional. Responses are likewise determined by the interval between boundaries called 'thresholds'. Thus, computation of threshold for every adjacent response category in a scale is expected. For any given set of response categories, the number of thresholds is one less than the actual number of categories. Each threshold has its own estimate of difficulty (endorsement/dilemma),

which is assumed to be equidistant with the rest of the threshold estimates for a particular item or for a set of categories (Andrich, 1978; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008).

5.4.2 Item Analysis Using the Rating Scale Model

The API was analysed at the micro level (item level) employing the Rating Scale Model. The 21 API items were all subjected to analysis using ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007). To judge the acceptability of the items, the residual-based fit statistic was used. The unweighted mean square (UMS) was particularly employed to indicate whether or not an item conforms to the Rasch Model. The choice for UMS as the indicator was due to its strength in detecting misfitting items, in providing stable Type I error rates (Smith, 2004; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008), and for being sample independent for polytomous data (Smith, et al., 2008). The critical range of 0.70 – 1.30 as used in many studies (Smith, et al., 2008) was used for the UMS value. Items that had UMS values outside the adopted range were removed one at a time as they violated the measurement requirements. Moreover, the category structure was examined for possible reversed or disordered thresholds as these hint at problems related to the process of responding to the item (Andrich, 1995). In the analysis of polytomous data, a t statistic is not employed as this is affected by the sample size (Smith, et al., 2008; Wu & Adams, 2007; Curtis, 2004).

As the Rasch Model was employed to further assess the measurement properties of the API at the item level and to support the decision on the choice of a particular model that is appropriate for the API, the analysis was done on three-factor model and one-factor model. For the three-factor structure, a separate analysis for each factor or construct was performed. All items that were hypothesised to reflect each construct were analysed. Using the same process, all items were also analysed under one-factor structure. In analysing the API, item fit to the Rasch Model was initially examined. After which, the ordering of thresholds was verified.

Under the three-factor structure, Rasch analysis was first carried out for the nine items (items 1, 2, 3, 4, 6, 13, 14, 15, and 16) that were hypothesised to reflect the ‘assessment purpose’. The results of the

initial analysis are presented in Table 5.2. As shown, only one item (item 16) exhibited misfit. Its UMS of 1.67 indicates underfit with respect to the model and implies a noise or unpredictable responses from the participants. Item 16 is about the use of assessment results to provide feedback to students. The practice asked in this item was expected to be executed by the participants, as teachers in Tawi-Tawi often use assessment results like test scores to provide information about student achievement. Perhaps, the item did not fit the Rasch model due to case or person misfit (Curtis, 2004). Thus, the item was discarded.

Table 5.2. Results of the initial analysis of the API items under the assessment purpose

Item	Estimate (Difficulty)	Error	UMS	t
1	- 0.290	0.044	0.99	- 0.1
2	- 0.393	0.044	0.83	- 3.1
3	- 0.472	0.044	0.88	- 2.1
4	- 0.022	0.043	0.88	- 2.1
6	0.020	0.043	1.08	1.3
13	0.148	0.042	1.03	0.5
14	0.529	0.041	0.80	- 3.6
15	- 0.177	0.043	0.83	- 3.0
16	0.656*	0.122	1.67**	9.6

*Separation Reliability = 0.98; Chi-Square Test of Parameter Equality = 432.56; df=8; Sig level=0.000; *Constrained; **Misfitting*

After the removal of item 16, the remaining eight items were recalibrated. The results are shown in Table 5.3. As can be seen from the table, all the remaining items were fitting the Rasch model as indicated by the acceptable UMS values. In terms of arrangement of threshold values, it was of increasing order as expected. The absence of disordered threshold indicates that the category structure works well as intended. Moreover, the value of separation reliability (0.98) was indicative of the small measurement error and high discriminating power (Alagumalai & Curtis, 2005; Ben, 2010). This further indicates that the items have more precise measurement and reliability (Wright & Stone, 1999). Hence, the eight items can be finally taken to measure the 'assessment purpose'.

Table 5.3. Results of the final item analysis of the API items under assessment purpose

Item	Estimate (Difficulty)	Error	UMS	t
1	- 0.236	0.050	1.11	1.9
2	- 0.351	0.051	0.94	- 1.0
3	- 0.438	0.051	0.87	- 2.3
4	0.065	0.049	0.87	- 2.4
6	0.112	0.049	1.17	2.7
13	0.259	0.049	1.05	0.9
14	0.697	0.047	0.94	- 1.0
15	- 0.109*	0.131	0.90	- 1.7

*Separation Reliability = 0.98; Chi-Square Test of Parameter Equality = 398.49; df=7; Sig level=0.000; *Constrained*

A separate Rasch analysis was done for the seven items (items 5, 7, 8, 9, 10, 11, and 12) that were hypothesised to reflect the 'assessment design'. The initial analysis involving all the items and the responses was performed. The results are provided in Table 5.4. As presented, all the items appeared to fit the Rasch model as indicated by the acceptable UMS values. Examination of the category structure also revealed no disordered thresholds, which indicates that the items and corresponding response categories were functioning well. A separation reliability of 0.99 also implies that the items were precise and reliable. Thus, all the seven items can be retained as indicators of the second construct – the 'assessment design'.

Table 5.4. Results of the initial and final item analysis of the API items under assessment design

Item	Estimate (Difficulty)	Error	UMS	t
5	0.188	0.037	1.25	3.9
7	- 0.825	0.041	1.00	0.1
8	0.088	0.037	1.27	4.3
9	0.717	0.035	0.96	- 0.7
10	- 0.609	0.040	0.94	- 1.0
11	0.174	0.037	0.85	- 2.7
12	0.268*	0.092	0.80	- 3.7

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 1119.27; df=6; Sig level=0.000; *Constrained*

The next group of items that were analysed was those of 'assessment communication'. Five items (items 17, 18, 19, 20, and 21) were calibrated using all the responses from the participants. The results of the initial analysis are given in Table 5.5. As can be gleaned from the table, one item (item 21) exhibited misfitting result as indicated by UMS value of 1.36. Item 21 is about writing comments on student report card. In the Tawi-Tawi/Philippine context, writing comments about student behaviour and performance on report card has been part of the practice, though not strictly required. Perhaps, some participants did not observe this practice or the current policy no longer requires it that led to varied responses and thus unpredictable answers from teacher respondents. The same characteristics were obtained for item 20, which is about writing comments on students' test papers. Teachers are usually expected to write comments on students' test papers as feedback but perhaps for similar reasons they provide responses that were also unpredictable. The items did not fit the Rasch Model possibly due to case or person misfit (Curtis, 2004). Hence, items 20 and 21 were deleted.

Table 5.5. Results of the initial item analysis of the API items under assessment communication

Item	Estimate (Difficulty)	Error	UMS	t
17	- 0.415	0.036	0.85	- 2.7
18	0.019	0.035	0.74	- 4.8
19	0.376	0.034	0.82	- 3.2
20	0.317	0.034	1.13	2.2
21	- 0.297*	0.070	1.36**	5.6

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 336.75; df=4; Sig level=0.000; *Constrained; **Misfitting*

After the removal of all misfitting items one at a time, final results were obtained. These results are presented in Table 5.6. As shown, three items appeared to fit the model as indicated by the acceptable UMS values. The absence of reversed thresholds and high separation reliability (0.99) also provide further support on the acceptable quality of these items. Thus, the three items can be retained to represent the third construct – the 'assessment communication'.

Table 5.6. Results of the final item analysis of the API items under assessment communication

Item	Estimate (Difficulty)	Error	UMS	t
17	- 0.725	0.046	1.06	1.0
18	0.041	0.045	0.84	- 2.9
19	0.684*	0.064	1.11	1.8

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 250.02; df=2; Sig level =0.000; *Constrained; **Misfitting*

Finally, the last series of Rasch analyses was performed for all the 21 items under the one-factor structure. Using the same process, the 21 items (items 1-21) were calibrated. The initial results are presented in Table 5.7. As the table shows, five items (items 8, 14, 15, 20, and 21) did not fit the Rasch model. In discarding the item, the most underfitting item or the item that has the most unpredictable response pattern was to be deleted first. Thus, item 21 was initially deleted.

Table 5.7. Results of the initial item analysis of the API items under assessment practices

Item	Estimate (Difficulty)	Error	UMS	t
1	- 0.459	0.038	1.02	0.3
2	- 0.539	0.039	0.79	- 3.8
3	- 0.599	0.039	0.75	- 4.7
4	- 0.252	0.037	0.79	- 3.9
5	0.097	0.036	1.21	3.4
6	- 0.219	0.037	0.85	- 2.6
7	- 0.918	0.040	1.04	0.7
8	- 0.001	0.036	1.69**	9.8
9	0.618	0.034	1.03	0.4
10	- 0.699	0.039	0.99	- 0.2
11	0.085	0.036	0.87	- 2.4
12	0.177	0.036	0.79	- 3.8
13	- 0.120	0.037	0.75	- 4.6
14	0.169	0.036	0.64**	- 7.1
15	- 0.370	0.038	0.69**	- 5.9
16	0.265	0.035	1.24	3.8
17	0.158	0.036	0.99	- 0.1
18	0.574	0.034	0.86	- 2.5
19	0.908	0.033	1.03	0.5
20	0.854	0.033	1.39**	5.9
21	0.272*	0.164	1.71**	10.1

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 3733.02; df=20; Sig level=0.000; *Constrained; **Misfitting*

After the removal of all misfitting items one at a time, the final calibration results were obtained. The results are shown in Table 5.8. As revealed, 16 items appeared to fit the Rasch Model as indicated by the UMS values. The acceptable quality of these items is further supported by the absence of reversed thresholds and by high separation reliability (0.99). Thus, these items can be employed to measure the main construct – the 'assessment practices'. As for the misfitting items (Items 5, 8, 16, 20, and 21), they were finally discarded as they violated the Rasch Model. Item 5 (I prepare table of specifications as my guide in constructing test) and item 8 (I use answer key when marking objective tests like multiple choice, true-false, and matching types) are part of the expected teachers' practices and it was unanticipated for these items to misfit the Rasch Model. Perhaps, the respondents answered the item with lack of attention (as sometimes happens in survey research) or with carelessness, which led to their unpredictable responses. Thus, case or person misfit can be invoked as possible reason for these items to misfit the Rasch Model.

Table 5.8. Results of the final item analysis of the API items under assessment practices

Item	Estimate (Difficulty)	Error	UMS	t
1	- 0.433	0.039	1.20	3.2
2	- 0.526	0.040	0.89	- 1.9
3	- 0.597	0.040	0.79	- 3.9
4	- 0.195	0.039	0.82	- 3.3
6	- 0.158	0.038	1.02	0.4
7	- 0.965	0.041	1.19	3.1
9	0.847	0.036	1.27	4.2
10	- 0.717	0.040	1.09	1.5
11	0.202	0.037	0.97	- 0.5
12	0.313	0.037	0.90	- 1.8
13	- 0.041	0.038	0.82	- 3.2
14	0.305	0.307	0.79	- 3.9
15	- 0.332	0.039	0.77	- 4.2
17	0.293	0.037	1.11	1.8
18	0.798	0.036	1.03	0.5
19	1.207*	0.148	1.27	4.3

Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 2796.28; df=15; Sig level=0.000;
**Constrained*

The Rasch analysis results have generally indicated that either the three-factor model or the one-factor model is an appropriate dimensional structure for the API. This finding supports the theoretical underpinning from which the API was developed. Moreover, the analysis results provide evidence that most of the API items have acceptable measurement properties. Thus, it can be concluded that the API possesses good psychometric properties.

5.4.3 Structural Analysis Using CFA

The API was analysed at macro level to determine its structure and the fit of the proposed measurement model to the data. In running the analysis, the LISREL 8.80 software (Jöreskog & Sörbom, 2006) was used. All the 21 API items were subjected to CFA. The items were first tested in terms of the three-factor structure upon which the final development of the API was based. The three-factor structure corresponded to the three key qualities of classroom assessment, i.e., 'assessment purpose', 'assessment design', and 'assessment communication'. Using this model, each key quality represents the latent factor (the unobserved factor) and the items serve as the manifest variables (the observed factors). The constructs and the items were mapped as follows: Assessment purpose - items 1, 2, 3, 4, 6, 13, 14, 15, and 16; assessment design – items 5, 7, 8, 9, 10, 11 and 12; and assessment communication – items 17, 18, 19, 20, 21. The conceptual structure of the three-factor model is presented in Figure 5.2 on the next page.

5.4.4.1 Model Fit

The API's hypothesised model was initially evaluated in terms of its overall fit to the data. The overall model fit was examined using a number of fit indices as discussed in Chapter 3. Table 5.9 presents the summary results of these indices. As can be gleaned from the table, two indices (SRMR and CFI) indicate acceptable model fit, PGFI indicates some degree of model complexity, and the other five indices (χ^2, χ^2/df, RMSEA, GFI, and AGFI) indicate poor model fit. These results generally indicate that the API's three-factor structure did not fit the data well. However, the proposed model appears to have some degree of acceptability as indicated by the two indices and therefore it cannot be totally rejected. Nevertheless,

other models need to be tested to determine the structure that fits the data better and to adopt the appropriate model for the assessment practices investigated in this study.

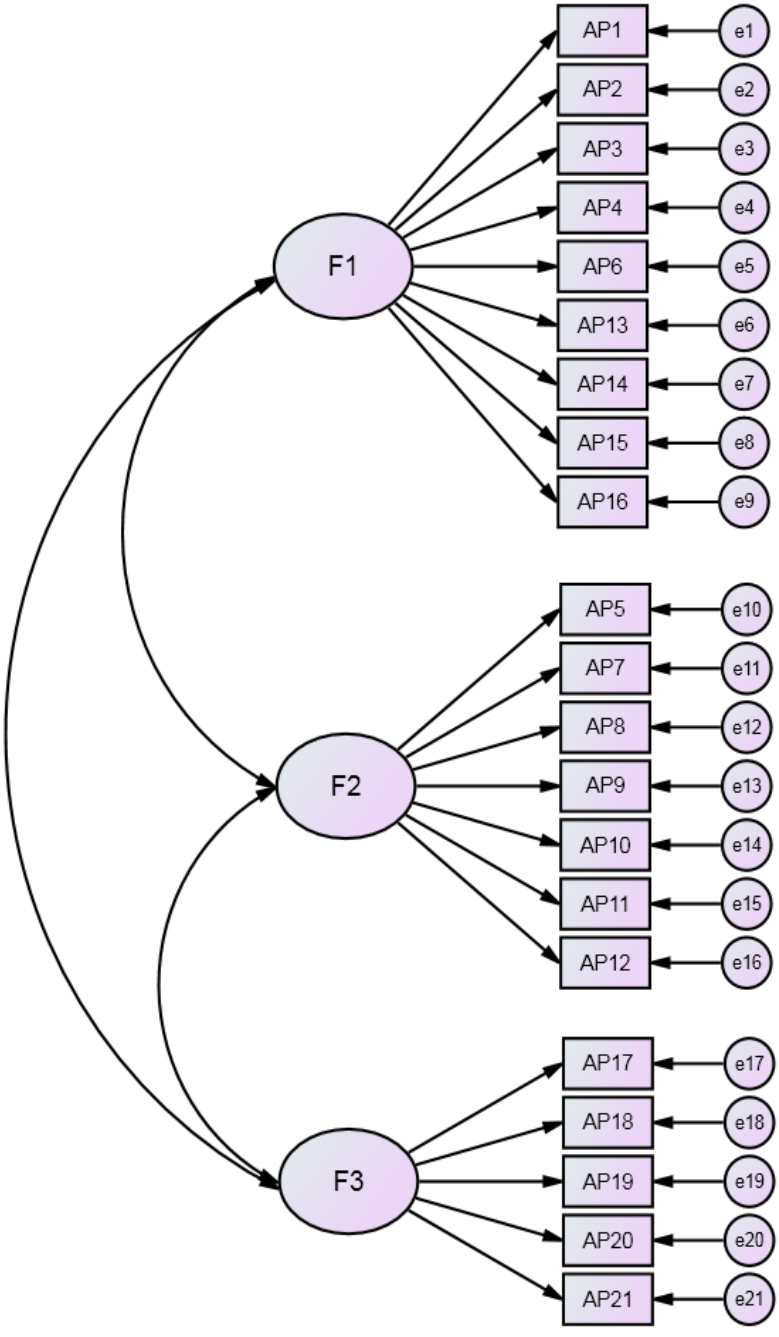


Figure 5.2. Structure of the three-factor model for API

Table 5.9. Summary results of fit indices for the three-factor API structure

Fit Index	Obtained Value	Remark
χ^2	1263.21 (P = 0.00 - Significant)	Poor fit
χ^2 /df	1263.21/186= 6.79	Poor fit
RMSEA	0.11	Poor fit
SRMR	0.07	Acceptable fit
GFI	0.81	Poor fit
AGFI	0.76	Poor fit
CFI	0.93	Acceptable fit
PGFI	0.65	Some model complexity

5.4.4.2 CFA of the Hypothesised Measurement Model

Apart from the examination of the overall model fit, it was necessary to evaluate the factor loadings of the API items to check whether or not the items reflect the factors they represent. In judging the acceptability of the factor loading, a threshold of 0.40 (Matsunaga, 2010) as used in many research studies was adopted for this scale. The items that had a factor loading of at least 0.40 were to be retained while those below the threshold were to be discarded. The factor loadings of the API items under the three-factor structure are presented in Table 5.10. As shown in the table, the 21 API items were all having factor loadings of more than 0.40. On this basis, the items appeared to reflect the three constructs in the three-factor structure. However, it cannot be concluded that the three-factor model is the appropriate structure for the API as the overall model fit was poor. Thus, examination of alternative models is warranted.

Table 5.10. Factor loadings of API items under the three-factor model

Factor	Item	Loading(se)*
Assessment Purpose (F1)	1	0.63(0.04)
	2	0.73(0.04)
	3	0.73(0.04)
	4	0.65(0.04)
	6	0.62(0.04)
	13	0.71(0.04)
	14	0.70(0.04)
	15	0.74(0.04)
	16	0.45(0.04)
Assessment Design (F2)	5	0.49(0.04)
	7	0.58(0.04)
	8	0.52(0.04)
	9	0.57(0.04)
	10	0.70(0.04)
	11	0.78(0.04)
	12	0.78(0.04)
Assessment Communication (F3)	17	0.79(0.04)
	18	0.91(0.03)
	19	0.78(0.04)
	20	0.43(0.04)
	21	0.50(0.04)

**n = 582*

5.4.4 CFA of the Alternative Models

Two models were further tested as alternative structures for the API. These were the one-factor model and the hierarchical model. Each of these models was evaluated using the same analytic technique/process.

The one-factor model was tested by loading all the 21 API items to one latent construct called the "assessment practices". The structure of the model is presented in Figure 5.3. The results of the fit indices for the one-factor structure are presented in Table 5.11. Moreover, the results on factor loading for all the items under this tested structure are presented in Table 5.12.

5.4.4.1 Model Fit of the Alternative One-factor Model

As revealed in Table 5.11, most of the indices indicate poor fit of the model to the data. Of the eight fit indices, only one index (SRMR) showed acceptable fit. Moreover, the PGFI value indicated that a one-factor structure is less parsimonious. Although similar results of poor model fit have been obtained, the fit indices results under the one-factor model appeared worse than those under the three-factor structure. Based on the results of fit indices, this tested model is not a better alternative structure for the API.

Table 5.11. Summary results of fit indices for the one-factor API structure

Fit Index	Obtained Value	Remark
χ^2	2036.80 (P = 0.00 - Significant)	Poor fit
χ^2 /df	2036.80.21/189= 10.78	Poor fit
RMSEA	0.14	Poor fit
SRMR	0.09	Acceptable fit
GFI	0.72	Poor fit
AGFI	0.66	Poor fit
CFI	0.89	Poor fit
PGFI	0.59	Some model complexity

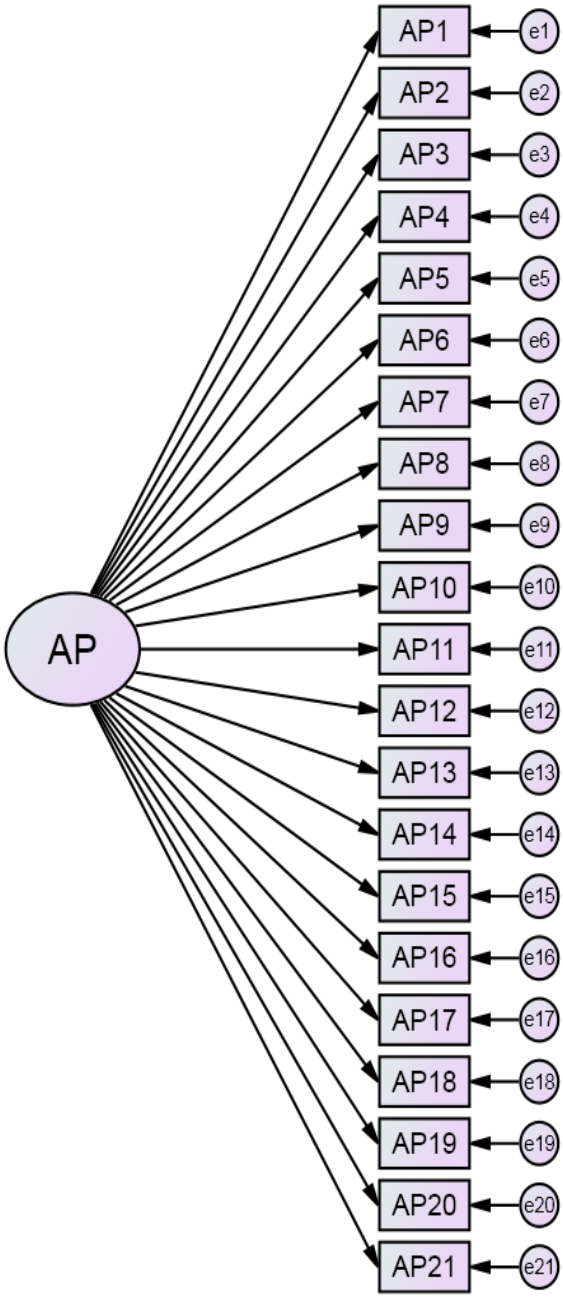


Figure 5.3. Structure of one-factor model for the API

5.4.4.2 CFA of the Hypothesised One-factor Measurement Model

Table 5.12 presents that all the items have factor loadings of more than 0.40, except for item 20 that has a factor loading of 0.39. Although the loadings appeared to indicate that the items reflect the construct/structure, it cannot be taken alone to judge the acceptability of the model. The factor loadings should be used together with the overall fit to determine the structure. Hence, as the overall model fit is poor, the one-factor model cannot be a structure for the API.

Table 5.12. Factor loadings of API items under the one-factor model

Factor	Item	Loading(se)*
Assessment Practices (AP)	1	0.55(0.04)
	2	0.66(0.04)
	3	0.67(0.04)
	4	0.60(0.04)
	5	0.51(0.04)
	6	0.62(0.04)
	7	0.56(0.04)
	8	0.47(0.04)
	9	0.54(0.04)
	10	0.62(0.04)
	11	0.70(0.04)
	12	0.72(0.04)
	13	0.73(0.04)
	14	0.72(0.04)
	15	0.72(0.04)
	16	0.47(0.04)
	17	0.55(0.04)
	18	0.59(0.04)
	19	0.53(0.04)
	20	0.39(0.04)
	21	0.43(0.04)

**n = 582*

The API structure was also tested using the hierarchical model. Under this model, the 21 API items were first loaded to three constructs as follows: Items 1, 2, 3, 4, 6, 13, 14, 15, and 16 were all loaded to 'assessment purpose'; items 5, 7, 8, 9, 10, 11, and 12 were loaded to 'assessment design'; and items 17, 18, 19, 20, and 21 were loaded to 'assessment communication'. After which, the three constructs were made to reflect the main construct called the 'assessment practices'. The structure of the hierarchical model is shown in Figure 5.4. The results of fit indices and factor loadings are presented in Tables 5.13 and 5.14, respectively.

5.4.4.3 Model Fit of the Alternative Hierarchical Model

Table 5.13 shows that similar to the initially proposed three-factor model most of the fit indices indicate poor fit of the hierarchical model to the data. Of the eight fit indices, only two indices (SRMR and CFI) showed acceptable fit. Moreover, the PGFI value indicated that hierarchical structure has some degree of complexity. Based on the results of fit indices, the hierarchical model is also not a better alternative structure for the API.

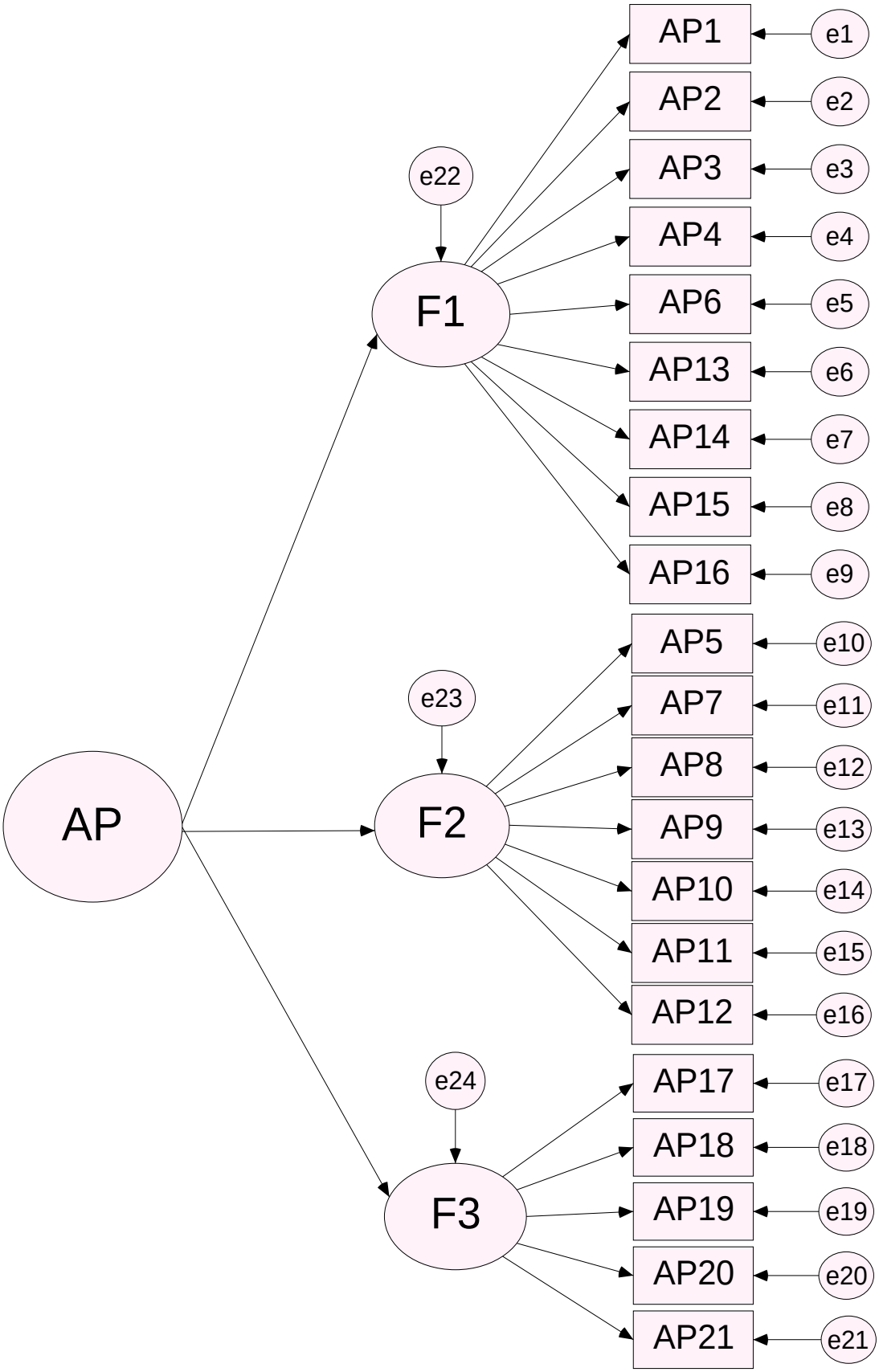


Figure 5.4 Structure of the hierarchical model for the API

Table 5.13. Summary of fit indices for the API hierarchical structure

Fit Index	Obtained Value	Remark
χ^2	1263.21 (P = 0.00 - Significant)	Poor fit
χ^2 /df	1263.21/186= 6.79	Poor fit
RMSEA	0.11	Poor fit
SRMR	0.07	Acceptable fit
GFI	0.81	Poor fit
AGFI	0.76	Poor fit
CFI	0.93	Acceptable fit
PGFI	0.65	Some model complexity

5.4.4.4 CFA of the Hypothesised Hierarchical Measurement Model

Table 5.14 presents the factor loadings of the items under the hierarchical model. As shown, all the factor loadings are above the adopted threshold of 0.40 and thus appeared to tap/represent the constructs/structure. However, as these results alone cannot be used to justify the acceptability of the tested model, the hierarchical structure of the API can also be challenged. In other words, the hierarchical model also appears inappropriate for the API. These results suggest that API may have other structures that warrant examination.

Table 5.14. Factor loadings of API items under the hierarchical model

First Factor	Second Factor	Item	Loading(se)*
Assessment Practices (AP)	Assessment Purpose	1	0.63(0.05)
		2	0.73(0.05)
		3	0.73(0.05)
		4	0.65(0.05)
		6	0.62(0.05)
		13	0.72(0.05)
		14	0.70(0.05)
		15	0.74(0.05)
		16	0.45(0.05)
	Assessment Design	5	0.49(0.06)
		7	0.58(0.06)
		8	0.52(0.06)
		9	0.57(0.06)
		10	0.70(0.07)
		11	0.78(0.07)
		12	0.78(0.07)
	Assessment Communication	17	0.79(0.04)
		18	0.91(0.04)
		19	0.78(0.04)
		20	0.43(0.04)
		21	0.50(0.04)

**n = 582*

The three tested models provided more or less the same results that none of them seems to be an appropriate structure for the API. Although differences on the values of fit indices and item loadings were noted, the same interpretation holds for all the models. However, on the basis of the theoretical proposition and model parsimony, the three-factor structure appeared to represent assessment practices. This is confirmed by the results of Rasch analysis.

5.4.5 Model Used in the Study

The Rasch analytic technique provided evidence that the three-factor model retains the most number of items and works well as the structure for the API or 'assessment practices'. The CFA results partly supports the Rasch analysis results. Although all the tested models under CFA exhibited poor fit to the data, the three-factor structure manifested a slight degree of acceptability as indicated by the acceptable values of few indices and by high magnitude of factor loadings. Rasch analysis results are consistent with the theoretical underpinning of the model in question, upon which the API was finally developed. Hence, the three-factor structure is adopted as the model for the API in this study.

5.5 Summary

This chapter dealt with the development and calibration of the API. The API was developed using the 'Keys to Quality Classroom Assessment' as the conceptual framework and was validated at item and structural levels using the Rating Scale Model and CFA, respectively. The rating scale and CFA analyses were carried out using ConQuest 2.0 and LISREL 8.80 software, respectively. The Rasch analysis results showed that either the three-factor or one-factor model could serve as the structure for the API. On the other hand, the CFA results indicated that all hypothesised models for the API exhibited poor fit to the data. However, the three-factor structure manifested certain degree of quality as indicated by acceptable values of few fit indices and by the high magnitude of the item loadings. Based on Rasch analysis results, the theoretical underpinning, and the objectives of this study, the three-factor structure was adopted as the model for the study.

Chapter 6: The Teaching Practices Scale

6.1 Introduction

The conceptualisation of this study posited that teacher assessment literacy impacts on other relevant variables through which student outcomes are ultimately affected. This paradigm was based from the view available in the literature that teacher characteristics such as knowledge influence teacher professional activities (OECD, 2010), which in turn affect student attributes (Wilkins, 2008). One of the attributes that were proposed to be affected by teacher assessment literacy is the teacher classroom practice. Thus, teacher instructional practices were considered in this study. Specifically, the possible role of teaching practices, as one of the mediating factors between teacher knowledge on assessment and student outcomes, was examined. The conceptual representation of the model involving the proposed directional relationships among teacher assessment literacy, teacher instructional practices, and student outcomes is presented in Figure 6.1 below.

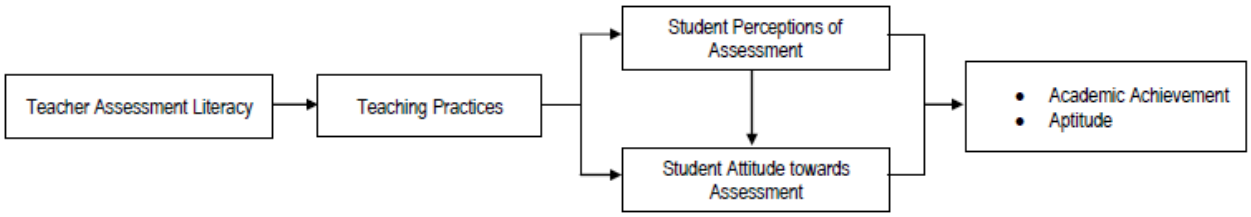


Figure 6.1. The relationship among teacher assessment literacy, teaching practices, and student outcomes

To examine the teaching practices and the relevant relationships as depicted in the figure above, it was necessary to employ the scale that encompassed constructs related to student learning and other outcomes and that fulfilled the intention of the study. The scale, herein referred to as the "teaching practices scale" (TPS), that was used in the study was adopted from the 2008 Teaching and Learning International Survey (TALIS) teacher questionnaire. As the study utilised the 2008 TALIS conceptual framework as one of the bases of its proposed model, adopting the scale from the TALIS teacher questionnaire was deemed

appropriate. The TALIS survey questionnaire is a well-established instrument and its measurement capacity had already been established. However, while the questionnaire had been rigorously developed and validated, there was a need for the TPS to be revalidated as it had been slightly modified and as it was applied to a new context and/or new group of respondents. Hence, the TPS underwent revalidation process to ensure its utility for the current study. This chapter deals with the description and revalidation of the TPS.

The chapter begins with the information about the TPS, its development, previous validation, and description. After which, the modification and pilot test of the scale in the research venue are discussed. The ensuing section is devoted mainly to the revalidation of the TPS, first at the item level and finally at the structural level. The chapter concludes by reiterating the key points.

6.2 The TPS: Its Development, Previous Validation, and Description

The TPS was part of the 2008 TALIS teacher questionnaire. It was employed to gather data about teaching practices that formed part of teachers' professional activities. The scale was intended to contribute to the examination of TALIS survey themes, which were based on conceptual framework that views teaching practices as a factor that affects student learning and outcomes (OECD, 2010).

The development and validation of the TALIS teacher questionnaire, of which TPS was a part, underwent a rigorous process that was executed by four entities: the TALIS Board of Participating Countries (BPC), the Organisation of Economic Cooperation and Development (OECD) Secretariat, the Instrument Development Expert Group (IDEG), and the TALIS National Project Managers (NPM). The TALIS BPC was created to set the goals for the development of the survey questionnaires, to review and approve the questionnaires at every stage of the validation process (pilot, field trial, and main survey), and to advise on the relevance and validity of the survey content and items with respect to the goals and context of the TALIS; the OECD Secretariat led the work of the IDEG on the drafting of survey questionnaires; the IDEG was formed to translate the goals established by the TALIS BPC into survey analysis and survey questionnaires. The IDEG's specific tasks were as follows:

> "to review the proposed indicators for the survey to ensure that the variables, indicators and themes provide a logical basis for instrument development, giving consideration to completeness and coherence; to review the catalogue of existing questions compiled from national and international studies in order to assess their suitability for measuring the variables within the TALIS analytic framework and to identify other possible sources of exemplary questions; to consider and advise on implications for the sample design arising from the questionnaire development and vice versa; to consider and advise on the extent to which the teacher questionnaire in the main survey should be the same as that in the experimental PISA link; to review and revise the questionnaires in the light of pilot and field trial (FT) results; to contribute to the drafting of the survey analysis plans; and to present proposed questionnaires and analysis plans to the BPC" (OECD, 2010, pp. 30-31).

The IDEG's main responsibility was to develop the questionnaire content for review by the TALIS NPM and finally by the TALIS BPC. The TALIS NPM was tasked to advise on the validity of the questions for every country, the significance of the questions in the intended analysis, and the clarity of the drafting and sequencing of the questions (OECD, 2010).

The TALIS questionnaire development and validation took about three years to complete. The TALIS BPC started the process by conducting a priority rating exercise to initially determine the main themes of the TALIS. After which, the IDEG under the leadership of the OECD Secretariat elaborated the themes into research questions, variables, and constructs. The IDEG initially drafted the questionnaire items. The drafted items were then reviewed by the TALIS BPC. The TALIS NPM further reviewed the second draft of the items/questionnaire. After all the reviews, the pilot version of the questionnaires was approved. The pilot test, which was intended to initially test the questionnaires, was conducted in five countries namely, Brazil, Malaysia, Norway, Portugal, and Slovenia. The OECD, the representatives of the international contractor, the International Association for the Evaluation of Educational Achievement (IEA), and the NPM reviewed the pilot results. From the review and statistical analysis of the results, some changes in relation to the length of the questionnaires, item wording, and suitability of some items were

done. Some items were reworded/simplified/modified and some were deleted as a result of feedback from the pilot participants. However, the pilot test results were found to be generally suited for the TALIS. Based on the results, the IDEG conducted consultations on proposed changes for the questionnaire FT. The FT version of the questionnaire was finally approved by the TALIS BPC and was administered to test the survey operations and further evaluate the questionnaire validity. The IDEG conducted a further meeting with the involvement of BPC to review the FT results and propose changes for the main survey (MS) questionnaires. After the final consultations with the BPC, the MS questionnaires were finalised (OECD, 2010).

The framework that guided the questionnaire development placed teaching practices as a factor between knowledge (content knowledge and pedagogical content knowledge) and student learning and outcomes. It considered teaching practices as some sort of a mediating variable that can be influenced by teacher knowledge and other characteristics and that can, in turn, impact on student learning and outcomes. It recognises that teachers possess the cognitive structure that gives rise to actual classroom practice through which students are made to learn (OECD, 2010).

The TPS measured the teaching practices that involved elements of instructional quality. It adopted three constructs that pertained to the teaching structure relevant to the components of direct instruction, supportive climate and individualised instruction, and cognitive activation. The constructs labeled as 'structure', 'student orientation', and 'enhanced activities' serve as basic dimensions of teaching practices – a domain-general version of the triarchic model which TALIS used. This model had its basis from the three basic second-order dimensions of instructional quality which Klieme, at al. (2006, as cited in OECD, 2010) proposed based on results from the TIMSS video study. The model had its theoretical foundation and empirical evidence from the work of Lipowsky, Rakoczy, Pauli, Drollinger-Vetter, Klieme, and Reusser (2009).

The TPS consists of 20 items covering the three constructs. The items are distributed as follows: Structure – 10 items (Struct1, Struct2, Struct3, Struct4, Struct5, Struct6, Struct7, Struct8, Struct9, and

Struct10); Student Orientation – 6 items (Stud1, Stud2, Stud3, Stud4, Stud5, and Stud6); and Enhanced Activities – 4 items (Enact1, Enact2, Enact3, and Enact4). All these items adopted a five-point Likert scale of “never or hardly ever”, “in about one-quarter of lessons”, “in about one-half of lessons”, “in about three-quarters of lessons” and “in almost every lesson”, which were coded 1, 2, 3, 4, and 5, respectively.

6.3 Modification and Pilot Test of the TPS in the Current Study

For purposes of this study, the TPS was slightly modified by splitting one of its items into two statements to reduce the cognitive load (see Table 6.1 for the items and their wording for the TALIS and modified versions; also see Appendix B). This was an attempt to make the TPS items appropriate for all the participants in the research context. After which, the researcher in consultation with his supervisors reviewed the items. Further judgment on the items was made by three experts from the MSU Tawi-Tawi for appropriateness and suitability (the experts judged the items as acceptable for Tawi-Tawi context and recommended that the instrument be administered). The TPS items were then organised into one section and formed part of the teacher survey. They were pilot tested together with the other items in the study’s teacher questionnaire to 45 MSU Tawi-Tawi elementary and secondary school teachers to obtain initial validity/reliability, test the survey operation, obtain feedback about the items, and to determine the time for questionnaire completion. The survey process, the amount of time to accomplish the questionnaire, and the feedback from the pilot participants were all noted in finalising and administering the instrument. Specifically, the feedback was mainly on the improvement of the scale structure. Besides, a Chronbach alpha of 0.86, which indicated acceptable reliability, was obtained for this instrument.

Table 6.1. The original and modified teaching practices scale

2008 TALIS Version		Modified Version	
Item Code	Item Wording	Item Code	Item Wording
Struct1	I present new topics to the class (lecture-style presentation).	Struct1	I present new topics to the class in a lecture-style presentation.
Struct2	I explicitly state learning goals.	Struct2	I explicitly state learning goals.
Struct3	I review with the students the homework they have prepared.	Struct3	I review with the students the homework they have prepared.
Struct4	I ask my students to remember every step in a procedure.	Struct4	I ask my students to remember every step in a procedure.
Struct5	At the beginning of the lesson I present a short summary of the previous lesson.	Struct5	At the beginning of the lesson, I present a short summary of the previous lesson.
Struct6	I check my students' exercise books.	Struct6	I check my students' exercise books.
Struct7	I work with individual students.	Struct7	I work with individual students.
Struct8	Students evaluate and reflect upon their work.	Struct8	Students evaluate and reflect upon their work.
Struct9	I check, by asking questions, whether or not the subject matter has been understood.	Struct9	I check, by asking questions, whether or not the subject matter has been understood.
Struct10	I administer a test or quiz to assess student learning.	Struct10	I administer a test or quiz to assess student learning.
Stud1	Students work in small groups to come up with a joint solution to a problem or task.	Stud1	Students work in small groups to come up with a joint solution to a problem or task.
Stud2	I give different work to the students that have difficulties learning and/or to those who can advance faster.	Stud2	I give different work to the students that have difficulties learning the subject matter.
		Stud3	I give different work to the students that can learn faster.
Stud3	I ask my students to suggest or to help plan classroom activities or topics.	Stud4	I ask my students to suggest classroom activities including topics.
Stud4	Students work in groups based upon their abilities.	Stud5	Students work in groups based upon their abilities.
Stud5	Students work individually with the textbook or worksheets to practice newly taught subject matter.	Stud6	Students work individually with the textbook or worksheets to practice newly taught subject matter.
Enact1	Students work on projects that require at least one week to complete.	Enact1	Students work on projects that require at least one week to complete.
Enact2	Students make a product that will be used by someone else.	Enact2	Students make a product that will be used by someone else.
Enact3	I ask my students to write an essay in which they are expected to explain their thinking or reasoning at some length.	Enact3	I ask my students to write an essay in which they are expected to explain their thinking or reasoning at some length.
Enact4	Students hold a debate and argue for a particular point of view which may not be their own.	Enact4	Students hold a debate and argue for a particular point of view which may not be their own.

6.4 Examination of the Item and Model Fit of the TPS

As the TPS had been modified and applied to the Philippine/Tawi-Tawi context, it was revalidated to ensure that it worked as intended in the study. Specifically, the scale was examined at the item and structural levels. The item and structural fit of the scale were evaluated using the Rasch Model, particularly the Rating Scale Model (Andrich, 1978), and the CFA, respectively (see Chapters 3 and 5 for details about Rating Scale Model and CFA). The item-level analysis was carried out using ConQuest software (v. 2.0) (Wu, Adams, Wilson, & Haldane, 2007). At the structural level, CFA was used as the instrument had been developed using a priori. The CFA was performed using LISREL 8.80 software (Jöreskog & Sörbom, 2006). The item and model fit were evaluated using similar process and indicators employed in the previous validation chapters. The item and structural analysis results are presented in the succeeding subsections.

6.4.1 Item Analysis Results Using the Rating Scale Model

The TPS items were analysed separately for each of the three identified constructs using the Rating Scale Model. The decision to split the analysis by construct was based from the underlying theory upon which the scale development was based. The purpose of the analysis was to determine whether or not the items functioned as hypothesised and if all the items under each construct fit the Rasch model. All the responses from the 582 participants for the concerned items were subjected to analysis using the ConQuest 2.0 software (Wu, Adams, Wilson, & Haldane, 2007). The results of the Rasch analysis are presented separately for each of the constructs in the following subsections.

6.4.1.1 Rasch Analysis Results of the TPS Items under the 'Structure' Construct

The TPS items under the construct of ‘structure’ were the first group to be subjected to Rasch analysis. The item statistics for the initial and final calibrations are presented in Table 6.2 and Table 6.3, respectively. As shown in Table 6.2, all the items in the initial run were within the acceptable UMS range of $\geq$ 70 and $\leq$ 1.30, except for Struct9 that is underfit (UMS = 1.38). Struct9 is about checking the understanding of the subject matter by asking questions. This item is part of teachers’ common practice and

it was unexpected that it did not fit the Rasch Model. Again, this was perhaps due to case or person misfit (Curtis, 2004). As a procedure, Struct9 was deleted as it exhibited UMS value beyond the acceptable maximum value of 1.30. All other items were then recalibrated. The results of the second and final calibration are in Table 6.3.

Table 6.2. Results of the initial item analysis of the 'structure construct' of the TPS

Item	Estimate (Difficulty/Endorsability/Dilemma)	Error	UMS	t
Struct1	0.19	0.03	0.99	- 0.1
Struct2	- 0.07	0.04	0.77	- 4.2
Struct3	- 0.18	0.04	1.00	- 0.0
Struct4	0.05	0.03	1.17	2.8
Struct5	- 0.42	0.04	0.94	- 1.0
Struct6	0.05	0.03	1.04	0.7
Struct7	0.89	0.03	1.14	2.3
Struct8	0.32	0.03	0.89	- 1.9
Struct9	- 0.30	0.04	1.38*	5.8
Struct10	- 0.528**	0.10	1.04	0.8

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 1160.00; df=9; Sig level=0.000; *Misfitting; **Constrained*

As can be gleaned from Table 6.3, the final analysis results revealed that nine items under the 'structure' construct fit the Rasch model as indicated by the acceptable UMS values. These results imply that the nine remaining items possess measurement capacity and reflect one single or dominant construct called 'structure'. Moreover, examination of the arrangement of thresholds for every item showed no disordered values, which indicate that the item categories function well as intended. Furthermore, the resulting value of separation reliability, which is 0.99 for this scale, discloses high discrimination and precision (Alagumalai & Curtis, 2005; Wright & Stone, 1999) and provides further indication that the scale is acceptable based on the Rasch model.

Table 6.3. Results of the final item analysis of the 'structure construct' of the TP

Item	Estimate (Difficulty/Endorsability/ Dilemma)	Error	UMS	t
Struct1	0.16	0.03	1.04	0.7
Struct2	- 0.11	0.04	0.75	- 4.6
Struct3	- 0.22	0.04	1.01	0.2
Struct4	0.01	0.04	1.10	1.6
Struct5	- 0.48	0.04	0.97	- 0.5
Struct6	0.02	0.04	1.05	0.8
Struct7	0.89	0.03	1.14	2.4
Struct8	0.30	0.03	0.93	- 1.2
Struct10	- 0.585*	0.10	1.18	3.0

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 1082.56; df=8; Sig level=0.000; *Constrained*

6.4.1.2 Rasch Analysis Results of the TPS Items under the 'Student Orientation' Construct

The TPS items under the construct of 'student orientation' were the next group analysed at the micro level using the same analytic technique. Six items that were hypothesised to indicate the construct were subjected to analysis. The results in Tables 6.4 and 6.5 present similar picture with the previous construct.

As shown in Table 6.4, the initial Rasch analysis disclosed that five of the six items under the construct of 'student orientation' fit the Rasch model as indicated by the UMS acceptable values while one item (Stud4) was underfit (UMS=1.32). Stud4 (I ask my students to remember every step in a procedure) exhibited unpredictable responses and misfit for similar reason as used to justify Struct9. The presence of the non-fitting item necessitated its deletion and recalibration of the remaining items. Thus, Stud4 was deleted and the rest of the items were recalibrated.

Table 6.4. Results of the initial item analysis of the 'student-oriented activity construct' of the TPS

Item	Estimate (Difficulty/Endorsability/Dilemma)	Error	UMS	t
Stud1	- 0.24	0.03	0.93	- 1.2
Stud2	0.19	0.03	0.91	- 1.7
Stud3	- 0.21	0.03	0.93	- 1.2
Stud4	0.33	0.03	1.32*	5.1
Stud5	- 0.06	0.03	0.98	- 0.4
Stud6	- 0.013**	0.07	1.20	3.3

*Separation Reliability = 0.98; Chi-Square Test of Parameter Equality = 265.90; df=5; Sig level=0.000; *Misfitting; **Constrained*

Table 6.5 provides the final results for all the remaining items under the 'student orientation' construct. As presented, all the remaining items were fitting the Rasch model well as indicated by the UMS values of between 0.70 and 1.30. By these results, it can be interpreted that the remaining five items have desirable measurement property and can represent the 'student orientation' construct. Moreover, the absence of disordered thresholds in all the items implies that the item categories functioned well as intended. The separation reliability value of 0.99, which indicates high degree of item discrimination and precision, also provides a strong support that the items could be retained and utilised to measure the concerned construct.

Table 6.5. Results of the final analysis of the 'student-oriented activity construct' of the TPS

Item	Estimate (Difficulty/Endorsability/Dilemma)	Error	UMS	t
Stud1	- 0.19	0.03	0.87	- 2.3
Stud2	0.27	0.03	0.92	- 1.4
Stud3	- 0.15	0.03	1.07	1.1
Stud5	0.01	0.03	1.03	0.5
Stud6	0.058*	0.06	1.21	3.3

*Separation Reliability = 0.98; Chi-Square Test of Parameter Equality = 130.23; df=4; Sig level=0.000; *Constrained*

6.4.1.3 Rasch Analysis Results of the TPS Items under the 'Enhanced Activities' Construct

The TPS items under the 'enhanced activities' construct constituted the last group that was analysed separately using the same technique/process. All the responses in the four items (Enact1, Enact2, Enact3, and Enact4) under this construct were subjected to analysis using the same statistical software. The initial and final analysis results are presented in Table 6.6. As shown, all the items had UMS values that were within the acceptable range of 0.70 to 1.30. This means that the items were fitting the Rasch Model and were functioning well as hypothesised. The absence of disordered categories likewise provided an indication that the hypothesised categories were also functioning well as intended. In addition, the obtained separation reliability value of 0.99 further revealed that the items had desirable degree of discrimination and precision. Thus, all items could be retained and considered for this construct.

Table 6.6. Results of the initial and final item analyses of the 'enhanced activity construct' of the TPS

Item	Estimate Difficulty/Endorsability/Dilemma)	Error	UMS	t
Enact1	- 0.23	0.03	0.93	- 1.3
Enact2	0.17	0.03	1.10	1.7
Enact3	- 0.44	0.03	1.02	0.3
Enact4	0.500*	0.05	0.92	- 1.4

Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 310.31; df=3; Sig level=0.000

6.4.4.4 Rasch Analysis Results of the TPS Items under the Proposed one-Construct of Teaching Practices

As the original structure of the TPS was a second-order three-factor model, all items under this scale were combined and all the relevant responses were subjected to Rasch analysis. This was to determine whether or not the items generally reflect a single or a dominant construct as implied in the hypothesised model. The results revealed that the TPS items reflect a single/dominant dimension. The specific item statistics are shown in Tables 6.7 and 6.8.

As provided in Table 6.7, the initial Rasch analysis disclosed that 19 of the combined 20 TPS items were fitting the Rasch model as indicated by their corresponding acceptable UMS values. Only item TP9 (Struct9) exhibited underfit as shown by UMS value of 1.49. Again, as a procedure, item TP9 was deleted and the remaining items recalibrated. After deleting TP9, the final item analysis results were obtained. These results are provided in Table 6.8.

Table 6.7. Results of the initial items analysis of the 'combined teaching practices construct' of the TPS

Item	Estimate (Difficulty/Endorsability/Dilemma)	Error	UMS	t
TP1	- 0.13	0.03	1.00	0.0
TP2	- 0.41	0.03	0.89	- 2.0
TP3	- 0.52	0.03	1.01	0.3
TP4	- 0.28	0.03	1.11	1.8
TP5	- 0.78	0.03	1.06	1.1
TP6	- 0.28	0.03	1.02	0.4
TP7	0.59	0.03	1.11	1.8
TP8	0.01	0.03	0.88	- 2.2
TP9	- 0.65	0.03	1.49*	7.3
TP10	- 0.89	0.03	1.15	2.5
TP11	0.04	0.03	0.70	- 5.7
TP12	0.40	0.03	0.93	- 1.2
TP13	0.07	0.03	0.93	- 1.1
TP14	0.52	0.03	1.14	2.3
TP15	0.19	0.03	0.77	- 4.2
TP16	0.23	0.03	0.92	- 1.3
TP17	0.26	0.03	0.80	- 3.7
TP18	0.62	0.03	1.21	3.4
TP19	0.08	0.03	0.99	- 0.1
TP20	0.930*	0.13	1.01	0.2

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 4228.13; df=19; Sig level=0.000; *Misfitting; **Constrained*

As can be seen in Table 6.8, the remaining 19 items possess UMS values that are within the adopted acceptable range of 0.70 – 1.30. By Rasch model, these items possess the desirable

measurement property and conform to the hypothesis that they represent a single/dominant dimension. Other positive indications such as the absence of disordered thresholds and high separation reliability of 0.99 further support the original hypothesis that TPS can be a scale with one dimension. Hence, the 19 TPS items can be retained and be taken to reflect one dimension – the teaching practices. This is a potential alternative when considering the data from this scale for further analysis.

Table 6.8. Results of the final item analysis of the 'combined teaching practices construct' of the TPS

Item	Estimate (Difficulty/Endorsability/Dilemma)	Error	UMS	t
TP1	- 0.17	0.03	1.06	0.9
TP2	- 0.45	0.03	0.91	- 1.6
TP3	- 0.57	0.03	1.04	0.7
TP4	- 0.33	0.03	1.14	2.3
TP5	- 0.84	0.03	1.15	2.5
TP6	- 0.32	0.03	1.06	1.0
TP7	0.57	0.03	1.10	1.7
TP8	- 0.03	0.03	0.91	- 1.5
TP10	- 0.95	0.03	1.29	4.6
TP11	0.01	0.03	0.72	- 5.4
TP12	0.38	0.03	0.93	- 1.3
TP13	0.03	0.03	0.92	- 1.4
TP14	0.50	0.03	1.14	2.4
TP15	0.16	0.03	0.80	- 3.6
TP16	0.20	0.03	0.96	- 0.6
TP17	0.24	0.03	0.82	- 3.3
TP18	0.61	0.03	1.24	3.8
TP19	0.05	0.03	1.05	0.8
TP20	0.919*	0.12	1.04	0.7

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 3901.12; df=18; Sig level=0.000; *Constrained*

6.4.2 Structural Analysis Using CFA

The TPS was also analysed at the macro level using LISREL 8.80 software (Jöreskog & Sörbom, 2006) to determine the hypothesised structure and the fit of the proposed measurement model to the data.

This was to provide other perspective on the hypothesised relationships between the items and the constructs and among the latent constructs. In running the CFA, only the items that were fitting the Rasch model were included as they are considered well-functioning items with respect to the hypothesised dimensions. The first analysis was performed using the original hypothesis that TPS had second-order three-factor structure. After which, the analysis on the alternative models that include first-order three-factor structure and one-factor structure was carried out. The relevant CFA results are provided in the succeeding sections/subsections.

6.4.3 The Second-Order Three-Factor Structure of the TPS

The second-order three-factor model of the TPS was examined to confirm the hypothesis that it is the appropriate structure for this scale. Under this model, the main factor – teaching practices – was hypothesised to be reflected by three endogenous latent constructs namely, 'structure', 'student orientation', and 'enhanced activities', which were also assumed to be reflected by individual items. The construct of 'structure' was represented by nine items labeled as Struct1, Struct2, Struct3, Struct4, Struct5, Struct6, Struct7, Struct8, and Struct10; the 'student orientation' was reflected by five items labeled as Stud1, Stud2, Stud3, Stud5, and Stud6; and the 'enhanced activities' was measured by four items labeled as Enact1, Enact2, Enact3, and Enact4. The conceptual representation of this model is shown in Figure 6.2. In evaluating this structure, a number of fit indices (see Chapter 3) for the overall model fit and the threshold of 0.40 for the item loadings were used. The CFA results are presented in Tables 6.9 and 6.10.

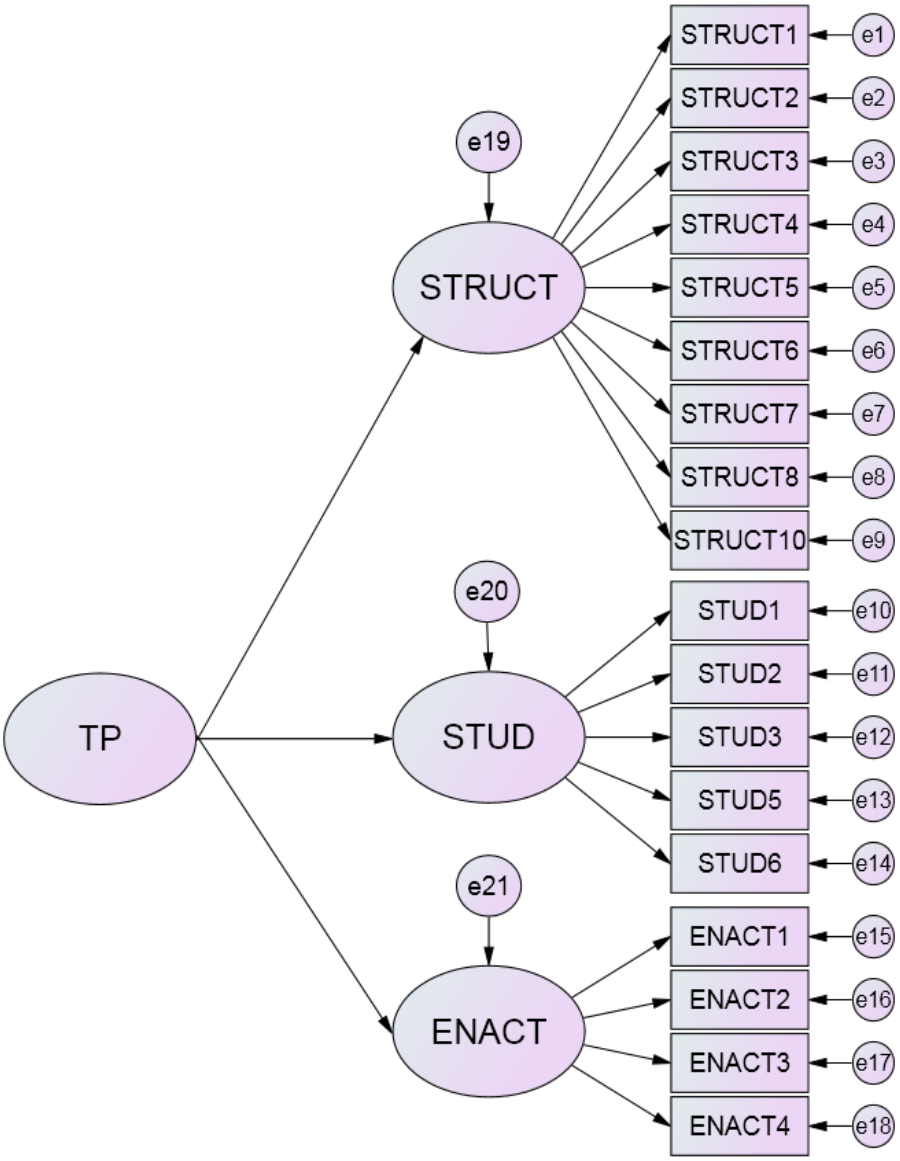


Figure 6.2. The second-order three-factor structure of the TPS

6.4.3.1 Model Fit

Examining the results in Table 6.9, the second-order three-factor structure of the TPS appeared to exhibit poor fit to the data as shown by poor results of most of the fit indices. Only three fit indices (RMSEA, SRMR, and CFI) showed acceptable results while the other four (χ^2, χ^2/df, GFI, and AGFI) indicated poor fit. However, the indices that indicate acceptable fit are known to be more dependable as they are not as sensitive to the sample size as those that showed poor fit. Thus, it can be argued that this structure has some merit as a possible model for the TPS. Moreover, the result of the PGFI implied that the model had some degree of parsimony despite being a hierarchical structure. Hence, this model to some extent can be adopted for the TPS.

Table 6.9. Summary results of fit indices for the hierarchical structure of the TPS

Fit Index	Obtained Value	Remark
X^2	908.74 (P = 0.00)	Poor fit
X^2 /df	908.74/132 = 6.88	Poor fit
RMSEA	0.10	Mediocre fit
SRMR	0.08	Acceptable fit
GFI	0.85	Poor fit
AGFI	0.80	Poor fit
CFI	0.91	Acceptable fit
PGFI	0.66	Some model complexity

6.4.3.2 CFA of the Hypothesised Measurement Model

In terms of the factor loadings, the resulting statistics appeared to support the aforementioned argument. All the loadings for the items as provided in Table 6.10 are significantly higher than the adopted threshold of 4.0. The nine items for the 'structure', five items for the 'student orientation', and the four items for the 'enhanced activities' had significant loadings above the threshold. By these results, the groups of items appeared to reflect the corresponding constructs. In addition, the magnitude of the relationships

between the main factor of teaching practices and the endogenous latent constructs of 'structure', 'student orientation', and 'enhanced activities' are higher, which can be interpreted that the teaching practices are well reflected by the hypothesised constructs. Thus, the structure and measurement model of this scale can be partly confirmed and can be possibly adopted for the TPS.

Table 6.10. Factor loadings of the teaching practices items under hierarchical model

Structure	Construct	Magnitude of Relationship with the Main Factor	Item	Loading(se)*
Teaching Practices (TP)	Structuring	0.80 (0.08)	Struct1	0.45(0.07)
			Struct2	0.59(0.07)
			Struct3	0.65(0.07)
			Struct4	0.56(0.06)
			Struct5	0.61(0.07)
			Struct6	0.59(0.07)
			Struct7	0.48(0.06)
			Struct8	0.60(0.07)
			Struct10	0.41(0.06)
	Student-Oriented	1.05 (0.07)	Stud1	0.65(0.05)
			Stud2	0.63(0.05)
			Stud3	0.61(0.05)
			Stud5	0.65(0.05)
			Stud6	0.56(0.05)
	Enhanced Activities	0.80 (0.07)	Enact1	0.60(0.05)
			Enact2	0.56(0.05)
			Enact3	0.61(0.06)
			Enact4	0.70(0.06)

**n = 582*

6.4.4 The CFA of the Alternative Models

The first-order three-factor and the one-factor structures were tested as alternative models for the TPS. This was to provide other potential models that can be used to appropriately represent the proper

structure of the concerned scale. In evaluating these models, similar technique, process, software, and indicators were employed. The results are shown and discussed in the ensuing sections/subsections.

6.4.4.1 The First-Order Three-Factor Structure of the TPS

The first-order three-factor model was examined for the TPS. Similar to the second-order three-factor structure, this model hypothesised that the construct of 'structure' was represented by nine items labeled as Struct1, Struct2, Struct3, Struct4, Struct5, Struct6, Struct7, Struct8, and Struct10; the 'student orientation' was reflected by five items labeled as Stud1, Stud2, Stud3, Stud5, and Stud6; and the 'enhanced activities' was measured by four items labeled as Enact1, Enact2, Enact3, and Enact4. The conceptual representation of this model is shown in Figure 6.3. The relevant CFA results are presented in Tables 6.11 and 6.12.

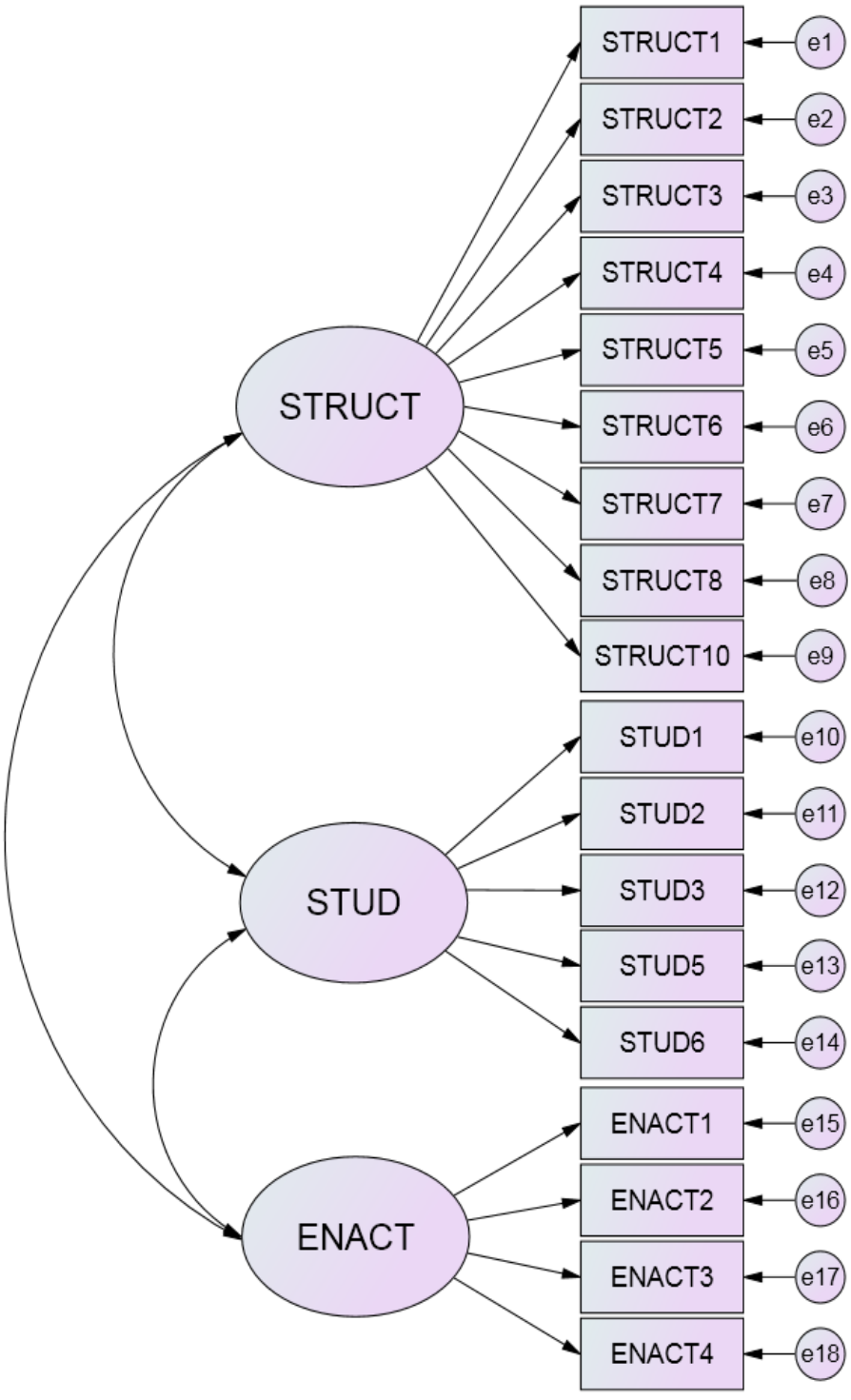


Figure 6.3. The first-order three-factor structure of the TPS

6.4.4.2 Model Fit of the first-order three-factor structure

The structural statistics in Table 6.11 revealed similar picture with the results of the second-order three-factor model. The overall fit of the structure to the data appeared to be poor as indicated by the results of most of the fit indices. Only three fit indices (RMSEA, SRMR, and CFI) showed acceptable results while the other five (χ^2, χ^2/df, GFI, and AGFI) indicated poor fit. However, the indices that indicate acceptable fit are known to be more dependable as they are considered more robust than those that showed poor fit and due to the other known criticisms of the latter. Thus, it can be argued that this structure has some degree of the merit as a possible alternative model for the TPS. Moreover, the result of the PGFI implied that the model had some degree of parsimony despite being a three-factor structure. Hence, this model to some extent can be adopted as alternative structure for the TPS.

Table 6.11. Summary of fit indices for the three-factor structure of the teaching practices

Fit Index	Obtained Value	Remark
X^2	908.74 (P = 0.00)	Poor fit
X^2 /df	908.74/132 = 6.88	Poor fit
RMSEA	0.10	Mediocre fit
SRMR	0.08	Acceptable fit
GFI	0.85	Poor fit
AGFI	0.80	Poor fit
CFI	0.91	Acceptable fit
PGFI	0.66	Some model complexity

6.4.4.3 CFA of the Hypothesised First-Order Three-Factor Measurement Model

In terms of the factor loadings as presented in Table 6.12, the resulting statistics appeared to support the argument that the structure has the merit to be a possible alternative model. All the item loadings are higher than the adopted threshold of 4.0 and they significantly loaded to the respective constructs as proposed. Moreover, the magnitude of the correlation coefficient between any two

hypothesised constructs is likewise high, which revealed that the three constructs are significantly related. Thus, the correlated structure and measurement model of this scale can be adopted as a possible alternative to the second-order three-factor model for the TPS.

Table 6.12. Factor loadings of the teaching practices items under the three-factor model

Structure	Construct	Correlation between Constructs	Item	Loading(se)*
Three-Factor Model	Structuring	0.84 (Structuring and Student-Oriented)	Struct1	0.45(0.04)
			Struct2	0.59(0.04)
			Struct3	0.65(0.04)
			Struct4	0.56(0.04)
			Struct5	0.61(0.04)
			Struct6	0.59(0.04)
			Struct7	0.48(0.04)
			Struct8	0.60(0.04)
			Struct10	0.41(0.04)
	Student-Oriented	0.84 (Student-Oriented and Enhanced Activities)	Stud1	0.65(0.04)
			Stud2	0.63(0.04)
			Stud3	0.61(0.04)
			Stud5	0.65(0.04)
			Stud6	0.56(0.04)
	Enhanced Activities	0.65 (Structuring and Enhanced Activities)	Enact1	0.60(0.04)
			Enact2	0.56(0.04)
			Enact3	0.61(0.04)
			Enact4	0.70(0.04)

**n = 582*

6.4.4.4 The One-Factor Structure of the TPS

The one-factor structure for the TPS was also examined in an attempt to provide other alternative model and to adopt more appropriate structure for this study. Under this proposed model, all the items were combined and labeled as TP1, TP2, TP3, TP4, TP5, TP6, TP7, TP8, TP10, TP11, TP12, TP13, TP15,

TP16, TP17, TP18, TP19, and TP20. These items were loaded to one main factor called the 'teaching practices (TP)'. The conceptual representation of this model is presented in Figure 6.4. To determine the acceptability of this model with respect to the data, the same technique, process, software, and indicators were used. The CFA results are shown in Tables 6.13 and 6.14.

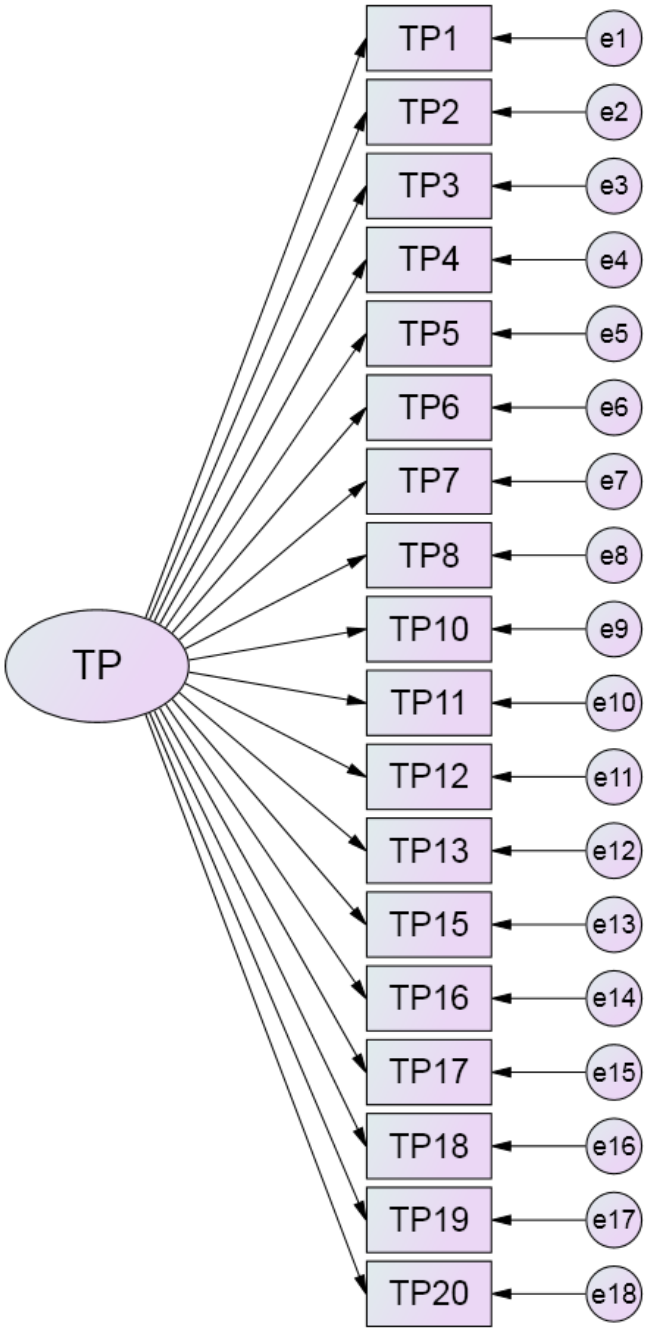


Figure 6.4. Structure of one-factor model of the TPS

6.4.4.5 Model Fit of the One-factor Structure

The CFA results presented in Table 6.13 below indicate that the one-factor structure has poor fit to the data. Of the fit indices employed in this study, only one (SRMR) exhibited acceptable fit while the rest ((χ^2, χ^2/df, RMSEA, GFI, AGFI, and CFI) indicated poor fit. Moreover, the PGFI value of 0.64, which revealed some degree of parsimony, appeared to provide evidence that the one-factor model is not different from the previous models in terms of the simplicity of the structure despite being a single dimension. Thus, it can be deduced that by CFA results, the one-factor model is not a better alternative to either of the models tested earlier.

Table 6.13. Summary results of fit indices for the one-factor structure of the teaching practices

Fit Index	Obtained Value	Remark
X^2	1063.63 (P = 0.00)	Poor fit
X^2 /df	1063.63/135 = 7.88	Poor fit
RMSEA	0.12	Poor fit
SRMR	0.08	Acceptable fit
GFI	0.81	Poor fit
AGFI	0.77	Poor fit
CFI	0.89	Poor fit
PGFI	0.64	Some model complexity

6.4.4.6 CFA of the Hypothesised One-factor Measurement Model

In terms of the factor loadings, the results appeared to negate the poor fit of the model to the data. As can be seen in Table 6.14, of the 18 TPS items analysed under the one-factor structure, only one (TP10) exhibited a weak loading of 0.34 while the other 17 items had loadings above the threshold of 0.40. However, as these results need to be interpreted in the light of the overall fit of the model, the factor loadings do not warrant the appropriateness of the structure. Thus, the judgment that this model is not appropriate for the TPS holds.

Table 6.14. Factor loadings of teaching practices items under one-factor model

Structure	Item	Factor Loading(se)*
	TP1	0.41(0.04)
	TP2	0.51(0.04)
	TP3	0.57(0.04)
	TP4	0.51(0.04)
	TP5	0.53(0.04)
	TP6	0.59(0.04)
	TP7	0.53(0.04)
	TP8	0.61(0.04)
One-Factor Model	TP10	0.34(0.04)
	TP11	0.64(0.04)
	TP12	0.59(0.04)
	TP13	0.58(0.04)
	TP15	0.63(0.04)
	TP16	0.55(0.04)
	TP17	0.60(0.04)
	TP18	0.45(0.04)
	TP19	0.51(0.04)
	TP20	0.54(0.04)

**n = 582*

6.4.5 Model Used in the Study

The Rasch and CFA analyses of the three models tested for the TPS provided more or less a picture of the appropriate structure that can be adopted in this study. The analysis results revealed that both the first-order and the second-order three-factor structures appeared appropriate for the TPS while the one-factor model appeared to be a weak alternative structure. As this chapter is just to confirm the appropriateness of the originally hypothesised TPS structure and as the two other models failed to provide better structure for this scale, this study adopted the second-order three-factor model for the TPS.

6.5 Summary

This chapter dealt with the revalidation of the TPS. The TPS was revalidated at the item (micro) and structural (macro) levels using the Rating Scale Model and CFA, respectively. The rating scale and CFA analyses were carried out using ConQuest 2.0 and LISREL 8.80 software.

Three models were tested for the TPS. The first one was the second-order three-factor structure – the originally hypothesised model on which the scale was developed and calibrated. By Rasch and CFA results, this original structure appeared to be working well as intended, thus confirming its appropriateness as the structure for the TPS. The other two models, the first-order three-factor and the one-factor structures, were examined as possible alternatives to the original structure. However, the results of the same analytic techniques revealed that the alternative models failed to provide better structure than the original model. Hence, the study adopted the second-order three-factor model as the structure for the TPS.

Chapter 7: The Student Perceptions of Assessment Scale

7.1 Introduction

Student perception of assessment is one particular attribute that is considered important due to its relevance to the teaching-learning process. In fact, it is regarded as a vital source of information about the subjective qualities of the assessment tasks, such as classroom tests (Zeidner, 1987). The information about student views on assessment can help inform, guide, and improve educational practices and student learning (Struyven, Dochy, & Janssens, 2005). Through student perceptions, teachers are able to gather evidences about how students react to specific assessment methods and activities and how this reaction influences their approaches to learning. This evidence, in turn, provides the basis for assessment practices to be properly tailored to meet students' interests and improve learning. The potential of student perceptions to enhance student learning formed part of the consideration to include it in this study.

Specifically, the decision to include student perceptions of assessment was due to its relevance to the purpose of this study. Other notable reasons were the inadequacy of research on this topic and an attempt to explore its relationships with other education variables. The insufficient research studies on students' perceptions of assessment (Dorman & Knightley, 2006) warrant more similar undertakings to help ascertain this student characteristic. Moreover, as the web of influence between this attribute and other factors is still not clear (Struyven, et al., 2005), an attempt to establish the relationships was deemed relevant.

The study posited that teachers' assessment literacy affects assessment and instructional practices, which, in turn, influence student perceptions of assessment. This proposition is represented in Figure 7.1. As depicted in the figure, only one-way causal relationship was proposed as the study was mainly concerned with the influence of teachers' assessment literacy on other variables. The proposition

stemmed from the argument raised in the earlier chapter that assessment knowledge influences assessment practices. In addition, assessment practices had been proposed to affect students' experiences of learning, which may arise in response to student perceptions (Maclellan, 2001).

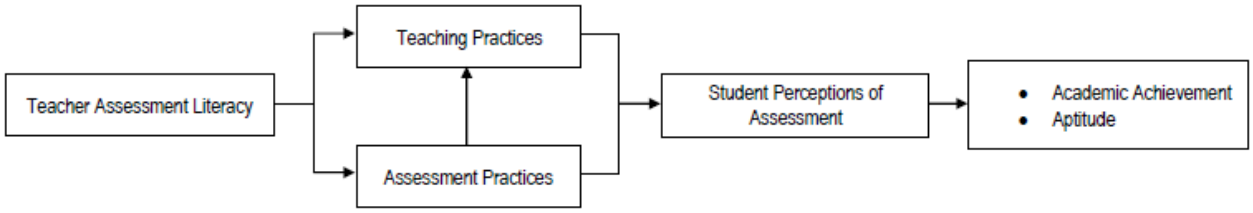


Figure 7.1. The relationship among teacher assessment literacy, assessment practices, teaching practices, student perceptions of assessment and student outcomes in this study

To answer the research questions concerning 'student perceptions of assessment' and to explore the relationships as posited, a relevant instrument was needed. As a result, a search for the relevant scale was conducted. In looking for the questionnaire, the criteria that the scale should suit the intention of the study and be applicable to the research context were used as bases. However, no appropriate instrument was found from the available literature. Thus, items for the sought scale, herein referred to as the 'Student Perceptions of Assessment Scale (SPAS)', were modified from the existing and closely related questionnaire. As a modified scale, the SPAS was subjected to a rigorous validation process to ensure its measurement capacity and utility. This chapter deals with the modification, description, and validation of the SPAS.

The chapter begins with the modification and description of the SPAS. After which, the pilot test of the scale in the research venue is discussed. The ensuing section is devoted to the validation of the SPAS, first at the item level and finally at the structural level. The chapter concludes by reiterating the essential points.

7.2 The SPAS: Its Modification and Description

The SPAS was designed to measure the general perceptions of students on assessment tasks. It was specifically aimed to capture the perceptions on test and assignment. The perceptions on test were intended to draw students' views about the teacher-made or classroom-based tests while those of assignment were to elicit responses on their opinions about other tasks such as seatwork, homework, student demonstration, project, and the like. The test and assignment had been identified as the coverage of the scale as they are the two most common assessment tasks executed by teachers in the research locale.

As no instrument that fully serves the purpose was found at the time of the study, the SPAS was formed using the most relevant available scale, the Students' Perceptions of Assessment Questionnaire (SPAQ) (Cavanagh, Waldrip, Romanoski, Dorman, & Fisher, 2005); Waldrip, Fisher, & Dorman, 2008), as a guide or basis. The SPAQ is an established scale that had been validated using both CTT and Rasch analytic techniques. It measures students' perceptions of assessment tasks in the science subject. It has five constructs namely, congruence with planned learning, authenticity, student consultation, transparency, and diversity. These constructs are defined by Cavanagh, et al. (2005, p. 3) as follows:

a) Congruence with planned learning – Students affirm that assessment tasks align with the goals, objectives and activities of the leaning program;
b) Authenticity – Students affirm that assessment tasks feature real life situations that are relevant to themselves as learners;
c) Student consultation – Students affirm that they are consulted and informed about the forms of assessment tasks being employed;
d) Transparency – The purposes and forms of assessment tasks are affirmed by the students as well-defined and made clear; and
e) Accommodation of student diversity – Students affirm they all have an equal chance of completing assessment tasks.

The SPAQ originally contained 30 items that were equally distributed among the stated constructs. A number of these items were modified to constitute the SPAS. However, in selecting the items for modification, those on student consultation were not included, as they were believed to be irrelevant to the

context of this study. In the research locale, the education department through the national curriculum prescribes the assessment forms to be used and classroom teachers mostly decide the assessment tasks or activities. Also, the SPAQ items were considered in terms of their relevance to the contexts of test and assignment. Thus, all the selected SPAQ items were reworded to make them capture the general perceptions on test and assignment regardless of the specific subject area. As a result, 25 modified items initially formed the SPAS. Fifteen of these items, labeled as PTEST1, PTEST2, PTEST3, PTEST4, PTEST5, PTEST6, PTEST7, PTEST8, PTEST9, PTEST10, PTEST11, PTEST12, PTEST13, PTEST14, and PTEST15, were on 'perceptions of test (PTEST)', and the other ten items, labeled as PASS1, PASS2, PASS3, PASS4, PASS5, PASS6, PASS7, PASS8, PASS9, and PASS10, were on 'perceptions of assignment (PASS)'. All the SPAS items adopted the original response categories, a four-point Likert scale of "almost never", "sometimes", "often", and "almost always", which were coded 1, 2, 3, and 4, respectively. The original and the modified items are presented in Table 7.1.

Table 7.1. The original and modified versions of the SPAS items

Original Version (Waldrip, Fisher, & Dorman, 2008; Cavanagh, Waldrip, Romanoski, Dorman, & Fisher, 2005)		**Modified Version**	
Item #	Wording/Statement	Item Code	Wording/Statement
1	My assessment in science tests what I know.	PTEST1	Tests in my subject measure what I know.
2	How I am assessed is similar to what I do in class.	PTEST2	How I am tested is the same with what I do in class.
3	I am assessed on what the teacher has taught me.	PTEST3	I am tested on what the teacher has taught me.
4	I find science assessment tasks are relevant to what I do outside of school.	PTEST4	My tests are related to what I do outside of school.
5	a) Assessment in science tests my ability to apply what I know to real-life problems; b) I am asked to apply my learning to real life situations.	PTEST5	Tests in my subject measure my ability to apply what I learn to real life situations.
6	Assessment in science examines my ability to answer every day questions.	PTEST6	Tests in my subject measure my ability to answer every day questions.
7	I am aware how my assessment will be marked.	PTEST7	I am aware how my tests will be marked.

8	I know what is needed to successfully accomplish a science assessment task.	PTEST8	I understand what is needed to successfully complete the test.
9	I am told in advance when I am being assessed.	PTEST9	I am told in advance when I am being tested.
10	I am told in advance on what I am being assessed.	PTEST10	I am told in advance on what I am being tested.
11	I am clear about what my teacher wants in my assessment tasks.	PT EST11	I understand what my teacher wants in my test.
12	I have as much chance as any other student at completing assessment tasks.	PTEST12	I have as much chance as any other student at completing the test.
13	I complete assessment tasks at my own speed.	PTEST13	I complete the test at my own speed.
14	I am given assessment tasks that suit my ability.	PTEST14	I am given the test that suits my ability.
15	When I am confused about an assessment task, I am given another way to answer it.	PTEST15	When I am confused about the test, I am given another way to answer it.
16	My assignments/tests are about what I have done in class.	PASS1	My assignments, including project, are about what I have done in class.
17	I find science assessment tasks are relevant to what I do outside of school.	PASS2	My assignments, including project, are related to what I do outside of school.
18	I am aware how my assessment will be marked.	PASS3	I am aware how my assignments will be marked.
19	I know what is needed to successfully accomplish a science assessment task.	PASS4	I understand what is needed to successfully complete my assignment tasks.
20	I am clear about what my teacher wants in my assessment tasks.	PASS5	I understand what my teacher wants in my assignments, including project.
21	I have as much chance as any other student at completing assessment tasks.	PASS6	I have as much chance as any other student at completing my assignments, including project.
22	I complete assessment tasks at my own speed.	PASS7	I complete my assignments, including project, at my own speed.
23	I am given assessment tasks that suit my ability.	PASS8	I am given assignment tasks that suit my ability.
24	When I am confused about an assessment task, I am given another way to answer it.	PASS9	When I am confused about an assignment task, I am given another way to do it.
25	When there are different ways I can complete the assessment.	PASS10	I can complete assignment activity when I am given different ways to do it.

7.3 Pilot Test of the SPAS

After the modification process, the SPAS items were subjected to the review that was carried out in three stages. The first review was done by the researcher himself and his supervisors. After which, the

items were judged by three experts from MSU Tawi-Tawi who were familiar with the classroom assessment situation in the research locale. To further ensure the face and content validity of the scale, the relevance of the items to test and assignment was finally evaluated by 14 Filipino teacher colleagues at the University of Adelaide from which a content validity index (CVI) was computed.

The CVI is a method of establishing content validity in which "a panel of experts is asked to rate each scale item in terms of its relevance to the underlying construct" (Polit & Beck, 2006, pp. 490-491). The concept of CVI stresses that in a scale of four, a rating of three or four by expert indicates that the content is valid and consistent with the conceptual framework (Lynn 1996, as cited in Parsian & Dunning, 2009). Thus, for any item to be retained, a CVI of 3/4 and 4/4 should be obtained. In other words, a CVI of a scale item can be computed by adding the ratings at the relevant and very relevant levels and dividing it with the total number of raters/experts. A CVI value at the relevant level is the threshold for accepting/retaining the item (Parsian & Dunning, 2009). This method was further used for the SPAS, as the literature on perceptions of test and assignment, which could have been used as guide to develop the SPAS, was not available in the literature at the time of the study. The judgment on the relevance of the SPAS items (CVI results) is shown in Table 7.2.

After the review and the computation of CVI, the items were organised into one section and formed part of the study's Student Questionnaire. The questionnaire was pilot tested to the 30 MSU Tawi-Tawi elementary and secondary school students to obtain further feedback. There were two parts of the pilot process. The first part was the administration of the instrument to the selected pilot respondents. This was carried out to obtain the initial reliability, to test the survey operation, and to determine the time for questionnaire completion. The second part was the interview that involved five selected students from the targeted class levels. This was conducted to further determine the suitability of the items in terms of the level of difficulty of the words used, the length of the statements and of the questionnaire as a whole. All feedback from the pilot participants were noted in finalising and administering the instrument. After the initial validation/pilot test of the SPAS, 11 items on perceptions of test and 7 items on perceptions of assignment

were retained (see Appendix B). Moreover, a Chronbach alpha of 0.77 that indicated acceptable reliability of the scale was obtained. The data from the 18 final SPAS items were used to establish the construct validity of the scale using the Rating Scale Model and confirmatory factor analysis (CFA).

Table 7.2. Face and content validity of the SPAS

Construct/Item	Likert Scale				Total	CVI
Perceptions of Test (PTEST)	Not Relevant[a]	Somewhat Relevant[b]	Relevant[c]	Very Relevant[d]		
1. Tests in my subject measure what I know.			6 (43%)	8 (57%)	14 (100%)	14/14 = 1 (Ok)
2. How I am tested is the same with what I do in class.			6 (42%)	7 (58%)	13 (100%)	13/13 = 1 (Ok)
3. I am tested on what the teacher has taught me.		1 (7%)	5 (36%)	8 (57%)	14 (100%)	13/14 = 0.93 (Ok)
4. My tests are related to what I do outside of school.	1 (8%)		5 (38%)	7 (54%)	13 (100%)	12/13 = 0.92 (Ok)*
5. Tests in my subject measure my ability to apply what I learn to real life situations.		1 (7%)	4 (29%)	9 (64%)	14 (100%)	13/14 = 0.93 (Ok)
6. Tests in my subject measure my ability to answer every day questions.			6 (42%)	7 (58%)	13 (100%)	13/13 = 1 (Ok)
7. I am aware how my tests will be marked.		1 (7%)	4 (29%)	9 (64%)	14 (100%)	13/14 = 0.93 (Ok)
8. I understand what is needed to successfully complete the test.		1 (7%)	6 (43%)	7 (50%)	14 (100%)	13/14 = 0.93 (Ok)
9. I am told in advance when I am being tested.	1 (7%)		4 (29%)	9 (64%)	14 (100%)	13/14 = 0.93 (Ok)
10. I am told in advance on what I am being tested.		1 (7%)	6 (43%)	7 (50%)	14 (100%)	13/14 = 0.93 (Ok)
11. I understand what my teacher wants in my test.		1 (7%)	4 (29%)	9 (64%)	14 (100%)	13/14 = 0.93 (Ok)
12. I have as much chance as any other student at completing the test.		1 (7%)	6 (43%)	7 (50%)	14 (100%)	13/14 = 0.93 (Ok)
13. I complete the test at my own speed.		1 (8%)	7 (54%)	5 (38%)	13 (100%)	12/13 = 0.92 (Ok)*
14. I am given the test that suits my ability.		3 (21%)	4 (29%)	7 (50%)	14 (100%)	11/14 = 0.79 (Not Ok)**
15. When I am confused about the test, I am	1 (8%)		6 (42%)	6 (50%)	13 (100%)	12/13 = 0.92

given another way to answer it.					(Ok)*
Perceptions of Assignment (PASS)					
1. My assignments, including project, are about what I have done in class.		6 (43%)	8 (57%)	14 (100%)	14/14 = 1 (Ok)
2. My assignments, including project, are related to what I do outside of school.	1 (7%)	6 (43%)	7 (50%)	14 (100%)	13/14 = 0.93 (Ok)
3. I am aware how my assignments will be marked.		7 (50%)	7 (50%)	14 (100%)	14/14 = 1 (Ok)
4. I understand what is needed to successfully complete my assignment tasks.	1 (7%)	5 (36%)	8 (57%)	14 (100%)	13/14 = 0.93 (Ok)
5. I understand what my teacher wants in my assignments, including project.	1 (7%)	4 (29%)	9 (64%)	14 (100%)	13/14 = 0.93 (Ok)
6. I have as much chance as any other student at completing my assignments, including project.		6 (43%)	8 (57%)	14 (100%)	14/14 = 1 (Ok)
7. I complete my assignments, including project, at my own speed.		7 (50%)	6 (50%)	13 (100%)	13/13 = 1 (Ok)
8. I am given assignment tasks that suit my ability.	3 (21%)	3 (21%)	8 (57%)	14 (99%)	11/14 = 0.79 (Not ok)**
9. When I am confused about an assignment task, I am given another way to do it.	2 (14%)	5 (36%)	7 (50%)	14 (100%)	12/14 = 0.86 (Not ok)**
10. I can complete assignment activity when I am given different ways to do it.	2 (15%)	4 (31%)	7 (54%)	13 (100%)	11/13 = 0.85 (Not ok)**

Items judged to be inapplicable in Tawi-Tawi context and were thus deleted;deleted items based on CVI value; a – not measuring the construct; b – somewhat measuring the construct; c – measuring the construct; d – really measuring the construct*

7.4 Item Analysis Using the Rating Scale Model

The SPAS was initially analysed at the micro level to verify the functioning of the items and to confirm at finer level the appropriateness of the scale. This was a needed process to ensure that SPAS possesses good psychometric properties. Individual items are considered as the backbone of any

instrument and performing item analysis using a more recommended and appropriate technique is a way to ensure that the SPAS worked well as intended or hypothesised.

As mentioned in the earlier section, the Rating Scale Model was employed to analyse SPAS at the item level. The purpose of this analysis was to determine whether or not the items functioned as hypothesised and if all the items under each construct fit the Rasch Model. All the responses from the 2,077 student participants were subjected to analysis using the ConQuest 2.0 software (Wu, Adams, Wilson, & Haldane, 2007). In doing the analysis, the SPAS items were first grouped according to construct and separate analyses for each of the two originally proposed constructs were done. This was to further evaluate the appropriateness of the two-factor model. After which, all items were combined and were examined whether or not they also represent a single or a dominant dimension called the 'students' perceptions of assessment'. The results of the analysis are presented in the relevant subsections below.

7.4.1 Rasch Analysis Results of the SPAS Items under the 'Perceptions of Test (PTEST)' Construct

The SPAS items under the 'PTEST' construct were the first group to be subjected to Rasch analysis. Eleven items were analysed for this construct. The item statistics for the initial and final calibration are presented in Table 7.3. As can be gleaned from the table, all the items in the first calibration fit the Rasch Model as indicated by the acceptable UMS values. The UMS of all the items had a minimum value of 0.95 and a maximum value of 1.11, which were within the acceptable range of 0.70 - 1.30. This revealed that all the 11 items have the capacity to measure the PTEST as a latent trait. In terms of the functioning of the response categories, no disordered thresholds and/or deltas were obtained for the items. This further disclosed that the response options worked well as intended. Moreover, the separation reliability value of 0.99 indicated that the items had a high degree of discrimination and precision (Alagumalai & Curtis, 2005; Wright & Stone, 1999). This served as additional evidence that all the items had desirable spread and accuracy in measuring the PTEST construct.

Table 7.3. Results of the initial and final item analyses of the PTEST construct' of the SPAS

Item	Estimate (Difficulty/Endorsability/Dilemma)	Error	UMS	t
PTEST1	0.05	0.02	0.97	- 1.1
PTEST2	0.13	0.02	0.95	- 1.6
PTEST3	- 0.29	0.02	1.08	2.5
PTEST4	- 0.08	0.02	0.97	- 1.0
PTEST5	- 0.05	0.02	0.99	- 0.4
PTEST6	0.28	0.02	1.11	3.4
PTEST7	- 0.24	0.02	0.95	- 1.5
PTEST8	0.07	0.02	1.01	0.2
PTEST9	0.13	0.02	1.02	0.6
PTEST10	- 0.16	0.02	1.06	1.8
PTEST11	0.169*	0.06	0.98	- 0.8

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 837.42; df=10; Sig level=0.000;*Constrained*

7.4.2 Rasch Analysis Results of the SPAS Items under the 'Perceptions of Assignment (PASS)' Construct

The SPAS items under the 'PASS' construct were the next group analysed at the micro level using the same analytic technique. Seven items that were hypothesised to indicate the construct were subjected to the analysis. The results are shown in Table 7.4. As presented in the table, the initial and final Rasch analysis disclosed that all the 7 items under this construct fit the Rasch model as indicated by the acceptable UMS values. The UMS of all the items had a minimum value of 0.93 and a maximum value of 1.16, which were within the adopted range of 0.70 – 1.30. This confirmed the proposition that the items could indeed reflect the hypothesised construct. In terms of response thresholds and/or deltas, no disordered values were observed. Again, these implied that the response categories functioned as designed. A separation reliability of 0.99 further indicated that the items had high discrimination and precision, which means that they have desirable psychometric properties in measuring 'PASS' as a latent

attribute. Hence, it can be deduced that the seven items can be adopted to measure a construct called 'perceptions of assignment (PASS)'.

Table 7.4. Results of the initial and final item analyses of the 'PASS construct' of the SPAS

Item	Estimate (Difficulty/Endorsability/Dilemma)	Error	UMS	t
PASS1	- 0.03	0.02	0.97	- 0.9
PASS2	0.64	0.02	1.16	4.8
PASS3	0.15	0.02	1.04	1.1
PASS4	- 0.33	0.02	0.93	- 2.4
PASS5	- 0.21	0.02	0.95	- 1.5
PASS6	0.03	0.02	0.97	- 1.0
PASS7	- 0.247*	0.04	0.97	- 0.8

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 1743.08; df=6; Sig level=0.000;*Constrained*

7.4.3 Rasch Analysis Results of the SPAS Items under a Single/Dominant Dimension

After the analysis of SPAS items under the two originally proposed constructs, the next step was to combine all the items and subject them to similar analysis. This was to determine further the possibility of all SPAS items to reflect a single or dominant dimension. All responses from the 2,077 student respondents were analysed using the same statistical software. The initial and final analysis results are presented in Table 7.5.

Table 7.5 showed that all the 18 SPAS items fit the Rasch model as revealed by the acceptable UMS values. The UMS values for all the items had a minimum of 0.90 and a maximum of 1.25, which were within the adopted range of 0.70 to 1.30. These results implied that the items indeed reflected a single or a dominant dimension called the 'students' perceptions of assessment'. In terms of response categories, the results appeared to also indicate that they functioned as hypothesised as disordered thresholds and/or deltas were not spotted. This means that both the items and the response categories were fitting the Rasch model and were working well as hypothesised. In addition, the obtained separation reliability value of 0.99 further revealed that the items had desirable degree of discrimination and precision in measuring the

proposed construct. Hence, the 18 SPAS items could be retained and could be adopted to reflect 'students' perceptions of assessment' as a single or dominant dimension.

Table 7.5. Results of the initial and final item analyses of the SPAS items under a single/dominant dimension

Item	Estimate Difficulty/Endorsability/Dilemma)	Error	UMS	t
PTEST1	0.00	0.02	0.93	- 2.2
PTEST2	0.08	0.02	0.92	- 2.6
PTEST3	- 0.34	0.02	1.07	2.3
PTEST4	- 0.13	0.02	0.94	- 1.8
PTEST5	- 0.10	0.02	0.97	- 1.1
PTEST6	0.22	0.02	1.02	0.7
PTEST7	- 0.29	0.02	0.93	- 2.2
PTEST8	0.02	0.02	1.05	1.5
PTEST9	0.07	0.02	1.05	1.5
PTEST10	- 0.21	0.02	1.02	0.7
PTEST11	0.11	0.02	0.93	- 2.1
PASS1	0.05	0.02	1.00	0.1
PASS2	0.74	0.02	1.25	7.6
PASS3	0.23	0.02	1.04	1.3
PASS4	- 0.26	0.02	0.90	- 3.3
PASS5	- 0.14	0.02	1.00	0.1
PASS6	0.11	0.02	0.95	- 1.5
PASS7	- 0.172*	0.08	1.00	- 0.1

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 2980.52; df=17; Sig level=0.000;*Constrained*

7.5 Examination of the Structure and Item Loadings of the SPAS Items

Further evaluation on the structure of the SPAS was done to ensure that it worked as conceptualised in the study. Specifically, there was a need to examine the structure of the scale to confirm the relationship between the proposed latent constructs and the item loading to verify the relationship between each construct and the corresponding items (construct validity). Similar construct validation process as carried out in the previous chapters was employed for this instrument. The structural fit of the SPAS was evaluated using CFA. The CFA technique was used as the instrument had been formed using a priori. The structural analysis was carried out through LISREL 8.80 software (Jöreskog & Sörbom, 2006) The CFA results are presented in the succeeding subsections.

7.5.1 Structural Analysis Using CFA

The SPAS was analysed initially at the macro level to determine the appropriateness of its hypothesised structure and the fit of the proposed measurement model to the data. This was to confirm the hypothesised relationships between the latent constructs and between the construct and the items. In running the CFA analysis, all the responses from the 2077 student participants were included. The first analysis was performed using the original hypothesis that the SPAS was having a two-factor correlated structure. After which, the analysis of a one-factor structure as the alternative model was carried out. The relevant CFA results are discussed in the ensuing sections/subsections.

7.5.1.1 The Correlated Two-Factor Structure of the SPAS

The correlated two-factor model of the SPAS was examined. This model hypothesised that the SPAS has two underlying constructs namely, 'perceptions of test' abbreviated as PTEST and 'perceptions of assignment' abbreviated as PASS. The PTEST was to be represented by 11 items (PTEST1, PTEST2, PTEST3, PTEST4, PTEST5, PTEST6, PTEST7, PTEST8, PTEST9, PTEST10, and PTEST11) while PASS was to be reflected by seven items (PASS1, PASS2, PASS3, PASS4, PASS5, PASS6, and PASS7). The conceptual representation of this model is shown in Figure 7.2. In evaluating this structure, a number of fit indices for the model fit and the threshold of 0.40 for the item loadings (see Chapter 3) were used. The CFA results are presented in Tables 7.6 and 7.7.

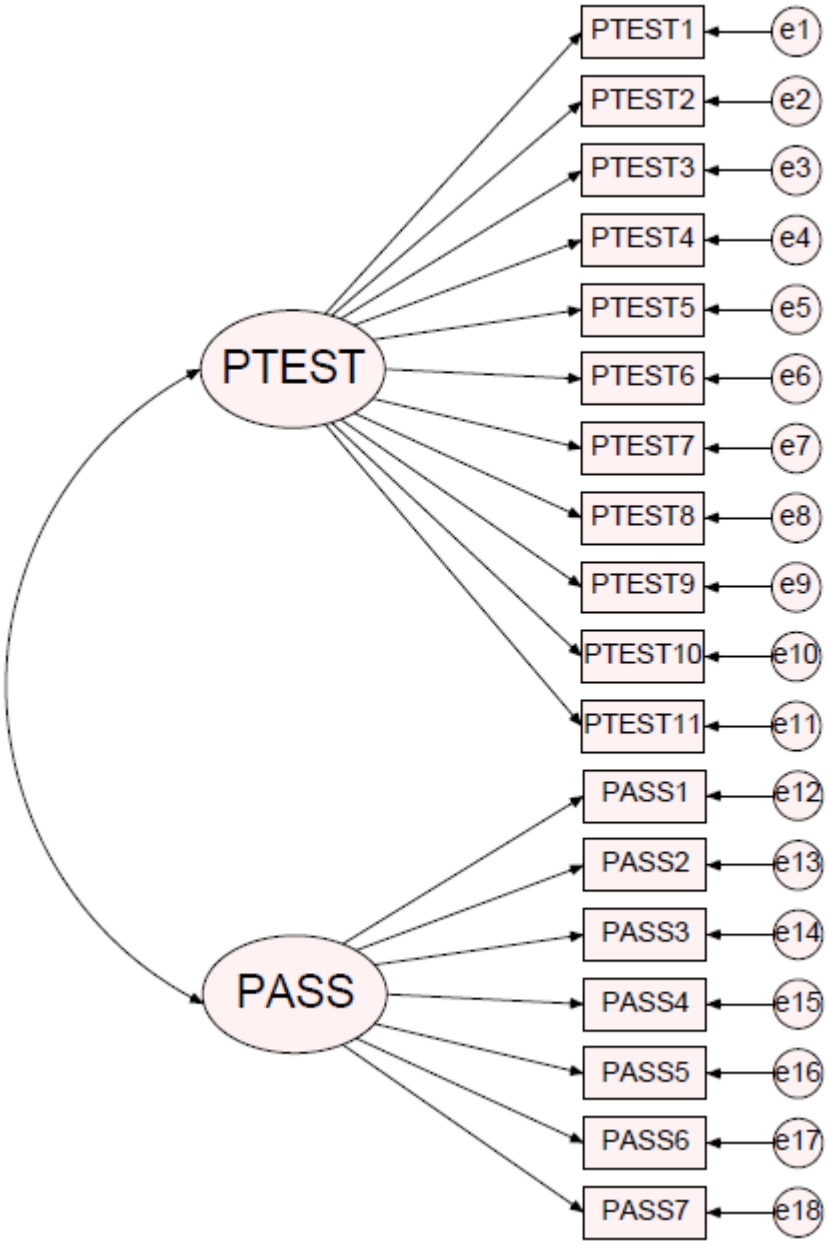


Figure 7.2. Structure of the two-factor model of the SPAS

7.5.1.2 Model Fit

Examining the results in Table 7.6, the correlated two-factor structure of the SPAS appeared to exhibit acceptable fit to the data as shown by the results of most of the fit indices. Only two fit indices (χ^2 and χ^2/df) showed poor results while the other five indices (RMSEA, SRMR, GFI, AGFI, and CFI) indicated acceptable fit. The poor results indicated by χ^2 and χ^2/df should not really affect the judgment about the

structure as these two indices are known to have a number of serious pitfalls such as sensitivity to sample size. In terms of parsimony, the model appeared desirably less complex as implied by the PGFI value of 0.73. Hence, it can be concluded that the correlated two-factor model is an acceptable structure for the SPAS.

Table 7.6. Summary of fit indices for the first-order two-factor structure of the SPAS

Fit Index	Obtained Value	Remark
X^2	1284.13 (P = 0.00)	Poor fit
X^2 /df	1284.13/134 = 9.58	Poor fit
RMSEA	0.07	Acceptable fit
SRMR	0.05	Acceptable fit
GFI	0.93	Acceptable fit
AGFI	0.91	Acceptable fit
CFI	0.91	Acceptable fit
PGFI	0.73	Some degree of parsimony

7.5.1.3 CFA of the Hypothesised Measurement Model

The resulting statistics in Table 7.7 appeared to confirm the hypothesis that the two constructs are correlated and are reflected by the modified items. The correlation coefficient of 0.96 supported the proposition that perceptions of test has, indeed, a positive relationship with perceptions of assignment. This result was expected as both test and assignment are commonly understood as forms of assessment. In terms of the functioning of the items, the results showed that the items measured the respective constructs. For the 'perceptions of test (PTEST)', all the items (PTEST1, PTEST2, PTEST3, PTEST4, PTEST5, PTEST6, PTEST7, PTEST8, PTEST9, PTEST10, and PTEST11) exhibited acceptable factor loadings that were within the range of 0.40 to 0.51. This means that all the eleven items really tapped the construct of PTEST. For the 'perceptions of assignment (PASS)', six items (PASS1, PASS3, PASS4, PASS5, PASS6, and PASS7) had acceptable loadings while one item (PASS2) had a loading of 0.29 that was below the

adopted threshold of 0.40. The factor loadings of the six functioning items were within the range of 0.44 to 0.51, which indicated that the said items reflected the construct of PASS. These results provided further support to acceptable model fit and confirmed the appropriateness of the correlated two-factor structure for the SPAS. The misfitting item PASS2 was perhaps unfamiliar to some students or some of them possibly failed to link their assignment with outside activities.

Table 7.7. Factor loadings of the SPAS items under the first-order two-factor model

Structure	Construct	Correlation between Constructs	Item	Loading(se)*
Two-Factor Correlated Model	Perceptions of Test (PTEST)	0.96	PTEST1	0.45(0.02)
			PTEST2	0.40(0.02)
			PTEST3	0.47(0.02)
			PTEST4	0.41(0.02)
			PTEST5	0.49(0.02)
			PTEST6	0.43(0.02)
			PTEST7	0.47(0.02)
			PTEST8	0.45(0.02)
			PTEST9	0.45(0.02)
			PTEST10	0.44(0.02)
			PTEST11	0.51(0.02)
	Perceptions of Assignment (PASS)	0.96	PASS1	0.46(0.02)
			PASS2	0.29(0.02)
			PASS3	0.44(0.02)
			PASS4	0.51(0.02)
			PASS5	0.44(0.02)
			PASS6	0.50(0.02)
			PASS7	0.44(0.02)

**n = 582*

7.5.2 The CFA of the Alternative Model

While the originally hypothesised structure worked well for the SPAS, it was also necessary to evaluate an alternative model to determine the possible existence of a better and less complex structure. Thus, the one-factor model was also tested for the SPAS. In evaluating this alternative model, similar technique, process, software, and indicators were employed. The results are shown and discussed in the following subsections.

7.5.2.1 The One-Factor Structure of the SPAS

The one-factor structure was examined to determine its possibility as a better model for the SPAS. Under this proposed model, all the items from the two constructs were combined. The same labels for all the items were used for coherence and proper identification purposes. These items were loaded to one main factor called the 'students' perceptions of assessment (SPA)'. The conceptual representation of this model is presented in Figure 7.3. The CFA results for this model are shown in Tables 7.8 and 7.9.

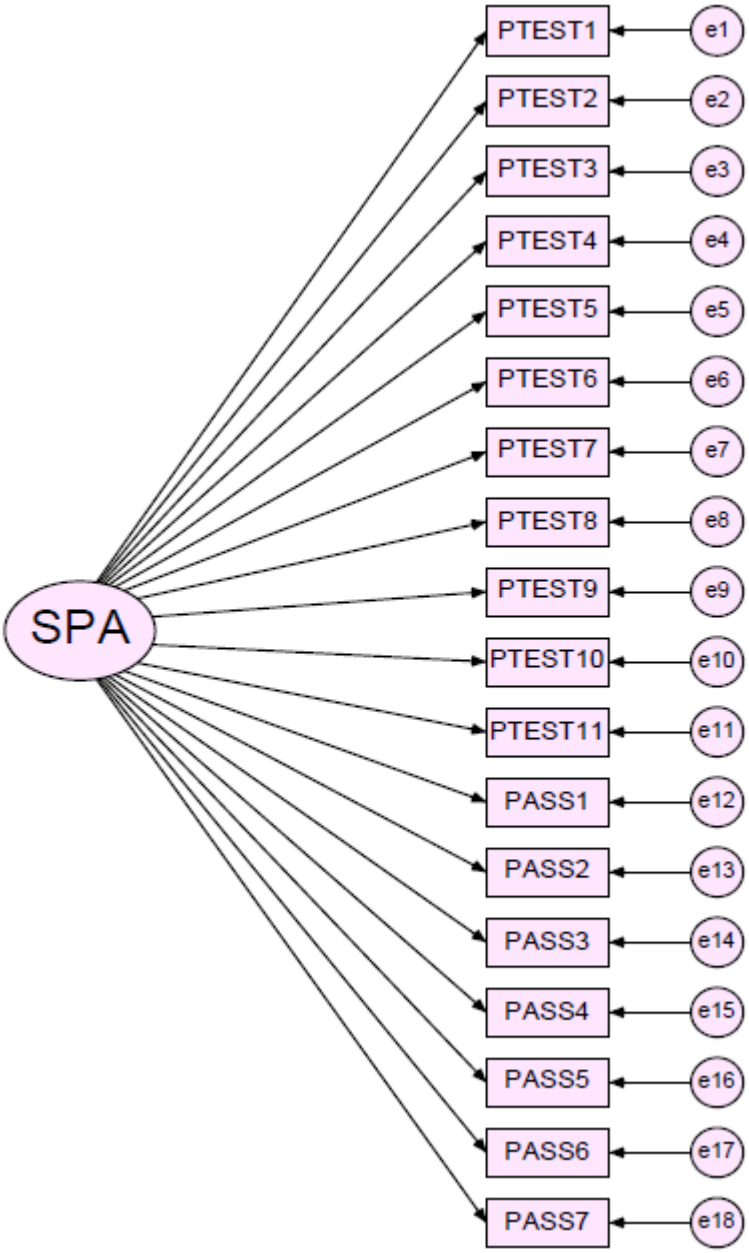


Figure 7.3. Structure of the one-factor model of the SPAS

7.5.2.2 Model Fit of the Alternative One-factor Structure

The CFA results presented in Table 7.8 below indicate that the one-factor structure has also acceptable fit to the data. Similar to what were obtained under the two-factor model, only two fit indices (χ^2 and χ^2/df) exhibited poor fit while the rest (RMSEA, SRMR, GFI, AGFI, and CFI) indicated acceptable fit. In terms of parsimony, the one-factor model is slightly better than the previous model as implied by the PGFI

value of 0.74. Thus, it can be deduced that the one-factor model is a possible alternative to the earlier model.

Table 7.8. Summary of fit indices for the one-factor structure of the SPAS

Fit Index	Obtained Value	Remark
X^2	1290.00 (P = 0.00)	Poor fit
X^2/df	1290.00/135 = 9.56	Poor fit
RMSEA	0.07	Acceptable fit
SRMR	0.05	Acceptable fit
GFI	0.93	Acceptable fit
AGFI	0.91	Acceptable fit
CFI	0.91	Acceptable fit
PGFI	0.74	Some degree of parsimony

7.5.2.3 CFA of the Hypothesised One-factor Measurement Model

Examining the factor loadings, similar picture appeared under the one-factor model. As can be seen in Table 7.9, all of the items appeared to reflect the main construct (students' perceptions of assessment) as indicated by the acceptable factor loadings of 0.40 to 0.51, except for item PTA2, which had a factor loading of 0.29. The results were in agreement with the acceptable overall fit of the model to the data. Hence, the one-factor model and the 17 functioning items can be adopted for the scale.

Table 7.9. Factor loadings of the SPAS items under one-factor model

Structure	Construct	Item	Loading(se)*
One-Factor	Student Perceptions of Assessment (SPA)	PTEST1	0.45(0.02)
		PTEST2	0.40(0.02)
		PTEST3	0.47(0.02)
		PTEST4	0.41(0.02)
		PTEST5	0.48(0.02)
		PTEST6	0.44(0.02)
		PTEST7	0.47(0.02)
		PTEST8	0.44(0.02)
		PTEST9	0.45(0.02)
		PTEST10	0.44(0.02)
		PTEST11	0.51(0.02)
		PASS1	0.45(0.02)
		PASS2	0.29(0.02)
		PASS3	0.43(0.02)
		PASS4	0.51(0.02)
		PASS5	0.43(0.02)
		PASS6	0.50(0.02)
		PASS7	0.43(0.02)

**n = 582*

7.5.3 Model Used in the Study

The Rasch analysis of the original and alternative models initially revealed that the structures reflected by these models were appropriate for the SPAS. The same results were obtained when the SPAS was analysed at the structural level using CFA. As the results of Rasch analysis were the final bases in judging the acceptability of the tested models, it was discerned that both one-factor and two-factor models were appropriate for the SPAS. However, the use of model parsimony as a further criterion in judging between the models led to the preference for one-factor structure over the other model. Thus, in this study, the one-factor structure was adopted for the SPAS.

7.6 Summary

This chapter dealt with the process of forming and validating the SPAS. The SPAS was formed by modifying the items of the SPAQ, a well-established and a closely related instrument. In validating the SPAS, CFA, using LISREL 8.80, was utilised to evaluate its structure and the fit of the measurement model to the data. To further examine the scale, the rating scale model, through ConQuest 2.0, was carried out to determine the appropriateness of the individual items with respect to the implied dimensions. There were two models tested for the SPAS. The first one was the originally proposed correlated two-factor structure and the second one was the possible one-factor model. By CFA results, the appropriateness of these models appeared acceptable. These results were confirmed by the findings from the rating scale model. As by Rasch results the two models were equally appropriate, model parsimony was used as a basis in accepting either of the models. Hence, the one-factor model was adopted in this study as the appropriate structure for the SPAS.

Chapter 8: The Student Attitude Towards Assessment Scale

8.1 Introduction

This study covered student attitude towards assessment. The consideration of this attribute stemmed from a number of assumptions. Firstly, teachers' assessment literacy was believed to influence student attitude through assessment activities that teachers carry out in the classroom. As proposed in the earlier chapter, assessment knowledge has a causal relationship with assessment practices. These practices, in turn, were assumed to affect student attributes, which include attitude towards assessment. Secondly, any assessment task was deemed to excite a feeling or reaction thereby possibly forming attitude towards the assessment itself. This is in line with some theories, such as those held by the neo-behaviorist view, that any stimulus or attitude object can form attitude (Naumann, Richter, Groeben, & Christmann, n.d.). Thirdly, attitude could predict actions or behavior under certain conditions; the attitude-behavior causal relation is even stronger when the persons have direct experience with the attitude object (Glasman & Albarracin, 2006). This case was viewed to be true to students who are frequently subjected to do assessment tasks in the classroom. In the course of doing assessment-related activities, students inevitably develop positive or negative feelings about assessment. And fourthly, understanding student attitudes was assumed to be vital in any effort to enhance learning. When information about student attitude is available, teachers are in a better position to excogitate strategies and provide the kind of environment that elicits students' interests and develop desirable approaches to learning. In the case of assessment, positive attitude towards assessment activities would be possibly developed thereby encouraging students to engage more and acquire deep and meaningful learning. Other consideration for including student attitude towards assessment was the inadequacy of research studies on this topic for which more relevant studies are warranted.

As has been highlighted in the previous chapters, the study posited that teachers' assessment literacy affects assessment and instructional practices, which, in turn, influence student attributes including attitude towards assessment. This proposition is represented in Figure 8.1. As depicted in the figure, only one-way causal relationship was proposed as the study was mainly concerned with the influence of teachers' assessment literacy on other variables. The proposition stemmed from the argument implied in the aforementioned assumptions and from the hypothesis that assessment knowledge influences assessment practices, which further affect student characteristics.

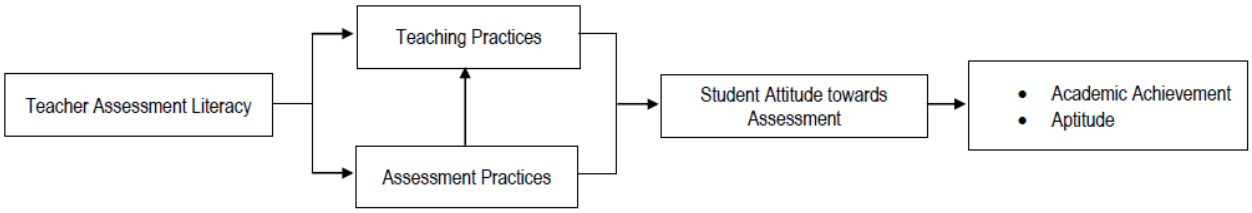


Figure 8.1. The relationship among teacher assessment literacy, assessment practices, teaching practices, student attitude towards assessment, and student outcomes in this study

To answer the research questions concerning the 'student attitude towards assessment' and to explore the relationships as shown in the figure above, a relevant instrument was needed. As a result, a search for the relevant scale was conducted. In searching for the questionnaire, the criteria that the scale should suit the intention of the study and be applicable to the research context were used as bases. However, no appropriate instrument was found from the available literature at the time of the study. Thus, items for the sought scale, herein referred to as the 'Student Attitude Towards Assessment Scale (SATAS)', were modified or developed using the related questionnaire as a guide or basis. As a modified or developed scale, the SATAS was subjected to a rigorous validation process to ensure its measurement capacity and utility for this study. This chapter deals with the modification/development, description, and validation of the SATAS.

The chapter begins with the modification/development and description of the SATAS. After which, the pilot test of the scale in the research venue is discussed. The ensuing section is devoted to the

validation of the SATAS, first at the item level and finally at the structural level. The chapter concludes by reiterating the salient points.

8.2 The SATAS: Its Development and Description

The SATAS was designed to measure the general attitude of students towards assessment as a means of achieving success in school. It was intended to capture the assessment attitude in an attempt to establish its link to teachers' assessment literacy and to highlight the view that understanding attitude has important implications for student learning.

As no instrument that measures assessment attitude, as intended, was available, the SATAS was formed using the existing questionnaire. This instrument was the 'Attitude Scale' developed by Mickelson (1990). The decision to use this scale as a guide was mainly based on the semblance of the situation it portrayed with the context of the research locale and on its relevance to the intention of the study. Originally, the 'Attitude Scale' was composed of 14 items that were hypothesised to represent two underlying constructs namely, 'abstract attitude' (8 items) and 'concrete attitude' (6 items). Mickelson (1990) adopted these constructs from Parkin's (1976, as cited in Mickelson, 1990) macrotheoretical constructs of dominant and subordinate value systems. Using this perspective, abstract attitude was taken to tap abstract beliefs about dominant values prevailing in the society while concrete attitude was taken to measure the subordinate value systems in which people see or experience the concrete realities. As stressed by Mickelson (1990), these constructs were appropriate and useful for understanding students or people who view education as an important means to future opportunities but who exhibit low academic achievements. In developing the SATAS, some of the 14 items were used as basis. However, all modified or developed items were intended to partly follow the abstract attitude, as it was believed that students with their age could hardly link assessment to outside realities and opportunities. Although students' life in school is punctuated by assessment activities, their feelings and views about assessment are confined to their world in the classroom. Thus, the underlying construct that SATAS intended to measure was more akin to the

'cognition-based attitude' (Wilson, Dunn, Kraft, & Lisle, 1989, as cited in Naumann, et al., n.d.). Under this framework, SATAS viewed that assessment attitude is part of the student cognitive process and is only stored in the students' minds. Being it so, this attribute affects students in the way they do assessment tasks. Hence, unpacking it could be a way to help turn student attitude into positive synergies.

The final form of SATAS contained seven items (ATTD1, ATTD2, ATTD3, ATTD4, ATTD5, ATTD6, and ATTD7) reflecting a single construct called, 'student attitude towards assessment'. These items adopted a four-point Likert scale of "strongly disagree", "disagree", "agree", and "strongly agree", which were coded 1, 2, 3, and 4, respectively. The source and the modified/developed SATAS items are presented in Table 8.1 below.

Table 8.1. Source and developed SATAS items

Source (Mickelson, 1990)		**Developed Items**	
Item	Wording/Statement	Item Code	Wording/Statement
1	Getting a good education is a practical road to success for a young black (white) man (woman) like me (Abstract Attitude).	ATTD1	Assessment helps me to become successful in my education (Abstract Attitude).
2	If everyone in America gets a good education, we can end poverty (Abstract Attitude).	ATTD2	If everyone in my school is given an effective assessment, we can gain good education (Abstract Attitude).
3	Achievement and effort in school lead to job success later on (Abstract Attitude).	ATTD3	Assessment in school leads to good academic achievement (Abstract Attitude).
4	Young white (black) women (men) like me have a chance of making it if we do well in school (Abstract Attitude).	ATTD4	I have a chance to be successful if I do well in my tests in school (Abstract Attitude).
5	School success is not necessarily a clear path to a better life (Abstract Attitude).	ATTD5	Success in school is not necessarily dependent on tests (Abstract Attitude).*
6	Based on their experiences, my parents say people like us are not always paid or promoted according to our education (Concrete Attitude).	ATTD6	Doing well in classroom tests is not always helpful in completing my education (Abstract Attitude).*
7	School success is not necessarily a clear path to a better life (Abstract Attitude).	ATTD7	Assessment, like tests, makes my education difficult (Abstract Attitude).*

**Items that are negatively worded*

8.3 Pilot Test of the SATAS

After the modification/development process, the SATAS items were reviewed following the same procedure that was used in validating the SPAS and as described in Chapter 7. The review was specifically

done by the researcher, researcher's supervisors, three experts from the Mindanao State University (MSU) in Tawi-Tawi, and by 14 teacher colleagues at the University of Adelaide to ensure face and content validity of the scale. A content validity index (CVI) was also computed (see Chapter 7, Section 7.3 for information about CVI). The judgment on the relevance of the items (CVI results) is shown in Table 8.2. The items were then organised into one section and formed part of the study's Student Questionnaire. As described in Chapter 7, the questionnaire was pilot tested to 30 MSU Tawi-Tawi elementary and secondary school students to obtain further feedback. The questionnaire was administered to obtain the initial validity and reliability, to test the survey operation, and to determine the time for questionnaire completion. The interview that involved five selected students from the targeted class levels was also conducted to further determine the suitability of the items in terms of the level of difficulty of the words used, the length of the statements and of the questionnaire as a whole. All feedback from the pilot participants were noted in finalising and administering the instrument. After the pilot test, a Chronbach alpha of 0.77 (acceptable reliability) was obtained and four SATAS items were retained. The data from the four final SPAS items were used to establish the psychometric properties and the construct validity of the scale using the Rasch Rating Scale Model and confirmatory factor analysis (CFA).

Table 8.2. Face and content validity of the SATAS items

Construct/Item	Likert Scale				Total	CVI
Perceptions towards Test	Not Relevant[a]	Somewhat Relevant[b]	Relevant[c]	Very Relevant[d]		
1. Assessment helps me to become successful in my education.			5 (36%)	9 (64%)	14 (100%)	14/14 = 1 (Ok)
2. If everyone in my school is given an effective assessment, we can gain good education.			3 (21%)	11 (79%)	14 (100%)	14/14 = 1 (Ok)
3. Assessment in school leads to good academic achievement.			3 (21%)	11 (79%)	14 (100%)	14/14 = 1 (Ok)
4. I have a chance to be successful if I do well in my tests in school.		1 (7%)	4 (29%)	9 (64%)	14 (100%)	13/14 = 0.93 (Ok)
5. Success in school is not necessarily dependent on tests.*	1 (7%)	1 (7%)	2 (14%)	10 (71%)	14 (99%)	12/14 = 0.86 (Not ok)**
6. Doing well in classroom tests is not always helpful in completing my education.*	2 (14%)		4 (29%)	8 (57%)	14 (100%)	12/14 = 0.86 (Not ok)**
7. Assessment, like tests, makes my education difficult.*	2 (14%)		4 (29%)	8 (57%)	14 (100%)	12/14 = 0.86 (Not ok)**

**Negatively worded items; **Deleted items based on CVI; a – not measuring the construct; b – somewhat measuring the construct; c – measuring the construct; d – really measuring the construct*

8.4 Examination of the Item and Structural Fit of the SATAS

Further evaluation of SATAS was done to ensure that it worked as conceptualised in the study. Specifically, there was a need to examine the fit of SATAS items using the Rasch model, particularly the Rating Scale Model (Andrich, 1978). This was to ascertain that individual SATAS items were functioning as intended. The item-level analysis was performed using ConQuest software (v. 2.0) (Wu, Adams, Wilson, & Haldane, 2007). Moreover, the structure of the scale was examined to confirm the relationship between the proposed latent construct and the corresponding items (construct validity). Similar construct validation process as carried out in the previous chapters was employed for this instrument. The structural fit of the SATAS was evaluated using CFA. The CFA technique was used as the instrument had been formed using a priori. The structural analysis was carried out through LISREL 8.80 software (Jöreskog & Sörbom, 2006)

(see Chapter 3 for details about Rasch Model and CFA, and Chapter 5 for details about the Rating Scale Model). The analysis results are presented in the succeeding subsections.

8.4.1 Item Analysis Results Using the Rating Scale Model

The SATAS was initially analysed at the micro level to verify further the functioning of the items and to confirm at finer level the appropriateness of the scale. The purpose of this analysis was to determine whether or not the items functioned as hypothesised and if all the items under a single construct fit the Rasch Model or the Rating Scale Model to be specific. All the responses from the 2077 student participants were subjected to analysis using ConQuest 2.0 software (Wu, Adams, Wilson, & Haldane, 2007). The results of the analysis are presented in the relevant subsections below.

8.4.1.1 Rasch Analysis Results of the SATAS Items under a Single/Dominant Dimension

The SATAS items under a single or dominant dimension were subjected to Rasch analysis. Four items were analysed for the hypothesised construct. The item statistics for the initial and final calibration are presented in Table 8.3. As can be spotted from the table, all the items in the first calibration fit the Rasch model as indicated by the acceptable UMS values. The UMS of all the items had a minimum value of 0.91 and a maximum value of 1.09, which were within the acceptable range of 0.70 - 1.30. This revealed that the four SATAS items have the capacity to measure the 'student attitude towards assessment' as a latent trait. In terms of the functioning of the response categories, no disordered thresholds and/or deltas were obtained for the items. This further disclosed that the response options worked well as intended. Moreover, the separation reliability value of 0.99 indicated that the items had a high degree of discrimination and precision (Alagumalai & Curtis, 2005; Wright & Stone, 1999). This served as additional evidence that all the items had desirable spread and accuracy in measuring the construct. Hence, it can be deduced that the proposed dimension and its reflecting items are very appropriate for the SATAS. Furthermore, the data gathered through this scale are deemed trustworthy and can be used for subsequent analysis.

Table 8.3. Results of the initial and final items analyses of the SATAS items under a single/dominant dimension

Item	Estimate (Difficulty/Endorsability/Dilemma)	Error	UMS	t
ATTD1	- 0.34	0.03	0.98	- 0.6
ATTD2	0.14	0.02	0.91	- 2.9
ATTD3	0.03	0.03	0.98	- 0.7
ATTD4	0.171*	0.04	1.09	2.9

*Separation Reliability = 0.99; Chi-Square Test of Parameter Equality = 199.25; df=3; Sig level=0.000; *Constrained*

8.4.2 Structural Analysis Using CFA

The SATAS was further analysed at the macro level to determine the appropriateness of its hypothesised one-factor structure and the fit of the proposed measurement model to the data. This was to confirm the hypothesised relationship between the latent construct and the items. In running the CFA analysis, all the responses from the 2077 student participants were included. The results are discussed in the ensuing sections/subsections.

8.4.2.1 The One-Factor Structure of the SATAS

The one-factor model of the SATAS was examined. This model hypothesised that the SATAS has one underlying construct called the 'student attitude towards assessment'. This construct was to be represented by four items namely, ATTD1, ATTD2, ATTD3, and ATTD4. The conceptual representation of this model is shown in Figure 8.2. In evaluating this structure, a number of fit indices for the model fit and the threshold of 0.4 for the item loadings were used (see Chapter 3 for details). The CFA results on the overall model fit and the fit of the measurement model to the data are presented in Tables 8.4 and 8.5.

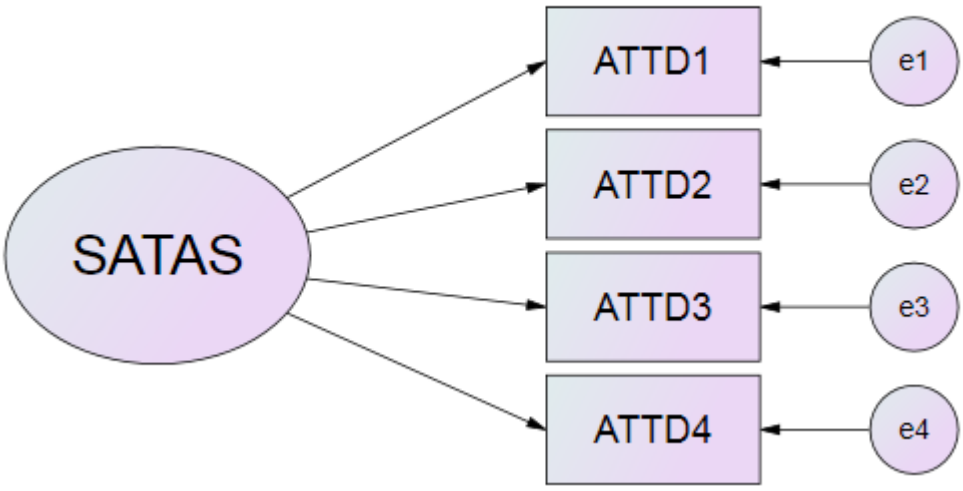


Figure 8.2. Structure of the one-factor model of the SATAS

8.4.2.2 Model Fit

Examining the results in Table 8.4, the one-factor structure of the SATAS appeared to exhibit a very good fit to the data as shown by the results of the adopted fit indices. These fit indices (χ^2, χ^2/df, RMSEA, SRMR, GFI, AGFI, and CFI) provided consistent results. It could be noted that GFI, AGFI, and CFI indicated perfect fit while RMSEA and SRMR were almost at the most desired level. However, the PGFI result of 0.20 was indicative of less parsimony. This result can perhaps be ignored as there is only one model tested for the scale. Hence, it can be concluded that the one-factor model is the appropriate structure for the SATAS.

Table 8.4. Summary results of fit indices for the one-factor structure of the SATAS

Fit Index	Obtained Value	Remark
X^2	2.50 (P = 0.29)	Good fit
X^2 /df	2.50/2 = 1.25	Good fit
RMSEA	0.01	Good fit
SRMR	0.01	Good fit
GFI	1.00	Excellent fit
AGFI	1.00	Excellent fit
CFI	1.00	Excellent fit
PGFI	0.20	Less parsimonious

8.4.2.3 CFA of the Hypothesised Measurement Model

The resulting statistics in Table 8.5 appeared to confirm the hypothesis that the four remaining items reflect SATAS' single construct. The four SATAS items (ATTD1, ATTD2, ATTD3, and ATTD4) exhibited acceptable factor loadings that were within the range of 0.57 to 0.68, well above the adopted threshold of 0.40. Thus, these items were retained to reflect the student attitude towards assessment.

Table 8.5. Factor loadings of the SATAS items under the one-factor model

Structure	Construct	Item	Loading(se)*
One-Factor	Student Attitude towards Assessment (SATA)	ATTD1	0.68(0.02)
		ATTD2	0.66(0.02)
		ATTD3	0.60(0.02)
		ATTD4	0.57(0.02)

**n = 582*

8.5 Model Used in the Study

The Rasch analysis of the originally proposed one-factor structure of the SATAS disclosed highly acceptable results. The same picture was obtained when the SATAS was analysed at the structural level

using CFA. As the results of both Rasch Model and CFA were indicative of a single/dominant dimension as the most appropriate structure, this study adopted the one-factor model for the SATAS.

8.6 Summary

This chapter dealt with the process of forming and validating the SATAS. This scale was formed by developing/modifying items using the existing 'attitude scale' as a guide. To validate the SATAS at the micro level, Rasch Model/Rating Scale Model was used through ConQuest 2.0 software. To further examine its utility at the macro level, CFA, using LISREL 8.80 software, was employed. Only one model (one-factor model) was tested for the SATAS. By Rasch analysis and CFA results, the appropriateness of this model appeared highly acceptable. Hence, the one-factor model was adopted in this study as the appropriate structure for the SATAS.

Chapter 9: Descriptive and Some Inferential Results

9.1 Introduction

In this study, teacher assessment literacy and relevant variables namely, assessment practices, teaching practices, assessment perceptions, assessment attitude, academic achievement, and aptitude were investigated. In addition, by combining factors at the student and teacher levels, the influence of teacher assessment literacy on student achievement and aptitude were examined. The effects of demographic factors such as gender, age range, academic qualification, years of teaching experience, and school type on the teacher-level variables and gender on the student-level variables were likewise explored. This was based on a model that has been developed and drawn from previous studies (refer to Chapter 2). This model was used to answer the general research questions advanced in Chapter 1 and the following specific questions:

1. What is the level of assessment literacy of the elementary and secondary school teachers?
2. What are the assessment practices of the elementary and secondary school teachers?
3. What are the teaching practices of the elementary and secondary school teachers?
4. What are the perceptions of the elementary and secondary school students on assessment?
5. What is the attitude of the elementary and secondary school students towards assessment?
6. What is the level of academic achievement of Grade 6 and Second Year high school students?
7. What is the level of general aptitude of Fourth Year high school students?

8. Is there any significant difference on the levels of elementary and secondary school teachers' assessment literacy, assessment practices, and teaching practices in terms of gender, age range, academic qualification, years of teaching experience, school level, and school type?
9. How does teacher assessment literacy interact with assessment practices, teaching practices, student perception of assessment, student attitude towards assessment, academic achievement, and aptitude?

Question 9 leads to the following specific questions under the two broad headings:

9.1 Teacher-level factors

9.1.1 What is the influence of gender, age range, academic qualification, years of teaching experience, and school type on teachers' assessment literacy, assessment practices, and teaching practices?

9.1.2 What is the influence of teachers' assessment literacy on their assessment and teaching practices?

9.1.3 What is the influence of teachers' assessment practices on their teaching practices?

9.1.4 What is the influence of teacher assessment literacy on student academic achievement and aptitude through assessment practices, teaching practices, student perceptions of assessment, and student attitude towards assessment?

9.2 Student-level factors

9.2.1 What is the influence of gender on student perceptions of assessment, student attitude towards assessment, academic achievement, and aptitude?

9.2.2 What is the influence of students' perceptions of assessment on their attitude towards assessment?

9.2.3 What is the impact of Grade 6 and Second Year high school students' perceptions of assessment and attitude towards assessment on their academic achievement?

9.2.4 What is the impact of Fourth Year high school students' perceptions of assessment and attitude towards assessment on their aptitude?

To answer questions 1-8 above, descriptive and inferential analyses were carried out. However, before carrying out subsequent analyses, it was important to extract descriptive information from the dataset to provide the profile of samples with respect to the demographic factors considered in this study. This is to provide a complete picture of the data for each of the factors and to allow proper interpretation of relevant results.

This chapter describes the sample in terms of the distribution of the following: student and teacher gender, age range of teachers, the academic qualification of teachers, and years of teaching experience of the teachers, school type where the sample were drawn, and the school level taught by teachers. The chapter also includes a description of the steps carried out in the scaling process – data preparation, and the steps undertaken to transform raw scores into measures, and also including measures taken to handle missing data. The level of analysis employed in this study is discussed, and the descriptive and inferential analysis results are also provided. The chapter concludes with a summary that reiterates the key points/findings.

9.2 Descriptive Information about the Sample

9.2.1 Student Gender

As gender is one of the factors examined in this study, it is important to present its distribution. Table 9.1 on the next page shows the distribution of the student respondents by gender.

Table 9.1. Distribution of student respondents by gender

Student gender	Frequency	Percent
Female	1239	59.7%
Male	838	40.3%
Total	2077	100%

It can be observed that there are more female than male students in the sample. This could be attributed to the fact that Tawi-Tawi has a bigger female population than male population according to the 2010 Philippine Census (www.census.gov.ph). A graphical representation of Table 9.1 is provided in Figure 9.1 to get an easier grasp of the student sample distribution.

The student questionnaire was administered to the students described above for the data needed in this study. The data collected became part of the raw data for this study. It contains students' demographic information and data for each of the scale in the questionnaire intended for student participants.

The student participants in this study came from different schooling levels: Grade 6 Elementary (primary), 2nd Year High School, and 4th Year High School. A breakdown of how many female and male students participated in each schooling level is provided in Table 9.2. A trend similar to the one presented in Figure 9.1 can be observed, that there are more female student participants than male.

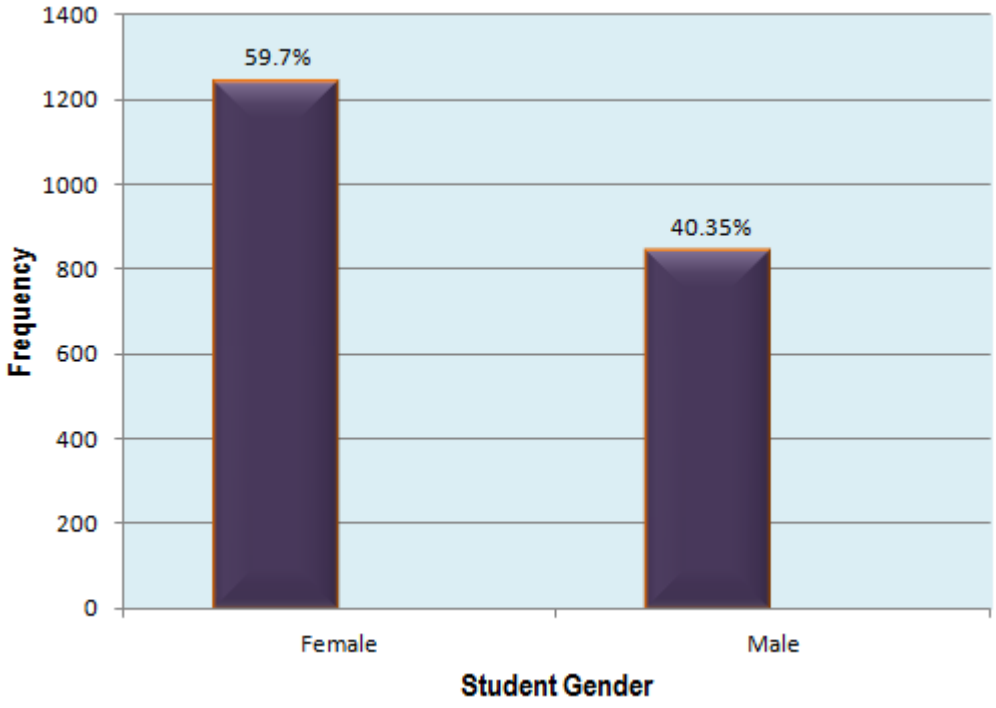


Figure 9.1. Distribution of student respondents by gender

Figure 9.2 shows a clearer picture of the female and male student participant distribution by schooling level. It is of interest to note that the male to female ratio is between 3:6 and 4:6 (or 2:3). Again, this is consistent with the reported trend of Tawi-tawi male and female population distribution.

Table 9.2. Gender distribution of students by schooling level

Grade/Year Level	Gender		Total
	Female	Male	
Grade 6	537 (58.7%)	378 (41.3%)	915 (100%)
2nd Year HS	331 (64.3%)	184 (35.7%)	515 (100%)
4th Year HS	371 (57.3%)	276 (42.7%)	647 (100%)
Total	1239	838	2077

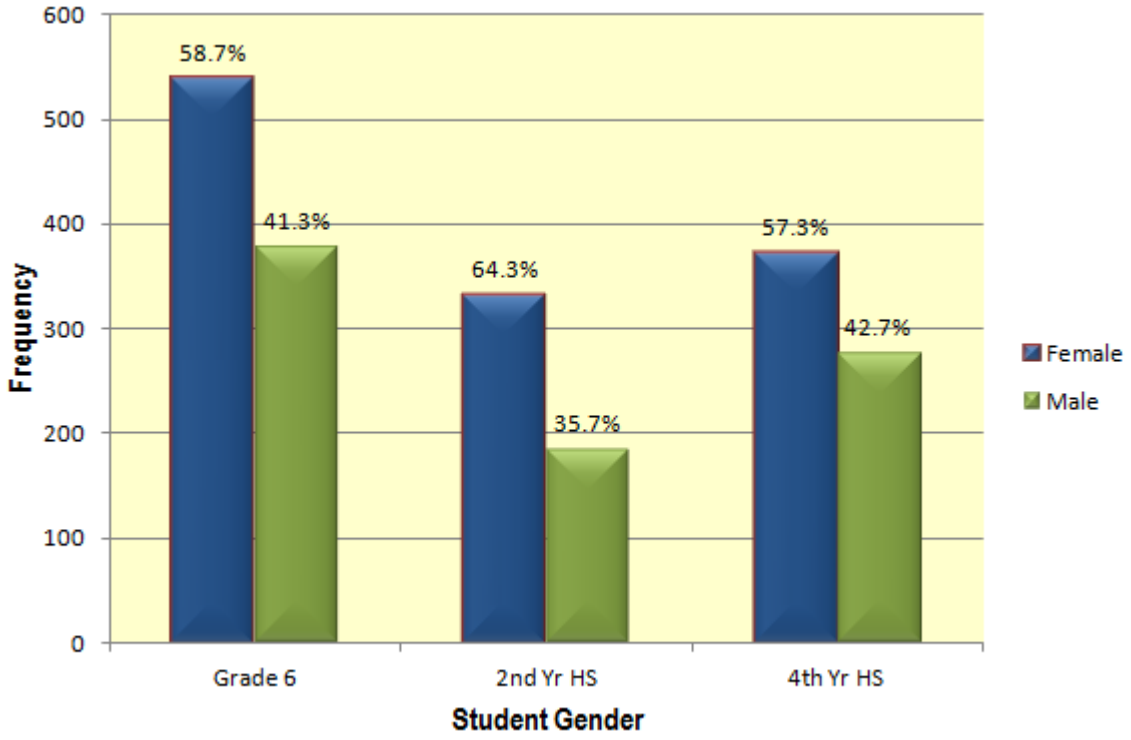


Figure 9.2. Gender distribution of students by schooling level

9.2.2 Teacher Gender

It was also important to take into account the gender distribution of the teacher sample due to the same reason that gender was one of the variables examined in this study. Table 9.3 shows the distribution of male and female teachers in the study sample.

Table 9.3. Distribution of teacher respondents by gender

Gender	Frequency	Percent
Female	359	61.7%
Male	223	38.3%
Total	582	100%

Similar to the trend shown for student gender distribution, there are more female teachers in the teacher sample than males. The ratio between males and females are strikingly similar to those of the

students' – roughly 2:3. This was considered important to note as gender was hypothesised to have significant influence on some of the factors examined in this study.

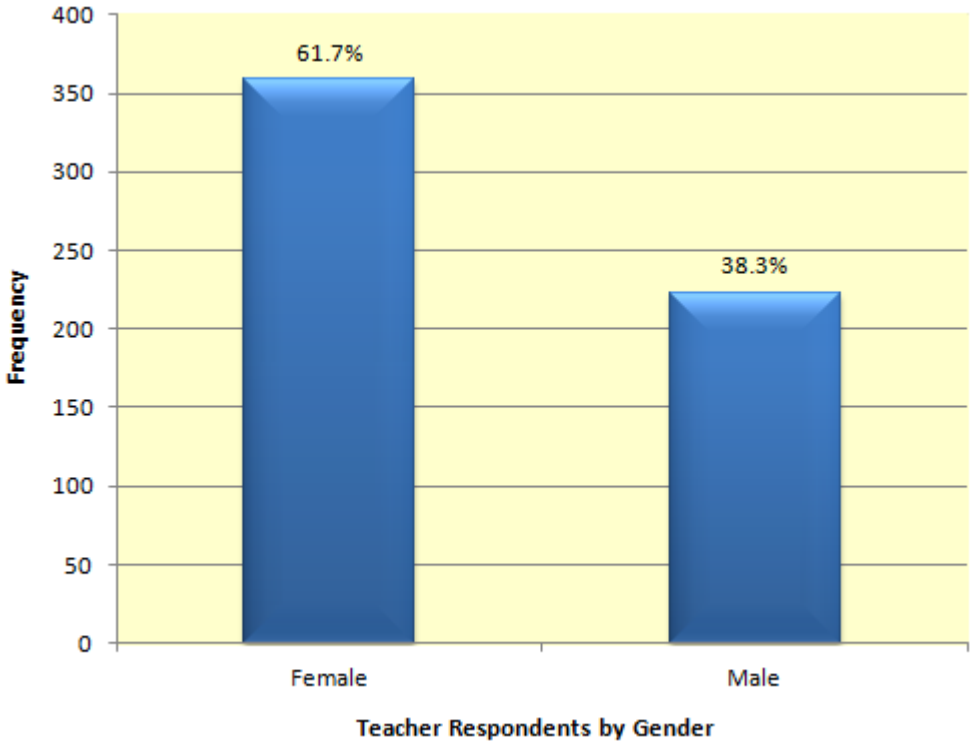


Figure 9.3. Distribution of teacher respondents by gender

The distribution of male and female teachers in the study sample can be best represented by a bar graph, which is shown in Figure 9.3.

9.2.3 Age Range of the Teacher Sample

The biological age of the teacher participants has also been considered. It was hypothesised that the age of a teacher will have an influence on his/her assessment literacy, teaching and assessment practices, and on student-level variables such as assessment perceptions, assessment attitude, academic achievement, and aptitude. Thus, it was important to examine this teacher demographic. Based on the age data collected, teachers were grouped according to the age range shown in Table 9.4. The age range starts at "under 25 years old" and tops at "60 years and above". In between are increments of 10 years. At

under 25 years old, teachers are still considered “new” or “inexperienced” as they would have just come out of university and just passed their teacher licensure examination. At 60 years and above, teachers in this category would have been teaching for at least 35 years, and could probably be thinking of retirement.

Table 9.4. Age distribution of teacher respondents

Teacher age range	Frequency	Percent
Under 25 years	42	7.2%
25-29 years	68	11.7%
30-39 years	191	32.8%
40-49 years	161	27.7%
50-59 years	101	17.4%
60 years and above	18	3.1%
Unidentified	1	0.2%
Total	582	100%

Age range increment of 10 years was used with the assumption that the span of 10 years would have given teachers enough time to progress their teaching career or “up-skill” themselves through post graduate studies, professional development programs, conferences, and seminars. This is on top of the teaching experience they have had during this period.

A pictorial representation of the distribution of the teacher respondents according to their age is shown in Figure 9.4.

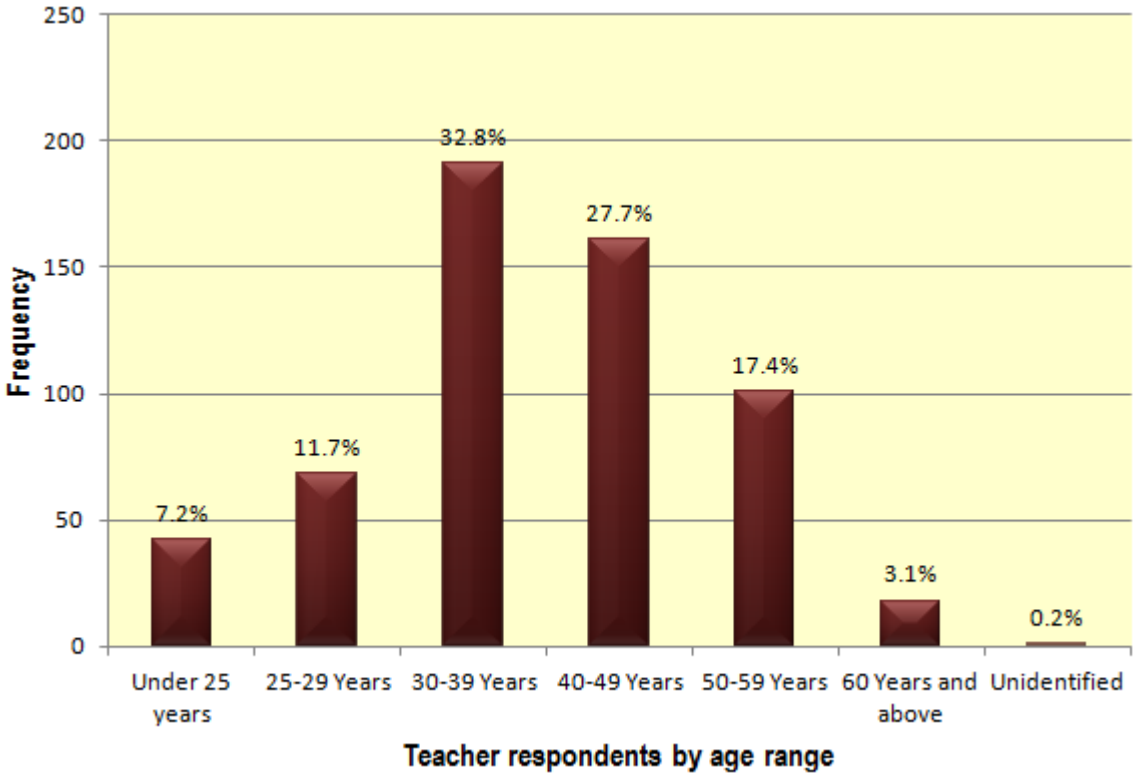


Figure 9.4. Distribution of teacher respondents by age

9.2.4 Academic Qualifications of the Teacher Sample

One of the most important factors considered in this study that could have significant influence on teachers' assessment literacy and assessment practices is their academic qualification. In the Philippines, prospective teachers will have to finish an undergraduate degree in education focusing on either elementary or secondary education, and pass a national licensure examination for teachers conducted by the Philippine Professional Regulation Commission. Teachers have the option to complete postgraduate degrees such as Masters or PhD, but due to heavy teaching loads, they often just brush this option aside. This has been the trend in the major places in the Philippines, and certainly true for Tawi-Tawi like what is shown in Table 9.5.

Table 9.5. Distribution of teacher respondents by academic qualification

Teacher academic qualification	Frequency	Percentage
Bachelor's Degree	493	84.7%
Postgraduate Degree/Units	89	15.3%
Total	582	100%

To get a clearer picture of this huge disparity between teachers with 'only' a bachelor's degree and those with postgraduate units (or those who have completed a postgraduate degree), a graphical representation is essential. This is shown in Figure 9.5.

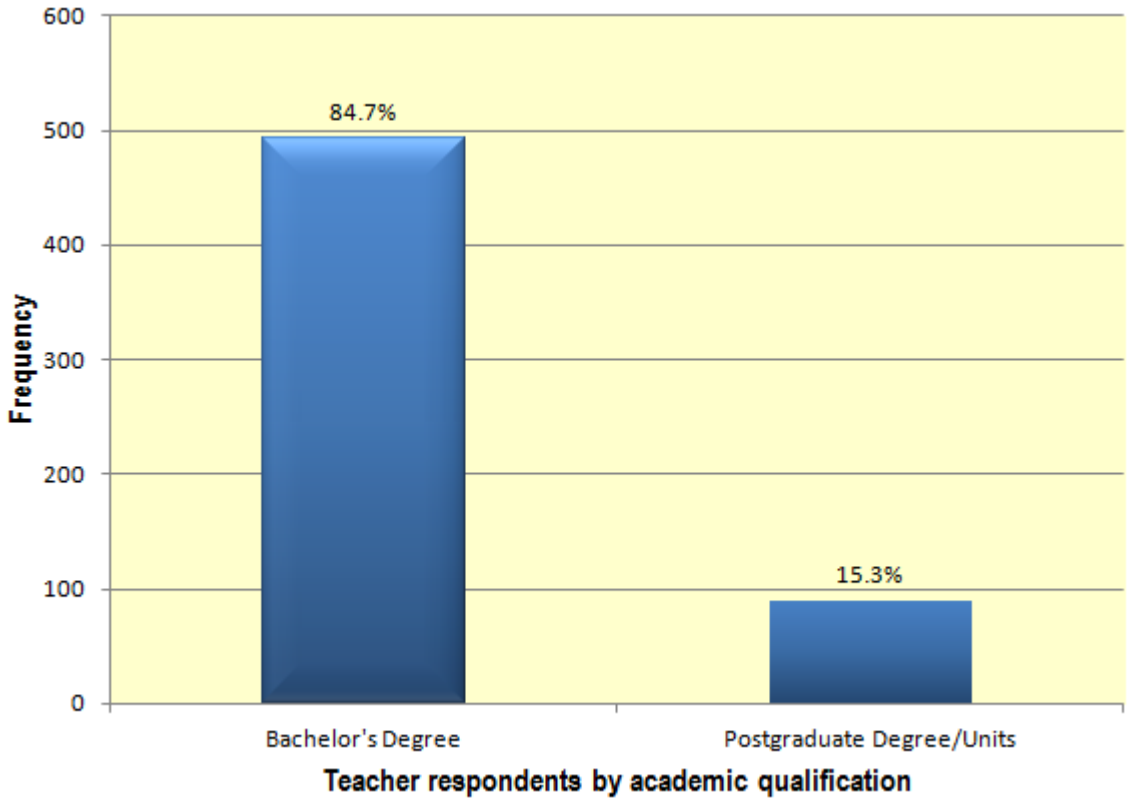


Figure 9.5. Distribution of teacher respondents by academic qualification

9.2.5 School Type

In the Philippines, both the government and private education sectors provide elementary, secondary, and tertiary education. The Philippines' Department of Education (DepEd) is the chief government agency responsible for providing elementary and secondary education, and is responsible for setting up the curricula. The private school education sector follows the DepEd-prescribed curricula, although they have the option to add or remove from it depending on which will give them the perceived 'high quality' education often to be believed by the general community. The majority of teachers who participated in this study came from public (government-owned) schools. Only very few came from private schools.

Table 9.6. Distribution of teacher respondents according to school type

School type	Frequency	Percent
Private	54	9.3%
Public	528	90.7%
Total	582	100%

The distribution of teacher respondent sample according to the school type where they teach is shown in Table 9.6. It can be observed that only less than 10% of the respondents teach in private schools. This is indicative of the fact that there are only very few private schools in Tawi-Tawi. This huge disparity can be more effectively represented graphically. This is shown in Figure 9.6.

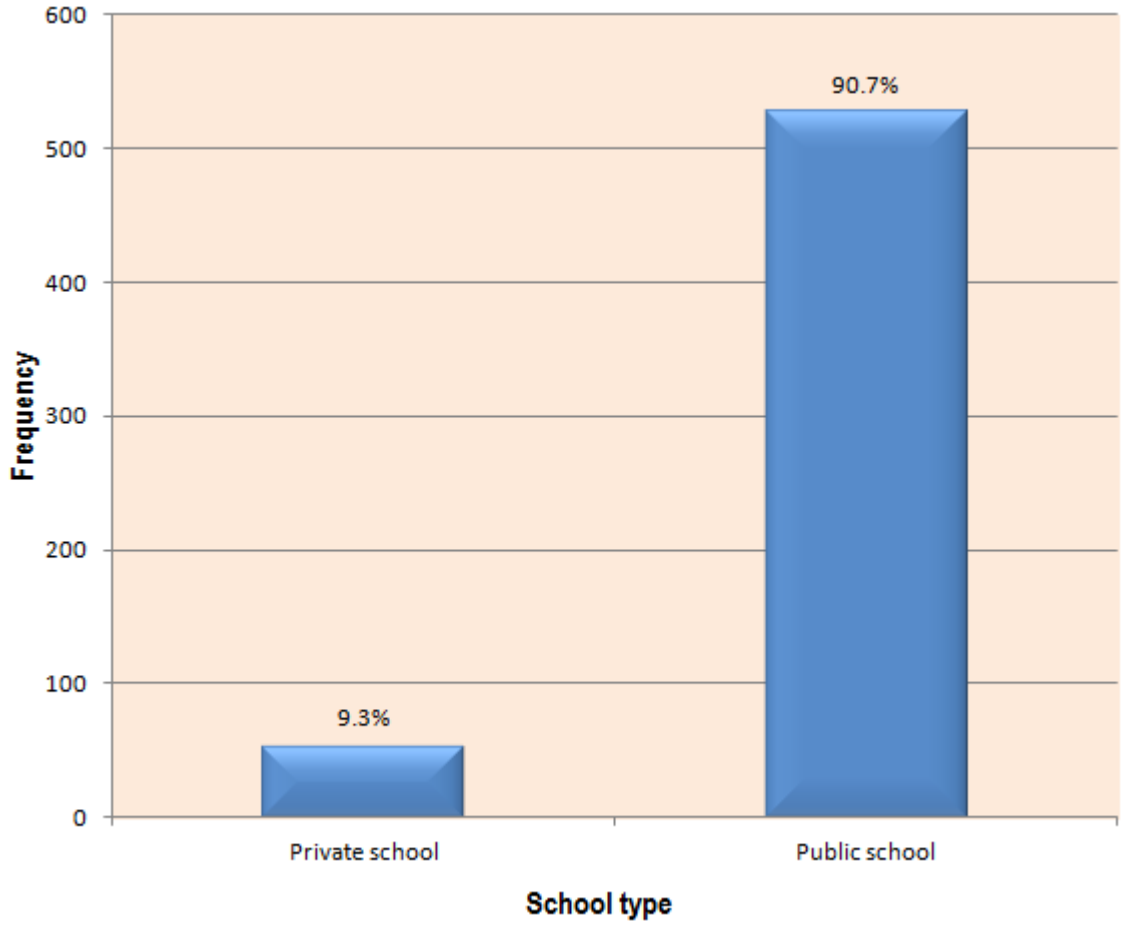


Figure 9.6. Distribution of teacher respondents according to school type

9.2.6 School Level

It was mentioned earlier in the chapter that students from Grade 6 Elementary, 2nd Year High School, and 4th Year High School levels participated in the present study. Teachers teaching in these levels were asked to participate. This was important because the impacts of the school level taught on teachers' assessment literacy, assessment and teaching practices were examined. The distribution of teacher respondents is shown in Table 9.7.

Table 9.7. Distribution of teacher respondents according to school level

Grade/Year Level	Frequency	Percent
Grade 6	321	55.2%
2nd Year HS	135	23.2%
4th Year HS	126	21.6%
Total	582	100%

It can be noted that over 50% of the teacher respondents are elementary school teachers. Respondents who are teaching in 2nd Year and 4th Year High School are distributed roughly equally at over 20% for each group.

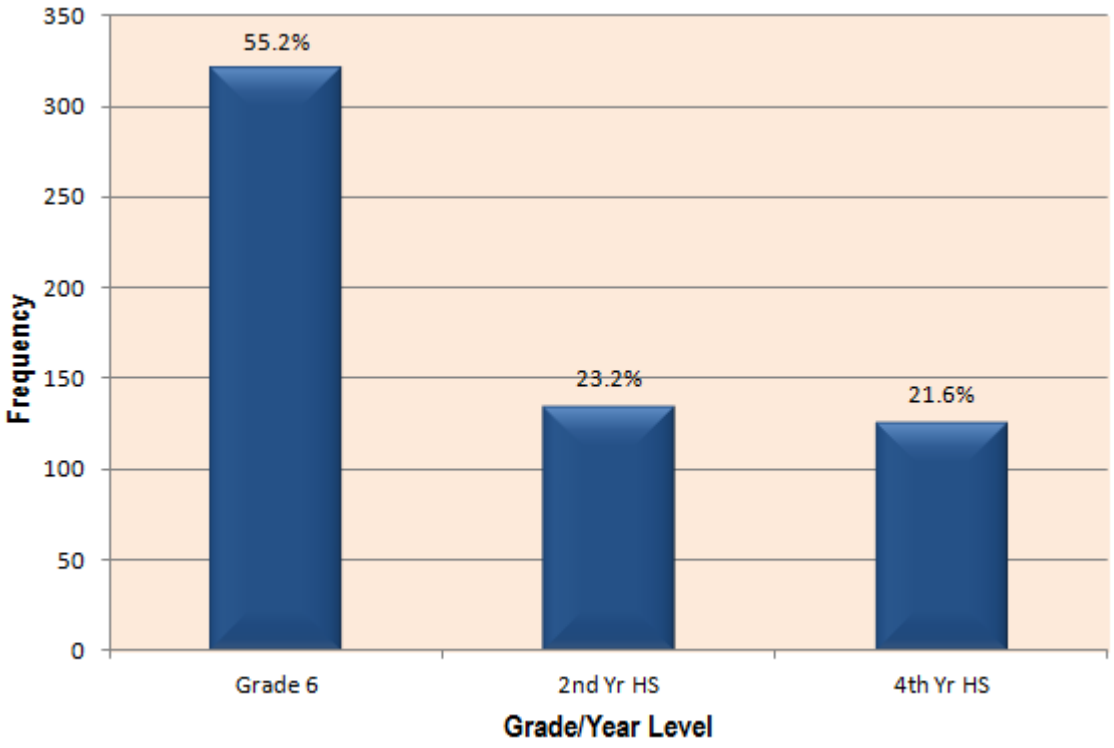


Figure 9.7. Distribution of teacher respondents by schooling level

A clear graphical representation of this distribution is shown in Figure 9.7. Perhaps this is indicative of a larger number of elementary schools compared to secondary schools in the province of Tawi-Tawi.

9.2.7 Years of Teaching Experience of the Teacher Sample

How long teachers have been teaching was examined in terms of its influence on their assessment literacy, assessment and teaching practices, and on students-level variables. The number of years of teaching experience was set at 5-year increments because the teachers' responses on this questionnaire item had a wide range. These responses are tabulated in Table 9.8.

Table 9.8. Distribution of teacher respondents according to years of teaching experience

Years of teaching experience	Frequency	Percent
1-5 Years	165	28.4%
6-10 Years	124	21.3%
11-15 Years	101	17.4%
16-20 Years	63	10.8%
21-25 Years	60	10.3%
26-30 Years	43	7.4%
More than 30 Years	26	4.5%
Total	582	100%

It can be observed that over 28% of the teacher participants are young teachers who have just finished their teaching degrees. This group combined with those who have between 6 and 10 years of teaching experience comprise around half of the total teacher respondents. Only very few teachers out of the 582 who participated have teaching experience of 30 years and over. The bar graph shown in Figure 9.8 clearly shows the distribution of teachers based on their length of teaching experience.

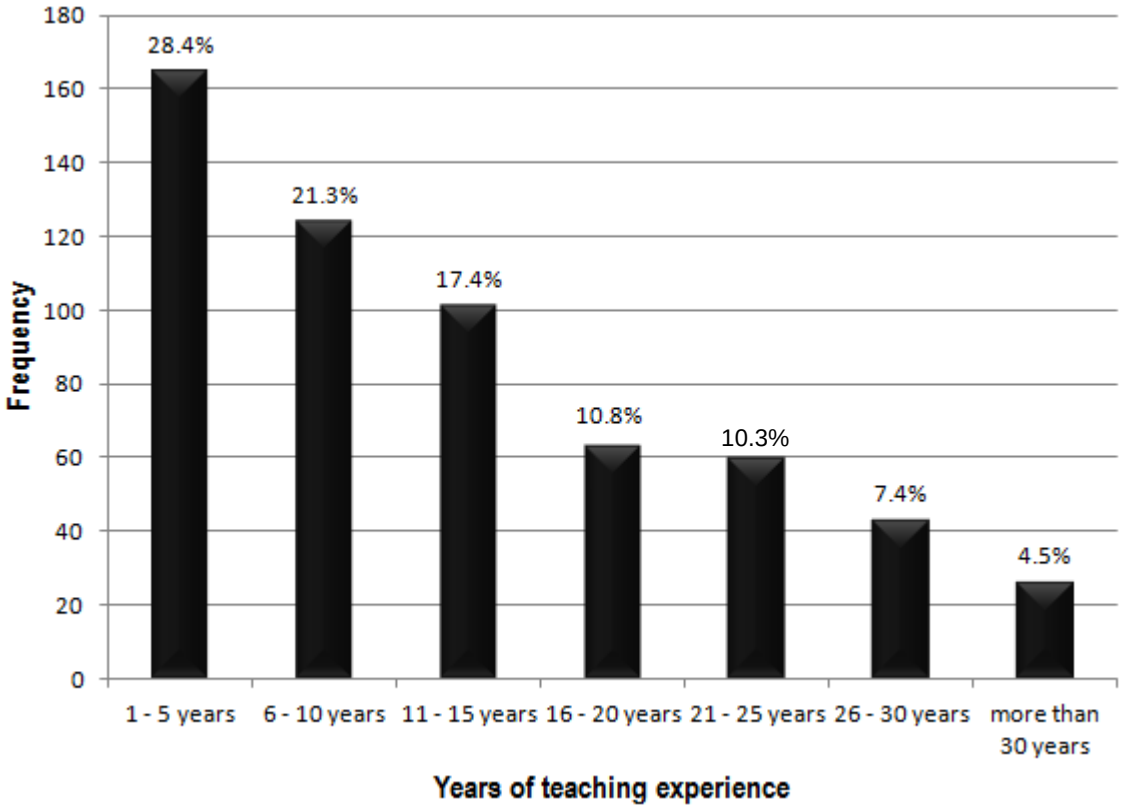


Figure 9.8. Distribution of teacher respondents according to years of teaching experience

9.3 The Data

The quantitative data used in the present study was collected using paper questionnaires. The questionnaires were distributed to teachers and students to fill out. The questionnaires for teachers are different from the student questionnaires. These questionnaires contain all the scales described and discussed in Chapters 4 to 8. Teacher interviews were also carried out to serve as the qualitative data that could support some of the findings from the analysis of the quantitative data.

Preparation of the collected data includes the entry of numerical data into a spreadsheet using Microsoft Excel, then exported to SPSS for data 'tidying' and carrying out descriptive analysis for descriptive information about the samples such as the ones presented above. Qualitative data were manually transcribed and written in text form using Microsoft Word. These became the raw dataset for the present study. The dataset constitutes nominal data from items used to extract descriptive information, and ordered

category data (in Likert form) from the different scales included in the questionnaire. Each scale is composed of items with a set of ordered response categories, which constitute the respondents' raw scores, which, according to Wright and Linacre (1989), are considered counts of observed events. These scores, however, cannot be used in the analysis, as they are not yet considered measures (Wright & Linacre, 1989) because they do not have a standard starting point. This implies that raw scores could have no starting point and have units of more than one kind. Wright and Stone (1999) describe measure as a count of "standard" units from a "standard" starting point to anchor a scale. This argument was used in this study to transform scores into measures before analysis could commence.

9.3.1 The Scaling Process

A number of ability estimation methods can be employed to transform scores into measures. These include 'Maximum Likelihood Estimation' (or MLE) by Lord (1980), 'Bayes Modal Estimation' (or BME) by Mislevy (1986), 'Marginal Maximum Likelihood Estimation' by Bock and Aitkin (1981), the 'Expected A-Posteriori' (or EAP) by Bock (1983), and the 'Weighted Likelihood Estimation' (or WLE) by Warm (1989). The WLE was employed in this study due to its attribute that minimizes estimation bias, and also to be consistent with what has been used in large-scale studies such as the Programme for International Student Assessment (PISA). WLE was carried out using ConQuest 2.0 computer program.

WLE values were then further transformed to *W* scores (developed by Woodcock and Dahl in 1971). The WLE obtained from ConQuest can be transformed into *W* scores by using the formula

W = 9.1024(WLE Logits) + 500

Converting WLE to *W* scores has several advantages. Wright and Panchapakesan (1969) enumerate them:

1. Dealing with negative values are eliminated by the centering constant at 500.
2. The need for decimal values in many applications is eliminated by the multiplicative scaling

constant of 9.1024.

3. The signs of the item difficulty and person ability scales are set so that low values imply either low item difficulty or low person ability. Conversely, high values imply either high item difficulty or high person ability.
4. Distances along the *W* scale have probability implications that are more convenient to remember and to use than distances along the logits scale.

Transforming all the WLE values was carried out using the mathematical function within the Microsoft Excel spreadsheet program. The final dataset ready for analysis was then exported to SPSS.

9.3.2 Addressing Missing Values and Missing Data

In any large-scale survey, it is very difficult to avoid having missing responses and missing data. According to Kline (1998), missing data occurs in many areas of research. This research certainly has missing data. However, though there are some missing data, it is very minimal (less than 1%). Nevertheless, missing values in datasets can affect inferences and reporting of study results. A number of quantitative researchers including Muthén, Kaplan and Hollis (1987), and Schafer and Graham (2002) suggest some standard statistical techniques to handle data with missing values. These include 'listwise deletion' approach (also known as the complete analysis approach), 'available case methods' approach, and 'imputation', which entails filling in missing values with estimated scores. However, these methods have their own downsides when using them to address missing values and data in datasets. When using the listwise deletion approach, Darmawan (2003) pointed out that in multivariate settings where missing data occur in more than one variable, a considerable loss in sample size may occur especially when the number of variables with missing values is large. Using this method may also result to inefficiency if there is removal of large amounts of information.

Casewise and imputation methods pose disadvantages as well. Casewise methods tend to increase sample size and sample base for each variable changes depending on missing value patterns

(Darmawan, 2003). Imputation involves assigning values to missing data based on some values from other data cells, or substituting a reasonable estimate for a missing data (Little & Rubin, 1989). However, this distorts the covariance structure resulting to the estimated variance and covariance biasing towards zero (Darmawan, 2003). In addition, imputation removes data that may be unique to a particular individual respondent, and that the nonresponse bias is ignored (Patrician, 2002).

This study used the listwise deletion method since there is only a very small number of missing data. In addition, researchers such as Myers, Gamst and Guarino (2006), and Allison (2002) support this method because of its usability in handling a multitude of multivariate techniques including multiple regressions and structural equation modeling.

9.3.3 Level of Analysis

The data collected for this study intended to be used to answer the research questions advanced in Chapter 1 and in this chapter is nested on two levels – teacher level and student level. Teacher level factors include assessment literacy, assessment practices, teaching practices, and demographic variables such as gender, age range, academic qualification, years of teaching experience, and school type. Student level factors include perceptions of assessment, attitude towards assessment, academic achievement, aptitude, and gender.

To describe the variables and to obtain some comparisons among the factors tested in this study, descriptive and inferential analyses were carried out. Demographic factors were described in terms of frequency and percentage as presented in the early part of this chapter. Moreover, teacher-level and student-level variables were described using mean scores. For comparison involving two independent groups, t-test of independent samples was performed. For those involving at least three groups with independent and dependent variables, one-way ANOVA was used. The comparison was in terms of the significant differences between the means of the compared groups.

To obtain a general picture of how variables at each of the teacher and student levels interact with each other, a single level path analysis was carried out. Two separate analyses corresponding to the two levels were done. Independent analyses for the two levels were performed as they were distinctly of nested structure. In other words, factors at the two levels were not combined due to the hierarchical nature of the data and challenges in using path analysis to analyse multilevel data. As experts have stressed, combining data from different levels is problematic. Aggregation of data, according to Snijders and Bosker (1999) could potentially produce the following errors: shift of meaning, ecological fallacy, neglect of the original structure, and prevention from examining possible cross level interaction effects. Likewise, disaggregation of data could produce some distorting effects known as disaggregation bias. Snijders and Bosker (1999, p. 15) describe that disaggregation of data can result to

> … „the miraculous multiplication of the number of units"…disaggregation and treating the data as if they are independent implies that the sample size is dramatically exaggerated. For the study of between-group differences, disaggregation often leads to serious risks of committing type I errors.

In other words, both aggregation and disaggregation of data can produce bias and erroneous estimates, which could result to bigger measurement error (Darmawan, 2003).

Therefore, it was necessary to take into account the hierarchical nature of the collected data to minimise the errors caused by using a single level path analysis that includes drawing wrong conclusions. Multilevel analysis techniques take into consideration the nested nature of the collected data. Hierarchical linear modeling (HLM) was employed in this study to carry out multilevel analysis. Details of HLM are provided in Chapter 11.

9.4 Descriptive Analysis Results

9.4.1 Mean Score Distribution: 'Assessment Literacy'

The levels of assessment literacy of the elementary and secondary teachers who participated in this

study are presented in Table 9.9. These levels are indicated by the mean W-scores that were derived using the W-score formula described in the previous section. The scores take 500 as the mean or average level. Using this as a guide, it can be spotted that the elementary school teachers' general assessment literacy level was below average with a W-score of 491.66; in terms of the specific standards, their assessment literacy levels were all below average as indicated by W-scores of 495.23, 491.20, 491.84, 492.05; 491.95, 492.04, and 491.79 for Standards 1-7, respectively; Of these standards, they performed highest on Standard 1 (Choosing assessment methods appropriate for instructional decisions) with a mean W-score of 495.23 and lowest on Standard 2 (Developing assessment methods appropriate for instructional decisions) with a mean W-score of 491.20. Similar results appeared for the secondary school teachers who likewise obtained low assessment literacy on the whole (W-score = 492.88) and below average in all the tested standards (Mean W-scores of 494.73, 491.49, 492.57, 494.37, 492.69, 493.93, and 493.86 for Standards 1-7, respectively). Of the seven standards examined, the high school teachers performed highest on Standard 1 (W-score = 494.73) and lowest on Standard 2 (W-score = 491.49). These results provide empirical evidence that the sampled teachers in the province of Tawi-Tawi did not possess adequate literacy in the area of student assessment, as illustrated through the ALI and as measured in terms of the assessment standards adopted in this study. Moreover, it appeared that while Tawi-Tawi teachers, to a certain extent, possessed knowledge in selecting assessment methods as illustrated by their highest performance in Standard 1, they were nonetheless least ready in developing them as indicated by their lowest performance in Standard 2. This possibly suggests that some teachers were using assessment methods and tools that were readily available from other sources such as commercially produced textbooks and possibly from curriculum documents made available to them. Perhaps, some of the concerned teachers were having the assumption that commercially produced assessment tools including tests are valid and reliable. These findings and explanation are supported by the interview results.

Interviews were conducted to gather qualitative data that support the interpretation of the quantitative results. Thirty-four (34) teacher respondents, who were drawn from Grade 6, Second Year, and

Fourth Year high school levels and from public and private schools, were selected to participate. These teachers were asked on the assessment tools and the qualities of the assessment forms they employed. The interview questions were in relation to the assessment literacy, particularly with Standards 1 and 2, and the assessment practices. When asked on the qualities of the assessment tools they used, most of the interviewed teachers (33 or 97%) responded that they choose assessment methods and tools that are valid and reliable. However, when asked about their views on valid and reliable assessment forms, some teachers provided responses that were not in accordance with the concepts of validity and reliability. Moreover, some of them lack the understanding about methods of establishing the two basic qualities of any measuring instrument. Some of their responses are provided below:

Researcher: Do you choose assessment forms/methods that are valid and reliable? What is your view of a valid assessment tool? reliable assessment tool?

Teacher 4: Yes. So, valid and reliable, it depends on, on the scores of our students. I think I can say it is valid whenever half of the students passed and then invalid whenever my students failed, many failed. Reliability is the same with this validity.

Teacher 17: Yes. I think the ah, ah, in my part, I think it is valid if the student really answer ah, the questions and at the same time, if you are going to check that as a teacher if you are going to check that because and if you are going to shall I say if you are going to strict, strict way of facilitating them in taking the test because in some way, if we are, if we are just giving a quiz or a test or then if you are not going to facilitate them well all, all they have to do is they can cheat so, I think that is not valid if they are going to cheat so, if we are going to facilitate them well, then you have to check that immediately then that's the time you can make that you are assessing the student valid. My test is reliable if the students really understand my lesson I think, that is one, one thing that I can say.

Teacher 19: Yes. Validity ah…means if you measure what intend to measure, I think that is validity. Now, the reliability is that ah, if the result of the exam like for example you got the result of the exam and you want to try if it is reliable then you make a re-test, then if that test will have the same result I think that is reliable.

Teacher 26: Yes. I can ah, I can say my test is valid or reliable, if most of my pupils pass the exam.

Teacher 27: Yes. They are valid because my learners was able to answer the given test. I don't have any idea about reliability.

Teacher 28: Yes. Ah yes, it is valid and reliable because ah, most of the students obtain ah, high score whenever I gave, whenever we having a quiz or other assessment.

Teacher 30: Yes. Sometimes, if you are very strict, in giving this assessment this will be a reliable but we if not…if you are not very strict and we are not very strict to the pupils this will not, this will not a reliable. Yes, when the pupil got on that particular assessment, 70 percent above that mean, that means that the assessment is valid and reliable. When they get low meaning ah…this not reliable.

From the interview responses cited above, it appeared that some teacher respondents associate validity and reliability with the test score or with passing the test, and from their own operational definition.

Researcher: How do you establish validity and reliability?

Teacher 2: Ok so, in, to validate my test questions usually I, in my subject I usually formulate my test questions in terms of the levels of learning. The simple recall, the simple knowledge, comprehension, analysis, synthesis, evaluation, and so on and so forth, so I prepare my test because I know this type of test is to me it is very valid because it determines how much, ok, I have ah put, ah and output as the students their output also, there are also students who get lower ah lower grade or lower score but there are also excellent students. So I think most of the tests I if I given the test probably I know it is valid.

Teacher 8: I have no knowledge about this…

Teacher 9: ah in the idea of valid and reliability I have heard that but ah seems I forgot already but what I am using to let me consider or determine that an assessment is valid or reliable when ah certain student were able to answer the question like essay type in writing he was able to answer it or the assessment were conducted proper in the class without any cheating ah and then after that he will ask also orally to discuss the students orally by proving it to determine whether is valid or invalid so that is the point that I can consider the assessment is valid and reliable.

Teacher 14: The reliability, yes, I can check it using textbooks, I usually compare textbooks that's why we have a lot of textbooks. The reliability of the test depends on it. Validity I use, I use mean, mode and the medians in the basic and even the t-test.

Teacher 23: I…I have applied, but then it's not that more serious usually because ah…usually when I give test ah…I…I…I get it base from the book.

Teacher 27: No idea at all on this…

Teacher 30: That…on how ah, yes! by means of students who can get a passing score and by applying the statistical method on grade computation and the transmutation table.

From the interview responses under the second question, it appeared that some teachers either lacked the understanding of the methods of establishing validity and reliability, or were not concerned with these methods/concepts as their tests were taken from books/textbooks.

The general finding that teacher respondents were relatively low in their assessment literacy is consistent with what were revealed in the previous studies such as those conducted by Plake, Impara, and Fager (1993), and by Mertler (2003). In these studies, American teachers were found to exhibit low literacy in the area of student assessment, as revealed by their obtained scores. However, in terms of the highest and lowest performances on the specific standards, the results of this study appeared to be different from the earlier studies. For instance, in the study of Plake, Impara, and Fager (1993), which involved in-service

teachers, the respondents were strongest in Standard 3 (Administering, scoring, and interpreting the results of both externally-produced and teacher-produced assessment methods) and weakest in Standard 6 (Communicating assessment results to students, parents, other lay audiences, and other educators). Mertler's (2003) study also revealed that in-service teachers performed highest on Standard 3 but weakest on Standard 5 (Developing valid pupil grading procedures). Furthermore, the results of the study conducted in the Philippines by Balagtas, et al. (2010), which involved graduate students who have experienced teaching for a number of years, disclosed that the respondents were highest in Standard 2 and weakest in Standard 6. The differences in the strengths and weaknesses of these teachers on the tested standards can perhaps be attributed to specific teachers' background and/or context. Graduate student respondents involved in the study of Balagtas, et al. (2010) could have come from urban areas where the environment and exposure are very different from Tawi-Tawi's rural context.

Table 9.9. Levels of assessment literacy of elementary and secondary school teachers (Distribution of mean W-scores on assessment literacy by school level and standards tested)

ASLIT/Assessment Standards	Teacher Participants				Overall Assessment Literacy (ASLIT)	
	Elementary		Secondary		Elementary & Secondary	
	W-score	S.D.	W-score	S.D.	W-score	S.D.
Standard 1 (STAN1)	495.23	9.69	494.73	11.51	495.00	10.54
Standard 2 (STAN2)	491.20	8.72	491.49	10.34	491.33	9.48
Standard 3 (STAN3)	491.84	10.17	492.56	9.76	492.17	9.99
Standard 4 (STAN4)	492.05	9.22	494.37	8.70	493.09	9.06
Standard 5 (STAN5)	491.95	9.48	492.69	8.97	492.28	9.25
Standard 6 (STAN6)	492.04	9.52	493.93	9.51	492.89	9.55
Standard 7 (STAN7)	491.79	10.36	493.86	10.23	492.72	10.34
Overall Assessment Literacy (ASLIT)	491.66	5.89	492.88	6.31	492.21	6.11

Note: W-score has an assigned mean of 500; S.D. = Standard deviation; $N_{elementary}$ = 321; $N_{secondary}$ = 261; N_{Total} = 582

9.4.2 Mean Score Distribution: 'Assessment Practices'

Table 9.10 presents the levels of assessment practices of elementary and secondary school teachers. Similar to the description of assessment literacy, assessment practices are likewise indicated by

mean W-scores with an assigned mean of 500. However, the W-scores for assessment practices represent the frequency of their assessment practice. This means that a W-score of 500 indicates occasional practice, a W-score higher than 500 indicates frequent practice or constant practice while a score of below 500 implies rare or no practice at all. Examining the values in Table 9.10, the elementary and secondary school teachers appeared to consider assessment purpose, employ appropriate assessment methods, and communicate assessment results frequently. Specific results showed that elementary school teachers frequently practiced assessment with respect to the three constructs namely, purpose, design, and communication. Of these constructs, their foremost consideration was 'purpose' when employing assessment. This means that in using assessment and in doing assessment-related activities, their main consideration was the purpose of assessment (e.g. to determine the pace of instruction and to improve student learning). In addition, they indicated that they also considered frequently the assessment design and communication when undertaking assessment. Thus, in their assessment practices they often follow the procedure (e.g. using table of specifications to construct test, providing clear directions, and using rubrics to check their students' projects) in choosing and applying assessment methods/tools to get meaningful results. They likewise indicated that they communicate assessment results to students and parents as needed (e.g. providing feedback/comments and explaining about grades). Similar findings were drawn for the secondary school teachers. The high school teachers also reflected in their responses that they often practiced assessment by taking into consideration the purpose and the procedure in using it, and communicating its results. Assessment purpose also appeared as their primary consideration, which implies that 'purpose' is their main criterion when conducting assessment activities in the classroom. Their next consideration was assessment communication indicating that they likewise communicated assessment results to students and parents. They reported that they often followed appropriate procedure and used proper methods/tools when undertaking assessment. Considering these results, the elementary and high school teachers in the province of Tawi-Tawi, Philippines generally appeared to practice assessment by giving attention to its purpose, design, and results.

Table 9.10. Levels of assessment practices of elementary and secondary school teachers (Distribution of mean W-scores on assessment practices by school level and sub-factors tested)

ASPRAC/Sub-factors Tested	Teacher Participants				Overall Assessment Practices (ASPRAC)	
	Elementary		Secondary		Elementary & Secondary	
	W-score	S.D.	W-score	S.D.	W-score	S.D.
Assessment Purpose (PUR)	521.95	13.91	521.99	14.83	521.97	14.32
Assessment Design (DES)	512.29	10.67	513.45	9.93	512.81	10.35
Assessment Communication (COM)	512.27	18.89	514.36	18.99	513.21	18.95
Overall Assessment Practices (ASPRAC)	513.60	10.25	514.33	10.47	513.93	10.35

Note: W-score has an assigned mean of 500; S.D. = Standard deviation; $N_{elementary}$ = 321; $N_{secondary}$ = 261; N_{Total} = 582

Questions asked in the interview also attempted to elicit responses on assessment forms that teachers employed in the class. Teachers were asked on their most used assessment form and on the second most frequently employed tool in their respective classes. On the assessment form that they used most of the time, multiple choice appeared to be the most commonly used type as indicated by 22 (65%) of the 34 teachers interviewed. The second most frequently used types were completion or filling the blank, as indicated by 3 (9%) of the respondents and essay/rubrics as indicated also by 3 (9%) participants. Some of their responses are provided below:

Researcher: I understand that you use assessment types or strategies in your class to ascertain and improve your student learning. What is the assessment type that is used most of the time in your class? And if you are to rank them, what is the second frequently employed assessment form?

Teacher 2: Ok. Usually, in my…in my teaching during the previous years until now, I usually gave a selection type or multiple choice then I give ah… also essay that is compare and contrast between two terms in which students will be able

to determine two terms for example ah…ah differentiate between these words and the other words.

Teacher 9: almost ah every periodical grading period I use ah multiple choice and also the essay type.

Teacher 15: Usually, I gave activity related to high order thinking or they call it HOTS, like journal and also essay.

Teacher 20: Ah…in my class…in my class, in own experience in my class, I use the traditional assessment or strategies, usually want students to choose the response from…from the multiple choice. Yes, multiple choice first, true or false test, or matching type after.

Teacher 28: I used fill in the blanks type of assessment and then true or false…

The interview results above confirm the findings of the studies undertaken by Fleming and Chambers (1993 as cited in McMillan & Workman, 1998), Cross and Weber (1993 as cited in McMillan & Workman, 1998), McMillan, Myran, and Workman (2002), and Stiggins and Bridgeford (1985) that teachers practiced objective types of assessment such as those mentioned in the teachers' interview responses.

9.4.3 Mean Score Distribution: 'Teaching Practices'

The mean score distribution representing teachers' responses on their teaching practices is provided in Table 9.11. The interpretation of the results concerning teaching practices is similar with those in assessment practices. The only difference is that responses on teaching practices represent teachers' frequency of use with respect to the lessons. This means that a mean W-score of 500 indicates that teachers use a particular kind of activities in half of their lessons. For a W-score above 500, it implies the use of activities in three-quarters of the lessons or in all lessons while a W-score below 500 indicates the use of activities in about one-quarter of the lesson or no use at all. Using this as a guide, the results in Table

9.11 suggest that elementary school teachers used both direct transmission method and alternative approach in more than half of their lessons, as indicated by their *W*-scores in the relevant sub-factors. Their dominant teaching practices were on 'structuring activities' as shown by the corresponding mean W-score, although "student-oriented activities" and "enhanced activities" were also practiced in more than half of their lessons. Of the three sub-factors, the "enhanced activities" were the least used as the teachers obtained the lowest mean *W*-score on this sub-variable. A similar pattern was observed among high school teachers. They likewise practiced a mix of direct transmission and alternative methods in more than half of their lessons. Of the sub-factors, "structuring activities" were their most dominant and "enhanced activities" were their least practiced activities. These results pointed out that, although both tested methods were used in more than half of the lessons, the elementary and high school teachers in the province of Tawi-Tawi were more inclined to use the direct transmission method as indicated by their mean *W*-score in "structuring activities" than the alternative approach as represented by their mean *W*-scores in "student-oriented and enhanced activities". In other words, instructional activities were mostly prepared and structured by teachers. These results appeared to be consistent with the results of the 2008 TALIS (OECD, 2009), in which, on average, the 'structuring activities' were the most frequently employed teaching activities, followed by the 'student-oriented activities', and further tailed by the 'enhanced activities' across a number of countries that participated in the survey. In this international survey, the structuring activities appeared to be the dominant teaching activities of teachers in most countries and this finding is supported by what came out in this study.

Table 9.11. Levels of teaching practices of elementary and secondary school teachers (Distribution of mean W-scores on teaching practices by school level and sub-factors tested)

TPRAC/Sub-factors Tested	Teacher Participants				Overall Teaching Practices (TPRAC)	
	Elementary		Secondary		Elementary & Secondary	
	W-score	S.D.	W-score	S.D.	W-score	S.D.
Structured Activities (STRUCT)	509.66	8.30	508.98	6.53	509.36	7.56
Student-oriented Activities (STUDOR)	505.85	10.23	504.98	8.58	505.46	9.52
Enhanced Activities (ENACT)	501.35	9.46	502.39	7.70	501.81	8.72
Overall Teaching Practices (TPRAC)	506.16	7.75	506.07	6.47	506.12	7.20

Note: W-score has an assigned mean of 500; S.D. = Standard deviation; $N_{elementary}$ = 321; $N_{secondary}$ = 261; N_{Total} = 582

9.4.4 Mean Score Distribution: 'Student Perceptions of Assessment'

Similar steps can be taken to interpret the results on students' perceptions of assessment. As revealed in Table 9.12, Grade 6 pupils, Second Year and Fourth Year high school students had perceptions of assessment that were relatively higher than the mean score reported in SPAS. In terms of the specific constructs, all three groups of students had perceptions about test and assignment that were relatively higher than the average scores reported in SPAS. However, between these sub-factors, they reflected higher mean score towards test. This means that although the concerned students had perceptions that were more than the average in mean score towards assessment in general, they had higher self-reported perceptions of test. This result is expected as the education system in the Philippines considers test as one of the major assessment tools and as the students were more familiar with this traditional assessment mode.

Table 9.12. Levels of assessment perception of student respondents (Distribution of mean W-scores on student perception of assessment by sub-factors)

SPA/Sub-factors	Student Respondents							
	Grade 6		2nd Year		4th Year		Overall Levels (SPA)	
	W-score	S.D.	W-score	S.D.	W-score	S.D.	W-score	S.D.
Perception of Test (PTEST)	508.53	6.92	507.35	6.37	505.37	5.49	507.25	6.51
Perception of Assignment (PASS)	506.55	6.63	505.40	6.85	504.10	5.90	505.50	6.55
Overall Level (SPA)	507.70	6.36	506.51	5.94	504.83	5.04	506.51	5.99

Note: W-score has an assigned mean of 500; S.D. = Standard deviation; $N_{Grade\ 6}$ = 915; $N_{2nd\ Year}$ = 515; $N_{4th\ Year}$ = 647; N_{Total} = 2077

9.4.5 Distribution of Mean Responses on 'Student Attitude towards Assessment'

In terms of assessment attitude, similar trend was exhibited by the students. Students from the three-targeted classes were having attitude that was more than the average score towards assessment. This result is again expected for the same reason as that in assessment perceptions.

Table 9.13. Levels of attitude toward assessment of student respondents (Distribution of W-scores of attitude toward assessment of student respondents)

SATA	Student Respondents							
	Grade 6		2nd Year		4th Year		Overall Levels (SATA)	
	W-score	S.D.	W-score	S.D.	W-score	S.D.	W-score	S.D.
Student Attitude towards Assessment (SATA)	516.28	12.12	514.48	12.58	512.41	11.45	514.63	2.14

Note: W-score has an assigned mean of 500; S.D. = Standard deviation; $N_{Grade\ 6}$ = 915; $N_{2nd\ Year}$ = 515; $N_{4th\ Year}$ = 647; N_{Total} = 2077

9.4.6 Academic Achievement Data: NAT Standardised Scores

Table 9.14 shows the academic achievement (NAT scores) of Grade 6 and Second Year high school students. The scores in the table are not the mean W-scores but they are standardised scores with an assigned mean of 500 and a standard deviation of 100. The scores represent the students' general achievement taken to be the mean composite score from the core areas of math, science, English, and

Filipino. As can be spotted from the table, the Grade 6 Elementary and Second Year High School students obtained scores that were in the level of below average. This indicated that these students obtained low performances in the core areas tested in the NAT.

Table 9.14. Levels of academic achievement of Grade 6 and Second Year high school students and of aptitude of Fourth Year high school students (Distribution of W-scores on academic achievement (NAT) of Grade 6 and Second Year high school students and on aptitude (NCAE) of Fourth Year High School students)

Variables	Student Respondents					
	Standard Score	S.D.	Standard Score	S.D.	Standard Score	S.D.
Academic Achievement (ACHIV)	389.67	100.34	420.12	79.92		
Aptitude (APT)					482.95	102.68

Note: Standard score has a mean of 500 and S.D. of 100; S.D. = Standard deviation; $N_{Grade\ 6} = 915$; $N_{2nd\ Year} = 515$; $N_{4th\ Year} = 647$; $N_{Total} = 2077$

9.4.7 Aptitude Data: NCAE Standardised Scores

On the aptitude, the scores are likewise standardised scores with an assigned mean of 500 and a standard deviation of 100. The scores represent the general aptitude of students. The scores are also a composite score derived from the specific scores in mathematics, science, English and Filipino. As can be gleaned from Table 9.14, Fourth Year high school students also obtained below average score, indicating that their general aptitude was low in the core areas tested in the NCAE.

9.5 Inferential Results

9.5.1 T-test Results of Significant Differences on the Levels of Teacher Respondents' Mean Responses

The t-test results of significant differences are presented in Table 9.15. As can be gleaned, male teachers' mean score was significantly higher than those of female teachers in Standard 4 (t = - 2.076, $p<0.05$) (Using assessment results when making decisions about individual students, planning teaching,

developing curriculum, and school improvement), indicating that male teachers possessed more knowledge on this aspect of student assessment than their female counterpart; in terms of academic qualification, teachers with postgraduate qualification had significantly higher mean scores than those with bachelor degree in assessment literacy as a whole (t = - 2.254, $p<0.05$), in Standard 3 (t = -2.325, $p<0.05$), Standard 4 (t = - 2.076, $p<0.05$) , and in assessment communication (t = -2.060, $p<0.05$), implying that teachers with master's or doctoral units/degree had higher assessment literacy and tended to communicate assessment results more often that those without higher degree. The results also disclosed that high school teachers obtained significantly higher mean scores than the elementary school teachers in assessment literacy in general (t = - 2.399, $p<0.05$), and in Standard 4 (t = - 3.101, $p<0.05$), Standard 6 (t = - 2.391, $p<0.05$) , and Standard 7 (Recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information) (t = - 2.412, $p<0.05$), suggesting that secondary school teachers had higher assessment literacy as a whole and were more adept in Standards 4, 6, and 7 than the elementary school teachers. In terms of the results on school type, teachers from the private school obtained significantly higher mean scores in assessment literacy as a whole (t = 2.330, $p<0.05$) and in Standard 2 (t = 2.597, $p<0.05$) than teachers from the public school, denoting that private school teachers possessed higher assessment literacy than the public school teachers; however, public school teachers had significantly higher mean score than the private school teachers (t = - 2.270, $p<0.05$) in the use of structured activities, implying that the former group tended to be more structured in their teaching practices than their counterpart.

Table 9.15. t-Test results of significant differences on the variables tested by selected demographic factors at the teacher level

Compared Groups	N	Factor/ Variable	DF	Mean Difference	S.E. Difference	Computed t-value	P-level (p<0.05)
Female	359	STAN4	580	- 1.60	0.77	- 2.076	0.038*
Male	223						
Bachelor's Degree	493	ASLIT	580	- 1.58	0.70	- 2.254	0.025*
Postgraduate	89						
Bachelor's Degree	493	STAN3	580	- 2.66	1.15	- 2.325	0.020*
Postgraduate	89						
Bachelor'sDegree	493	STAN4	580	- 3.22	1.04	- 3.110	0.002**
Postgraduate	89						
Bachelor's Degree	493	COM	580	- 4.48	2.18	- 2.060	0.040*
Postgraduate	89						
Elementary	321	ASLIT	580	- 1.22	0.51	- 2.399	0.017*
Secondary	261						
Elementary	321	STAN4	580	-2.32	0.75	- 3.101	0.002**
Secondary	261						
Elementary	321	STAN6	580	- 1.90	0.79	- 2.391	0.017*
Secondary	261						
Elementary	321	STAN7	580	- 2.07	0.86	- 2.412	0.016*
Secondary	261						
Private	54	ASLIT	580	2.02	0.87	2.330	0.02*
Public	528						
Private	54	STAN2	580	3.50	1.35	2.597	0.010*
Public	528						
Private	54	STRUCT	580	- 2.10	0.93	- 2.270	0.026*
Public	528						

*Note: N = number of respondents; DF = degrees of freedom; S.E. = standard Error; *significant at p<0.05; **highly significant at p<0.05*

9.5.2 ANOVA Results of Significant Difference on the Levels of Teacher Respondents' Mean Responses

Tables 9.16 and 9.17 show the results of one-way ANOVA by age range. As can be seen, teachers whose age was below 25 years had significantly higher mean scores than those whose age was within 40 to 49 years in Standard 2 (F = 2.474, $p<0.05$), revealing that younger teachers tended to be more knowledgeable than the compared group in this specific standard. In other words, young teachers knew how to develop assessment methods more than those whose ages were from 40 to 49 years. Considering that teachers with more years of experience are expected to possess more knowledge as a result of their

experiences and learning while on the job, this result was not expected. However, perhaps teachers at the higher age range were inclined to use externally prepared methods or tended to use the same methods while new teachers were inclined to develop assessment methods themselves.

Table 9.16. One-way analysis of variance (ANOVA) results of significant difference on assessment literacy (Standard 2) by age range

Comparison	DF	SS	MS	Computed F-value	P-level (p<0.05)
Between groups	5	1098.64	219.73	2.474	0.031
Within groups	575	51063.58	88.81		
Total	580	52162.22			

*Note: DF=degrees of freedom; SS = sum of squares; MS = mean squares; *significant at p<0.05*

Table 9.17. Post Hoc Tests (Tukey) results of significant difference on assessment literacy (Standard 2) by age range

Comparison	Mean Difference	S.E.	P-level at p<0.05
Under 25 years vs. 40-49 years	4.72	1.63	0.045*

*Note: S.E. = standard error; *significant at p<0.05*

In terms of years of teaching experience, Tables 9.18 and 9.19 revealed the results. As can be spotted, teachers who had 1-5 years of teaching experience had significantly higher mean scores than those who had 11-15 years (F = 3.279, $p<0.05$; $p<0.01$) and 21-25 years of teaching experience (F = 3.279, $p<0.05$; $p<0.01$) in Standard 2, indicating that younger teachers possessed higher assessment literacy on this specific standard than the compared groups. Moreover, teachers who had 6-10 years of teaching experience had significantly higher mean score that those with 1-5 years of teaching experience in Standard 5 (F = 2.357, $p<0.05$), revealing that the former were more assessment literate in this standard than the latter. Teachers who had 6-10 years of teaching experience likewise obtained significantly higher mean score than those with 16-20 years of experience in Standard 7 (F = 2.343, $p<0.05$), indicating that the former group of teachers were more literate in this standard than the latter group.

Table 9.18. One-way analysis of variance (ANOVA) results of significant difference on assessment literacy (ASLIT, Standards 2, 5, and 7) by years of teaching experience

Comparison		DF	SS	MS	Computed F-value	P-level at p<0.05
ASLIT	Between groups	6	496.02	82.67	2.246	0.038*
	Within groups	575	21166.66	36.81		
	Total	581	21662.68			
STAN2	Between groups	6	1726.17	287.70	3.279	0.004**
	Within groups	575	50444.62	87.73		
	Total	581	52170.79			
STAN5	Between groups	6	1194.61	199.10	2.357	0.029*
	Within groups	575	48562.10	84.46		
	Total	581	49756.71			
STAN7	Between groups	6	1483.32	41.70	2.343	0.030*
	Within groups	575	60679.46	91.80		
	Total	581	62162.79			

*Note: DF=degrees of freedom; SS = sum of squares; MS = mean squares; *significant at p<0.05; **highly significant at p<0.05*

Table 9.19. Post Hoc Tests (Tukey) results of significant difference on assessment literacy (ASLIT, Standards 2, 5, and 7) by years of teaching experience

Variables	Comparison	Mean Difference	S.E.	P-level at p<0.05
ASLIT (No significant results after post hoc tests)				
	1-5 years vs. 11-15 years	4.26	1.18	0.006**
STAN2	1-5 years vs. 21-25 years	4.83	1.41	0.012*
STAN5	1-5 years vs. 6-10 years	- 3.40	1.09	0.032*
STAN7	6-10 years vs. 16-20 years	5.28	1.59	0.016*

*Note: S.E. = standard error; *significant at p<0.05; **highly significant at p<0.05*

Tables 9.20 and 9.21 further provided the ANOVA results concerning years of teaching experience and teaching practices. As revealed, teachers with more than 30 years of teaching experience obtained significantly higher mean score than those with 6-10 years of teaching experience in student-oriented

activities (*F* = 2.881, *p<0.05; p<0.01*), revealing that the former group of teachers were more adept in using alternative teaching practices than the latter group.

Table 9.20. One-way analysis of variance (ANOVA) results of significant difference on teaching practices (STUDOR) by years of teaching experience

Comparison	DF	SS	MS	Computed F-value	P-level at p<0.05
Between groups	6	1538.61	256.44	2.881	0.009**
Within groups	575	51173.08	89.00		
Total	581	52711.69			

*Note: DF=degrees of freedom; SS = sum of squares; MS = mean squares; **highly significant at p<0.05*

Table 9.21. Post Hoc Tests (Tukey) results of significant difference on teaching practices (STUDOR) by years of teaching experience

Comparison	Mean Difference	S.E.	P-level at p<0.05
6-10 years vs. More than 30 years	- 7.01	2.03	0.011*

*Note: S.E. = standard error; *significant at p<0.05*

9.6 Summary

This chapter highlighted descriptive information and inferential analysis results about both the teachers and students who participated in this study. The information includes student and teacher gender, age range of teachers, the academic qualification of teachers, and years of teaching experience of the teachers, school type where the sample were drawn, and the school level taught by teachers. The chapter also provided a description of the steps carried out in the data preparation and the scaling process, and the steps undertaken to transform raw scores into measures. Raw scores were transformed into measures by using the Weighted Likelihood Estimation (WLE). WLEs were further transformed into *W* scores for the advantages it offers in terms of data handling, and to be consistent with those used in large-scale studies. Listwise deletion method was employed to deal with missing values and data in the dataset. Single level path analysis and multilevel analysis techniques were also discussed. The descriptive analysis results

generally revealed that teachers had low assessment literacy. In terms of specific standards, their highest was on choosing assessment methods while their lowest was developing assessment methods. Moreover, teachers also appeared to frequently practice assessment with respect to purpose, design, and communication. Furthermore, they employed both direct transmission and alternative approaches in more than half of their lessons. However, they generally practice direct transmission more than the alternative approach.

The next chapter reports and discusses the results obtained from the statistical analysis using the single level path analysis.

Chapter 10: Path Analysis of the Teacher-level and Student-level Factors

10.1 Introduction

In Chapter 1, general research questions were put forward to examine the variables and their possible relationships. Specifically, the relationships among demographic factors, assessment literacy, assessment practices, teaching practices, student perceptions of assessment, and student attitude towards assessment, and their influence on academic achievement and aptitude were investigated. The specific research questions concerning these variables and relationships are as follows:

1. What is the level of assessment literacy of the elementary and secondary school teachers?
2. What are the assessment practices of the elementary and secondary school teachers?
3. What are the teaching practices of the elementary and secondary school teachers?
4. What are the perceptions of the elementary and secondary school students on assessment?
5. What is the attitude of the elementary and secondary school students towards assessment?
6. What is the level of academic achievement of Grade 6 and Second Year high school students?
7. What is the level of general aptitude of Fourth Year high school students?
8. Is there any significant difference on the levels of elementary and secondary school teachers' assessment literacy, assessment practices, and teaching practices in terms of gender, age range, academic qualification, years of teaching experience, school level, and school type?
9. How does teacher assessment literacy interact with assessment practices, teaching practices, student perceptions of assessment, student attitude towards assessment, academic achievement, and aptitude?

Question 9 leads to the following specific questions under the two broad headings:

9.1 Teacher-level factors

9.1.1 What is the influence of gender, age range, academic qualification, years of teaching experience, and school type on teachers' assessment literacy, assessment practices, and teaching practices?

9.1.2 What is the influence of teachers' assessment literacy on their assessment and teaching practices?

9.1.3 What is the influence of teachers' assessment practices on their teaching practices?

9.1.4 What is the influence of teacher assessment literacy on student academic achievement and aptitude through assessment practices, teaching practices, student perceptions of assessment, and student attitude towards assessment?

9.2 Student-level factors

9.2.1 What is the influence of gender on student perceptions of assessment, student attitude towards assessment, academic achievement, and aptitude?

9.2.2 What is the influence of students' perceptions of assessment on their attitude towards assessment?

9.2.3 What is the impact of Grade 6 and Second Year high school students' perceptions of assessment and attitude towards assessment on their academic achievement?

9.2.4 What is the impact of Fourth Year high school students' perceptions of assessment and attitude towards assessment on their aptitude?

As reflected in the questions above, the factors were grouped into two levels: the teacher level and the student level. At the teacher level, the factors were the teacher assessment literacy, assessment practices, teaching practices, and the demographic factors that contained gender, age range, academic qualification, years of teaching experience, and school type. The student-level factors consisted of student

perceptions of assessment, student attitude towards assessment, academic achievement, aptitude, and gender as a demographic part. To investigate the directional relationships among these factors and to answer Question 9 or specifically questions 9.1.1, 9.1.2, 9.1.3, 9.2.1, 9.2.2, 9.2.3, and 9.2.4, regression/path analysis was carried out. Separate analysis was done for each of the two levels. Factors at the teacher and student levels could not be combined in a single path analysis due to the limitations of this technique in handling the hierarchically structured data. Hence, models for the factors at the teacher level were analysed independently from those at the student level. Within each level, there were two models tested. One model involved only the main factors while the other one only included the specific sub-factors. In the analysis of each model, relationship between any pair of variables was first evaluated. After which, all factors were analysed simultaneously to obtain an overview of the relationships and interaction among factors at teacher and student levels.

This chapter reports on the processes and results of regression/path analysis that was carried out to determine the influence of factors at each of the teacher and student levels. Particularly, the chapter begins with the general descriptions of the structural equation modeling (SEM) and Linear Structural Relationships (LISREL) 8.80 to provide background on the statistical techniques and software employed in the analysis. From SEM and LISREL descriptions, the chapter proceeds by also describing the concepts and steps including the model building and testing of statistical assumptions, which this study adopted. It continues with the discussion and presentation of the analysis results. The chapter ends with a summary to emphasise the key points.

10.2 The Structural Equation Modeling (SEM)

The SEM is described as a statistical methodology that is composed of many techniques. It is "a comprehensive statistical approach to testing hypotheses about relations among observed and latent variables" (Hoyle, 1995, p.1). Other terms that are used interchangeably with SEM are 'covariance structure analysis', 'covariance structure modeling', or 'analysis of covariance structures' (Kline, 2011). Under the

SEM approach, the theoretical model or proposition posed by the researcher is expected to be quantified and validated using an empirical data (Schumacker & Lomax, 2010; Raykov & Marcoulides, 2006; Lei & Wu, 2007; Byrne, 1998; 2010). The hypothesised relationships among constructs or factors are usually examined to determine whether or not the theoretical model/proposition holds. If the data support the hypothesised relationships, the original model/proposition can be accepted and more complex models can be tested further. However, if the data do not support the researcher's theoretical model or assertion, then the model/hypothesis in question can be modified and retested or it can be rejected and new theoretical models are developed and evaluated (Schumacker & Lomax, 2010).

Authors of SEM have stressed that this multivariate statistical technique continues to be a preferred method of many researchers. Lei and Wu (2007) considered SEM's generality and flexibility as probable reasons for this preference. Other reason is due to SEM's capability to model and evaluate complex phenomena, making it the preferred method for confirming or disconfirming theoretical models in a quantitative fashion (Schumacker & Lomax, 2010). According to Marcoulides and Kyriakides (2010, p. 277), SEM has become popular as "it permits researchers to study complex multivariate relationships among observed and latent variables, whereby both direct and indirect effects can be evaluated". Byrne (1998; 2010) reinforced these reasons by stressing further that being more of a confirmatory rather than exploratory approach or by requiring relationships to be specified a priori, SEM can best be utilised for inferential purposes compared with other multivariate procedures that are descriptive in nature. Other point that this author emphasised as SEM's edge over traditional multivariate techniques is SEM's capability in explicitly providing estimates of error variance parameters.

The SEM typically has two parts or sub-models: the measurement model and the structural model. The *measurement model* defines the relationship between the unobserved or the latent factor with its corresponding observed indicators; it also provides information on the validities and reliabilities of these indicators (Diamantopoulos & Siguaw, 2000). The measurement model involves factor analytic models - the confirmatory factor analytic (CFA) model (described in Chapter 3) and the exploratory analytic (EFA) model.

However, in SEM, the measurement model is evaluated through the use of CFA (Lei & Wu, 2007; Hoyle, 1995). On the other hand, the *structural model* prescribes the relationship (association, direct effect, and indirect effect) between the latent factors and the observed variables that are not manifest indicators of the latent variables (Hoyle, 1995). Related to the structural model is the *multiple regression model* in which no latent variables are involved. The multiple regression model is described as "a structural model without latent variables and limited to a single outcome" (Hoyle, 1995, p. 3). Another related model is the *path model*. This model is an extension of multiple regression model as various multiple regression equations are simultaneously estimated (Lei & Wu, 2007). As such, it is also a structural model that "examines structure or casual models with observed variables" (Rintaningrum, Wilkinson, & Keeves, 2009, p. 46).

The SEM carries two important aspects of the procedure. First, the hypothesised causal or directional relationships under examination are represented by a series of structural or regression equations; and second, these equations can be modeled pictorially to enable clearer conecptualisation of the proposition or theory (Byrne, 1998; 2010). Related to these two aspects is the concept of communicating SEM hypothesis and results through a path diagram. A *path diagram* is a graphical representation of the SEM's hypothesis or theory that the researcher wishes to evaluate (Raykov & Marcouldies, 2006; Hoyle, 1995). The three basic components of a path diagram are rectangles, ellipses, and arrows. The description of each of these parts are given by Hoyle (1995, p. 11) as follows:

- *Rectangles* are used to indicate observed variables, which may be indicators of latent variables in the measurement model or independent or dependent variables in the structural model;
- *Ellipses* are used to indicate latent variables, independent and dependent variables as well as errors of prediction in the structural model and errors of measurement in the measurement model; and
- *Arrows* are used to indicate association and are of two sorts. Straight one-headed arrows are used to indicate directional relationship, from predictor to outcome; and curved double-headed arrows are used for non-directional association.

The tested factors or variables represented in the path diagram generally can be classified into two with respect to directional influences or association. The factors to which the straight one-headed arrows are pointing to are called 'endogenous' variables; these variables are sometimes termed as dependent or result variables. The other factors to which no straight one-headed arrows are pointing to them are labeled as 'exogenous variables'; these factors that have only one-headed arrows departing from them are analogous to independent or source variables (Lei & Wu, 2007).

A number of authors of SEM books (e.g. Bollen & Long, 1993; Hoyle, 1995; Diamantopoulos & Siguaw, 2000; Raykov & Marcoulides, 2006; Kline, 2010; Schumacker & Lomax, 2010) have provided sequential steps of SEM application (also discussed in Chapter 3 under CFA). Figure 1 shows these steps. The first of these steps is the *model specification*. In this initial step, the researcher forms or states a theoretical model based on existing theories, empirical results, and professional knowledge (Schumacker & Lomax, 2010). The researcher usually begins by stating the model in the form of diagram or a series of equations (Kline, 2011). The second step is the *identification*. This step aims to determine whether the model can be estimated using the collected data (Ben, 2010). In other words, it "concerns whether a single, unique value for each and every free parameter can be obtained from the observed data" (Hoyle, 1995, p. 4). *Estimation* is the third step. This is the process in which estimates of free parameters in the model are obtained from the gathered data (Hoyle, 1995). There are a number of estimation methods but more complicated methods are usually used depending on the distributional properties of the variables under study (Bollen & Long, 1993). The fourth step is the *model testing*. This process is done to evaluate how well the model fits with the data (Schumacker & Lomax, 2010; Hoyle, 1995). The fit is determined by checking the values of a number of analysed indices (Ben, 2010). The last step is the re-specification. This process is performed when the researcher wishes to modify the original model to obtain better fit (Schumacker & Lomax, 2010).

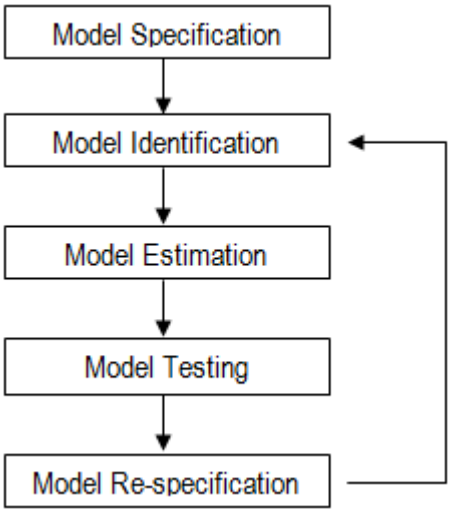


Figure 10.1. Basic steps in SEM
(Adapted from Diamantopolous & Siguaw, 2000)

In this study, the SEM's structural sub-model particularly the multiple regression/path model (single level analysis) was employed following the relevant concepts and steps to examine the relationships (direct and indirect effects) of factors at the teacher level and student level. To do this, LISREL (Version 8.80) was used. There were a number of other available SEM/path analysis software such as AMOS by Arbuckle (2007) and Mplus by Muthen and Muthen (2007). However, this study utilised LISREL 8.80 for the reasons that follow. First, LISREL is the most widely used software (Matsunaga, 2010) and a number of guides that demonstrate the use of LISREL for path analysis are readily available (Ben, 2010). Second, it is considered superior due to its robustness in standard error calculation and parameter estimation (Byrne, 1998; von Eye & Fuller, 2003, as cited in Matsunaga, 2010). Third, the development and etiological roots of SEM are strongly tied to LISREL (Brown, 2006). And fourth, LISREL 8.80 was readily available to the researcher. The next section describes LISREL and its steps as employed in this study.

10.3 The Use of LISREL 8.80 Software

The term LISREL stands for 'Linear Structural Relationships' and is a computer program that was initially used to do covariance structure analysis. One of the first widely used software, it was developed by Karl G. Joreskog and Dag Sorbom in the 1970's and subsequently updated by them a number of times

(Diamantopoulos & Siguaw, 2000; Kline, 2011). The LISREL software has become strongly associated with SEM (Scientific Software International, n.d.).

LISREL 8.80 is a software package that can handle all stages of analysis, starting from data entry and management to the evaluation of various structural equation models. It has two forms of command – the classic LISREL syntax that is based on matrix algebra and SIMPLIS (Simple LISREL), which is a LISREL programming language that uses words/statements for variables and equations (Kline, 2011). Included with this software is the PRELIS (Pre-LISREL), a program that is used to pre-process data and check data characteristics prior to their input into LISREL (Diamantopoulos & Siguaw, 2000). Du Toit, du Toit, Mels and Cheng (n. d., p. 1) describe PRELIS as a "32-bit application for manipulating data, transforming data, generating data, computing moment matrices, computing asymptotic covariance matrices, performing multiple linear, censored, logistic and probit regression analyses, performing exploratory factor analyses, etc."

LISREL 8.80 can be run in two ways: using its Windows-based point-and-click graphical user interface (GUI) or using PRELIS syntax files (Jöreskog, 2005, as cited in Ben, 2010). When using the Windows GUI, the LISREL user needs to initially import and convert a raw data file into a PRELIS system file data format (*.psf). LISREL is able to import and convert data sets from a number of formats including the common ones such as SPSS files (*.sav) and Excel files (*.xls). After converting a data file into the *.psf format, the user can form and specify a path diagram by simply dragging and dropping the variable names into a drawing panel. Once the path diagram has been created, SIMPLIS syntax and project files, which specify the structural equation model, are generated by clicking the "create SIMPLIS syntax" button. Using the generated SIMPLIS syntax file, LISREL then fits the specified model to the data corresponding to the created *.psf file. Pressing the 'run' button of the software generates a path diagram, which is a graphics file with a PTH extension. The path diagram can either display the estimates or the standardised solutions (Ben, 2010). Despite the easy use of the Windows GUI, this procedure of running LISREL 8.80 was not used for similar reasons as given by Ben (2010). GUI's specification of model parameters is rather restricted

because it only uses its default settings. Although the generated SIMPLIS syntax can be edited, this just adds complexity to the whole process of running the software. Hence, writing a PRELIS syntax file, which is a less restrictive procedure of running the application to test the models, was used in this study. Detailed steps in using PRELIS syntax file can be found in du Toit, et al. (n.d.) and detailed examples on how the procedure is carried out are available from http://www.ssicentral.com.

Using PRELIS system files to test the hypothesised models requires five general steps. First, the raw data (in either SPSS or Excel format) is converted to ASCII (or text) format. Second, a PRELIS2 (*.pr2) command file is created to read and transform the raw data. The descriptive statistics, including the skewness and kurtosis, and the desired matrix (correlation or covariance matrix) for analysis can be produced through the PRELIS2 command file. The method to treat the missing data can also be requested using this file (its default is listwise deletion, which was used in this study since only one case contained missing value after transforming the raw data into W-scores). Third, a model is specified by making a sketch of a diagram that shows the paths. Fourth is the creation of a LISREL syntax file (*.spl) that shows the relationships of variables based on the sketched diagram (the model of interest). This is where the modeling parameters are set and the desired outputs (such as path diagram) are requested. The fifth and final step is the interpretation of the LISREL output file (*.out) for model fit to the data (Ben, 2010).

Described in the following are the relevant concepts and steps, which this study utilised in carrying out regression/path analysis of the data that reflect the factors at the teacher and student levels.

10.4 Models and Representation in Quantitative Research

In quantitative research the use of models to represent phenomena and their relationships is part of the practice. This is especially true when one is interested in employing the modeling technique in the examination of variables of interest. As described in Section 10.2, a model can either be in statistical/mathematical form or graphical form. A model that is in statistical form appears as a series of structural equations that define the relations among variables. When in graphical form, it appears as a path

diagram, which can pictorially present the variables and their relations as contained in the theoretical model under investigation. These forms of the model were employed in the study.

According to Lohmöller (1989), a model building involves three steps: specification, estimation, and evaluation. These have been described earlier but for emphasis, these steps are restated as follows: a model specification requires careful selection and definition of the phenomena under examination and explication. At the initial stage of model building, it is important to be clear about the variables and the kind of association in the theoretical model. Model estimation requires translation of hypotheses into mathematical expressions that can be compared with a set of data (Neale, Heath, Hewitt, Eaves & Fulker, 1989). At this step, analysis on the degree of consistency of the theoretical model with the data can be carried out through statistical testing (Byrne, 1998; 2010). Model evaluation requires checking of the fit indices to determine if the theoretical model is consistent with the data. In this final stage, a judgment is done on whether or not the model fits the data well (i.e., if the data support the hypothesised model) (Neale et al., 1989). Again, these steps were followed in creating the models in this study.

10.5 Testing for Normality of Data and Multicollinearity

Prior to conducting regression/path analysis, it was important to test whether the observed data were normally distributed and whether extreme collineariy among variables was nonexistent. These conditions are necessary, especially when testing hypothesis such as in the case of regression analysis or general linear models (Field, 2009) and when using maximum likelihood as a method of estimation in SEM (Schumacker & Lomax, 2010; Kline, 2011). These important assumptions are briefly described below.

The assumption of normality requires that the observations or variable data need to be normally distributed. The data are normally distributed when it follows the so-called symmetrical and mesokurtic distribution, which in graphical form, appears as a bell-shaped curve. Normality of data can be checked through graphs such as normal probability plots and histogram. However, it is always important to check it through skewness and kurtosis, especially if the samples are 200 or more (Field, 2009). *Skewness* indicates

a distribution's 'asymmetry' and *kurtosis* is its 'peakedness' or 'flatness' (Asaad & Hailaya, 2001). The values for skewness and kurtosis that can be used to judge data's departure from normality are suggested by Kline (2011). These values (absolute) are <3 for skewness and <8 for kurtosis.

Another important assumption that needs to be checked is the absence of extreme collinearity or multicollinearity between two or more variables. Multicollinearity exists when there are strong inter-correlations among the predictors; it poses a problem in multiple regression at it makes difficult to determine the contribution or importance of a predictor (Stevens, 2009; Field, 2009). Other problems caused by multicollinearity include severe limitation in the size of multiple correlation and increase in the variances of regression coefficients; when these increases in the variances occur, the prediction equation will become unstable (Stevens, 2009).

There are ways to diagnose multicollinearity as suggested by experts. One of these is by using and examining the value of variance inflation factors (VIF) of each variable to be included in the regression. VIF "indicates whether a predictor has a strong linear relationship with the other predictors" (Field, 2009, p. 224). VIF values in excess of 10 indicate serious multicollinearity and that the variable is possibly redundant (Kline, 2011). Thus, it has been suggested that to reduce or address multicollinearity, the model should be respecified by removing one or more variables that are strongly correlated with other independent variables. However, O'Brien (2007) suggested taking caution in doing this process because it may do more harm than good. He further suggested that the relevant rules concerning the VIF should be interpreted in the context of other factors that impact on the stability of the estimates of the regression coefficient.

The tests for normality and multicollinearity were carried out using SPSS 16.0 (SPSS, Inc., 2007a). Both skewness and kurtosis values, plus the histogram, for each of the variables in the study were obtained. None of the variables showed skewness greater than 3 and a kurtosis greater than 8. Hence, all the data distributions for the variables were considered normally distributed to a sufficient degree for further analysis to be carried out. The test for multicollinearity was done by including all the factors. It was found out that the main factors were multicollinear with their corresponding sub-factors. This is what Kline (2011,

p. 51) described as “extreme collinearity when composite variables and constituent variables are analyzed together.” As a result, separate multicollinearity diagnosis was done for both the main factors and the sub-variables. The resulting VIF values were all below 10 indicating that either main factors or sub-factors were not multicollinear with each other.

10.6 Model Specification

Prior to doing path analysis, the theoretical model was developed. It was then necessary to present it in a graphical form through a path diagram for better conceptualisation of the hypothesis and to serve as guide in the analysis. As the LISREL's Windows-based GUI was not used in carrying out path analysis, it was necessary to draw the path diagram for the single-level analysis using other means. In this study, AMOS (Arbuckle, 2007) was used, as the path diagram and other figures can be easily and neatly drawn using its user-friendly interface.

10.7 Model Trimming

Model trimming was part of the path analysis process, and in LISREL, it involves removing the variables that do not show significant paths in the model. In determining whether or not the paths are significant, the *t*-value and the regression coefficients (also known as the *beta* value) are examined. The critical *t*-value used in this study to indicate a significant path is 1.96. Any value that is less than this was considered not significant. The variables that showed non-significant paths were removed from the model. This procedure was carried out separately for the factors at the teacher level and student level and for each of the models that involved main factors and sub-factors.

10.8 Univariate Regression Analysis

Regression analysis was used to determine the explanatory relationship between the variables examined in this study. It was employed specifically to estimate the relative explanatory or predictive power of the independent variable (X) or to identify the best predictors (X_n) of the dependent variable (Y)

(Jöreskog, & Sörbom, 1993; Stevens, 2009). Analysis through regression follows the general linear equation of the form,

$$Y = \mathcal{B}_0 + \mathcal{B}_1 X_1 + \mathcal{B}_2 X_2 + \ldots + \mathcal{B}_n X_n + \varepsilon \quad (10.1)$$

Regression represented in equation 11.1 is also called *univariate multiple* regression or simply *multiple regression* (Jöreskog, & Sörbom, 2006). Equation 10.1 contains dependent variable, also called outcome or criterion variable, as represented by Y, independent variable, also called regressor or predictor variable, as represented by *X*, the constant (intercept) as signified by $\mathcal{B}_0$, the standardised regression coefficients or the *beta* values for the independent variables as symbolised by $\mathcal{B}_1$, $\mathcal{B}_2$ and $\mathcal{B}_n$, and the residual or error term as denoted by ε. In simple regression, $\mathcal{B}_1$ represents the slope or gradient of the regression line and can sometimes be taken to be equivalent with the correlation coefficient. This gradient represents the change in the outcome (*Y*) resulting from a unit change in the predictor (*X*). If $\mathcal{B}_1$ is zero, then the expected change in the outcome would be zero – meaning the model is *bad* as the predictor does not explain or influence the dependent variable at all (Field, 2005, p. 150, as cited in Ben, 2010; Field, 2009, p. 204). Equation 10.1 comes from the general regression equation, which generally takes the form of

$$Outcome_i = (Model) + \varepsilon_i \quad (10.2)$$

This general equation simply means that the outcome variable can be predicted or explained by whatever model we fit to the observed data plus some kind of error. When equation 10.2 contains only one explanatory or predictive variable, it is referred to as *simple linear regression* or just *simple regression*, which can be mathematically represented by

$$Y_i = (\mathcal{B}_0 + \mathcal{B}_1 X_1) + \varepsilon_i \quad (10.3)$$

where Y_i is the outcome variable, $\mathcal{B}_0$ is the constant (intercept of the line), $\mathcal{B}_1$ is the standardised regression coefficient of the predictor (gradient or slope of the line), and *e* is the residual or error term.

However, if two or more predictors are fed into the equation, it becomes multiple regression in the form of equation 10.1 in which more than one independent variable are modeled to predict the dependent variable (Field, 2009).

LISREL 8.80 was used to carry out regression analysis. The resulting values reported are the standardised regression coefficients ($\mathcal{B}_1$, as explained above) to indicate the influence of predictor on outcome variable and *t*-values to indicate the significance. The *t*-statistic tests for the null hypothesis that the value of $\mathcal{B}_0$ is zero. If the *t*-value is significant (- *1.96* ≥ ***t*** ≥1.96), then this means that $\mathcal{B}_0$ is significantly different from zero and that the predictor (*X*) contributes significantly in estimating the value of the outcome (*Y*) (Field, 2009; Ben, 2010).

The variables that were subjected to regression analysis were clustered into teacher-level variables and student-level variables. As mentioned in section 10.1, the teacher-level factors included teacher assessment literacy (ASLIT), assessment practices (ASPRAC), teaching practices (TPRAC), and the demographic factors that contained gender or sex (TSEX), age range (AGE), academic qualification (ACAD), years of teaching experience (EXYR), and school type (SCHTYPE). The ASLIT contained seven sub-factors labeled standards, from STAN1 to STAN7; the ASPRAC consisted of three dimensions namely, purpose (PUR), design (DES), and communication (COM); and TPRAC also composed of three sub-constructs called structured or structuring activities (STRUCT), student-oriented activities (STUDOR), and enhanced activities (ENACT). On the demographic factors, TSEX was obviously composed of males (TMALE) and females (TFMALE); AGE was of six groups corresponding to different age ranges and thus labeled as AGE1 to AGE6; ACAD was of two categories covering Bachelor's qualification (UNDERGRAD) and master's/doctoral qualification (POSTGRAD); years of teaching experience was composed of seven categories corresponding to seven ranges of years of teaching experience and were labeled as EXYR, thus covering EXYR1 to EXYR7; and SCHTYPE contained two classifications, the public school (PUB) and the private school (PRIV). The student-level factors consisted of student perceptions of assessment (SPA),

student attitude towards assessment (SATA), academic achievement (ACHIV), student aptitude (APT), and gender (SSEX) as a demographic part. The SPA covered two sub-factors namely, perceptions of test (PTEST) and perceptions of assignment (PASS); the SATA was a one-construct variable; the ACHIEV and APT were taken as general (main) variables; and SSEX was composed of males (SMALE) and females (SFMALE). The analysis was done using the following stages/steps:

1. Variables at the teacher level were first grouped into two: the group that includes only the main factors (ASLIT, ASPRAC, and TPRAC) plus the demographic factors (TSEX, AGE, ACAD, EXYR, and SCHTYPE) (Model 1), and the group that only involves sub-factors (STAN1, STAN2, STAN3, STAN4, STAN5, STAN6, STAND7, PUR, DES, COM, STRUCT, STUDOR, and ENACT) plus the same demographic factors (Model 2);
2. A similar way of grouping variables at the student level was done. However, grouping was further divided between two groups of student participants due to different outcome variables that each intended to predict. One student group was composed of Grade 6 and Second Year high school students for whom ACHIV was the outcome variable. The other group was composed of Fourth Year high school students for whom APT was the dependent variable. For these groups of students, similar explanatory variables and models were tested. That is, for Grade 6 and Second Year students, model 1 includes SPA, SATA, and ACHIV plus the lone demographic factor (SSEX) and model 2 covers PTEST, PASS, SATA, ACHIV and SSEX; the same variables were analysed for Fourth Year high school students but APT was used instead of ACHIV;
3. At the teacher level, the group containing the main and demographic factors was separately analysed first, followed by the analysis of the group covering only the sub-factors plus the same demographic factors. This process was also applied to student-level factors between the two groups of student respondents.

The rationale for taking the steps above was that the main factors and the sub-variables could not

be combined in one analysis as they were multicollinear, and thus making the estimation of the influence of the individual variable on the predicted outcome difficult. Other obvious reason was to examine the specific explanatory relationships among sub-factors to be able to pinpoint independent variables that can actually impact on dependent variable. The results of regression analysis are presented in section 10.9.

10.9 Results of Regression Analysis

10.9.1 Teacher-level Factors (Model 1)

The possible influence of demographic factors on the main variables of the study was explored using the data from the responses of 581 elementary and secondary school teachers. This was to answer Question 9.1.1 (What is the influence of gender, age range, academic qualification, years of teaching service, and school type on teachers' assessment literacy, assessment practices, and teaching practices?), Question 9.1.2 (What is the influence of teachers' assessment literacy on their assessment and teaching practices?), and Question 9.1.3 (What is the influence of teachers' assessment practices on their teaching practices?) as mentioned earlier in this chapter. The regression results are presented in Tables 10.1 and 10.2 below.

Table 10.1. Standardised regression coefficients and t-values from regression analysis on the influence of demographic factors on the main variables of the study at the teacher level

Main Variables	Demographic Factors			
	AGE	ACAD	SCHTYPE	EXYR
ASLIT	- 0.35 (-1.36)	1.21 (4.68)*	- 0.99 (- 3.71)*	- 0.029 (- 0.12)
ASPRAC	- 1.15 (- 2.58)*	1.25 (2.76)*	0.67 (1.42)	0.72 (1.67)
TPRAC	- 0.56 (- 1.91)	- 0.023 (- 0.078)	0.50 (1.64)	0.58 (2.07)*

*Note: t-value in parenthesis; *significant at p<0.01*

The significant relationships (with asterisk) between the demographic factors and main variables of the study at the teacher level can be expressed in the form of the following equations:

$$ASLIT = \mathcal{B}_0 + 1.21(ACAD) - 0.99(SCHTYPE) + error \quad (10.4)$$

$$ASPRAC = \mathcal{B}_0 - 1.15(AGE) + 1.25(ACAD) + error \quad (10.5)$$

$$TPRAC = \mathcal{B}_0 + 0.58(EXYR) + error \quad (10.6)$$

Equation 10.4 indicates the relationships between teacher assessment literacy and the two demographic factors: the academic qualification and the school type. The academic qualification was categorised as bachelor (undergraduate) or master's/doctoral (postgraduate) and the school type was pertaining to the two main classification of schools in the Philippines: the public (government-funded) and the private (privately funded) schools. The result indicates that the level of assessment literacy of elementary and secondary school teachers who participated in this study was positively influenced by their educational attainment. This can be interpreted that the higher was the academic qualification of teachers, the higher was the possibility for them to be more literate in the area of student assessment. This result is expected as teachers with more academic qualifications are deemed more competent due to their more exposure and familiarity with assessment concepts and processes. At the postgraduate level, students are usually required to take and pass one advanced course in educational measurement and evaluation as part of the academic requirements of their postgraduate education degrees, and perhaps, this could be one of the reasons for their higher assessment literacy. On the influence of school type on assessment literacy, the result reveals negative impact. This effect connotes that to be in the private school, teachers could be more literate in assessment. From this result, it can be discerned that the kind of environment in the private school provides more avenues for teachers to be better prepared in assessing their students than that in the public school. This can perhaps be attributed, among others, to the close supervision, rigid in-service training, and strict but supportive policies, which most of the private institutions tend to practice.

On the relationship between assessment practices and demographic factors, equation 10.5 shows the significant results. As can be gleaned from the results, teacher assessment practices were negatively influenced by age range but positively impacted by their academic qualification. The reverse influence of age range on assessment practices denotes that the younger teachers were having higher mean scores in assessment practices. This result appears to be contrary to the notion that the more matured the teachers are, the more knowledge and experience they should have and the higher their scores should be in assessment practices. However, this can possibly be explained by two observations. First, most teachers, if

not all, who were in the older age ranges completed their academic degrees under the old pre-service teacher education curriculum that offered less exposure on student assessment. Under the old curriculum, only one assessment subject that focused on testing was offered (CMO No. 11, s. 1999; Balagtas, Dacanay, Dizon, & Duque, 2010), making teachers' assessment knowledge/skills limited to testing practices. Balagtas, et al. (2010, p. 3) reported that "teachers have expressed their unpreparedness to the demands of the system especially that their academic preparation was just more on the utilization of traditional assessment." Second, teachers in the younger range obviously earned their qualification under the new teacher education curriculum in which more exposures on student assessment are offered. Some teacher education institutions in the country have started offering two assessment subjects as prescribed by the Commission on Higher Education (CMO No. 30, s. 2004) and integrate assessment in some of their professional subjects, thus making young graduates more familiar with assessment. In addition, the introduction of the performance-based grading system in 2004 by DepEd (DepEd Order No. 33, s. 2004) that requires the use of rubrics and portfolio and other alternative methods provided additional opportunity for young teachers to gain knowledge about these new methods. Consequently, their assessment knowledge is up to date thereby making them more aware about the appropriate practices of assessment. As for the academic qualification, the positive relationship implies that the possession of higher educational qualification influences better assessment practices. This result follows the common expectation that better qualification should lead to better professional practices, such as those related to student assessment.

The association between the teaching practices and demographic factors is represented in equation 10.6. As denoted in this equation, only years of teaching experience had a significant contribution to teaching practices. The positive influence of years of teaching experience on teachers' instructional practices highlights that longer teaching service tended to make teachers more competent in their teaching practices. Again, this is expected, as teachers tend to learn and improve while in the course of doing their professional job.

Table 10.2 presents the regression results on the relationships among the main variables examined at the teacher level. The significant relationship is represented by equation 10.7. From the equation, teaching practices appear to be negatively influenced by assessment literacy although the standardised regression coefficient is low. This can be interpreted that the more literate the teachers were, the more they exhibited poor teaching practices. This result is quite contrary to the theory or assertion in the literature that teachers who are knowledgeable in assessment are in a position to integrate assessment with teaching enabling them to utilise appropriate teaching methods (McMillan, 2000) and thus resulting to improved practices. It was assumed that teachers employed appropriate teaching practices based on the information provided by some assessment activities or results on which relevant assessment knowledge is needed. There are a number of probable explanations for this kind of outcome. One possible explanation is that, perhaps, teachers just practiced the scenarios depicted in the survey items without using assessment to base their decisions to employ those practices. In addition, it can be noted that items under the teaching practices questionnaire were in the form of Likert-type scale in which teachers were requested to indicate the frequency of their teaching practice on the specific questions presented to them. As such, it could be that teachers just reported frequent practice on most of the situations depicted in the items without actually doing them in the class. Also, teachers' responses on teaching practices could be in the form of their perceptions or beliefs, which cannot always be expected to agree with their knowledge on assessment. These finding and explanations somehow agree with the view of Mullens and Kasprzyk (1999) who stated that the reliability and validity of the data resulting from the self-rated or self-reported responses can sometimes be questioned. The positive relationship between assessment practices and teaching practices appear to confirm the view that better assessment practices should lead to better teaching practices. This is consistent with Popham's (2009) assertion that better classroom assessment activities will impact on classroom's day-to-day instructional activities.

Table 10.2. Standardised regression coefficients and t-values from regression analysis on the relationships among the main factors at the teacher level

Main Factors	ASPRAC	TPRAC
ASLIT	- 0.041 (- 0.57)	- 0.095 (- 2.04)*
ASPRAC		0.25 (9.07)*

*Note: t-value in parenthesis; *significant at p<0.01*

$$TPRAC = \mathcal{B}_0 - 0.095(ASLIT) + 0.25(ASPRAC) + error \quad (10.7)$$

10.9.2 Teacher-level Factors (Model 2)

The sub-factors at the teacher level were separately analysed to obtain the overview of the more specific relationships and to address Questions 9.1.1, 9.1.2, and 9.1.3. The regression results of the analysis are presented in Tables 10.3, 10.4, 10.5, and 10.6

Table 10.3. Standardised regression coefficients and t-values from regression analysis on the relationships among sub-factors of teacher assessment literacy

Sub-Factors	Demographic Factors				
	TSEX	AGE	ACAD	SCHTYPE	EXYR
STAN1	0.86(1.95)	- 0.45(- 0.99)	0.42(0.92)	0.53(1.11)	- 0.28(- 0.62)
STAN2	0.46(1.18)	- 0.41(- 1.00)	0.73(1.80)	- 1.50(- 3.55)*	- 0.10(- 0.26)
STAN3	0.29(0.70)	- 0.82(- 1.90)	1.71(3.98)*	- 0.31(- 0.69)	0.26(0.64)
STAN4	0.90(2.44)*	- 0.18(- 0.48)	2.17(5.69)*	- 1.08(- 2.72)	- 0.24(- 0.64)
STAN5	0.30(0.78)	0.33(0.82)	0.33(0.82)	- 1.12(- 2.67)*	- 0.31(- 0.79)
STAN6	- 0.25(- 0.63)	- 0.14(- 0.33)	1.42(3.44)*	- 1.17(- 2.71)*	0.031(0.077)
STAN7	0.081(0.19)	- 0.40(- 0.90)	0.73(1.63)	- 0.66(- 1.42)	- 0.28(- 0.65)

*Note: t-value in parenthesis; *significant at p<0.01*

$$STAN2 = \mathcal{B}_0 - 1.50(SCHTYPE) + error \quad (10.8)$$

$$STAN3 = \mathcal{B}_0 + 1.71(ACAD) + error \quad (10.9)$$

$$STAN4 = \mathcal{B}_0 + 0.90(TSEX) + 2.17(ACAD) + error \quad (10.10)$$

$$STAN5 = \mathcal{B}_0 - 1.12(SCHTYPE) + error \quad (10.11)$$

$$STAN6 = \mathcal{B}_0 + 1.42(ACAD) - 1.17(SCHTYPE) + error \quad (10.12)$$

Equations 10.8 to 10.12 indicate the significant relationships between five assessment standards and demographic factors. STAN2 (Developing assessment methods appropriate for instructional decisions), STAN5 (Developing valid pupil grading procedures), and STAN6 (Communicating assessment results to students, parents, other lay audiences, and other educators) have an adverse relationship with SCHTYPE. These results indicate that being in the private school influences teachers' assessment literacy on the three assessment standards than when in the public school. As mentioned earlier, perhaps this result is spurred by the kind of training and policies that the private schools implement for their teachers. Moreover, STAN3 (Administering, scoring, and interpreting the results of both externally-produced and teacher-produced assessment methods), STAN4 (Using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement), and STAN5 appear to be positively associated with teachers' academic qualification. As the teachers gained master's/doctoral units/degrees, they tended to be more knowledgeable on STAN3, STAN4, and STAN5. Topics in these three standards are common and for some teachers the topics are part of their assessment routine. Besides, these topics are usually reviewed/re-taught and made part of the activities/reading at the postgraduate level. Thus, teachers can be expected to gain better assessment literacy as they pursue higher academic degrees in their field. The results likewise reveal that teachers' sex affects their literacy on STAN4. Specifically, males appeared to have the inclination to be more literate in using assessment results.

Table 10.4. Standardised regression coefficients and t-values from regression analysis on the relationships among sub-factors of assessment practices

Sub-Factors	Demographic Factors				
	TSEX	AGE	ACAD	SCHTYPE	EXYR
PUR	- 0.84(- 1.43)	- 1.43(- 2.33)*	1.62(2.63)*	1.58(2.45)*	0.78(1.31)
DES	0.028(0.065)	- 1.22(- 2.73)*	0.68(1.51)	0.30(0.64)	0.94(2.20)*
COM	- 1.41(-1.82)	- 0.44 (-0.54)	2.84(3.47)*	- 0.31(- 0.37)	0.39(0.50)

*Note: t-value in parenthesis; *significant at p<0.01*

$$PUR = \mathcal{B}_0 - 1.43(AGE) + 1.62(ACAD) + 1.58(SCHTYPE) + error \quad (10.13)$$

$$DES = \mathcal{B}_0 - 1.22(AGE) + 0.94(EXYR) + error \quad (10.14)$$

$$COM = \mathcal{B}_0 + 2.84(ACAD) + error \quad (10.15)$$

The significant directional association among the sub-factors of assessment practices is indicated in equations 10.13, 10.14, and 10.15. As shown, PUR is negatively impacted by AGE but positively predicted by ACAD and SCHTYPE. These mean that the younger the teachers, the more their mean scores were likely higher in assessment practices pertaining to assessment purpose; the more they had higher qualification, the more they had higher mean scores in doing assessment activities for appropriate purpose; and being in the public school, they tended to practice assessment by taking into consideration the purpose of doing it. Moreover, DES is also negatively affected by AGE but positively influenced by EXYR. These associations mean that the younger the teachers were, the more they practiced appropriate assessment design. On the other hand, the longer the teachers spent time in the teaching service, the more they were inclined to design their assessment activities appropriately as implied in their higher mean scores. These results appear to be contradictory. However, these can perhaps be interpreted in this way: young teachers whose ideas about assessment were still fresh usually appeared to be more energetic and assertive in their practice. This is usually the observation among teachers who are new to the profession. Those who have been in the service but finished their degrees years ago gradually become familiar with the assessment principles and eventually managed to properly design their assessment activities. Assessment practices in the Philippines are mostly prescribed by the DepEd and any change in the curriculum and policy at the national level would compel teachers to upgrade themselves. For new teachers who have been exposed to newer methods, they encounter little problem in implementing the DepEd's new assessment requirements. However, for old teachers they need to catch up by having more trainings and practice, and in the course of doing the required assessment they gain competence. As for the sub-variable COM, it has a positive relationship with ACAD. This means that the higher was the academic qualification of teachers, the more they communicated assessment results appropriately to students and other stakeholders. This is anticipated as better qualification is deemed contributory to appropriate professional practice.

Table 10.5. Standardised regression coefficients and t-values from regression analysis on the relationships among sub-factors of teaching practices

Sub-Factors	Demographic Factors				
	TSEX	AGE	ACAD	SCHTYPE	EXYR
STRUCT	0.65(2.25)*	- 0.92(- 3.05)*	0.16(0.52)	1.10(3.46)*	0.77 (2.64)*
STUDOR	0.35(0.95)	- 0.47(-1.21)	- 0.40(-1.01)	0.075(0.18)	0.98(2.62)*
ENACT	0.57(1.63)	- 0.90(- 2.48)*	- 0.28(-0.76)	0.29(0.77)	0.64(1.84)

*Note: t-value in parenthesis; *significant at p<0.01*

$$STRUCT = \mathcal{B}_0 + 0.65(TSEX) - 0.92(AGE) + 1.10(SCHTYPE) + 0.77(EXYR) + error \quad (10.16)$$

$$STUDOR = \mathcal{B}_0 + 0.98(EXYR) + error \quad (10.17)$$

$$ENACT = \mathcal{B}_0 - 0.90(AGE) + error \quad (10.18)$$

Table 10.5 presents the significant results of regression analysis on the relationships among sub-factors of teaching practices. These relationships are modeled in equations 10.16, 10.17, and 10.18. Equation 10.16 implies that STRUCT has a positive relationship with TSEX, SCHTYPE, and EXYR but negatively predicted by AGE. These results disclose that male teachers tended to teach in a more structured way following the direct transmission method than their female counterpart; in addition, teachers in public schools inclined to adopt structured teaching practices than those in the private institutions and that the more years of teaching experience the teachers had the more they employed well-structured instructional activities. In terms of age range, the younger the teachers, the more they were structured in their teaching. In the case of STUDOR as represented by equation 10.17, only years of teaching experience is impacting it. The positive relationship shown by the equation denotes that as teachers had more years of teaching experience, their teaching practices tended to be more student-oriented. This can be taken to mean that teachers with more years of professional experience can vary and adapt their teaching activities to student needs. In other words, they were more likely to engage students as they gained more experience. Furthermore, equation 10.18 indicates negative association between ENACT and AGE. This implies that the younger the teachers, the more they had the inclination to use enhanced activities in their teaching practices. In other words, younger teachers appeared to use alternative approach while older teachers were more inclined to employing direct transmission approach in their teaching.

Table 10.6 below further shows the regression analysis results on the relationships among sub-factors at the teacher level. As can be seen, there are six equations (equations 10.19 to 10.24) that

Table 10.6. Standardised regression coefficients and t-values from regression analysis indicating the relationships among sub-variables at the teacher level

Sub-Factors	Sub-Factors					
	STAN5	STAN6	STAN7	PUR	DES	COM
PUR	0.14(2.02)*	- 0.11(-1.67)	0.029(0.48)			
DES	0.099(2.02)*	- 0.074(-1.58)	0.077(1.76)			
COM	0.0064(0.072)	- 0.31(- 3.62)*	0.089(1.11)			
STRUCT	0.0015(0.044)	- 0.011(- 0.34)	- 0.039 (-1.33)	0.097(3.72)*	0.039(1.04)	0.085(4.90)*
STUDOR	- 0.066(-1.54)	- 0.018(- 0.43)	- 0.061(-1.61)	0.10(3.12)*	0.030(0.63)	0.12(5.52)*
ENACT	- 0.023(- 0.58)	- 0.019(- 0.48)	- 0.076(-2.13)*	0.029 (0.94)	0.063(1.42)	0.11(5.12)*

*Note: t-value in parenthesis; *significant at p<0.01*

$$PUR = \mathcal{B}_0 + 0.14(STAN5) + error \quad (10.19)$$

$$DES = \mathcal{B}_0 + 0.099(STAN5) + error \quad (10.20)$$

$$COM = \mathcal{B}_0 - 0.31(STAN6) + error \quad (10.21)$$

$$STRUCT = \mathcal{B}_0 + 0.097(PUR) + 0.085(COM) + error \quad (10.22)$$

$$STUDOR = \mathcal{B}_0 + 0.10(PUR) + 0.12(COM) + error \quad (10.23)$$

$$ENACT = \mathcal{B}_0 - 0.076(STAN7) + 0.11(COM) + error \quad (10.24)$$

indicate significant relationships. Equation 11.19 points to positive relationship between PUR and STAN 5. This implies that teachers' practices on assessment purpose are positively affected by their knowledge on the development of valid pupil grading procedures. In other words, as teachers became more knowledgeable in developing valid grading procedures they tended to be aware about the purpose of giving assessment activities in the class. Similarly, DES is modeled to have a positive relationship with STAN5 in equation 10.20. The same meaning holds that teachers' practices on assessment design are positively impacted by their knowledge of proper grading practice. As the teachers were getting more competent in the development of valid grading procedures, they were more likely to execute proper or appropriate

assessment process in carrying out certain assessment activities. This result attests to the fact that grading procedure is a required process that teachers in Tawi-Tawi and the Philippines are expected to be familiar with. Giving grades that serve to indicate student achievement is part of their accountability and their knowledge on how the grades are derived can somehow affect the way they design assessment activities or instruments such as test. However, the negative relationship between COM and STAN6 in equation 10.21 is quite troubling. It is indicated that teachers' practices on assessment communication is adversely affected by their knowledge on "communicating assessment results to students, parents, other lay audiences, and other educators." How can one's better knowledge about assessment communication make him communicate assessment results poorly or vice versa? Again, this goes back to the issue of consistency between what teachers indicated in their survey responses and what they actually do and know about this standard. What teachers indicated in their responses on assessment practices concerning communication were perhaps their beliefs while their responses on the concerned assessment standard reflected their knowledge. Beliefs and knowledge of any person, and teachers for that matter, cannot always be expected to be the same. Other reasons associated with self-reported responses as mentioned earlier possibly could help explain this inconsistent result. On the relationship depicted in equation 10.22, STRUCT is positively predicted by PUR and COM. This means that teachers' structured teaching practices are positively influenced by their assessment practices concerning assessment purpose and communication. Similar relationship exists for sub-variables in equation 10.23. Teachers' student-oriented instructional practices are also positively affected by their assessment practices concerning purpose and communication. In other words, teachers with sufficient knowledge on the purpose of using particular assessment methods and on how to communicate assessment data or information were more likely to be student-oriented in executing their instructional activities in the class. Finally, equation 10.24 reveals that ENACT is negatively affected by STAN7 (Recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information) but positively influenced by COM. This implies that teachers who had sufficient knowledge about assessment ethics were likely not to use enhanced activities in their teaching; conversely,

teachers who had knowledge about communicating assessment results tended to use elaborated activities to develop critical or higher order thinking skills.

From the results and discussion of directional relationships among main variables at the teacher level, teacher assessment literacy appeared not to influence assessment practices and to negatively affect teaching practices. However, deeper examination of the associations among sub-variables partly revealed otherwise. In fact, two of the three sub-variables of assessment practices are positively influenced by one assessment standard and two of the three sub-factors of teaching practices are positively impacted by two sub-factors of assessment practices. Hence, it can be deduced that assessment literacy somehow positively affects assessment practices; it likewise appears that assessment literacy impacts on teaching practices through assessment practices, though indirect effect needs to be examined to confirm this observation. However, it is evident from the results that assessment literacy has no direct link to teaching practices.

Discussed in the next subsections are the models and factors at the student level. As mentioned earlier, in the analysis of SEM, student participants/responses were divided into two groups (Grade 6 and Second Year high school students constituting the first group and Fourth Year high school students composing the second group). The reason for the grouping was because different outcome variables were tested for the two groups. For the first group, academic achievement as measured by NAT was the outcome variable. For the second group, aptitude as measured by NCAE was the dependent variable.

10.9.3 Student-level Factors (Model 1 for Grade 6 and Second Year high school students)

Relationships among main and sub-variables at the student level were also examined to answer Question 9.2.1 (What is the influence of gender on student perception of assessment, student attitude towards assessment, academic achievement, and aptitude?), Question 9.2.2 (What is the influence of students' perceptions of assessment on their attitude towards assessment?), Question 9.2.3 (What is the impact of Grade 6 and Second Year high school students' perceptions of assessment and attitude towards

assessment on their academic achievement?), and Question 9.2.4 (What is the impact of Fourth Year high school students' perceptions of assessment and attitude towards assessment on their aptitude?) as presented in Chapters 1, 9 and in the early part of this chapter. The regression results are presented in Tables 10.7, 10.8.

Table 10.7. Standardised regression coefficients and t-values from regression analysis indicating the relationships among variables at the student level (Grade 6 and Second Year high school)

Main Factors	Main Factors		
	SPA	SATA	ACHIV (NAT)
SSEX	- 0.28(-1.69)	- 1.17(- 4.22)*	- 8.30(-3.32)*
SPA		1.01(22.79)*	1.44(3.10)*
SATA			- 0.11(- 0.44)

*Note: t-value in parenthesis; *significant at p<0.01*

$$SATA = \mathcal{B}_0 - 1.17(SSEX) + 1.01(SPA) + error \quad (10.25)$$

$$ACHIV = \mathcal{B}_0 - 8.30(SSEX) + 1.44(SPA) + error \quad (10.26)$$

The significant results in Table 10.7 are represented by equations 10.25 and 10.26. Equation 10.25 specifically shows that SATA is negatively influenced by SSEX but positively impacted by SPA. This means that female students tended to have higher mean scores on attitude than male students towards assessment; conversely, assessment perception equates attitude towards assessment. That is, as students gained high mean scores in perceptions of assessment they were likely to also obtain high mean scores in attitude towards assessment. In the case of equation 10.26, it indicates that ACHIV is influenced by SSEX and SPA. The equation implies that female students tended to have higher achievement score than their male counterpart. Also, as the students obtained high mean scores in their perceptions of assessment, their achievement scores likewise increased. This is in agreement with the conventional view that students' positive behavior towards academic activities tends to increase their achievement in school.

10.9.4 Student-level Factors (Model 2 for Grade 6 and Second Year high school students)

Table 10.8 presents the significant results of regression on the association among main and sub-variables combined. The modeled relationships are shown in equations 10.27 and 10.28. From equation 10.27, SATA appears to be negatively affected by SSEX while positively predicted by PTEST and PASS. This signifies that Grade 6 and Second Year High School female students tended to obtain higher mean scores in their attitude towards assessment than their male counterpart; moreover, as the concerned students' mean scores in perceptions of test and assignment increased, their corresponding mean scores in attitude towards assessment tended to improve. As regards to equation 10.28, the modeled relationship between ACHIV and SSEX is negative and between ACHIV and PASS is positive. This implies that female students tended to obtain higher achievement scores than their male counterpart and that student achievement scores as measured by the National Achievement Test (NAT) were influenced by the student perceptions of assessment (e.g. as the mean scores in assessment perceptions increased, the achievement score also increased).

Table 10.8. Standardised regression coefficients and t-values from regression analysis indicating the relationships among main and sub-variables at the student level (Grade 6 and Second Year high school students)

Main Factors	Main Factors			
	PTEST	PASS	SATA	ACHIV
SSEX	- 0.20(-1.12)	- 0.31(- 1.76)	- 1.21(- 4.37)*	- 8.21(- 3.28)*
PTTEST			0.72(17.49)*	0.11(0.28)
PTASS			0.31(7.64)*	1.34(3.56)*
SATA				- 0.069(- 0.29)

*Note: t-value in parenthesis; *significant at p<0.01*

$$SATA = \mathcal{B}_0 - 1.21(SSEX) + 0.72(PTEST) + 0.31(PTASS) + error \quad (10.27)$$

$$ACHIV = \mathcal{B}_0 - 8.21(SSEX) + 1.34(PASS) + error \quad (10.28)$$

10.9.5 Student-level Factors (Model 1 for Fourth Year High School Students)

The regression analysis results for the student-level factors concerning fourth year high school students are displayed in the tables below. It can be seen from equation 11.29 that SATA is negatively

predicted by SSEX and at the same time positively influenced by SPA. Similar to the results of Grade 6 and Second Year high school students, female Fourth Year high school students tended to have higher mean scores in attitude towards assessment than their male counterpart; moreover, as the students' perceptions of assessment tended to increase in scores, their attitude towards assessment also increased in scores. Examining equation 10.30, the student aptitude (APT) is positively impacted by students' attitude towards assessment. The analysis results of the combined main and sub-factors at the student level (Fourth Year high school students) are presented in Tables 10.9 and 10.10.

Table 10.9. Standardised regression coefficients and t-values from regression analysis indicating the relationships among main factors at the student level (Fourth Year high school students)

Main Factors	Main Factors		
	SPA	SATA	APT (NCAE)
SSEX	0.14(0.72)	- 0.93(- 2.19)*	- 1.13(-0.28)
SPA		0.78(9.25)*	0.065(0.077)
SATA			1.04(2.79)*

*Note: t-value in parenthesis; *significant at p<0.01*

$$SATA = \mathcal{B}_0 - 0.93(SSEX) + 0.78(SPA) + error \quad (10.29)$$

$$APT = \mathcal{B}_0 + 1.04(SATA) + error \quad (10.30)$$

Table 10.10. Standardised regression coefficients and t-values from regression analysis indicating the relationships among main and sub-variables at the student level (Fourth Year high school students)

Main Factors	Main Factors			
	PTEST	PASS	SATA	APT
SSEX	0.18(0.82)	0.094(0.41)	- 0.93(-2.19)*	- 1.00(- 0.25)
PTEST			0.43(5.62)*	- 1.47(-1.96)
PASS			0.37(5.16)*	1.45(2.10)*
SATA				1.04(2.79)*

*Note: t-value in parenthesis; *significant at p<0.01*

$$SATA = \mathcal{B}_0 - 0.93(SSEX) + 0.43(PTEST) + 0.37(PASS) + error \quad (10.31)$$

$$APT = \mathcal{B}_0 + 1.45(PASS) + 1.04(SATA) + error \quad (10.32)$$

Similar to the analysis of the main factors, two equations appear to indicate significant regression results. Equation 10.31 reveals that SATA is negatively affected by SSEX but positively influenced by

PTEST and PASS. This means that female Fourth Year high school students tended to have higher mean scores in their attitude towards assessment than the male students; in addition, as Fourth Year high school students' perceptions toward test and assignment increased (or decreased) in terms of scores, their scores in attitude towards assessment correspondingly tended to increase (or decrease). In the case of variable APT (equation 10.32), the result shows that it has a positive relationship with PASS and SATA, which means that scores in perceptions of assignment and attitude towards assessment tended to predict Fourth Year high school students' aptitude scores. In other words, high scores in PASS and SATA could indicate high scores in aptitude.

As mentioned early in this chapter, factors at the teacher level and the student level were analysed separately to avoid multicollinearity and to examine the specific relationships among the factors within each of the analysed groups. Moreover, the factors were not combined due to the problems associated with the SEM, specifically the aggregation of student level factors to teacher level factors or disaggregation of teacher data to student data. Hence, analysis was either within teacher level or student level only. Nevertheless, the indirect effects of some variables on other variables within each level were determined through path analysis to obtain the whole picture of the set of relationships at each level. The next sections describe briefly path analysis and present path analysis results.

10.10 Path Analysis

Path analysis is described as an extension of the multiple regression, as it involves a number of multiple regression equations to be estimated simultaneously. It is considered a more effective technique in modeling mediation, indirect effects, and other complex relationships among variables. In path analysis, structural relations among variables are modeled. As path analysis involves the evaluation of hypothesis about directional influences or causal relations, it is sometimes called as 'causal modeling' (Lei & Wu, 2007). A path model can serve as a representation of the relationships among a number of variables (or causal relationships), which may be independent, intermediary or dependent variables (Ben, 2010). A *direct*

effect is simply the direct influence that one variable has on another variable (Schumacker & Lomax, 2010). It is a total effect of one variable on another which is not transmitted through mediating variables (Alwin & Hauser, 1975). An *indirect effect* "represents the influence of an independent variable on a dependent variable as mediated by one or more intervening variables" (Diamantopoulos & Siguaw, 2000, pp. 69 – 70). It is part of the variable's total effect that is transmitted through intervening variables (Alwin & Hauser, 1975). It can be calculated by multiplying the parameter estimates of the mediating variables (Diamantopoulos & Siguaw, 2000).

10.10.1 Results of Path Analysis

To obtain an overview of the relationships among the teacher level and student level factors, two models corresponding to main factors and sub-variables were finally created for each level. Each of these models is described in the succeeding sections.

10.10.1.1 Teacher Level – Model 1

Model 1 for teacher level involved only the main factors and the demographic variables. In this model, the influence of assessment literacy, assessment practices, and demographic factors on teaching practices were tested. A path diagram of this model is presented in Figure 10.2. As can be seen from the figure, assessment literacy, assessment practices, and years of experience had direct effects on teaching practices. Other demographic variables like age range and academic qualification also exerted influence through assessment practices. The direct and indirect effect estimates are given in Tables 10.11 and 10.12.

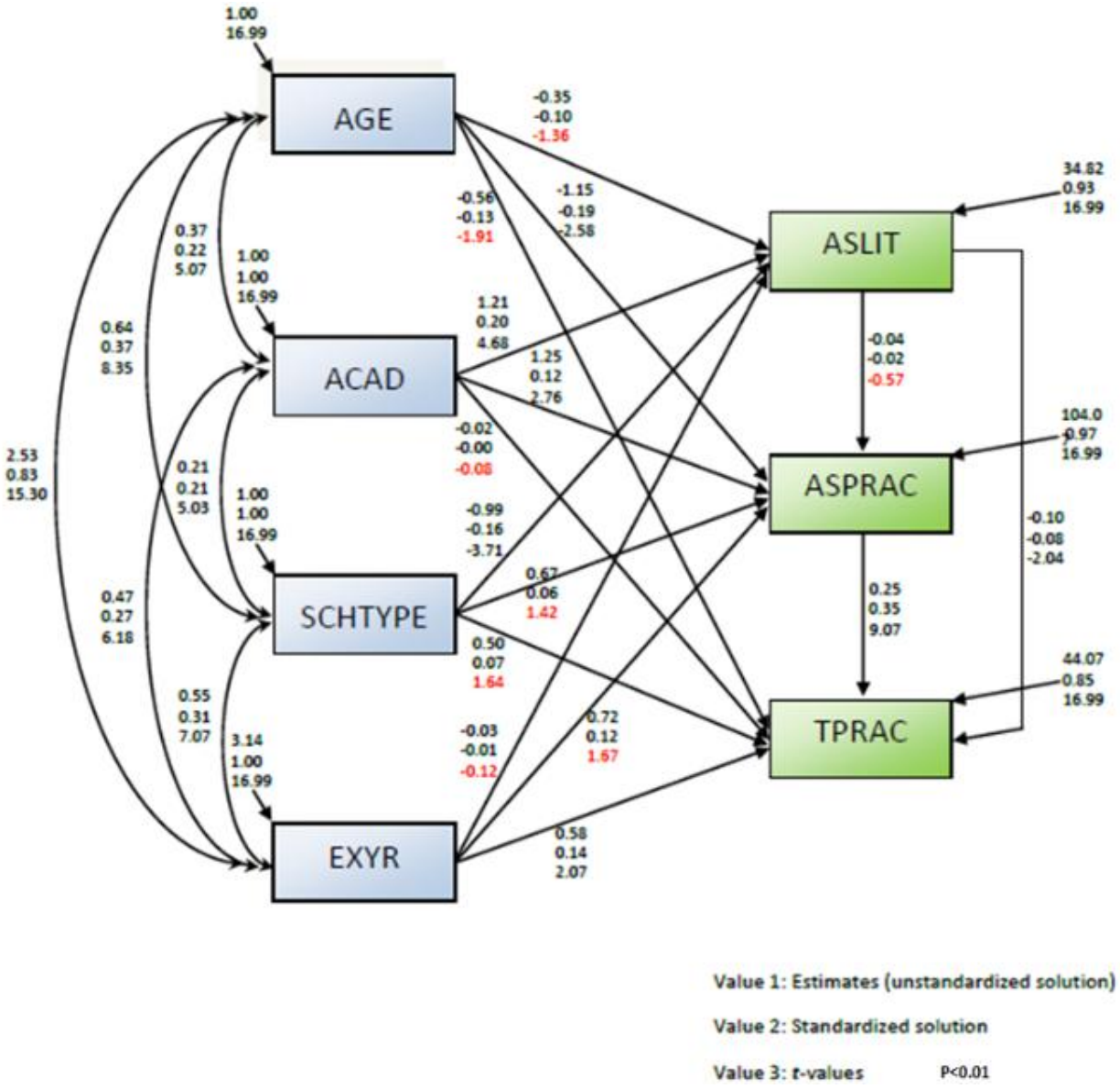


Figure 10.2. Direct and indirect effects of teacher-level factors on teaching practices (Model 1 for Teachers)

Table 10.11. Summary of direct effects on teaching practices

Direct Effects	TPRAC
ASLIT	- 0.08 (- 2.04)
ASPRAC	0.35 (9.07)
EXYR	0.14 (2.07)

Note: Regression Coefficient (Beta) – values outside the parentheses; t - values – values inside the parentheses; n = 581; P<0.01

It is shown in Table 10.11 that assessment literacy (ASLIT, - 0.08, *t* = - 2.04 at *p<0.01*) had a negative influence on teaching practices (see Figure 10.2). The path coefficient (- 0.08) indicates the extent

of influence that assessment literacy exerted on teaching practices. This means that for every increase by 0.08 in assessment literacy there is a corresponding decrease by the same value in the teaching practices. As mentioned earlier, this finding appears to be contrary to the view that assessment literacy contributes to instructional practices. The possible explanations provided in section 10.9 could help justify this result. On the other hand, assessment practices (ASPRAC, 0.35, t = 9.07 at $p<0.01$) and years of teaching experience (EXYR, 0.14, t = 2.07 at $p<0.01$) had positive effects on teaching practices. These results indicate that a change of 0.35 in the assessment practices and 0.14 in years of teaching experience would create a change of similar respective magnitude in teaching practices. This direct effect is expected as assessment practices and experience on the job are viewed as contributing factors to instructional practices. Equation 10.33 summarises the direct effects of teacher-level factors on teaching practices as shown in Figure 10.2.

$$TPRAC = \mathcal{B}_0 + 0.14(EXYR) - 0.08(ASLIT) + 0.35(ASPRAC) + error \qquad (10.33)$$

The indirect effects of age range and academic qualification on teaching practices are given in Table 10.12. As can be gleaned from the table, age range (-0.19 x 0.35 = -0.07) had a negative indirect effect on teaching practices. This means that teachers' age negatively influenced their teaching practices through their assessment practices and this path explains about 7% of the variance. However, one-way ANOVA results revealed no significant differences on the assessment and teaching practices of teachers by age range. Conversely, the academic qualification (0.12 x 0.35 = 0.04) had a positive indirect effect on the teaching practices through assessment practices. This path explains about 4% of the variance in the direct relationship between assessment practices and teaching practices. This denotes that as teachers gained better academic qualification, their assessment practices tended to improve, which thus results in the improvement of their teaching practices.

Table 10.12. Summary of indirect effects on teaching practices

Indirect Effects	TPRAC
AGE through ASPRAC	- 0.07 (7%)
ACAD through ASPRAC	0.04 (4%)

10.10.1.2 Teacher Level – Model 2

Due to the number of sub-factors and the complex relationships involved in Model 2 for teachers, the direct and indirect effects are presented through Table 10.13 instead of a figure. The significant paths and the corresponding coefficients in terms of estimates (unstandardised solution) standardised solution, and t-value are shown.

Table 10.13. Direct and indirect effects on sub-factors of teaching practices (Model 2 for Teachers)

Path	Coefficients		
	Estimates (Unstandardised solution)	Standardised Solution	t-value
SCHTYPE to STAN2	- 1.50	- 0.16	- 3.55
SCHTYPE to STAN5	- 1.12	- 0.12	- 2.67
SCHTYPE to STAN6	- 1.17	- 0.12	- 2.71
SCHTYPE to PUR	1.58	0.11	2.45
SCHTYPE to STRUCT	1.10	0.15	3.46
ACAD to STAN3	1.71	0.17	3.98
ACAD to STAN4	2.17	0.24	5.69
ACAD to STAN6	1.42	0.15	3.44
ACAD to PUR	1.62	0.11	2.63
ACAD to COM	2.84	0.15	3.47
TSEX to STAN4	0.90	0.10	2.44
TSEX to STRUCT	0.65	0.09	2.25
AGE to PUR	- 1.43	- 0.17	- 2.33
AGE to DES	- 1.22	- 0.20	- 2.73
AGE to STRUCT	- 0.92	- 0.21	- 3.05
AGE to ENACT	- 0.90	- 0.18	- 2.48
EXYR to DES	0.94	0.16	2.20
EXYR to STRUCT	0.77	0.18	2.64
EXYR to STUDOR	0.98	0.18	2.62
STAN5 to PUR	0.14	0.09	2.02
STAN5 to DES	0.099	0.09	2.02
STAN6 to COM	- 0.31	- 0.16	- 3.62
STAN7 to ENACT	- 0.076	- 0.09	- 2.13
COM to STRUCT	0.085	0.21	4.90
COM to STUDOR	0.12	0.24	5.52
COM to ENACT	0.11	0.23	5.12
PUR to STRUCT	0.097	0.18	3.72
PUR to STUDOR	0.10	0.16	3.12

As can be seen, there are a number of direct and indirect effects on the sub-factors of teaching practices. These results are as follows: six factors (SCHTYPE, TSEX, AGE, EXYR, PUR, and COM) exerted direct impact on teaching practices concerning structuring activities (STRUCT); three factors (EXYR, PUR, and COM) had direct effect on teaching practices concerning student-oriented activities (STUDOR); three factors (AGE, STAN7, and COM) had direct influence on teaching practices concerning enhanced activities (ENACT); five factors (SCHTYPE, ACAD, AGE, STAN5, and STAN6) appear to have indirect effects on STRUCT and STUDOR; and three factors (ACAD, SCHTYPE, and STAN6) had indirect impact on ENACT. These sub-variables and the associated effects are summarised in Tables 10.14 and 10.15.

Table 10.14. Summary of direct effects of teacher-level demographic sub-factors on the sub-variables of teaching practices

Direct Effects	STRUCT	STUDOR	ENACT
TSEX	0.09 (2.25)		
AGE	- 0.21 (- 3.05)		- 0.18 (- 2.48)
SCHTYPE	0.15 (3.46)		
EXYR	0.18 (2.64)	0.18 (2.62)	
STAN7			- 0.09 (- 2.13)
PUR	0.18 (3.72)	0.16 (3.12)	
COM	0.21 (4.90)	0.24 (5.52)	0.23 (5.12)

Note: Regression Coefficient (Beta) – values outside the parentheses; t - values – values inside the parentheses; n = 581; P<0.01

Table 10.14 shows the direct effect results of teacher-level demographic factors and sub-factors on the sub-variables of teaching practices. As can be gleaned from the table, gender (TSEX, 0.09, t = 2.25, $p<0.01$), school type (SCHTYPE, 0.15, t = 3.46, $p<0.01$), years of teaching experience (EXYR, 0.18, t = 2.64, $p<0.01$), assessment purpose (PUR, 0.18, t = 3.72, $p<0.01$), and assessment communication (COM, 0.21, t = 4.90, $p<0.01$) had direct positive effects on structuring activities (STRUCT). The respective path coefficients indicate the extent of change that the concerned factors/sub-factors transported to STRUCT. These results mean that: a) male teachers (gender coded as 0 and 1 for females and males, respectively) tended to teach using structuring activities more than the female teachers; b) teachers in the public school

(school type coded as 0 and 1 for private and public schools, respectively) tended to employ structuring activities in their teaching than those in the private schools; c) teachers with more years of teaching experience (years of teaching experience coded from 1 to 7 corresponding to increasing year ranges) tended to adopt structured instructional activities; d) teachers' use of assessment purpose influenced the use of structuring activities in their teaching; and e) teachers' use of assessment communication impacted on the use of structuring activities in their instruction. It can be noted that most of the teacher respondents came from the public school and a number of them have been in the teaching service for many years as revealed from the demographic data. This suggests that teachers were more familiar with and had used the direct transmission approach of teaching for a long time. Thus, it is possible that their views as reflected in the associated variables influenced the practice of structuring activities. On the other hand, teachers' age range (AGE, - 0.21, t = - 3.05, $p<0.01$) had a negative effect on teachers' structuring practices. This result indicates that younger teachers (AGE coded from 1 to 6 corresponding to the increasing age ranges) also tended to employ structuring activities in their teaching. Moreover, years of teaching experience (EXYR, 0.18, t = 2.62, $p<0.01$), assessment purpose (PUR, 0.16, t = 3.12, p<0.01), and assessment communication (COM, 0.24, t = 5.52, $p<0.01$) had also positive effects on student-oriented activities (STUDOR). The respective coefficients indicate the units of change that the variables exerted on STUDOR. These results mean that teachers with more years of teaching experience also tended to employ student-oriented activities in their instruction. Also, assessment purpose and communication impacted on this kind of teaching activities. This implies that teacher respondents did not only use structuring activities but also student-oriented activities in their classroom teaching. Furthermore, age range (AGE, - 0.18, t = - 2.48, $p<0.01$) and assessment literacy in Standard 7 (STAN7, - 0.09, t = - 2.13, $p<0.01$) appeared to have negative effects while assessment communication (COM, 0.23, t = 5.12, $p<0.01$) appeared to have positive effect on enhanced activities (ENACT). The extents of impact that these variables had on ENACT are indicated by their respective path coefficients. These results imply that younger teachers tended to adopt enhanced activities in their teaching while teachers' knowledge of assessment ethics tended to avoid the

use of enhanced activities. It could be that young teachers tended to employ enhanced activities as they have graduated under the new pre-service teacher education curriculum that offers revised and enhanced subjects on teaching methods. However, the finding that knowledge on assessment ethics negatively affected the use of enhanced activities is quite unanticipated. Perhaps, some teacher respondents viewed that there were issues associated with the use of enhanced activities or with the assessment of enhanced activities. As for the assessment communication, teachers who employed this aspect of assessment practices tended to use enhanced activities. These results of direct effects can be summarised in equation form as follows:

$$STRUCT = \mathcal{B} + 0.09(TSEX) - 0.21(AGE) + 0.15(SCHTYPE) + 0.18(EXYR) + 0.18(PUR) + 0.21(COM) + error \quad (10.34)$$

$$STUDOR = \mathcal{B} + 0.18(EXYR) + 0.16(PUR) + 0.24(COM) + error \quad (10.35)$$

$$ENACT = \mathcal{B} - 0.18(AGE) - 0.09(STAN7) + 0.23(COM) + error \quad (10.36)$$

Table 10.15. Summary of indirect effects of teacher-level demographic and sub-factors on sub-variables of teaching practices

Indirect Effects	STRUCT	STUDOR	ENACT
AGE through PUR	- 0.03 (3%)	- 0.03 (3%)	
ACAD through PUR	0.02 (2%)	0.02 (2%)	
ACAD through COM	0.03 (3%)	0.04 (4%)	0.03 (3%)
ACAD through STAN6 and COM	- 0.005 (0.5%)	- 0.006 (0.6%)	- 0.006 (0.6%)
SCHTYPE through PUR	0.02 (2%)	0.02 (2%)	
SCHTYPE through STAN5 and PUR	- 0.002 (0.2%)	- 0.002 (0.2%)	
SCHTYPE through STAN6 and COM	0.004(0.4%)	0.005(0.5%)	0.004 (0.4%)
STAN5 through PUR	0.02(2%)	0.01(1%)	
STAN6 through COM	- 0.03(3%)	- 0.04(4%)	- 0.04(4%)

Table 10.15 presents the results of indirect effects of teacher-level demographic factors and sub-variables on the sub-constructs of teaching practices. It can be seen from the table that STRUCT and STUDOR have the most number of indirect effects, indicating that these sub-factors are more associated

with a number of tested teacher variables. On the contrary, the ENACT has the least of the indirect effects, indicating that this sub-factor is less affected by other teacher factors.

The sub-factor STRUCT has a total of nine indirect effects of which five are positive and four are negative. Factors that transported positive effects include academic qualification (ACAD), school type (SCHTYPE), and assessment standard 5 (STAN5). The academic qualification exerted positive effects through assessment purpose (0.11x0.18=0.02 or 2%) and assessment communication (0.15x0.21=0.03 or 3%) indicating that about 2% of the relationship between assessment purpose and structuring activities and 3% of the association between assessment communication and structuring activities is due to teachers' academic qualification. This suggests that as teachers gained higher academic qualification, they tended to employ assessment practices concerning purpose and communication, which further influenced their instructional practices involving structured activities. The school type likewise transported positive effects through assessment purpose (0.11x0.18=0.02 or 2%) and through assessment standard 6 (STAN6) and assessment communication (-0.12x-0.16x0.21=0.004 or 0.4%). Similarly, this factor influenced the association between assessment purpose and structuring activities by about 2% and the relationships among assessment standard 6, assessment communication, and structuring activities by about 0.4%. This could mean that: a) public school teachers' assessment practices involving purpose impacted on their structuring activities in the class; and b) public school teachers' knowledge and practice on communicating assessment results or information positively affected their structuring activities, although the percentage of influence is quite low. Similar interpretation can be made for the assessment standard 5 through assessment purpose (0.09x0.18=0.02 or 2%). That is, assessment standard 5 impacted structuring activities through assessment purpose and the extent of influence was about 2%. This indicates that 2% of the relationship between the two related factors could be attributed to assessment standard 5. This implies that teachers' assessment knowledge on developing valid grading procedure tended to make them employ assessment practices concerning purpose and further influenced their instructional approach involving structuring activities.

However, academic qualification through assessment standard 6 and assessment communication (0.15x-0.16x0.21=-0.005 or 0.5%); school type through assessment standard 5 and assessment purpose (-0.12x0.09x0.18=-0.002 or 0.2%); age range through assessment purpose (-0.17x0.18=-0.03 or 3%); and assessment standard 6 through assessment communication (-0.16x0.21=-0.03 or 3%) exerted negative indirect effect on structuring activities. The resulting coefficients or the respective percentages indicate the extents of influence that the involved variables transported to structuring activities. For the negative effect of academic qualification through assessment literacy in standard 6 and assessment communication, the result could possibly mean that teachers with bachelor degree or with minimum academic qualification were not ready to grasp and interpret the important implications of their own assessment results to their teaching practices and perhaps this is the reason why it failed to influence their structuring activities. For the negative effect of school type, this can perhaps be explained by their literacy on standard 5 (developing a valid grading procedure). It has been shown in the previous results that school type impacted directly on structuring activities and provided positive indirect effect on the said teaching activities through the assessment purpose. This means that perhaps their low literacy in developing valid grading procedure impacted on the way they structured their teaching tasks. As for the negative influence of teachers' age range through assessment purpose, the result could mean that young teachers tended not to associate or employ assessment purpose with the way they structured their activities in the class. It could be that assessment purpose was not their main basis in deciding what kind of tasks to be provided, thus the link between their age and the associated factors was negative. Lastly, the negative influence of assessment standard 6 (communicating assessment results) through assessment communication can perhaps be explained by possible reasons provided in section 10.9. It could also indicate that teachers' low literacy in standard 6 negatively impacts on the way they interpret and make decisions about their students and teaching activities.

Examining the results in Table 10.15, similar factors and patterns of indirect effects can be observed for student-oriented activities (STUDOR). The only differences with structuring activities are on

some of the coefficients/percentage of coefficients, indicating the differences in the extents of change that the involved factors transported on STUDOR. However, the differences are very minimal. Hence, similar interpretations can be made for both STRUCT and STUDOR. With regard to the indirect effects on enhanced activities, two factors appear to exert positive influence while another two variables transmitted negative effects. Specifically, teachers' academic qualification through assessment communication (0.15x0.23=0.03 or 3%) and school type through assessment standard 6 and assessment communication (-0.12x-0.16x0.23=0.004 or 0.4%) appeared to have positive direct influence on ENACT. The resulting coefficients indicate the amount of change that can be attributed to these two influencing factors. These results could mean that high academic qualification could be a factor that was likely to influence teachers' decision to employ enhanced activities in their teaching. In addition, the positive effect of school type could mean that public school teachers tended to associate their assessment practices concerning communication with ENACT, although the percentage of coefficient is quite low. On the negative effects, the academic qualification through assessment standard 6 and assessment communication (0.15x-0.16x0.23=-0.006 or 0.6%), and assessment standard 6 through assessment communication (-0.16x0.23=-0.04 or 4%) are the influencing variables. These results can perhaps be attributed to the problems associated with the conflicting results between teachers' literacy on the communication of assessment results and their assessment practices concerning communication, as pointed earlier. It could also be that teachers' low literacy in interpreting and communicating assessment results was a factor in influencing this negative effect. However, this warrants further investigation to unpack the information concerning these variables and their relationship.

The path analysis results concerning the directional relationships among the main and sub-factors at the teacher level provide a general picture that some demographic factors exert influence on a number of teacher variables. In addition, the extent of influence of assessment literacy on teaching practices through assessment practices can be traced through specific factors tested in this study.

10.10.1.3 Student Level – Model 1 for Grade Six and Second Year High School Students

As described earlier, student factors were analysed for two groups of student participants: Grade 6 and Second Year high school students were combined to constitute one group and Fourth Year high school students were to compose another group. The grouping was made as these student groups had different outcome variables. For the first group, achievement was the outcome variable while the second group had aptitude as the dependent variable. Moreover, for each group, two models were tested. One model involved the analysis of the relationships among student-level main factors and the other model included the analysis of the relationships among the main and sub-factors at the student level. The reason for adopting the two models was to avoid the problem associated with multicollinearity and to obtain the overall picture of the directional relations among the factors and sub-factors at the student level.

The analysis at each of the two models for each student group involved the examination of the possible influence of the student-level demographic factor, the student gender (SSEX), on the main and sub-variables, and to investigate the effects of these variables on academic achievement (ACHIV) for Grade 6 and Second Year high school students and on aptitude (APT) for Fourth Year high school students. The analysis involved responses from 2077 student participants of which 1,430 were Grade 6 and Second Year high school students and 647 were Fourth Year high school students. The results of the path analysis for the two models for each group are provided below.

Figure 10.3 presents the results of path analysis of model 1 factors for Grade 6 and Second Year high school students. As can be seen from the figure, student gender (SSEX, - 0.09, t = - 4.32, $p<0.01$) exerted negative effect on student attitude towards assessment (SATA). The path coefficient (-0.09) indicates the strength of influence SSEX transported to SATA. This result could indicate that female students tended to have higher mean scores in attitude towards assessment than male students. Moreover, the student perception of assessment (SPA, 0.51, t = 22.79, $p<0.01$) positively impacts on SATA. The resulting path coefficient (0.51) signifies the strength of the associated effect. This could be interpreted that as SPA scores increased (or decreased), SATA scores also increased (or decreased). In other words, as

Grade 6 and Second Year high school students gained higher mean scores on perception towards assessment they correspondingly tended to obtain high mean scores on attitude towards assessment. This result can be expected as perception is deemed to influence attitude. On the effects on academic achievement, SSEX and student perception of assessment (SPA) appear to exert direct influence. However, no indirect effects can be seen from the figure. The direct effects of the involved demographic and main factors on academic achievement are summarised in Table 10.16.

It can be gleaned from Table 10.16 that there are two direct effects on student academic achievement (ACHIV). The first effect came from the student gender (SSEX, - 0.09, t = - 3.32, $p<0.01$), which negatively impacted on ACHIV. The path coefficient (-0.09) indicates the extent of negative influence of SSEX on ACHIV. This means that for every increase in SSEX by this value, there is a corresponding decrease by the same value in ACHIV. This result could indicate that female students (SSEX coded as 0 and 1 for females and males, respectively) tended to obtain and influenced higher academic achievement than male students. Conversely, the student perception of assessment (SPA, 0.10, t = 3.10, $p<0.01$) transported positive effect on ACHIV. The associated path coefficient (0.10) indicates the extent of effect. This result could mean that as the scores of Grade 6 and Second Year high school students in assessment perception increased (or decreased), their academic achievement or NAT scores tended to correspondingly increase (or decrease). This result is consistent with the view that perception affects performance on any task. The direct effects as shown in Figure 10.3 are summarised in the following equation.

$$ACHIV = \mathcal{B}_0 - 0.09(SSEX) + 0.10(SPA) + error \quad (10.37)$$

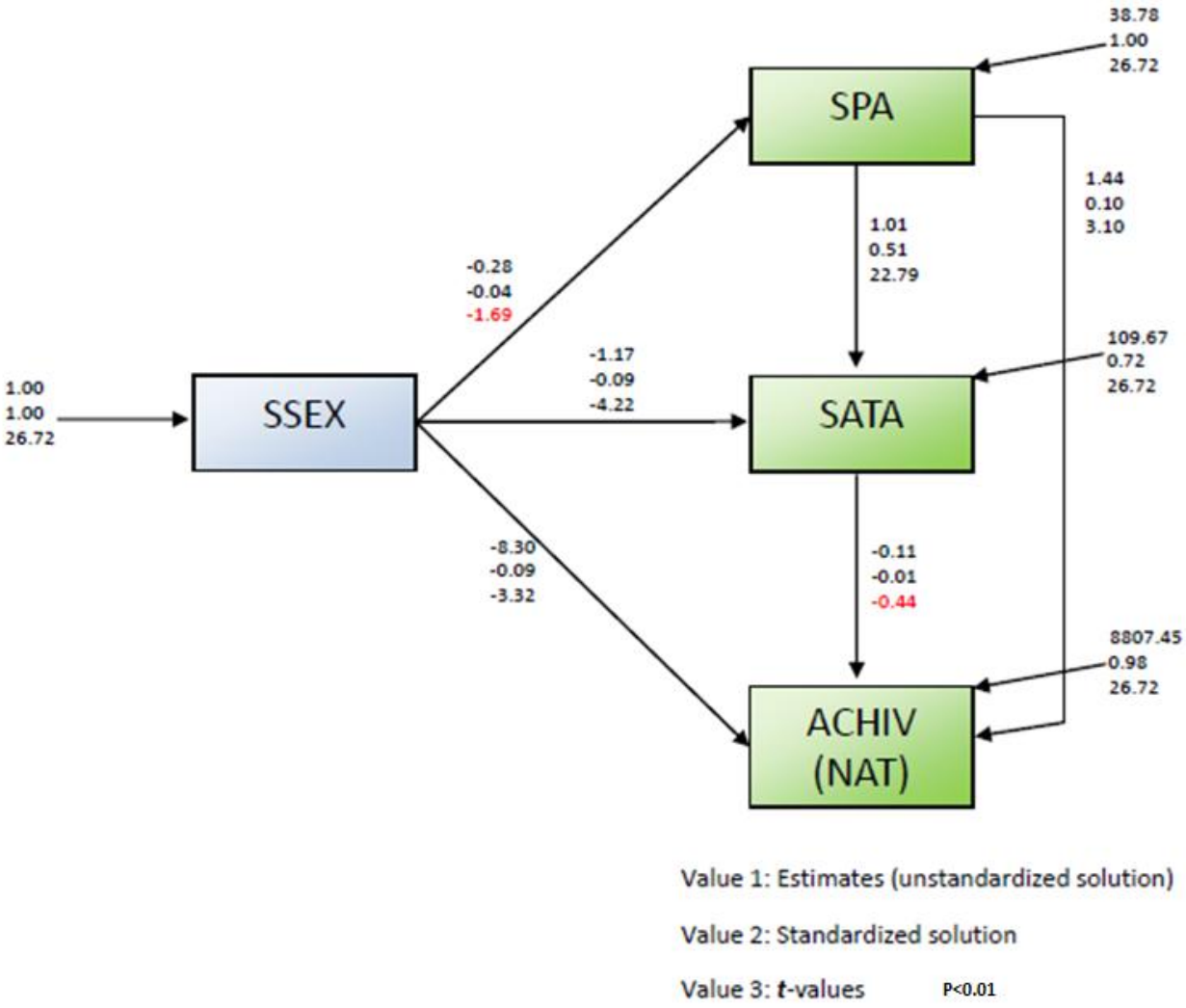


Figure 10.3. Direct and indirect effects of student-level demographic and main factors on academic achievement (Model 1 for Grade 6 and Second Year high school students)

Table 10.16. Direct effects of student-level demographic and main factors on academic achievement (Model 1 for Grade 6 and Second Year high school students)

Direct Effects	Academic Achievement (ACHIV)
SSEX	- 0.09 (- 3.32)
SPA	0.10 (3.10)

Note: Regression Coefficient (Beta) – values outside the parentheses; t - values – values inside the parentheses; n = 1,430; P<0.01

10.10.1.4 Student Level – Model 2 for Grade Six and Second Year High School Students

On the effects of demographic, main, and sub-factors on academic achievement (ACHIV), Figure 10.4 presents the results. As depicted in the figure, there are direct effects but there are no indirect effects of the tested factors on ACHIV.

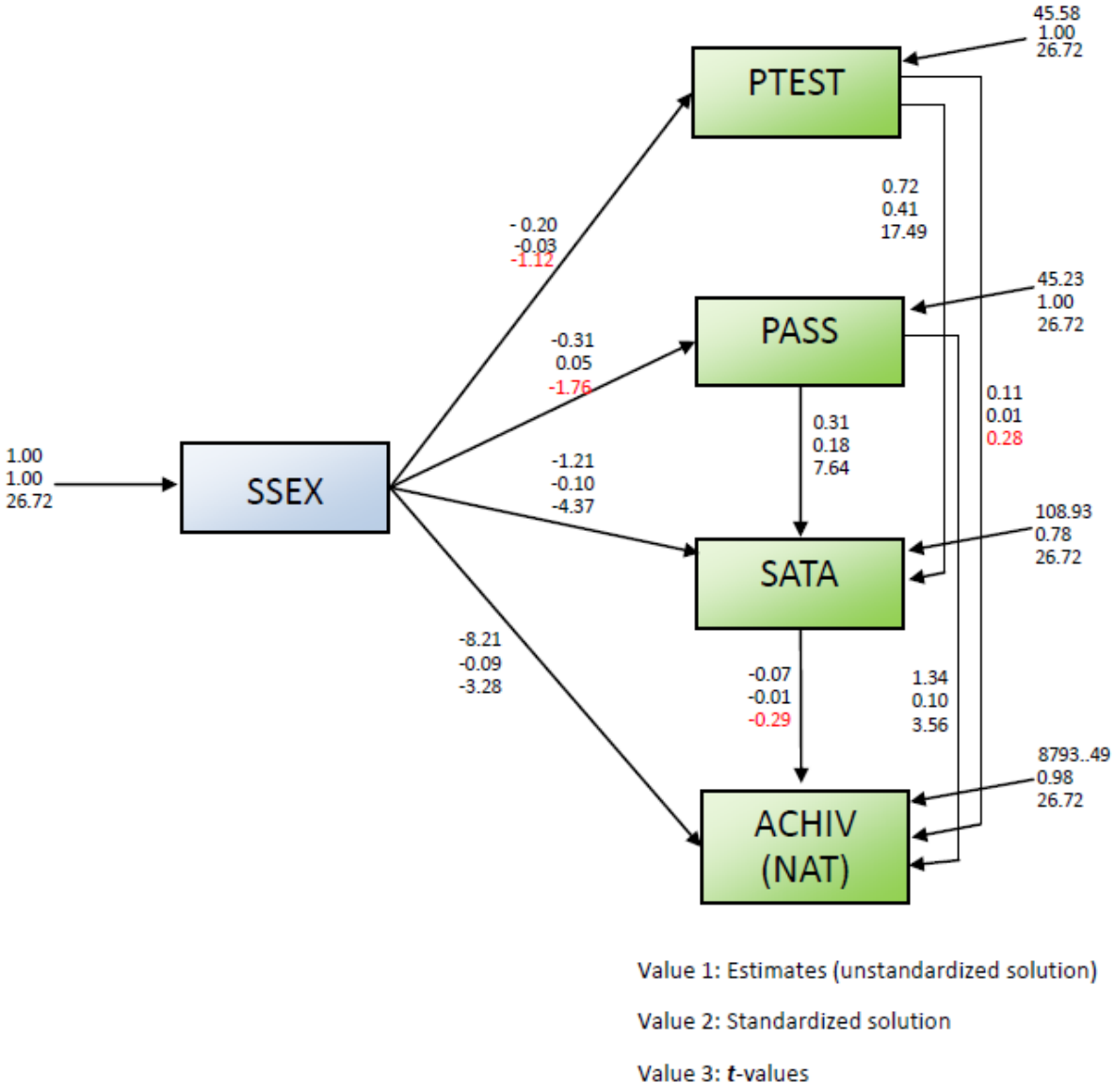


Figure 10.4. Direct and indirect effects of student-level demographic, main and sub-factors on academic achievement (Model 2 for Grade 6 and Second Year high school students)

The student gender (SSEX, - 0.10, t = - 4.37, $p<0.01$) again appears to exert a negative influence on student attitude towards assessment. This reveals that female students tended to have high mean score

on attitude towards assessment than male students. Besides, perception of test (PTEST, 0.41, *t* = 17.49, *p<0.01*) and perception of assignment (PASS, 0.18, *t* = 7.64, *p<0.01*) had positive direct influence on student attitude towards assessment. The respective path coefficients signify the extent of change that can be attributed to the two sub-constructs. This could indicate that as the mean scores on perceptions of Grade 6 and Second Year high school students towards test and assignment became high, their mean score on attitude towards assessment also tended to become high. Furthermore, the student gender (SSEX, - 0.09, *t* = - 3.28, *p<0.01*) and perception of assignment (PASS, 0.10, *t* = 3.56, *p<0.01*) appear to directly affect academic achievement. The direct effect results on this outcome variable is summarised in Table 10.17.

Table 10.17. Direct effects of student-level factors on academic achievement (Model 2 for Grade 6 and Second Year high school students)

Direct Effects	Academic Achievement (ACHIV)
SSEX	- 0.09 (- 3.28)
PASS	0.10 (3.56)

Note: Regression Coefficient (Beta) – values outside the parentheses; t - values – values inside the parentheses; n = 1,430; P<0.01

As can be observed from the table above, the SSEX had a negative effect with a path coefficient of -0.09 on ACHIV. This result could mean that female Grade 6 and Second Year high school students tended to obtain higher academic achievement or high NAT scores than their male counterpart. In addition, of the two sub-factors of SPA, the perception of assignment (PASS) appears to positively influence academic achievement with a path coefficient of 0.10. Examining the path coefficient of SPA as a single factor in Table 10.16, it appears that PASS was the main sub-factor influencing the effect of SPA on academic achievement. This could indicate that as the concerned students had high mean scores on perception towards assignment, their academic achievement tended to increase. This possibly suggests that students' view of assignment had more influence than their view about the test. This is unexpected taking into account the Philippine context, and Tawi-Tawi context for that matter, where assessment predominantly involves testing. This is especially so as the outcome variable is the academic achievement that reflects test (NAT) scores. Perhaps, the concerned students thought that doing assessment was more

contributory as a preparation for obtaining high scores in the test than the experience of taking the test itself. The direct effects on student academic achievement as depicted in Figure 10.4 can be presented in the form of equation 10.38.

$$ACHIV = \mathcal{B}_0 - 0.09(SSEX) + 0.10(PTASS) + error \quad (10.38)$$

10.10.1.5 Student Level – Model 1 for Fourth Year High School Students

The model 1 for Fourth Year high school students is shown in Figure 10.5. The figure presents the results of path analysis on the direct and indirect effects of student-level main factors on aptitude. As shown, there are direct and indirect relationships between the tested factors and the aptitude as the dependent variable.

For the effect of student gender (SSEX) on the main factors for this student group, the path shows that it (SSEX, - 0.08, t = - 2.19, $p<0.01$) has directional relation with student attitude towards assessment (SATA), though the effect (-0.08) is negative. This could be interpreted that female Fourth Year high school students tended to have high mean score on attitude towards assessment than the male students. Also, the student perception of assessment (SPA, 0.34, t = 9.28, $p<0.01$) appears to positively influence SATA, which likewise suggests that as Fourth Year high school students obtained high mean score on perception of assessment, their mean score on attitude towards assessment also tended to increase. However, the effect of SPA on aptitude (APT) is not significant. Moreover, there are one direct effect and two indirect effects on APT. These are summarised in Tables 10.18 and 10.19.

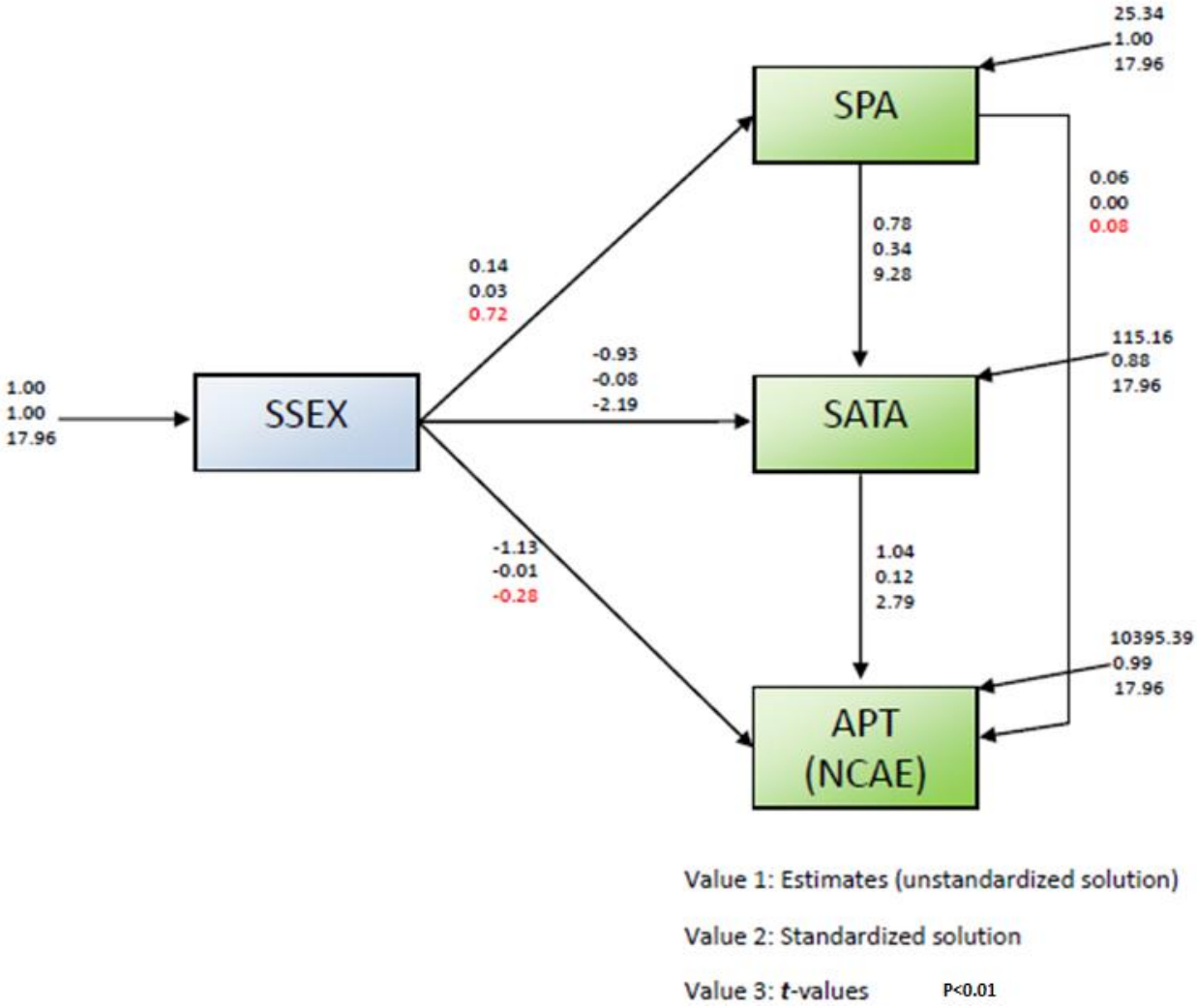


Figure 10.5. Direct and indirect effects of student-level demographic and main factors on aptitude (Model 1 for Fourth Year high school students)

Table 10.18. Direct effect of student-level main factors on aptitude (Model 1 for Fourth Year high school students)

Direct Effect	Aptitude (APT)
SATA	0.12 (2.79)

Note: Regression Coefficient (Beta) – values outside the parentheses; t - values – values inside the parentheses; n = 647; P<0.01

It can be seen from Table 10.18 that the attitude of Fourth Year high school students towards assessment directly affected their aptitude. The path coefficient of 0.12 indicates the strength of influence that SATA exerted on APT. This implies that as the mean score of Fourth Year high school students in

SATA increased (or decreased), their aptitude score also tended to correspondingly increase (or decrease). In equation form, the direct effect of SATA on APT can be represented as follows:

$$APT = \mathcal{B}_0 + 0.12(SATA) + error \quad (10.39)$$

Table 10.19. Indirect effects of student-level main factors on aptitude (Model 1 for Fourth Year high school students)

Indirect Effect	Aptitude (APT)
SSEX through SPA	- 0.01 (1%)
SPA through SATA	0.04 (4%)
Total Indirect Effects	2

Table 10.19 presents the indirect effects of student gender (SSEX) and student perception of assessment (SPA) on aptitude (APT). As shown, SSEX exerted negative influence on APT through SPA (-0.08x0.12=0.01 or 1%). The path coefficient of 1% indicates the extent of influence that SSEX had on the relationship between SPA and APT. This could mean that female Fourth Year high school students tended to have higher mean score on SPA, which thus influenced their aptitude. On the other hand, the SPA had a positive indirect effect on APT through the variable SATA (0.34x0.12=0.04 or 4%), although SPA had no direct effect on APT. The path coefficient (0.04) indicates the extent of change that SPA had on the relationship between SATA and APT. This means that as Fourth Year high school students obtained high mean score on SPA, it was likely that they obtained higher mean score on SATA, which thus led to their higher APT or NCAE scores. This result is expected as perceptions are deemed to affect attitude towards any academic activity.

10.10.1.6 Student Level – Model 2 for Fourth Year High School Students

Figure 10.6 shows the results of direct and indirect effects of model 2 student-level factors on aptitude (APT). As can be spotted, the same results as in model 1 are revealed for the effects of student gender (SSEX) on other student-level variables in model 2.

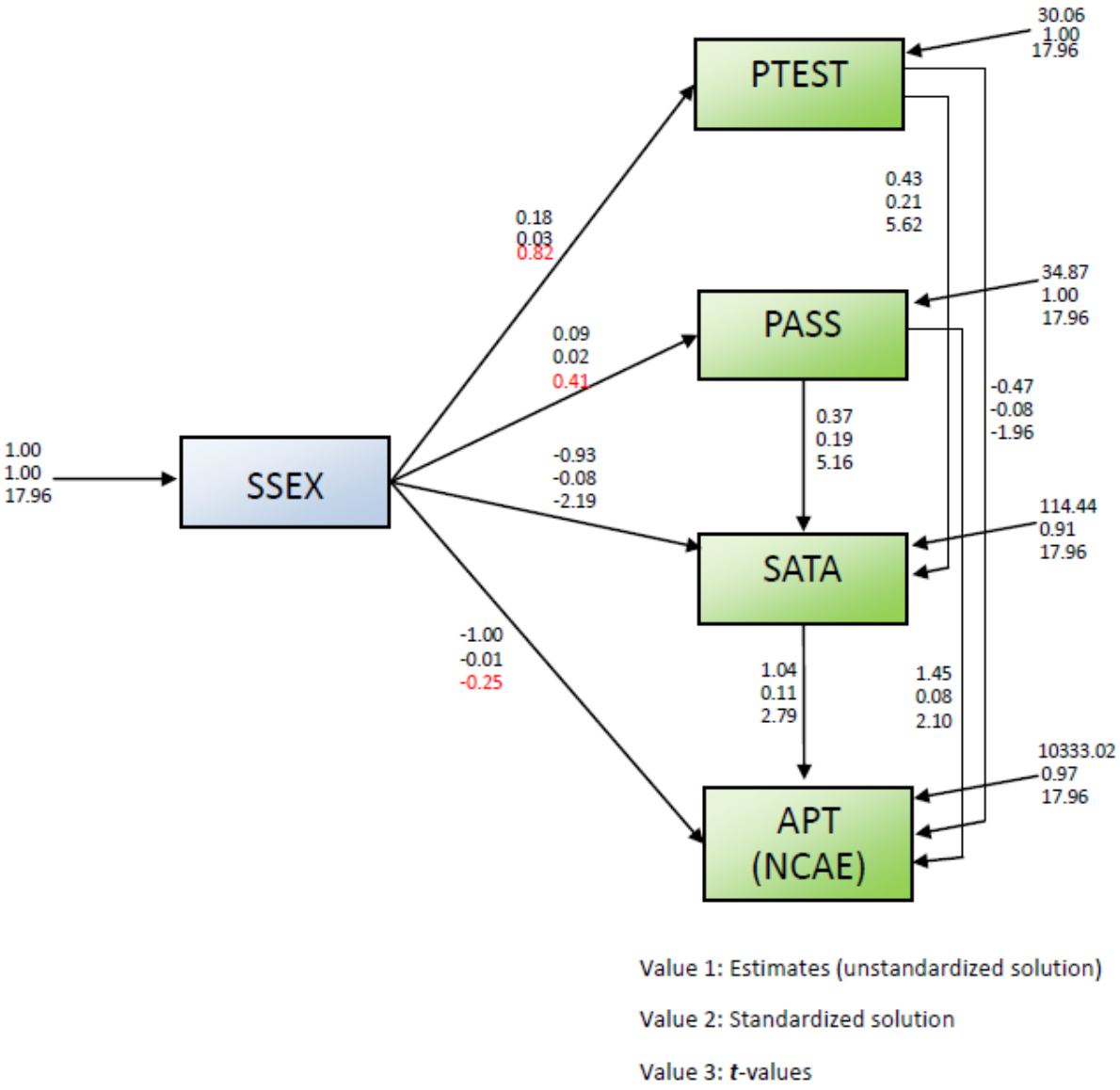


Figure 10.6. Direct and indirect effects of student-level demographic, main, and sub- factors on aptitude (Model 2 for Fourth Year high school students)

The SSEX exerted negative direct effect on student attitude towards assessment (SATA). The path coefficient of – 0.08 indicates the extent of effect SSEX exerted on SATA. This means that female Fourth Year high school students tended to obtain higher mean score on SATA than their male counterpart. Moreover, the figure shows that there are two direct effects and three indirect effects of student-level factors on APT under model 2. The factors exerting direct impact include SATA and perceptions of assignment (PASS) and the variables that had indirect effects include SSEX, PASS, and perceptions of test (PTEST). The results of direct and indirect results are summarised in Tables 10.20 and 10.21.

Table 10.20. Direct effects of student-level factors on aptitude under Model 2 (Fourth Year high school students)

Direct Effects	Aptitude (APT)
SATA	0.11 (2.79)
PASS	0.08 (2.10)

Note: Regression Coefficient (Beta) – values outside the parentheses; t - values – values inside the parentheses; n = 647; P<0.01

It can be spotted from Table 10.20 that SATA had a positive direct impact on APT with a path coefficient of 0.11. The path coefficient indicates the extent of impact that SATA exerted on APT. This means that as Fourth Year high school students gained higher mean score on assessment attitude, their aptitude or scores in the NCAE tended to increase. In other words, if the mean score of the concerned students in SATA could be increased, there is a tendency that their aptitude scores also get improved. Besides, the table shows that PASS had positive direct effect on APT with a coefficient of 0.08. The coefficient (0.08) signifies the strength of the effect that PASS transported to APT. This means that as mean scores of Fourth Year high school students in PASS increased, their APT scores tended to correspondingly improve. This result is anticipated as perception towards assignment can possibly affect scores in any test. The results on direct effect as depicted in Figure 10.6 are summarised in equation 10.40.

$$APT = \mathcal{B}_0 + 0.11(SATA) + 0.08(PTASS) + error \quad (10.40)$$

Table 10.21. Indirect effects of student-level factors on aptitude under model 2 (Fourth Year high school students)

Indirect Effects	Aptitude (APT)
SSEX through SATA	- 0.009 (0.9%)
PASS through SATA	0.02 (2%)
PTEST through SATA	0.02 (2%)

As for the indirect effects, Table 10.20 presents the results. It can be gleaned from the table that SSEX exerted negative indirect effect on APT through SATA (-0.08x0.11=-0.009 or 0.9%). The path/percentage indicates the extent of change that SSEX had on the relationship between SATA and APT.

This means that female Fourth Year high school students tended to obtain higher scores in SATA, which thus possibly influenced their scores in APT. On the other hand, the PASS had positive indirect effect on APT through SATA (0.19x0.11=0.02 or 2%). The path indicates that 2% of the relationship between SATA and APT was attributed to PASS. This result means that as Fourth Year high school students obtained high scores in PASS, their scores in SATA tended to be high, which could possibly influenced their APT scores. In addition, PTEST transported indirect effect on APT through SATA (0.21x0.11=0.02 or 2%), although PTEST had no direct effect on APT. Two-percent of the relationship between SATA and APT could be influenced by PTEST. This also means that as Fourth Year high school students obtained high scores in PTEST, they could likely obtain high scores in SATA, and thus possibly influenced their improved scores in APT.

From the path analysis results, it can be discerned that academic achievement and aptitude are affected by other student-level factors such as gender, assessment attitude, and assessment perceptions, especially those that pertain to perceptions of assignment.

10.11 Summary

This chapter dealt with regression/path analysis of the teacher-level factors and student-level factors. The analysis was done based on relevant research questions advanced in Chapters 1 and 9. The analysis commenced with regression to find the directional relationships between the factors at each of the teacher and student levels. This was followed by path analysis in which all factors in every model at each level were analysed simultaneously to examine any direct and indirect effects on the dependent variables. There were two models considered for each level. These models correspond to grouping of factors in which main factors and demographic variables were made to compose one model and all sub-factors and the same demographic variables were made to constitute another model. Within the student level, further grouping was made between Grade 6 and Second Year high school students as one group and Fourth Year High School students as another group. The teacher level and the student level factors were analysed

separately to avoid multicollinearity and bias as possibly caused by aggregation and disaggregation of data. Moreover, the students were grouped into two as they had different outcome variables.

Regression/path analysis results revealed that demographic factors such as gender, age range, academic qualification, years of teaching experience, and school type exerted influence on either teacher assessment literacy, assessment practices and teaching practices as main factors or including their corresponding sub-factors. Among the main variables, assessment literacy negatively influenced teaching practices while assessment practices positively impacted on teaching practices. No relationship was disclosed between assessment literacy and assessment practices. However, analysis of sub-variables at the teacher level showed that Standard 5 (Developing valid pupil grading procedures), an assessment literacy sub-variable, positively impacted on assessment purpose and design (sub-variables of assessment practices) while Standard 6 (Communicating assessment results to students, parents, other lay audiences, and other educators) and Standard 7 (Recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information), sub-variables of assessment literacy, negatively influenced assessment communication (sub-factor of assessment practices) and enhanced activities (sub-factor of teaching practices), respectively. Moreover, the assessment purpose appeared to positively influence structuring activities and student-oriented activities (sub-variables of teaching practices) while the assessment communication appeared to positively impact on all sub-factors of teaching practices. From these results, it can be generally traced that teacher assessment literacy somehow affected assessment practices, which, in turn, impacted on teaching practices. At the student level, gender appeared to negatively affect the assessment attitude and academic achievement of Grade 6 and Second Year high school students. These students' perceptions of assessment also appeared to positively influence their attitude towards assessment. Their specific perceptions of assignment likewise exerted positive impact on their academic achievement. Similarly, gender appeared to negatively influence the attitude of Fourth Year high school students towards assessment. Their perceptions of assessment positively impacted on their assessment attitude. In addition, the aptitude of these students was positively influenced by their

perceptions of assignment and attitude towards assessment. The results generally indicated that other student-level factors such as gender, assessment perceptions, and assessment attitude could affect academic achievement, aptitude, or both.

It was pointed earlier that relationships among the teacher-level factors were examined separately from those of the student-level factors. This was due to the nested or hierarchical nature (student-level factors nested within teacher-level factors) of the data collected for this study and to the challenges associated with SEM in analysing multilevel data. To address the multilevel nature of the data and SEM limitations, and to properly investigate the possible effect of teacher assessment literacy on the outcome variables through the intervening variables at the teacher and student levels, further analysis was carried out employing multilevel technique. The next chapter (Chapter 11) highlights the challenges associated with SEM in multilevel data analysis and deals with hierarchical linear modeling (HLM) analysis.

Chapter 11: Multilevel Analysis of the Tested Factors

11.1 Introduction

This study has the broad aim of examining teacher assessment literacy and its impact on student achievement and aptitude through the mediating and moderating variables at the teacher and student levels. Specifically, the study attempted to answer Question 9 (How does teacher assessment literacy interact with demographic factors, assessment practices, teaching practices, student perceptions of assessment, student attitude towards assessment, student achievement, and student aptitude?) as posed in Chapters 1, 9 and 10. To address the study's aim and the relevant research questions, it was necessary to subject the data to further analysis.

In Chapter 11 the questions concerning the relationships among variables at the teacher level and student level were addressed. Factors were grouped by level, and each of these levels was analysed separately using multiple regression/path analysis to examine the directional influences as hypothesised or reflected in the relevant research questions. In this chapter, the justification was made that the teacher-level factors were not combined with student-level factors in a single-level procedure due to the limitations of the structural equation modeling, particularly the multiple regression/path analysis, in handling the gathered data. The data collected in this study had the hierarchical characteristics, which is typical of any educational data. The attributes at the student level were deemed nested within the characteristics at the teacher level and when all the factors from these levels are combined, the analysis should take the nature of these data into account.

The criticisms in using single-level methods such as multiple regression/path analysis to study multilevel phenomena are on their limitations in taking into account the structure or clustering levels of the

variables under study. In these traditional linear methods, two approaches are usually carried out to deal with multilevel data: aggregation and disaggregation of data. The aggregation approach involves the process of raising the low-level data to the high-level data; conversely, the disaggregation method involves the process of bringing down the high-level variables to the low-level variables (Osborne, 2000; Lee, 2000; Beretvas, 2004; Guo, 2005). These methods are considered problematic in treating hierarchical data as they often lead to misleading and erroneous results (Raudenbush & Bryk, 1986; Snijders & Bosker, 1999). The aggregation approach has a number of issues that include the loss of information and important variation among the low-level variables (Osborne, 2000; Guo, 2005). On the other hand, the disaggregation strategy tends to violate the independence assumption as members of a group at the low level assume the same scores (Beretvas, 2004; Osborne, 2000).

Bryk and Raudenbush (1992) and Raudenbush and Bryk (2002) listed three most commonly encountered difficulties in analysing multilevel data when using single-level analytic methods. These difficulties are aggregation bias, misestimated standard errors, and heterogeneity of regression. The aggregation bias occurs when a variable assumes different meanings, and therefore, has different effects at different levels of aggregation (Lee, 2000; Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). A variable that is aggregated/disaggregated becomes a high- or low-level unit, respectively, and shifts in meaning resulting to different effects and interpretations (Snijders & Bosker, 1999). Moreover, the misestimated standard error happens when the dependence among individual responses within the group or classification is not taken into account (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002; Lee, 2000). This misestimated standard error can lead to serious risks of committing type 1 error for the between-group differences (Snijders & Bosker, 1999). Furthermore, the heterogeneity of regression takes place when the relationships between individual characteristics & outcomes vary across groups (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). The variation in the relationships can perhaps be attributed to group-level variables (Lee, 2000), which, when not considered in the analysis, may lead to invalid inferences (Raudenbush & Bryk, 1986). Hence, in consideration of the drawbacks of single-level methods

and the nature of data gathered in this study, the multilevel analysis employing hierarchical linear modeling (HLM) was carried out.

This chapter deals with the procedure and the results of the HLM analysis of the tested factors. Specifically, it begins with the description of the HLM and HLM software (version 6.08) to provide the background on the statistical technique and software employed in the analysis. The chapter continues with the presentation of the proposed model and analysis framework. After which, the results of the two-level model are presented and discussed. The chapter ends with a summary of key points.

11.2 Overview of HLM

The term *hierarchical linear model* (HLM) was adopted by Raudenbush and Bryk (1986) to refer to the analytic method that permits modeling of multilevel phenomena, such as those encountered in educational research. This method is an extension of multiple regression model (Snijders & Bosker, 1999; Ma, Ma, & Bardley, 2008). It is known in other terms as 'multiple linear models', 'mixed-effects models', 'random-effects models', 'random coefficient regression models', and 'covariance component models' (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). HLM is also synonymous with multilevel modeling or random coefficient modeling (Ciarleglio & Makuch, 2007).

The hierarchy reflected in the multilevel data is composed of units grouped at different levels (Goldstein, 2011). For instance, in the two-level model concerning students and teachers, individual students may be assigned as level-1 units while teachers can be designated as level-2 units. This notion considers students who occupy the first hierarchical level as clustered or nested within teachers who are located in the second level of the hierarchy (Roberts, 2004). Other group, such as school, in which students and teachers are also nested may be added as level-3 units making it a three-level model. This concept of hierarchy is applied in the context of HLM. In the case of two-level model, there are other terms that have been used to mean level-1 and level-2 units. Some authors have used 'individuals, within-group, low level,

and micro level' to refer to level 1, and 'group, between-group, high level, and macro level' to mean level 2. In this study, these terms are adopted and used interchangeably.

A number of experts have justified the use of multilevel modeling techniques such as HLM. In the context of this study, those expounded by Braun, Jenkins, & Grigg (2006) can be adopted. As stressed by these authors, conventional regression techniques either treat the group (e.g. teachers) as the unit of analysis, ignoring the variation among the individuals (e.g. students) within teachers, or treat the students as the unit of analysis, ignoring the nesting within teachers. As mentioned earlier, neither of these approaches is satisfactory and thus warrants the use of HLM (Osborne, 2000). Roberts (2004, p. 31) further provided three reasons why multilevel models are preferred. "First, statistical models that are not hierarchical sometimes ignore the structure of the data and as a result report underestimated standard errors (no between unit variation), thus resulting in increased type I error…second, multilevel techniques are much more statistically efficient than other techniques… and third, multilevel techniques assume a general linear model, and as such, can perform multiple types of analyses that provide more conservative estimates by allowing for correlated responses within clusters". Moreover, Luke (2004) emphasised that most of what we study, and especially in education, have multilevel character and so we should employ theories and analytic techniques that are also multilevel.

The HLM approach responds to the challenges associated with the single-level analysis as it takes into account the hierarchical character of data (e.g. educational data), which is often ignored in the traditional linear methods (Raudenbush & Bryk, 1986). It is, thus, a useful technique for analysing hierarchical, nested, or clustered data (Ciarleglio & Makuch, 2007; Woltman, Feldstain, MacKay, & Rocchi, 2012; Beretvas, 2004). Guo (2005) described HLM as a flexible and versatile method that can be used to answer questions in various research contexts. According to Braun, Jenkins, & Grigg (2006), HLM is more flexible because it involves two or more sets of linear regression equations that can incorporate predictor variables at each level of the data structure. The strength of HLM is further provided by Raudenbush and Bryk, (1986) who defined it as a powerful tool that permits separation of within-group (e.g. student

characteristics) for between-group (e.g. teacher characteristics) phenomena and allows simultaneous considerations of the effects of teacher factors not only on teacher means but also on structural relationships among students. This nature of HLM makes it more efficient in estimating for variance among variables at different levels than other existing analyses (Woltman, et al., 2012). As HLM appropriately provides direct effects from various levels and interaction effects between variables at different levels, this approach was employed in this study. Specifically, HLM was used to address three purposes: (a) improve estimation of individual effects; (b) model cross-level effects; and (c) partition variance-covariance components (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002).

However, like any other analytic techniques, HLM is not without limitations. HLM approach assumes only one dependent variable at the individual level of the hierarchy to be predicted by a number of independent variables at different levels (Hox, 2010; Kreft, de Leeuw, & Kim, 1990). This implies that it only allows one outcome variable to be analysed at any one time. This is an issue, especially in educational research, as there is often more than one outcome variable to be dealt with in educational phenomena (Kreft, et al., 1990). In addition, HLM is intended for observed variables, and although it does allow for latent variables, "it requires unrealistic assumptions about the underlying measurement model" (Scientific Software International, n.d., p. 1). However, the use of structural equation modeling, particularly the factor analytic method, should help address this second limitation (Goldstein, 2011). This is possible as the principal component scores (latent scores) for each construct involved in the models can already be calculated using other applications such as SPSS, MS Excel, and ConQuest (Ben, 2010).

11.3 Assumptions of HLM

According to Atkins (2010), every statistical model has assumptions, and testing these assumptions almost always involves examination of the residuals. The author stated that in HLM, residuals at different levels can be used to assess normality of error terms (student-level residuals and empirical Bayes residuals at teacher-level in this study) and equal variances (student-level residuals on fitted values, in this study).

The error terms need to be normally distributed to avoid biased standard errors at both within-group (e.g. students) and between-group (e.g. teachers) levels and consequently inaccurate computation of confidence intervals and hypothesis tests; errors in the within-group level also need to have equal variance to avoid inefficient estimates and biased standard errors in the between-group (e.g. teachers) level. (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). Further discussion on residuals in HLM is given in Section 11.8.

In addition to the examination of residuals, predictor and/or outcome variables need to be normally distributed to avoid biased HLM output (Woltman, et al. 2012). The predictor variables also need to be independent of their level-related errors and error terms should be independent of each other (Woltman, et. al., 2012; Richter, 2006). Moreover, the absence of multicollinearity (Woltman, et al., 2012), in which high or extreme correlation should not exist between two or more predictor variables, likewise needs to be examined. These independence and no-multicollinearity assumptions should be addressed for robust analysis and more accurate output. In this study, the HLM assumptions were checked while analysing and building the model and were deemed met.

11.4 Model Building in HLM

Generally, the initial step in building a multilevel model is to assess whether a multilevel model is needed in the first place. According to Luke (2004), this can be determined through empirical, statistical, and theoretical considerations. As expounded by this author, the empirical and statistical justifications are in relation to the variation in the outcome variable and the violation of independence assumption, respectively. When there are evidences that the variance in the dependent variable is attributed to the groups or factors at the macro level, and that the independence assumption is violated as is often the case when using single-level methods such as disaggregation of data, then a multilevel model is needed. Moreover, the theoretical justification is based on the study's theoretical framework or hypotheses. Luke (2004) further stressed that the theoretical propositions that involve constructs or variables operating and interacting at

different levels also warrant multilevel analysis of the data. The empirical and/or statistical evidence can be provided through the analysis of the unconstrained model while the theoretical justification is determined by the researcher. This study used these justifications in employing and running HLM analysis.

There are two general strategies when building a multilevel model: *top-down* and *bottom-up* strategies. The top-down approach begins with a complex model that includes the maximum number of variables. All these variables are included in the analysis and those that are found insignificant are successively removed from the model. As this approach starts with a large and complicated model, it requires longer computation time and is sometimes fraught with convergence problems. The opposite of this approach is the bottom-up strategy. The bottom-up procedure starts with a simple model and proceeds by adding variables one at a time. A variable that has insignificant effect is usually excluded from the model (Hox, 2010; Darmawan & Keeves, 2009). The advantage of the bottom-up approach is that it leads to a parsimonious model (Hox, 2010). In addition, the bottom-up strategy is more productive, allows identification of best predictors, and tends to avoid multicollinearity problems (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). As recommended by experts and on the basis of its advantages, this study adopted the bottom-up strategy in building and assessing HLM.

The 'bottom-up' strategy, with steps given by Lee (2000), can be adopted to build and evaluate HLM models. The first step is the creation and analysis of the fully unconditional model or what is commonly called as a *null* model. The null model is the simplest hierarchical linear model that contains no explanatory variables from any level of the hierarchical structure of the data (Darmawan & Keeves, 2009; Luke, 2004). This model is equivalent to a one-way analysis of variance (ANOVA) with random effects; it estimates and allows for the partitioning of variance in the dependent variable in each of the hierarchical levels (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). The purpose of running the null model is to obtain the empirical and/or statistical evidence to decide whether HLM is needed and to estimate other coefficients, such as those that can be used for model comparison (Roberts, 2004; Lee, 2000; Richter, 2006; Darmawan & Keeves, 2009). After running the null model, the next step is to set up

level-1 model. In this step, all level-1 predictors that are associated with the outcome variable are entered one at a time to estimate the unique contribution of each of them in the model (Roberts, 2006; Richter, 2006). The predictor variables that yield significant results are retained while those that are insignificant are typically excluded. Once the level-1 model is satisfactory, the potential explanatory variables for level 2 are then examined. In this step, the estimation of level-2 model predictors (e.g. the outcome is explored as a function of teacher characteristics), including cross-level interactions, is carried out (Lee, 2000; Luke, 2004; Richter, 2006). Similarly, all the variables are entered successively and those that have the significant contributions are included while those that are insignificant are discarded. Additional step and similar process can be done for the next higher level if analysing three-level model. Presented and described below are the general equations of the null, level-1, and level-2 models. The equations are up to level 2 as this study employed two-level hierarchical linear model (2L/HLM).

Null Model

The equation form of the null model is expressed in terms of the Level 1 and Level 2. These parts are as follows:

Level-1 Part of the Null Model.

The outcome variable is represented as a function of a predictor mean plus a random error. This is presented in the equation,

$$Y_{ij} = \beta_{0j} + r_{ij} \quad (11.1)$$

Where:

Y_{ij} represents the outcome variable;
β_{0j} is the level-1 coefficient; and
r_{ij} is the level-1 random effect.

The indices *i* and *j* denote level-1 units (e.g. students) and level-2 units (e.g. teachers), respectively, where there are

i = 1, 2, … N, students within J teachers; and

j = 1, 2, …, J teachers.

Level-2 Part of the Null Model.

In this model the level-1 coefficient, β_{0j}, becomes an outcome variable as shown in the following equation.

$$\beta_{0j} = \gamma_{00} + u_{ij} \tag{11.2}$$

Where:

γ_{00} is a level-2 coefficient; and

u_{ij} is a level-2 random effect.

When the level-1 and level-2 predictors are included, the above equations take the form of multiple linear regression equation where *Y* is the outcome (or dependent) variable and the *X*'s and W's are the predictors (or independent) variables.

Level-1 model:
$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \cdots + \beta_{qj}X_{qij} + r_{ij}$$
$$= \beta_{0j} + \sum_{q=1}^{Q} \beta_{qi}X_{qij} + r_{ij} \tag{11.3}$$

Where:

β_{qj} (q = 0, 1,..., Q) are level-1 coefficients;
X_{1ij} , X_{2ij} , X_{Qij} are level-1 predictors for case *i* in unit *j*; and
r_{ij} is the level-1random effect.

Level-2 model:
$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + \cdots + \gamma_{0S_0}W_{S_{0j}} + u_{ij}$$
$$= \gamma_{00} + \sum_{S=1}^{S_0} \gamma_{0s} W_{sj} + u_{0j} \tag{11.4}$$

Where:

γ_{os} (o = 0, 1,..., S_o) are level-2 coefficients;

W_{sj} are level-2 predictors; and

u_{0j} is a level-2 random effect.

The equations for level-1 and level-2 models as applied to the 2L/HLM tested in this study are further illustrated in the following sections.

11.5 HLM 6.08 Software

There are a number of software packages that can be employed to analyse multilevel model or data. One of these applications is the HLM (version 6.08) software (Raudenbush, Bryk & Congdon, 2009), which was employed in this study.

One of the leading statistical packages for hierarchical linear modeling, HLM has the capability to fit models to outcome variables that generate a linear model with predictor variables to which variations at each hierarchical level can be attributed; the software does not only estimate model coefficients but also predicts random effects associated with each sampling unit at every level (Raudenbush, Bryk & Congdon, n.d. as cited in Ben, 2010). Moreover, this application provides more information when compared with other programs; it gives a variety of tests/estimates such as t-test, chi-square test, reliability estimate, deviance statistic, and *p*-values which minimises the effort of the user from calculating some of them. Furthermore, the many examples provided and the educational character of HLM manual makes the use of this software more easy and applicable (Kreft, et al., 1990), especially in educational research. As de Leeuw (1992, p. xv) stated, "the program HLM, by Bryk and Raudenbush, was the friendliest and most polished of these products, and in rapid succession a number of convincing and interesting examples were published."

Through the years, HLM has progressed in the development of its capabilities and functionalities. The latest version of this software is highly compatible with the latest Windows operating systems. In addition, it provides, among other things, a wide choice of estimation options and can already handle three- and four-level models (Garson, n.d.). Also, since HLM reads data under a particular format from an external source, its importing capabilities have also been enhanced by being able to read data not only from a plain

text (ASCII) format but also from data saved in the latest SPSS/PASW and other statistical software (Ben, 2010).

In this study, HLM 6.08 was used to run and analyse the 2L/HLM. Specifically, the software was employed to estimate the effects of level-1 variables on the outcome variable (student level), and to estimate the effects of the level-2 variables (teacher level) on the coefficients of level-1 variables, and on the response variable in level 1.

11.6 Data and Variables Analysed in HLM

The data subjected to HLM analysis were taken from the responses of 582 teachers and 2,077 students from Grade 6 (elementary level), Second Year and Fourth Year (secondary level) high school classes. These data were collected during the school year (S. Y.) 2010-2011. The teachers and students involved in this study came from the public and private elementary and secondary schools in the province of Tawi-Tawi, Philippines.

The data gathered through questionnaires were first obtained in the form of raw scores. These raw scores were transformed into measures, except for the categorical variables. To transform raw scores into measures, the ability estimation technique introduced by Warm (1989) called the weighted likelihood estimation (WLE) was used. The calculation of WLE was performed using ConQuest 2.0. The WLE scores were further converted to W scores (Woodcock, 1999) using the formula, $W = 9.1024$ (WLE logits) + 500. The computation for the W scores was carried out using Microsoft Excel. In other words, the calculated W scores (principal component scores) as measures for each of the factors or variables involved in the HLM analysis were standardised scores with a mean of 500. These standardised scores made possible the direct comparison of coefficients of the different variables within the model. As for the categorical variables, they were treated as *dummy variables*. The dummy variables are described in a separate subsection. Moreover, the academic achievement scores and the aptitude scores were the secondary data taken from the results of the National Achievement Test (NAT) and the National Career Assessment Examination, respectively,

which were conducted during the school year 2010-2011. During this school year, the NAT was administered to Grade 6 (elementary level) and Second Year high school while the NCAE was administered to Fourth Year high school. The academic achievement and aptitude data were also standardised scores with a mean of 500 and a standard deviation of 100. As these scores were from the standardised tests and as such tests are usually designed to provide results that are nearly normal (Raudenbush & Bryk, 1986), this adds weight to the results of normality test that the outcome measures for this study met the assumption of normal distribution. As for the other continuous variables tested in this study, it was found from the results of the normality test that they were nearly normally distributed and as such the assumption of normality is deemed satisfied.

11.6.1 Dummy Variables and Coding

A *dummy variable* is a variable that indicates or represents an attribute variable. It is thus sometimes labeled as *indicator variable* (Skrivanek, 2009). Hardy (1993) described it as a dichotomous variable that the researcher usually creates from an originally qualitative variable. It can be used, among other applications, in the analysis of qualitative data from survey and in the representation of categories and value levels (Garavaglia & Sharma, 2004; Baker, 2006). The use of dummy variables is pointed out by Hardy (1993, p. 2) as follows:

> "When independent variables of interest are qualitative (i.e., "measured" at only the nominal level), we require a technique that allows us to represent this information in quantitative terms without imposing unrealistic measurement assumptions on the categorical variables...Defining a set of dummy variables allows us to capture the information contained in a categorization scheme and then to use this information in a standard estimation. In fact, the set of independent variables specified in a regression equation can include any combination of qualitative and quantitative predictors."

Dummy variables serve as a powerful and useful tool for analysis (Polissar & Diehr, 1982), such as in the case of regression and/or HLM analyses. Using dummy variables in regression allows characterisation of

subsets of observations, “easy interpretation and calculation of the odds ratios, and increases stability and significance of the coefficients”. It also makes it easier to use the model as a decision tool (Garavaglia & Sharma, 2004, p. 1). Moreover, employing dummy variables in the HLM analysis is useful for the interpretation of the results (Ben, 2010). The data gathered in this study were obtained from a cross-sectional survey and as such there is a likely occurrence of heteroscedasticity (errors or residuals at each level of the hierarchy have unequal variances). Dummy variables can be used in the cross-sectional data to estimate differences between groups and to assess whether group membership moderates the effects of other predictors (Hardy, 1993). Thus, in multilevel analysis, using dummy variables would allow a separate level 1 variance for the nominal/categorical variable from which they were created (Goldstein, 2011).

The procedure of creating dummy variables is to adopt the so-called *dummy coding*, which is a way of representing variables or factors using the binary coding of zeros and ones (Field, 2009). The code “1” indicates the “presence” (e.g. the attribute is present or there is a membership) and “0” indicates the absence (e.g. absence of the attribute or non-membership) in a particular category (Hox, 2010; Skrivanek, 2009; Baker, 2006). The code “0” is also used to indicate a reference or baseline category against which all other categories are compared. For instance, a dummy variable can be created for the nominal variable, “Gender”. This dummy variable can either be “Boy’’ or “Girl’’. A male respondent (“Boy”) can be assigned a code of “1’’ and a female respondent (“Girl” or NOT “Boy”) can be assigned a code of “0’’. In this example, the attribute is whether “Boy” or “Not Boy” or if “Boy” or “Not Boy” is belonging to a group. Moreover, the group that is composed of girls is a reference group against which the male group can be compared. The number of dummy variables or predictors is equal to the number of categories minus 1 (Field, 2009; Richter, 2006).

The dummy variables subjected to HLM analysis in this study were created from the demographic or nominal variables using the relevant procedure and the codes as described above. Specifically, the dummy variables were coded as follows: for teachers’ gender (TSEX) and students’ gender (SSEX), they were composed of females (TFMALE and SFMALE for teachers and students, respectively) who were

assigned a code of 0 and males (TMALE and SMALE for the respective groups) who were given a code of 1; teachers' age range (AGE) was of six categories corresponding to age ranges of under 25 (AGE1) years, 25-29 years (AGE2), 30-39 years (AGE3), 40-49 years (AGE4), 50-59 years (AGE5), and years of 60 and above (AGE6). For this demographic factor, five dummy variables were created and designated with the same codes of zeroes and ones; for teachers' academic qualification (ACAD), groups were divided between bachelors (UNDERGRAD) that were coded 0 and postgraduates (POSTGRAD) that were coded 1; teachers' years of experience (EXYEAR) were of seven groups corresponding to the assigned ranges of years of experience that included 1-5 years (EXY1), 6-10 years (EXY2), 11-15 years (EXY3), 16-20 years (EXY4), 21-25 years (EXY5), 26-30 years (EXY6), and above 30 years (EXY7). For this demographic factor, six dummy variables were created using the same coding scheme; and for the school type (SCHTYPE), groups were between the private schools (SCH_PRIV) that were coded 0 and the public schools (SCH_PUB) that were coded 1.

11.6.2 Mediating and Moderating Variables

Based on this study's theoretical framework and questions, intervening and/or moderating factors were involved in the analysis of directional relations among the tested variables. As such, these types of variables are described below to provide conceptual background.

The *mediating* and *moderating* variables are variables that generally affect the link between factors. They are considered as tools that can be utilised to enhance a deeper and more refined understanding of the directional relationship between independent and dependent variables (Wu & Zumbo, 2008). Specifically, a *mediating* variable is a third variable that intervenes between predictor and criterion variables (Baron & Kenny, 1986). It is described as a 'bridge' or a 'mechanism' through which one variable influences or affects another variable (Rose, Holmbeck, Coakley, & Franks, 2004; Wu & Zumbo, 2008). For this reason, it is also called *mediator* or *intervening* variable. The effect of mediating variable on another variable is also called *indirect effect*, *surrogate effect*, *intermediate effect*, or *intervening effect* (MacKinnon,

Lockwood, Hoffman, West, & Sheets, 2002). According to Wu and Zumbo (2008, p. 373), a mediator is a temporary or a relatively less stable construct; it is a "responsive variable that changes within a person". Thus, characteristics such as practices and perceptions can be mediating variables. On the other hand, a *moderating* variable (or a moderator) is a third variable that affects the direction and/or strength of the relationship between independent and dependent variables (Baron & Kenny, 1986; Rose, et al., 2004). It enhances, weakens, or modifies the strength and direction of the relationship between variables (Kim, Kaye, & Wright, 2001; Wu & Zumbo, 2008). The effect of moderating variable on another variable is called an *interaction effect*. However, while moderation effect suggests a causal relationship, interaction effect does not necessarily be causal in nature. In other words, a moderation effect can be an interaction effect but an interaction effect does not need to be a moderation effect. In addition, a moderating variable is typically an innate attribute, a relatively stable trait, or a relatively unchangeable background, environmental or contextual variable (Wu & Zumbo, 2008). Thus, factors such as gender and school type can be considered moderating variables. Table 11.1 below presents the variables that were subjected to HLM analysis in this study.

Table 11.1. List of variables used in the two-level HLM

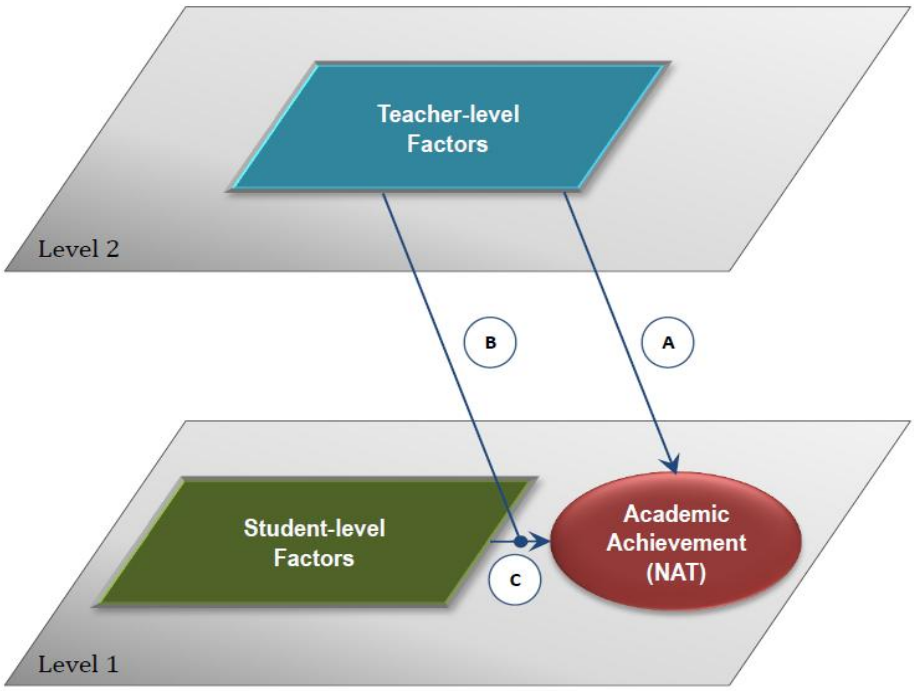
Hierarchical Level	Variable	Description
Level 2 (Teacher Characteristics)	TSEX	Teachers' gender: Male/Female
	AGE	Age range: Age1 (under 25 years), Age 2 (25-29 years), Age 3 (30-39 years), Age 4 (40-49 years), Age 5 (50-59 years), and Age 6 (60 years & above)
	ACAD	Academic qualification: Bachelors/undergraduate and Postgraduate
	EXYR	Years of teaching experience: Years of experience 1 (1-5 years), Years of experience 2 (6-10 years), Years of experience 3 (11-15 years), Years of experience 4 (16-20 years), Years of experience 5 (21-25 years), Years of experience 6 (26-30 years), and Years of experience 7 (above 30 years)
	SCHTYPE	School Type: Private/Public
	ASLIT	Assessment Literacy: Standard 1 – Standard 7
	ASPRAC	Assessment Practices: Purpose, Design, & Communication
	TPRAC	Teaching Practices: Structure, Student Orientation, & Enhanced Activity
Level 1 (Student Characteristics)	SSEX	Students' gender: Male/Female
	SPA	Student Perceptions of Assessment: Perceptions of Test & Perceptions of Assignment
	SATA	Student Attitude Towards Assessment
	ACHIEV	Academic Achievement
	APT	Aptitude

11.7 The Model and Analysis Framework

The HLM that was examined in this study was a two-level model. The 2L/HLM was considered on the basis of this study's theoretical framework in which teacher attributes and student characteristics were proposed to constitute two separate levels. Nevertheless, the possible three-level HLM (3L/HLM) was tested because of the presence of school type as one of the categorical variables. As has been tested in the

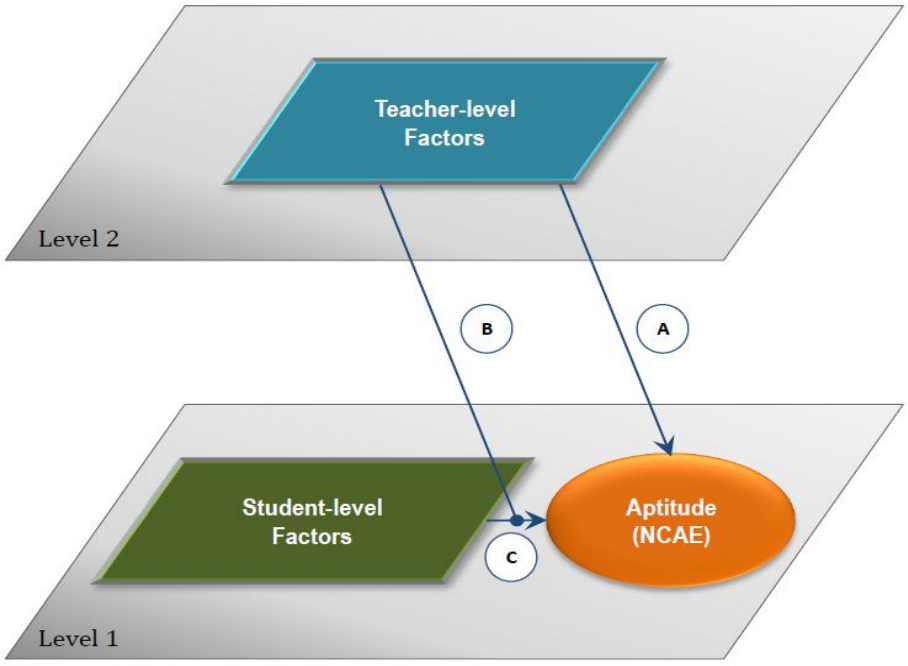
previous studies, school type can be used as a grouping variable under a school level, the third possible hierarchical level in this study. However, when the 3L/HLM was run, the results showed that the three-level model was not the structure of the data. Hence, the study proceeded with the analysis of the originally proposed 2L/HLM and the school type was made part of the teacher level.

Under the 2L/HLM considered in this study, the same levels as used in Chapter 11 were analysed. These levels were the teacher and student groups. The variables under each level are presented in Table 11.1. However, while simultaneously using teacher and student levels, two separate HLM analyses were carried out. One analysis involved teacher characteristics (level-2 model) and student characteristics (level-1 model) of Grade 6 and Second Year high school students. The other analysis included the same teacher attributes (level-2 model) but with student characteristics (level-1 model) of Fourth Year high school students. These two independent analyses were performed, as there were different outcome variables for the two groups of students. The group involving Grade 6 and Second Year high school students was having “academic achievement” as the dependent variable while the group involving Fourth Year high school students was having “aptitude” as its response variable. The conceptual models for the 2L/HLM for the two analyses are depicted in Figures 11.1 and 11.2. As shown in these figures, the teacher-level characteristics were hypothesised to mediate and/or moderate, and directly impact on the outcome variables while the student characteristics were proposed to directly affect the dependent variables.



A – Level 2 Variables effect on the Outcome variable
B – Cross level interaction between Level 2 Variable and Level 1 Variable
C – Level 1 Variables effect on the Outcome Variable

Figure 11.1. Two-level HLM with academic achievement as the outcome variable



A – Level 2 Variables effect on the Outcome variable
B – Cross level interaction between Level 2 Variable and Level 1 Variable
C – Level 1 Variables effect on the Outcome Variable

Figure 11.2. Two-level HLM with aptitude as the outcome variable

11.8 Model Building and Analysis Using HLM 6.08 Software

The analysis of multilevel model such as 2L/HLM using HLM 6.08 software usually involves three general steps. These are as follows: (1) import the data set into HLM 6.08 to create a multivariate data matrix (MDM) file; (2) run the analysis using this MDM file, and (3) evaluate the fitted model based on relevant estimates and residual file. The MDM file is constructed from raw data saved in common formats such as SPSS. Typically for a 2L/HLM, two raw data files corresponding to the two levels are required and recommended as input. In other words, one data file is prepared for level 1 and another data file is prepared for level 2. These two data files are linked by the level-2 ID variable (the Teacher ID in this study). The MDM file produced is used as input in all subsequent analyses. The MDM file can be viewed as a "system file" in a standard computing package that contains both the summarised data and the names of all the variables (Ben, 2010).

There is an important issue with regard to the treatment of variables when analysing HLM using HLM 6.08 (also true to other versions). This pertains to the information about the variables. In the analysis of multilevel model such as HLM, information related to all variables from different levels is central and it is essential that the meaning of these variables is well understood. The meaning depends on the locations or what is called *centering* of the involved factors. Thus, it is also vital to be familiar with the proper choice of centering to allow meaningful interpretation of results and numerical stability in estimating HLM (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002).

Bryk and Raudenbush (1992) and Raudenbush and Bryk (2002) listed possible ways of centering variables. The two commonly considered centering options are *the group mean centering,* in which the mean of the group is subtracted from the corresponding individual scores, and the *grand mean centering*, in which the overall mean is subtracted from all values of a variable (Hox, 2010). The group mean centering is usually considered when the analysis is supported by a strong theory. Specifically, this centering option is recommended when the research hypotheses involve the analysis of relationship between level-1 predictors and when determining the moderating effects of level 2 predictors on the strength of first-level relationship

(Hox, 2010; Luke, 2004). Moreover, it is used for level-1 variables to allow examination of the independent effects of level-1 and level-2 predictors (Woltman, et al., 2012). On the other hand, the grand mean centering is also recommended as it a useful option. This option facilitates meaningful interpretation of the multilevel analysis results. Particularly, it helps solve the problem of difficult interpretation of the intercept as the expected value of the outcome variable (Hox, 2012).

The estimation method also needs to be decided. HLM uses a full maximum likelihood or restricted maximum likelihood method together with an empirical Bayesian method to estimate the fixed and random effects in the model (Guo, 2005). The full maximum likelihood is an estimation that follows the target criterion to maximize the combined likelihood of the fixed parameters and variance components; the restricted maximum likelihood is a method in which the target criterion is the likelihood of the estimates for the variance components only (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). The restricted maximum likelihood has been found more robust in providing estimates. However, when analysing large samples, its difference with the full maximum likelihood is negligible. One advantage of using full maximum likelihood is that it allows hypothesis testing or model comparison using deviance statistic (Richter, 2006; Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002; Darmawan & Keeves, 2009).

The HLM can be evaluated by looking at the values of the chi-square test (χ^2), *intraclass correlation coefficient* (ICC), *reliability* of intercepts and slopes, and *deviance statistic*. The χ^2 is a significance test and indicates the existence of variance in the dependent variable that is attributable to level-2 factors. But the ICC that is examined through the null model is formally determined to decide whether HLM is needed. The empirical/statistical evidence can be provided by the ICC, a measure of the proportion of variance in the outcome variable that is attributed to level-2 units (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002; Guo, 2005; Richter, 2006). The ICC can be calculated by dividing the variance of level-2 units by the sum of the variances of level-1 and level-2 units (Guo, 2005). The high ICC indicates that much variance in the outcome variable is due to groups and as such a multilevel model is needed (Guo, 2005; Luke, 2004). According to Lee (2000), an ICC greater than 10% of the total variance in the outcome warrants the use of

HLM. Guo (2005) cited the ICC threshold of 0.25 and above as evidence for the multilevel model analysis to proceed. In addition, a high ICC is indicative of the existence of the nested structure of the data and, therefore, a violation of the independence assumption, further implying that HLM analysis needs to be carried out (Luke, 2004). The other important part of the output is the reliability of intercepts and slopes. Reliability here is viewed in terms of the degree of variability present between groups compared to total variability (i.e., between-group variance plus error variance) (Atkins, 2010). The concept of reliability is reiterated in Section 11.9. The threshold for reliability is 0.05. Any reliability that is below this threshold causes a variety of numerical difficulties and so parameter being estimated should be fixed (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002; Darmawan & Keeves, 2009). Furthermore, there is a need to look at *deviance statistic*. Deviance statistic is defined as a measure of fit between the data and the model (Luke, 2004; Darmawan & Keeves, 2009). It is a transformation of the likelihood statistic and is obtained by multiplying the natural log of the likelihood by minus two (-2LL); the deviance can be utilised to compare models (Luke, 2004; Roberts, 2006; Darmawan & Keeves, 2009), in which the significant reduction in the value of deviance indicates better fit of the model (Darmawan & Keeves, 2009).

In regression equation analysis, the *error* term (or *residual)* is a common sight. This is sometimes denoted by "*e*'' or "*R*''. The residual expresses the part of the dependent variable "*Y*'' that cannot be approximated by a linear function of that dependent variable (Snjider & Bosker, 1999). In other words, in multilevel modeling, residuals represent the unexplained variance in each level of the model.

The fit of HLM is also evaluated through the analyses of level-1 and level-2 residual files for tenability of relevant assumptions. Level-1 residual file contains: (a) the level-1 errors which show the differences between the observed and the fitted values, (b) fitted values for each level-1 unit, (c) the observed values of all predictors included in the model, and (d) selected level-2 predictors necessary for exploring possible relationships between such predictors and level-1 residuals. Level-2 residual file includes a number of important information such as the fitted values for each level-1 coefficient, which are the values predicted on the basis of the level-2 model. This residual file also includes information about the

discrepancies between the level-1 coefficients and the fitted values using the ordinary least squares (OLS) and the empirical Bayes (EB) estimates of the level-2 residuals (Ben, 2010; Atkins, 2010).

Analysis of 2L/HLM in this study follows the procedure as mentioned in Sections 11.4 to 11.8. For the level-1 model, 'academic achievement' for the group involving Grade 6 and Second Year high school students and 'aptitude' for the group involving Fourth Year high school students were designated as the outcome variables. For the predictor variables in this model, all student-level variables were tested for the two groups. The level-2 model for both groups includes all factors related to teacher characteristics. The results of the 2L/HLM are presented and discussed in Section 11.9.

11.9 The Results of the Two-level Model

In this section, the results of the 2L/HLM are presented and discussed. The 2L/HLM for the group involving Grade 6 and Second Year high school students (hereafter referred to as Group 1) is presented first, followed by the same model for the group involving Fourth Year high school students (hereafter referred to as Group 2).

The initial step undertaken in the analysis of 2L/HLM for both student groups was running the *null model* using the respective dependent variables. These variables were group mean centered, as the independent effects of level-1 and level-2 predictors on the outcome variable and as the cross-level interactions were examined. In addition, the full maximum likelihood was used as the estimation method to allow the use of deviance statistic for the later comparison between the null model and any complex model created in the process. The null model is expressed by the following equations:

Null Model

The equation form of the null model reflecting the Level 1 and Level 2 parts are adopted from equations 11.1 and 11.2. These equations are given as follows:

Level-1 Part of the Null Model.

With reference to equation 11.1, the academic achievement or aptitude is represented as a function of the teacher mean plus a random error:

$$Y_{ij} = \beta_{0j} + r_{ij}$$

Where:

Y_{ij} represents the academic achievement or aptitude among students *i* in teacher *j*;

β_{0j} is the estimated mean academic achievement or aptitude in teacher *j*; and

r_{ij} is the student-level error term or level-1 random effect (i.e. the deviation of student *ij*'s score from the teacher mean).

In the above equation, the indices *i* and *j* denote students and teachers, respectively, where

i = 1, 2, … N, students within J teachers; and

j = 1, 2, …, J teachers.

Level-2 Part of the Null Model.

With reference to equation 11.2, each teacher mean, β_{0j}, is viewed as varying randomly around a grand mean across all teachers:

$$\beta_{0j} = \gamma_{00} + u_{ij}$$

Where:

γ_{00} is the grand mean academic achievement or aptitude in teacher *J*; and

u_{ij} is the teacher-level error term or the random teacher effect (i.e., the deviation of teacher *j*'s mean from the grand mean).

This is under the assumption that the error term or random effect associated with teacher *j*, u_{0j}, has a normal distribution with a mean of zero and variance τ_{π}.

Running the null model provides a point estimate and confidence interval for the grand mean, γ_{00}. Also, as the null model does not contain any predictor variable, it only captures the error terms from the two levels accounting for the total criterion variance in the model (Richter, 2006). The variance at each level is represented by the following parameters: σ^2 for level-1 model, and τ_{π} for level-2 model (Bryk &

Raudenbush, 1992; Raudenbush & Bryk, 2002). The null model also permits the estimation of the proportions of variations within teachers, and among teachers, as represented by the following mathematical expressions, respectively:

$\sigma^2/(\sigma^2 + \tau_\pi)$ the proportion of variance within teachers (11.5)

$\tau_\pi/(\sigma^2 + \tau_\pi)$ the proportion of variance among teachers (11.6)

Equation 11.6 defines the ICC. Moreover, the average reliability for the least square estimates for each level-1 coefficient across a set of level-2 units (Raudenbush et al., 2004) is an indicator that could be used to assume (or not assume) the presence of random effect for a particular coefficient. The reliability represents the degree to which the teacher-level units can be discriminated between using the ordinary least squares estimates of β_{0j} (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). Furthermore, reliability measures the ratio of the true score (parameter variance) relative to the observed score (total variance of the sample mean) (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). The reliability estimate for the student sample mean for each teacher group can be calculated using the following equation:

Reliability (β_{0j}) = $\tau_\pi/(\tau_\pi + \sigma^2/n_{jk})$ (11.7)

Based on Equation 11.7, the average of the reliabilities across teachers or teacher groups may be viewed as measures of reliability of the teacher means (Raudenbush, Bryk, Cheong, & Congdon, 2004). As mentioned earlier, a "no random effect'' is assumed for a particular coefficient when reliability falls below the threshold of 0.05.

11.9.1 Group 1 (Grade 6 and 2nd Year Students) Results

The HLM results of the Group 1 null model are presented in Table 11.2. As shown, the between-teacher variance exhibits statistical significance (u0j = 6772.06, χ^2 (330) = 12808.09, p<0.01) implying that the mean academic achievement of Grade 6 and Second Year high school students varied across the

teacher groups. The ICC of 0.73 (6772.06/6772.06 + 2454.89 = 0.73 or 73%) provided the empirical and statistical evidence that the data for this sample group were indeed of hierarchical structure, which thus warrant the analysis of 2L/HLM. The ICC further indicates that about 73% of the variability in the student academic achievement is attributed to teacher characteristics while the remaining 27% is due to student characteristics.

The table also presents the reliability estimate. The reliability indicates the extent to which the mean of the dependent variable can be discriminated among level-2 units. According to Ma et al. (2008, p.78), in a null model, "the reliability is a good indicator of how well each school's sample mean (teacher's sample mean in this study) estimates the unknown parameter, β_{0j}." The reliability of 0.96 provided by the null model implies that teacher's sample mean can estimate well the student outcome (academic achievement).

Table 11.2. Null model results for the 2L/HLM for Group 1 (Grade 6 and 2nd Year Student Sample)

***Final estimation of fixed effects*:**

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. DF	P-value
For INTRCPT1, B0 INTRCPT2, G00	403.03	4.61	87.36	330	0.000

***Final estimation of variance components*:**

Random Effect	Reliability	Standard Deviation	Variance Component	DF	Chi-square	P-value
INTRCPT1, U0	0.96	82.29	6772.06	330	12808.09	0.000
Level-1, R		49.55	2454.89			

Deviance = 54175.88

Statistics for current covariance components model: Number of estimated parameters = 3

As the results of the null model indicated the existence of variance at the teacher and student levels, it was warranted that the multilevel analysis of the data should proceed. Hence, predictors were added to each level to further examine the variance accounted for by the independent variables and to build

the final 2L/HLM.

As mentioned earlier, building up the final model followed the process as described in the previous sections. After running the null model, the variables were entered into the equations. The bottom-up process was followed in entering the predictors. The level-1 variables were first entered one at a time. This was followed by the successive entry of level-2 variables. In including the variables in the analysis, care was exercised to avoid possible problems such as multicolinnearity. As the data analysed under HLM were in standardised form, which has a mean of 500, it was necessary to define the location of the variables for better interpretation of the results. For the rationale stated in Section 11.8 of this chapter, the level-1 continuous variables were successively added into the equation as group-mean centered predictors. On the other hand, the level-2 continuous variables were entered as grand-mean centered predictors. All categorical variables in both levels were un-centered. Besides, the full maximum likelihood was the estimation method used. This was to allow model comparison and/or examination of model fit. Predictors that were found to be non-significant based on the *t*-ratios were removed from the model with the next potential predictor filling in the place of the one removed. The equation was then re-analysed. Predictors with *t*-ratios greater than two, or 1.96 to be specific, were included in the model. This process was repeated until only the significant effects were left in the equation. The results of the final model analysis for Group 1 are presented in Table 11.3.

Table 11.3. Results of the 2L/HLM analysis for Group 1 (Grade 6 and 2nd Year Student Sample)

***Final estimation of fixed effects*:**

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. DF	P-value
For INTRCPT1, B0					
INTRCPT2, G00	364.88	14.74	24.75	328	0.000
AGE6, G01	58.31	27.82	2.10	328	0.037
SCHTYPE_PUB_1, G02	43.62	15.50	2.82	328	0.006
For SSEX_SMALE_1 Slope, B1	-8.14	1.52	-5.37	4976	0.000
INTRCPT2, G10					
For SPA Slope, B2	0.51	0.16	3.22	4976	0.002
INTRCPT2, G20					
For SATA Slope, B3	0.32	0.07	4.41	4976	0.000
INTRCPT2, G30					

***Final estimation of variance components*:**

Random Effect	Reliability	Standard Deviation	Variance Component	DF	Chi-square	P-value
INTRCPT1, U0	0.96	80.51	6481.57	328	12630.32	0.000
Level-1, R		49.05	2406.33			

Statistics for current covariance components model:

Deviance = 54068.75
Number of estimated parameters = 8

Model Comparison Test:

Chi-square statistic = 107.13435
DF = 5
P-value = 0.000

The table above provides the results of the 2L/HLM analysis for Group 1. Specifically, it shows the variables that finally composed level-1 and level-2 models and that directly impacted on the outcome variable. As revealed, teachers with 60 years of age and above (AGE6, 58.31) and school type (SCHTYPE, 43.62) are the level-2 variables that directly affected the academic achievement. The AGE6 shows a significant (P<05) positive effect on academic achievement. This suggests that teachers who were in the age range of 60 years and above tended to positively influence the NAT scores of Grade 6 and Second Year high school students. This result is expected, as old teachers tend to improve in their teaching skills because of their experiences on the job. In addition, old teachers are usually more familiar with the National Achievement Test (NAT) having been in the system for many years and as such they are in a better position

to prepare students for the NAT. On the effect of school type, the result discloses a significant (P<0.01) moderation on academic achievement. This could indicate that teachers in the public school (coded 1) tended to positively enhance the NAT scores of their students. This is perhaps due to the increasing professional support and training that have been provided by the Department of Education (DepEd) in recent years. Moreover, the student gender (SSEX, - 8.14), student perceptions of assessment (SPA, 0.51), and student attitude towards assessment (SATA, 0.32) are the level-1 variables that exerted influence on academic achievement. The SSEX reveals a significant (P<0.01) negative effect on academic achievement. This can be interpreted that female 6th grade and 2nd year high school students (coded 0) tended to increase the academic achievement (NAT scores) than their male counterpart. Besides, the SPA shows a significant (P<0.01) positive influence on academic achievement. This could indicate that as the mean scores of Grade 6 and Second Year high school students increased, their NAT scores tended to improve. Similarly, the SATA shows a significant (P<0.01) positive effect on academic achievement, which suggests that as the concerned students obtained high mean scores in their attitude towards assessment, their NAT scores tended to increase. These results are consistent with the view that positive perceptions and attitude tend to elicit positive actions or at least affect individuals in the way they approach their activities, like learning activities in the case of students.

Table 11.3 also presents the estimate of the deviance statistic (54068.75). When compared with the deviance statistic of the null model (54175.88), there is a significant (χ^2 (5) = 107.13435, p<0.01) reduction of 107.13 points. This implies that the 2L/HLM with predictors (the final model) is a better model in terms of fit with the data when compared with the null model. This provides support for the acceptance of the final model. So far, what have been presented are only the relationships of level-1 and level-2 predictors with the outcome variable. To answer the research question, the relationships among predictors, there was a need to look at the interactions between level-1 and level-2 explanatory variables. Table 11.4 shows these results.

It is evident from Table 11.4 that school type (level-2 variable) had interaction effects with student

sex (SSEX, -18.63), student perceptions of assessment (SPA, -12.71), and student attitude towards assessment (SATA, 0.85) (level-1 variables). This means that the type of school where students enrolled moderated the effects of level-1 predictors on the academic achievement. In other words, being in the public school modified the magnitude and the direction of the relationships between level-1 independent and dependent variables. Specific results reveal that the school type influenced the negative effect of gender factor and academic achievement. This means that female Grade 6 and Second Year high school students (coded 0) in the public school (coded 1) tended to increase academic achievement more than the other students. Moreover, the school type moderated the positive relationship between student perceptions of assessment and academic achievement. This could indicate that assessment perceptions of Grade 6 and Second Year high school students in the public school tended to moderate the general relationship between assessment perceptions and the NAT scores of students in Group 1. Additionally, the school type likewise appeared to moderate the positive effect of student attitude towards assessment on academic achievement. This could also indicate that Grade 6 and Second Year high school students in the private school tended to moderate the relationship between assessment attitude and NAT scores in this group. More relevant details are provided in the subsection on cross-level interaction effects.

Table 11.4. Results of interaction effects between level-1 and level-2 predictors for Group 1 (Grade 6 and 2nd Year Student Sample)

***Final estimation of fixed effects*:**

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. DF	P-value
For INTRCPT1, B0					
INTRCPT2, G00	369.52	14.94	24.73	329	0.000
SCHTYPE_PUB, G01	40.17	15.70	2.56	329	0.011
For SSEX_SMALE_1 slope, B1	-18.63	3.96	-4.70	4970	0.000
INTRCPT2, G10					
SCHTYPE_PUB, G11	12.05	4.28	2.81	4970	0.005*
For SPA Slope, B2	-12.71	3.67	-3.47	4970	0.001
INTRCPT2, G20					
SCHTYPE_PUB, G21	12.46	3.84	3.249	4970	0.002*
For SATA Slope, B5	0.85	0.21	4.04	4970	0.000
INTRCPT2, G50					
SCHTYPE_PUB, G51	-0.63	0.22	-2.82	4970	0.005*

***Final estimation of variance components*:**

Random Effect	Reliability	Standard Deviation	Variance Component	DF	Chi-square	P-value
INTRCPT1, U0	0.96	81.14	6583.09	329	13032.40	0.000
Level-1, R		48.71	2373.11			

Statistics for current covariance components model:

Deviance = 54008.86
Number of estimated parameters = 14

Model Comparison Test:

Chi-square statistic = 79.54279
DF = 5
P-value = 0.000

**Cross-level Interaction effects (P<0.01/P<0.05)*

Based on the results in Tables 11.3 and 11.4, the final 2L/HLM for Group 1 can be specified by the following equations:

Level-1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(SSEX) + \beta_{2j}(SPA) + \beta_{3j}(SATA) + r_{ij} \quad (11.8)$$

Level-2 model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(AGE6) + \gamma_{02}(SCHTYPE) + u_{0j} \quad (11.9)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(SCHTYPE) + u_{1j} \quad (11.10)$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(SCHTYPE) + u_{2j} \quad (11.11)$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31}(SCHTYPE) + u_{3j} \quad (11.12)$$

The final model is represented by the equation resulting from substituting Equations 11.9 to 11.12 into Equation 11.8:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(AGE6) + \gamma_{02}(SCHTYPE) + \gamma_{10}(SSEX) + \gamma_{11}(SSEX)(SCHTYPE) + \gamma_{20}(SPA) + \gamma_{21}(SPA)(SCHTYPE) + \gamma_{30}(SATA) + \gamma_{31}(SATA)(SCHTYPE) + u_{0j} + u_{1j}(SSEX) + u_{2j}(SPA) + u_{3j}(SATA) + r_{ij} \quad (11.13)$$

The final two-level model for Group 1 (6th Grade and 2nd Year high school students) represents five direct effects, three cross-level interaction effects, and a random error. Five variables were found to be statistically significant ($P<0.01$/$P<0.05$) to influence academic achievement (see Table 11.3). These variables that represented the direct effects are teachers' age range of 60 years and above (AGE6, Y_{01}) and school type (SCHTYPE, Y_{02}) at level-2, and three level-1 variables namely, student gender (SSEX, Y_{10}), student perceptions of assessment (SPA, Y_{20}), and student attitude towards assessment (SATA, Y_{30}). The cross-level interactions involve school type (SCHTYPE) and the three level-1 predictors: SSEX (Y_{11}), SPA (Y_{21}), and SATA (Y_{31}). The random error is represented in the equation by the terms "u_{0j} + u_{1j}(*SSEX*) + u_{2j}(*SPA*) + u_{3j}(*SATA*) + r_{ij}". These relationships are depicted in Figure 11.3.

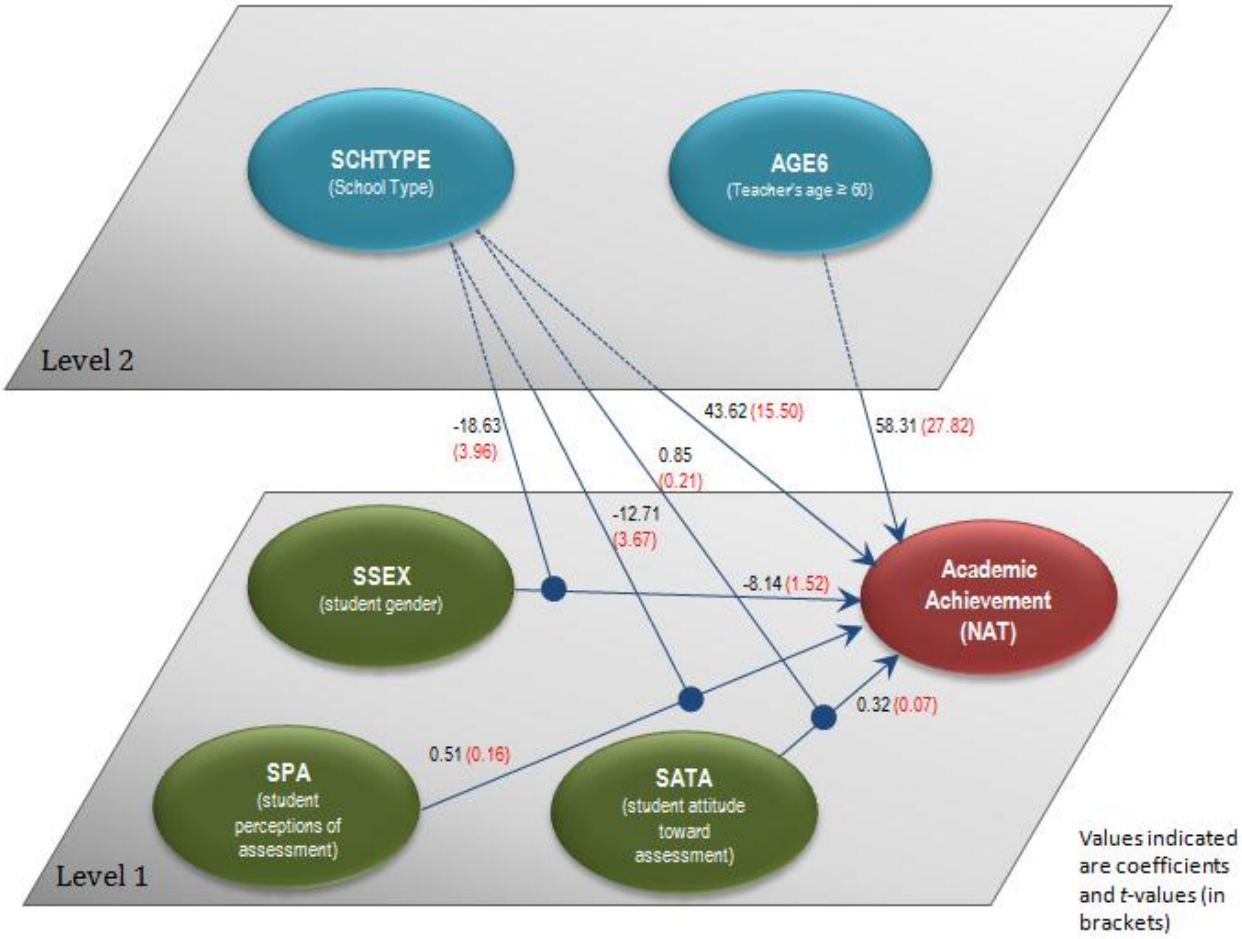


Figure 11.3. Final Two-level Model for Group 1 (6th Grade and 2nd Year Student Sample)

11.9.2.1 Cross-level Interaction Effects

Parts of the equation for the final model for the group of 6th Grade elementary and 2nd Year high school students can be drawn to show several cross-level interaction effects. These are as follows:

a. Student gender (SSEX) and school type (SCHTYPE) on Academic Achievement

$$Y_{ij} = \gamma_{00} + \gamma_{10}(SSEX) + \gamma_{11}(SSEX)(SCHTYPE) + r_{ij}$$

Where: $\gamma_{00} = 369.52;\ \gamma_{10} = -18.63; and\ \gamma_{11} = 12.05$

b. Students' perceptions of assessment (SPA) and school type (SCHTYPE) on Academic Achievement

$$Y_{ij} = \gamma_{00} + \gamma_{20}(SPA) + \gamma_{21}(SPA)(SCHTYPE) + r_{ij}$$

Where: $\gamma_{00} = 369.52;\ \gamma_{20} = -12.71; and\ \gamma_{21} = 12.46$

c. Students' attitude toward assessment (SATA) and school type (SCHTYPE) on Academic Achievement

$$Y_{ij} = \gamma_{00} + \gamma_{30}(SATA) + \gamma_{31}(SATA)(SCHTYPE) + r_{ij}$$

Where: $\gamma_{00} = 369.52;\ \gamma_{30} = 0.85; and\ \gamma_{31} = 0.21$

Using the equations above, the coordinates for the graphs to show the different cross-level interaction effects can be calculated.

For student gender (SSEX) and school type (SCHTYPE) on academic achievement, the information used to calculate the coordinates to graphically represent the cross-level interaction effect were the following:

a. SSEX (Female students=0; Male students=1)

b. SCHTYPE (Private schools=0; Public schools=1)

Using the above as guide, the calculated coordinates were as follows:

i. Female students and public schools (SSEX=0; SCHTYPE=1)

$$Y_{ij} = 369.52 - 18.63(0) + 12.05(0)(1) = 369.52$$

ii. Male students and public schools (SSEX=1; SCHTYPE=1)

$$Y_{ij} = 369.52 - 18.63(1) + 12.05(1)(1) = 362.94$$

iii. Female students and private schools (SSEX=0; SCHTYPE=0)

$$Y_{ij} = 369.52 - 18.63(0) + 12.05(0)(0) = 369.52$$

iv. Male students and private schools (SSEX=1; SCHTYPE=0)

$$Y_{ij} = 369.52 - 18.63(1) + 12.05(1)(0) = 350.89$$

Figure 11.4 shows the graphed coordinates.

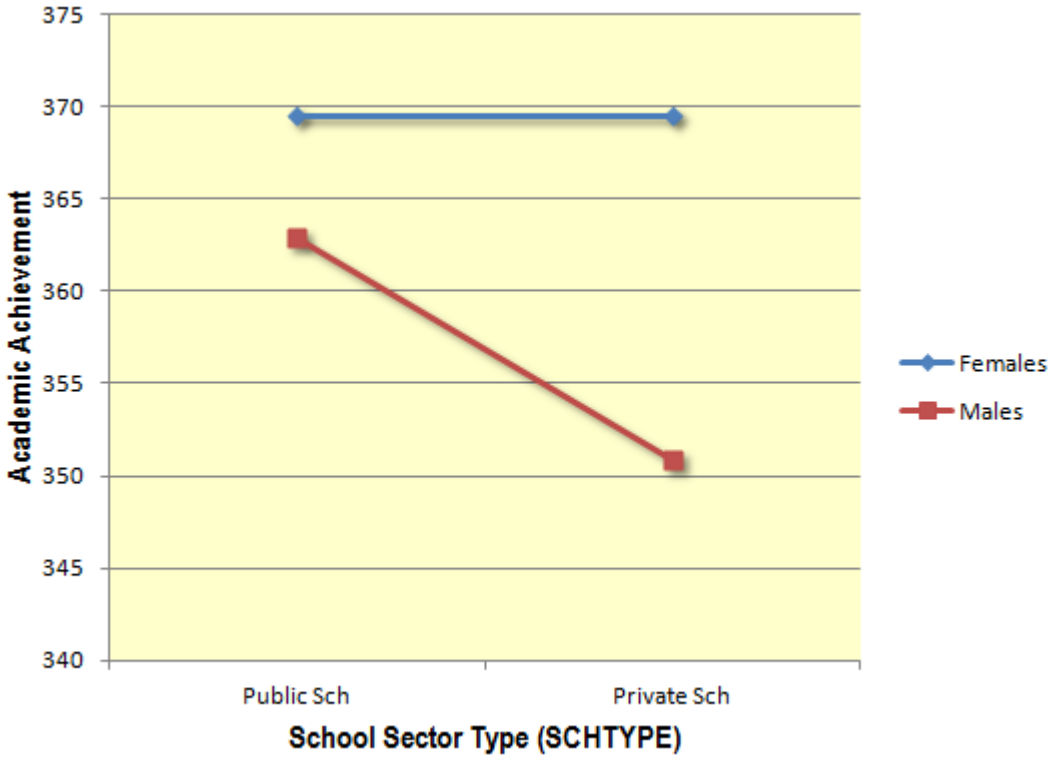


Figure 11.4. Cross-level interaction effect of school type on the slope of student gender on academic achievement

For student perceptions of assessment (SPA) and school type (SCHTYPE) on academic achievement, the information used to calculate the coordinates to graphically represent the cross-level interaction effects were:

a. One standard deviation above the average on SPA,

b. Average on SPA,

c. One standard deviation below the average on SPA,

d. SCHTYPE (Private schools=0; Public schools=1)

Using the above as guide, the calculated coordinates were as follows:

i. High SPA and public schools (SPA=1; SCHTYPE=1)

$$Y_{ij} = 369.52 - 12.71(1) + 12.46(1)(1) = 369.27$$

ii. Low SPA and public schools (SPA=-1; SCHTYPE=1)

$$Y_{ij} = 369.52 - 12.71(-1) + 12.46(-1)(1) = 369.77$$

iii. Average SPA and public schools (SPA=0; SCHTYPE=1)

$$Y_{ij} = 369.52 - 12.71(0) + 12.46(0)(1) = 369.52$$

iv. High SPA and private schools (SPA=1; SCHTYPE=0)

$$Y_{ij} = 369.52 - 12.71(1) + 12.46(1)(0) = 356.81$$

v. Low SPA and private schools (SPA=-1; SCHTYPE=0)

$$Y_{ij} = 369.52 - 12.71(-1) + 12.46(-1)(0) = 382.23$$

vi. Average SPA and private schools (SPA=0; SCHTYPE=0)

$$Y_{ij} = 369.52 - 12.71(0) + 12.46(0)(1) = 369.52$$

The following figure shows the graphed coordinates.

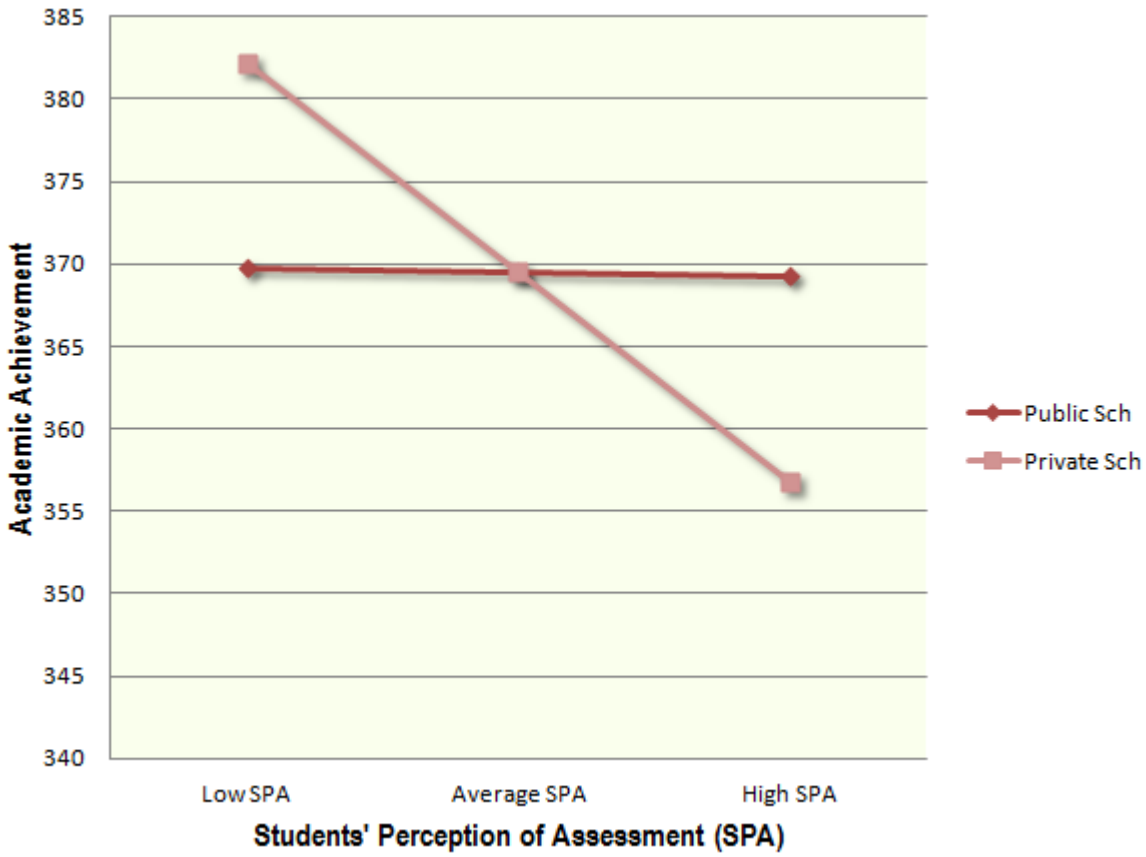


Figure 11.5. Cross-level interaction effect of school type on the slope of student perceptions of assessment on academic achievement

For student attitude toward assessment (SATA) and school type (SCHTYPE) on academic achievement, the information used to calculate the coordinates to graphically represent the cross-level interaction effects were as follows:

a. One standard deviation above the average on SATA,
b. Average on SATA,
c. One standard deviations below the average on SATA,
d. SCHTYPE (Private schools=0; Public schools=1)

Using the above as guide, the calculated coordinates were:

i. High SATA and public schools (SATA=1; SCHTYPE=1)

$$Y_{ij} = 369.52 + 0.85(1) + 0.21(1)(1) = 370.58$$

ii. Low SATA and public schools (SATA=-1; SCHTYPE=1)

$$Y_{ij} = 369.52 + 0.85(-1) + 0.21(-1)(1) = 368.46$$

iii. Average SATA and public schools (SATA=0; SCHTYPE=1)

$$Y_{ij} = 369.52 + 0.85(0) + 0.21(0)(1) = 369.52$$

iv. High SATA and private schools (SATA=1; SCHTYPE=0)

$$Y_{ij} = 369.52 + 0.85(1) + 0.21(1)(0) = 370.37$$

v. Low SATA and private schools (SATA=-1; SCHTYPE=0)

$$Y_{ij} = 369.52 + 0.85(-1) + 0.21(-1)(0) = 368.67$$

vi. Average SATA and private schools (SATA=0; SCHTYPE=0)

$$Y_{ij} = 369.52 + 0.85(0) + 0.21(0)(1) = 369.52$$

Figure 11.6 shows the graphed coordinates.

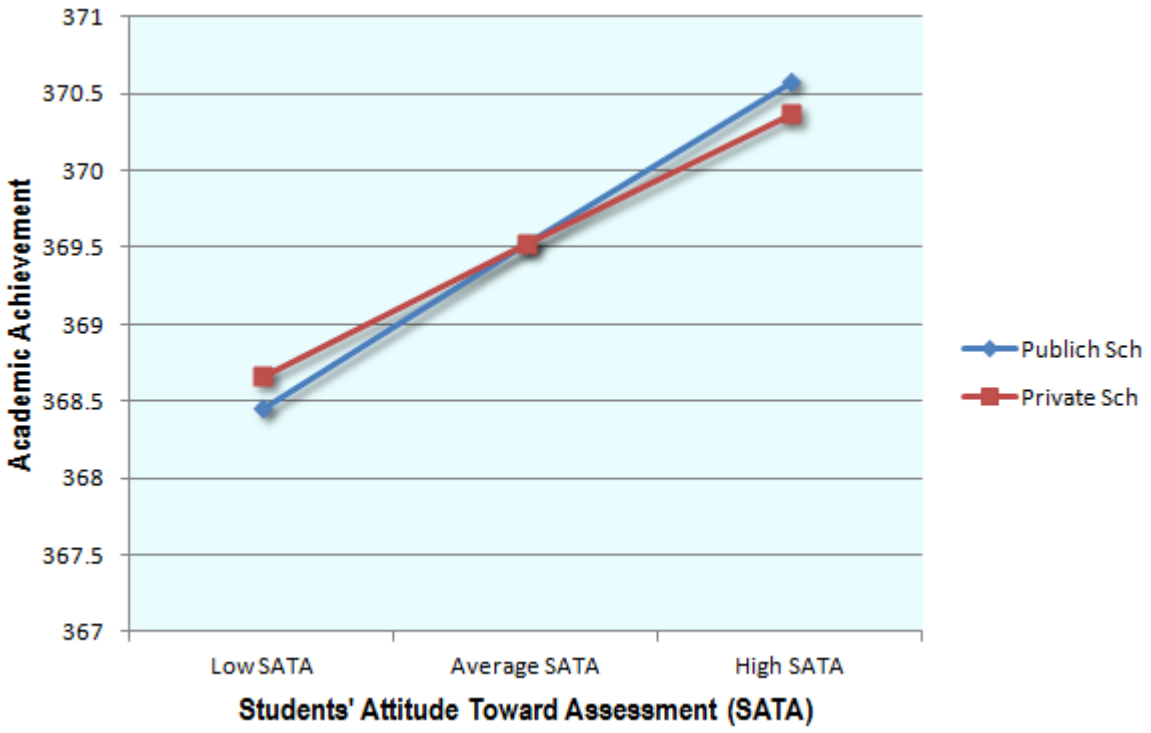


Figure 11.6. Cross-level interaction effect of school type on the slope of student attitude towards assessment on academic achievement

It has been shown in Table 11.4 that there are cross-level interaction effects involving the school type (SCHTYPE, 12.05 with SSEX; 12.46 with SPA; and -0.63 with SATA) and the three level-1 variables namely, student gender (SSEX, -18.63), student perceptions of assessment (SPA, -12.71), and student attitude towards assessment (SATA, 0.85). These are illustrated in Figures 11.3, 11.4, and 11.5, respectively. It can be observed in Figure 11.3 that there are two lines with different slopes. Each line represents the student sex in public and private schools with respect to academic achievement. The position of the line and the horizontal slope for females in Group 1 suggests that female 6th grade and 2nd year high school students tended to obtain higher and stable NAT scores. Conversely, the line for males had a negative slope implying that the boys tended to acquire low NAT scores when compared to girls and that the NAT scores tended to decrease in private school. A nearly similar picture can be seen in Figure 11.4. As this figure shows, there are two intersecting lines that have different slopes. Each line represents the relation between student perceptions of assessment and academic achievement in the two school

types. The line and the almost horizontal slope for students in the public school indicate that the concerned students' perceptions of assessment had only a slight change. In other words, the mean scores in assessment perceptions of Grade 6 and Second Year high school students in the public school appeared nearly stable with respect to their NAT scores. This suggests that the concerned students' views about assessment (covering tests and assignments) were almost the same for public school students. On the other hand, the line and the negative slope for the private institution imply that assessment perceptions of students in this school type tended to decrease with respect to their academic achievement. This means that the mean scores in assessment perceptions of Grade 6 and Second Year high school students in the private school tended to decline with respect to their scores in the NAT. This could indicate that assessment perceptions among these students tended to change or were different among them. This particular result reveals disparity in the relationships between assessment perceptions and academic achievement in the public and private schools. In the case of Figure 11.5, it likewise shows two lines with different slopes. Each of these lines represents the relationship between attitude towards assessment and academic achievement of students in the public and private institutions. As can be observed from the figure, though of different orientations, both lines have positive slopes indicating positive relationships between the two concerned factors in the two school types. However, the positions of the line and the steepness of the slopes in the figure imply that 6th grade and 2nd year high school students in public school tended to obtain higher mean scores with respect to their NAT scores than those in the private school.

Figures 11.3, 11.4, and 11.5 generally indicate that students in the public school tended to have stable and more positive results involving the relationships of student gender, assessment perceptions, and assessment attitude with academic achievement when compared with the results of those in the private school. The possible explanation for this is that perhaps private schools adopt different assessment methods than public schools, which lean more towards the use of testing as an assessment tool thereby making the students in this school type more accustom to test, and in the process develop positive behaviour towards test.

Table 11.5. Estimation of variance components for the final Two-level Model for Group 1 (6th Grade and 2nd Year Student Sample)

Model	Estimation of Variance Components	
	Between Students (n=1,430)	**Between Teachers** (n=581)
Null Model	2454.89	6772.06
Final Model	2406.33	6481.57
Variance at each level		
Between Students	2454.89/(2454.89 + 6772.06) = 0.2661 = 26.61%	
Between Teachers	6772.06/(2454.89 + 6772.06) = 0.7339 = 73.39%	
Proportion of variance explained by final model		
Between Students	(2454.89 – 2406.33)/2454.89 = 0.0198 = 1.98%	
Between Teachers	(6772.06 – 6481.57)/6772.06 = 0.0429 = 4.29%	
Proportion of total variance explained by final model		
(0.0198 x 0.2661) + (0.0429 x 0.7339) = 0.0368 = 3.68%		

Table 11.5 presents the estimated variance components and the proportions of variance explained by the final two-level model for Group 1 (Grade 6 and Second Year high school students).The results of the calculations for variance at each level in the null model (see Table 11.2) indicated that most of the variance (about 73%) was attributable to teacher characteristics. It was also revealed that about 27% of the variance was accounted for by student attributes. These portions of variance were shown and discussed earlier. In comparison to the null model, the final model that includes the level-1 and level-2 predictors for academic achievement, explains about 1.98% of the variance at the student level (level 1) and about 4.29% at the teacher level (level 2). Considering the amount of variance explained by the final model at each level in relation to the amount of available variance to be explained at each level, the total variance that the final two-level model could explain is about 3.68%.

The resulting total variance, albeit small in value, indicates that the final model involved factors that could explain the outcome variable (academic achievement/NAT scores). However, it also implies that there are still other variables not covered in the final model that can predict the academic achievement. This suggests that the final model needs to be improved. This can be addressed in relevant future research undertakings.

11.9.2 Group 2 (Fourth Year Students) Results

The 2L/HLM analysis results for the Group 2 null model are shown in Table 11.6. As can be spotted, there is a significant (u0j = 7593.71, χ^2 (94) = 4542.90, P<0.01) between-teacher variance indicating that the mean aptitude of Fourth Year high school students varied across teacher groups. The ICC of 0.70 (7593.71/7593.71.06 + 3271.22 = 0.70 or 70%) provided the empirical and statistical evidence, which justifies the multilevel nature of the data for this group. As such, HLM analysis should be carried out to determine the relationships among the tested factors from the two levels.

Table 11.6. Null Model results for the 2L/HLM for Group 2 (4th Year Student Sample)

***Final estimation of fixed effects*:**

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. DF	P-value
For INTRCPT1, B0 INTRCPT2, G00	495.52	9.06	54.71	94	0.000

***Final estimation of variance components*:**

Random Effect	Reliability	Standard Deviation	Variance Component	DF	Chi-square	P-value
INTRCPT1, U0	0.98	87.14	7593.71	94	4542.90	0.000
Level-1, R		57.19	3271.22			

Statistics for current covariance components model:

Deviance = 23369.64
Number of estimated parameters = 3

The ICC likewise indicates that 70% of the variability in the student aptitude is due to teacher characteristics while 30% is from student characteristics. The table also presents the reliability estimate of 0.98. This strongly indicates that enough amount of variance exists among the between-group variables thereby making the estimation of the outcome variable tenable. In other words, the reliability implies that the teacher-level means can estimate well the student-level outcome variable (aptitude). The clear indication of the nesting of data from the null model allows further analysis of the individual contribution of the predictors

from the two hierarchical levels. Thus, the 2L/HLM was analysed employing the same procedure as described in the previous sections and as used in the evaluation of 2L/HLM for Group 1. The results of the 2L/HLM analysis are presented in Table 11.7.

Table 11.7. Two-level model (2L/HLM) for Group 2 (4th Year Student Sample)

***Final estimation of fixed effects*:**

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. DF	P-value
For INTRCPT1, B0					
INTRCPT2, G00	494.96	10.54	46.94	91	0.000
AGE1, G01	-105.98	37.40	-2.83	91	0.006
ACAD_POSTGRAD_1, G02	44.16	18.24	2.42	91	0.018
EXYR4, G03	-59.73	26.01	-2.30	91	0.024
For SPA Slope, B1	-1.10	0.28	-3.92	2099	0.000
INTRCPT2, G10					
For SATA Slope, B2	0.94	0.12	7.92	2099	0.000
INTRCPT2, G20					

***Final estimation of variance components*:**

Random Effect	Reliability	Standard Deviation	Variance Component	DF	Chi-square	P-value
INTRCPT1, U0	0.97	79.71	6353.13	91	4187.08	0.000
Level-1, R		56.31	3170.30			

Statistics for current covariance components model:

Deviance = 23290.10
Number of estimated parameters = 8

Model Comparison Test:

Chi-square statistic = 79.54279
DF = 5
P-value = 0.000

As presented in Table 11.7, three level-2 variables namely, age range (AGE1, -105.98), academic qualification (ACAD, 44.16), and years of teaching experience (EXYR4, -59.73), and two level-1 variables namely, student perceptions of assessment (SPA, -1.10) and student attitude towards assessment (SATA, 0.94), were directly impacting the outcome variable (Aptitude). The variable AGE1 (below 25 years), ACAD (bachelors or postgraduate), and EXYR4 (16-20 years of teaching experience) account for the 70% variance while the variables SPA and SATA account for the other 30%. From these results, it was revealed

that AGE1 shows a significant (P<0.01) negative effect on student aptitude. This means that the group of youngest teachers tended to negatively influence the NCAE scores of Fourth Year high school students. Perhaps, young teachers were not yet in a position to provide guidance as they had just joined the teaching profession. In addition, young teachers may lack the expertise in the area of aptitude assessment and were not ready to prepare students for the aptitude test. Thus, they may be incapable in providing advice on matters involving career choice and training on aptitude test. As for the ACAD, it shows a significant (P<0.05) positive effect on the aptitude scores of students in this group. This means that teachers' academic qualification tended to enhance the NCAE scores of the concerned students. This is expected as high academic degree is believed to boost teachers' capability in performing their job and in providing advice, training or review such as those related to scholastic aptitude. Moreover, the variable EXYR4 reveals a significant (P<0.05) negative influence on aptitude. This indicates that teachers who were having 16 to 20 years of teaching experience tended to weaken the NCAE scores of Fourth Year high school students. The possible explanation for this result is that perhaps this group of teachers was not yet familiar with NCAE as the newly implemented aptitude examination. From the side of the students, the SPA discloses a significant (P<0.01) negative impact on student aptitude. This denotes that as the mean scores of Fourth Year high school students in assessment perceptions increased (decreased), their NCAE scores tended to decrease (increase). However, the variable SATA shows a significant (P<0.01) positive effect on aptitude. This suggests that as the mean scores of the concerned students increased (decreased), the NCAE scores tended to correspondingly increase (decrease). This result is consistent with the view that positive attitude usually elicits positive behaviour. The variables that compose the final model for Group 2 appear to exhibit mediocre fit with the data when compared with the null model. However, although the reduction in the deviance value is so small, it is still a significant difference. Thus, the model is deemed justified.

The variables shown in Table 11.7 were further subjected to interaction analysis to evaluate the moderating effects of level-2 predictors on the relations between level-1 independent and dependent

variables. The results are given in Table 11.8. As can be spotted from the table, only teachers' academic qualification (ACAD, 0.53) had an interaction effect on student characteristics, particularly the student attitude towards assessment (SATA, 1.17). This means that teachers' academic qualification modified the relationship between SATA and aptitude. In other words, teachers' academic qualification tended to strengthen the effect of SATA on aptitude. This is expected as teachers with higher academic qualification is deemed competent in preparing students for the test or examination, such as the one related to aptitude.

Table 11.8. Interaction effect results between level-1 and level-2 predictors for Group 2 (4th Year Student Sample)

***Final estimation of fixed effects*:**

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. DF	P-value
For INTRCPT1, B0					
INTRCPT2, G00	494.96	10.54	46.94	91	0.000
AGE1, G01	-105.98	37.40	-2.83	91	0.006
ACAD_POSTGRAD, G02	44.17	18.24	2.42	91	0.018
EXYR4, G03	-59.73	26.01	-2.30	91	0.024
For SPAS Slope, B1	-1.32	0.37	-3.56	2093	0.001
INTRCPT2, G10					
AGE1, G11	-0.74	1.27	-0.58	2093	0.562
ACAD_POSTGRAD, G12	0.57	0.60	0.95	2093	0.341
EXYR4, G13	0.16	0.77	0.21	2093	0.831
For SATA Slope, B2	1.17	0.16	7.34	2093	0.000
INTRCPT2, G20					
AGE1, G21	0.10	0.51	0.19	2093	0.852
ACAD_POSTGRAD, G22	0.53	0.26	-2.05	2093	0.040*
EXYR4, G23	-0.43	0.34	-1.27	2093	0.206

***Final estimation of variance components*:**

Random Effect	Reliability	Standard Deviation	Variance Component	DF	Chi-square	P-value
INTRCPT1, U0	0.97	79.71	6353.70	91	4199.23	0.000
Level-1, R		56.22	3161.14			

Statistics for current covariance components model: Deviance = 23284.28 Number of estimated parameters = 14

Model Comparison Test: Chi-square statistic = 85.36 DF = 11 P-value = 0.000

**Cross-level interaction effect (P<0.01)*

Based on the results presented in Tables 11.6, 11.7, and 11.8, the final 2L/HLM for the fourth year group can be specified by the equations as follows:

Level-1 model: $$Y_{ij} = \beta_{0j} + \beta_{1j}(SPA) + \beta_{2j}(SATA) + r_{ij} \quad (11.14)$$

Level-2 model: $$\beta_{0j} = \gamma_{00} + \gamma_{01}(AGE1) + \gamma_{02}(ACAD) + \gamma_{03}(EXYR4) + u_{0j} \quad (11.15)$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad (11.16)$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(ACAD) + u_{2j} \quad (11.17)$$

The final model is represented by the equation resulting from substituting Equations 11.15 to 11.17 into Equation 11.14:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(AGE1) + \gamma_{02}(ACAD) + \gamma_{03}(EXYR4) + \gamma_{10}(SPA) + \gamma_{20}(SATA) + \gamma_{21}(ACAD)(SATA) + u_{0j} + u_{1j}(SPA) + u_{2j}(SATA) + r_{ij} \quad (11.18)$$

The final two-level model for Group 2 (4th year high school students) represents five direct effects, one cross-level interaction effect, and a random error. Five variables were found to be statistically significant (*P<0.*01/*P<0.*05) to influence aptitude (see Table 11.7). These variables representing the direct effects are teachers' age range of below 25 years (AGE1, Y_{01}), academic qualification (ACAD, Y_{02}), and teaching experience of 16 to 20 years (EXYR4, Y_{03}) at level-2, and two level-1 variables namely, student perceptions of assessment (SPA, Y_{10}) and student attitude towards assessment (SATA, Y_{20}). The cross-level interaction involves academic qualification (ACAD) and one level-1 predictor, SATA (Y_{21}). The random error is represented in the equation by the terms "$u_{0j} + u_{1j}(SPA) + u_{2j}(SATA) + r_{ij}$". These relationships are shown in Figure 11.7.

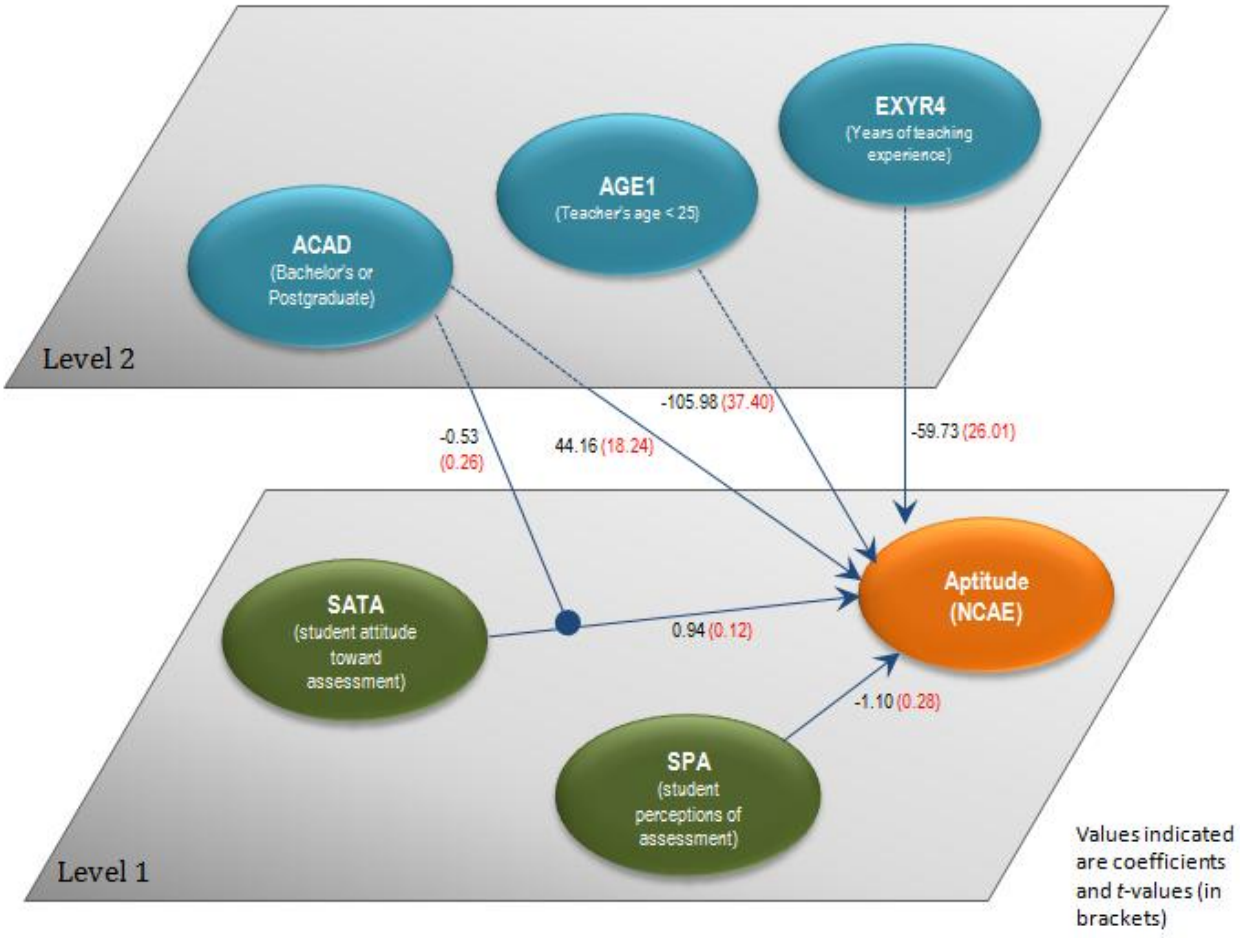


Figure 11.7. Final Two-level Model for Group 2 (4th Year Student Sample)

11.9.2.1 Cross-level Interaction Effect

A part of the equation for the final model for Group 2 (4th year high school students) can be drawn to show a cross-level interaction effect. For the teacher's academic qualification (ACAD) and student attitude toward assessment (SATA) on Aptitude, these are as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{20}(SATA) - \gamma_{21}(ACAD)(SATA) + r_{ij}$$

Where: $\gamma_{00} = 494.96;\ \gamma_{20} = 1.17; and\ \gamma_{21} = -0.43$

The information used to calculate the coordinates used to graphically represent the cross-level interaction effect between SATA and ACAD were the following:

a. One standard deviation above the average on SATA,

b. Average on SATA,

c. One standard deviation below the average on SATA,

d. ACAD (Bachelor's degree=0; Postgraduate degree=1)

Using the above as guide, the calculated coordinates were as follows:

i. High SATA and bachelor's degree (SATA=1; ACAD=0)

$$Y_{ij} = 494.96 + 1.17(1) - (-0.43)(0)(1) = 496.13$$

ii. Low SATA and bachelor's degree (SATA=-1; ACAD=0)

$$Y_{ij} = 494.96 + 1.17(-1) - (-0.43)(0)(-1) = 493.79$$

iii. Average SATA and bachelor's degree (SATA=0; ACAD=0)

$$Y_{ij} = 494.96 + 1.17(0) - (-0.43)(0)(0) = 494.96$$

iv. High SATA and postgraduate degree (SATA=1; ACAD=1)

$$Y_{ij} = 494.96 + 1.17(1) - (-0.43)(1)(1) = 496.56$$

v. Low SATA and postgraduate degree (SATA=-1; ACAD=1)

$$Y_{ij} = 494.96 + 1.17(-1) - (-0.43)(1)(-1) = 493.36$$

vi. Average SATA and postgraduate degree (SATA=0; ACAD=1)

$$Y_{ij} = 494.96 + 1.17(0) - (-0.43)(1)(0) = 494.96$$

The following figure shows the pictorial representation of the interaction effect between SATA and ACAD.

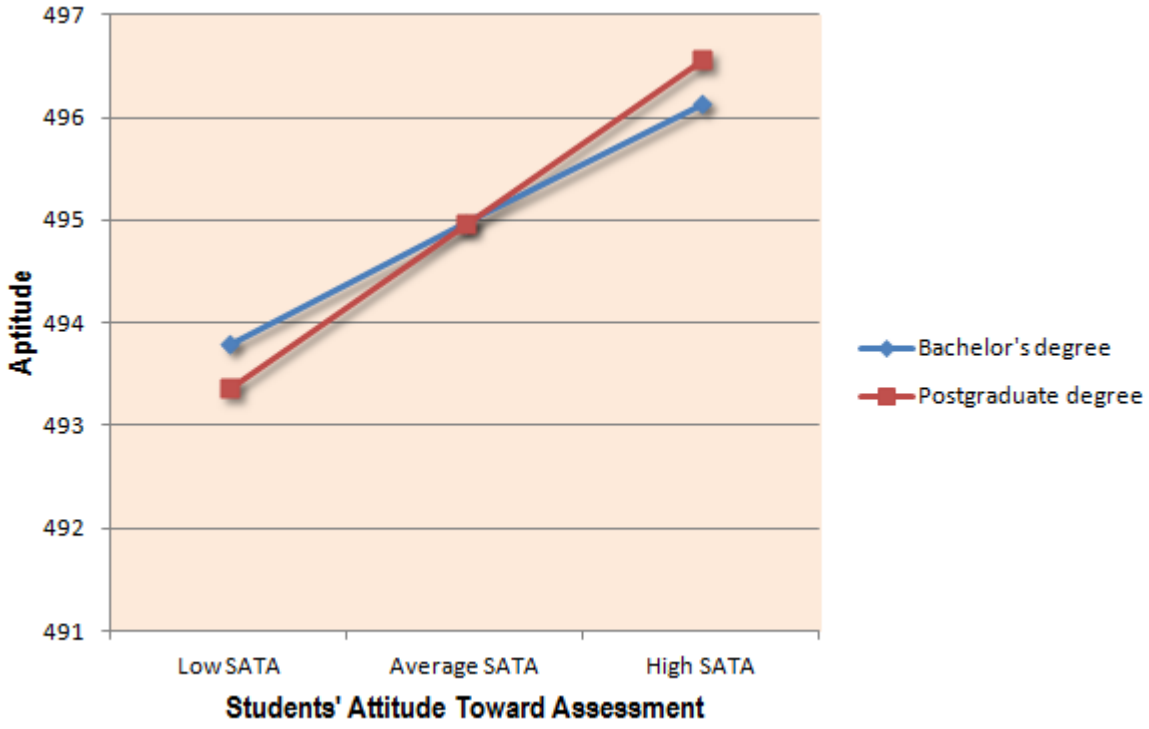


Figure 11.8. Cross-level interaction effect of academic qualification on the slope of student attitude towards assessment

It has been shown in Table 11.8 that there is a cross-level interaction effect involving the academic qualification (ACAD, 0.53) and one level-1 variable, the student attitude towards assessment (SATA, 1.17). This is illustrated in Figure 11.8. It can be observed in this figure that there are two lines with different slopes. Each line represents the relation between SATA and aptitude with respect to ACAD. Though of different positions, both lines have positive slopes indicating that ACAD strengthened the positive relationships between the two concerned factors. However, the positions of the line and the inclination of the slopes imply that teachers with higher academic qualification had the tendency to enhance the positive relationship between SATA and aptitude. Again, this was expected as higher academic qualification is believed to make teachers more competent in preparing students for the test.

Table 11.9. Estimation of variance components for the final Two-level Model for Group 2 (4th Year Student Sample)

Model	Estimation of Variance Components	
	Between Students (n=647)	Between Teachers (n=581)
Null Model	3271.22	7593.71
Final Model	3170.30	6353.13
Variance at each level		
Between Students	3271.22/(3271.22+ 7593.71) = 0.3011 = 30.11%	
Between Teachers	7593.71/(3271.22 + 7593.71) = 0.6989 = 69.89%	
Proportion of variance explained by final model		
Between Students	(3271.22 – 3170.30)/3271.22 = 0.0309 = 3.09%	
Between Teachers	(7593.71 – 6353.13)/7593.71 = 0.1634 = 16.34%	
Proportion of total variance explained by final model		
(0.0309 x 0.3011) + (0.1634 x 0.6989) = 0.1235 = 12.35%		

Table 11.9 shows the estimated variance components and the proportions of variance explained by the final two-level model for Group 2 (Fourth Year high school students).The results of the computations for variance at each level in the null model (see Table 11.6) indicated that about 70% of the variance was accounted for by teacher-level variables while about 30% of the variance was due to student-level variables. These percentages of variance were shown and discussed in the relevant section. Compared to null model, the final model, which includes the level-1 and level-2 predictors for aptitude, explains about 3.09% of the variance at the student level (level 1) and about 16.34% at the teacher level (level 2). Taking into account the amount of variance explained by the final model at each level and the amount of available variance to be explained at that level, the total variance that can be explained by the final two-level model is about 12.35%.

The resulting total variance indicates that the final model contained factors that could explain the outcome variable (aptitude). However, it also denotes that there are other variables not covered in the final model that can explain student aptitude (NCAE scores). This suggests that the final model needs further improvement. This can be further tested and improved with the inclusion of related variables in future research studies.

The results of the 2L/HLM analyses for both Groups 1 and 2 generally confirmed the results of SEM (Chapter 10) that age range, academic qualification, years of teaching experience, and school type (teacher-level variables) can affect other teacher and student variables including the outcome variables (academic achievement and aptitude). In addition, student sex, assessment perceptions, and assessment attitude (student-level factors) can predict either or both dependent variables (academic achievement/NAT scores and/or aptitude/ NCAE scores).

11.10 Summary

This chapter highlighted the limitations of SEM and the strengths of HLM in analysing multilevel data. It likewise dealt with the concepts of HLM, and the procedure and results of HLM analysis.

The HLM analysis was carried out to address the relevant research questions. Specifically, it was executed to appropriately examine the directional relations among factors operating at two hierarchical levels (teacher and student levels) and to investigate the effects of these factors on the outcome variables. To run the analysis, the HLM 6.08 software was employed. The steps taken as part of the procedure in building HLM include the creation of the null model, evaluation of level-1 predictors, examination of level-2 variables, and the interaction between level-1 and level-2 factors. Moreover, two independent HLM analyses for two groups of students were carried out.

The results of the HLM analysis revealed that the two-level hierarchical linear model was the model that reflected the data from the two groups of sample. The final two-level model for Group 1 (6th Grade and 2nd Year high school students) constituted five direct effects and three cross-level interaction effects. The five variables that were found to be statistically significant to directly influence academic achievement include teachers' age range of 60 years and above and school type at the teacher level, and student gender, student perceptions of assessment, and student attitude towards assessment at the student level. The group of teachers who were 60 years and above, school type, student perceptions of assessment, and student attitude towards assessment all showed significant positive effects on academic

achievement. Conversely, the student gender revealed a significant negative effect on academic achievement. The cross-level interactions for this group involved school type at the teacher level and the three student-level predictors namely, gender, student perceptions of assessment, and student attitude towards assessment. This indicated that the school type modified the magnitude and the direction of the relationships between level-1 predictors and academic achievement. Specific results revealed that the school type influenced the negative effect of gender factor on academic achievement. However, the school type moderated the positive relationships between student perceptions of assessment and academic achievement and between student attitude towards assessment and academic achievement. On the other hand, the final two-level model for Group 2 (4th year high school students) comprised five direct effects and one cross-level interaction effect. The five variables that were found to be statistically significant to directly influence aptitude include teachers' age range of below 25 years, academic qualification, and teaching experience of 16 to 20 years at teacher level, and student perceptions of assessment and student attitude towards assessment at student level. Specific results disclosed that academic qualification and student attitude towards assessment showed significant positive effects on aptitude. In contrast, teachers' age range of below 25 years, teaching experience of 16 to 20 years, and student perceptions of assessment revealed significant negative effects on student aptitude. The cross-level interaction for this group involved academic qualification and student attitude towards assessment. The result indicated that teachers with higher academic qualification had the tendency to enhance the positive relationship between student attitude towards assessment and aptitude.

The results of the 2L/HLM analyses for both Groups 1 and 2 generally confirmed the results of SEM (Chapter 10) that factors such as age range, academic qualification, years of teaching experience, and school type can affect other teacher and student variables, including the outcome variables. In addition, student gender, student assessment perceptions, and student assessment attitude can predict both or either or both of the dependent variables (academic achievement/NAT scores and aptitude/NCAE scores).

Chapter 12: Conclusion

12.1 Introduction

This study examined the assessment literacy of the elementary and secondary school teachers in the province of Tawi-Tawi, Philippines. It attempted to establish the directional influence of teacher assessment literacy on student academic achievement and aptitude through the mediating factors namely, assessment practices, teaching practices, assessment perceptions, and assessment attitude. In addition, the study explored the influence of demographic factors such as teacher's gender, age range, academic qualification, years of teaching experience, and school type on teacher assessment literacy and other teacher-level factors, and student gender on student-level variables. These objectives are reflected in the research questions stated in Chapters 1 and 9. To answer the research questions, this study adopted a framework (see Chapter 2) and employed methods in the analysis of the resulting data from the responses of teacher and student participants (see Chapter 3). From the results of data analyses, findings and implications were drawn. These findings and implications are presented in this chapter. Specifically, this concluding chapter highlights the design of the study and provides the summary of the findings, the relevant implications, and the limitations of the study.

12.2 The Design of the Study

This research study was generally concerned with the assessment literacy and how it contributes to academic achievement and aptitude. To address this, a number of questions that involved variables or factors at the teacher and student levels were posed. At the teacher level, the variables include teacher assessment literacy, assessment practices, teaching practices, and demographic factors such as gender, age range, academic qualification, years of teaching experience, school level, and school type. At the student level, the factors include student perceptions of assessment, student attitude towards assessment,

academic achievement, aptitude, and gender as a demographic factor. These factors, including the proposed relationships among them, were investigated using the responses of the 582 elementary and secondary school teachers and 2,077 elementary and secondary school students from the three targeted levels: Grade 6 Elementary, Second Year High School, and Fourth Year High School. These responses were collected during the School Year (S.Y.) 2010-2011.

The factors were gauged through carefully selected, modified, developed, and validated scales/instruments. The appropriateness of the instruments/scales was based on the objectives of the study and the research questions stated in Chapters 1, 9, and 10. The instruments/scales employed in this study include:

- Assessment Literacy Inventory (ALI) by Mertler and Campbell (2005);
- Assessment Practices Inventory (API), which was developed based on the available framework/literature and teacher questionnaire of PCAP-CCME (2010) and TIMSS (IEA, 1999), and Practices of Assessment Inventory (PAI) (Brown, et al., 2009);
- Teaching Practices Scale (TPS) that was adopted from the 2008 Teaching and Learning International Survey (TALIS) Teacher Questionnaire (OECD, 2009a; 2010);
- Student Perceptions of Assessment Scale (SPAS) that was adapted from the Students' Perceptions of Assessment Questionnaire (SPAQ) (Cavanagh, Waldrip, Romanoski, Dorman, & Fisher, 2005; Waldrip, Fisher, & Dorman, 2008); and
- Student Attitude towards Assessment Scale (SATAS) that was developed using the 'Attitude Scale' by Mickelson (1990) and relevant information from the literature as guide.

The students' academic achievement and aptitude were the secondary data (standardised scores) obtained from the 2010-2011 National Achievement Test (NAT) and the National Career Assessment Examination (NCAE) (DepEd-NETRC, 2013). The validity and reliability of the scales were established through Rasch scaling employing ConQuest 2.0 (Wu, Adams, Wilson & Haldane, 2007) and confirmatory factor analysis (CFA) using LISREL 8.80 (Jöreskog & Sörbom, 2006). Moreover, the psychometric utilities

of the scales were established using the data gathered from the teacher and student respondents.

The data from the validated scales and from the NAT and NCAE were utilised to examine the factors considered in this study, and to answer the research questions. In the investigation process, the descriptive and inferential statistical techniques and the corresponding specialised software were employed. Specifically, the frequency, mean, standard deviation, and percentage were used to describe the levels of assessment literacy, assessment practices, teaching practices, assessment perceptions, assessment attitude, academic achievement, aptitude, and the distributions of the demographic factors. The t-test of independent samples and one-way Analysis of Variance (ANOVA) were employed to determine the significant differences on the levels of teachers' and students' abilities, endorsement, and report. The analyses of descriptive statistics, t-test, and ANOVA were carried out using SPSS 16.0 (SPSS, Inc., 2007a). To examine the directional relations among the variables at each of the teacher and student levels, the structural equation modeling (SEM) was utilised. The SEM was performed using LISREL 8.80 (Jöreskog & Sörbom, 2006). To further determine the directional influence of teacher assessment literacy on academic achievement and aptitude through the intervening variables at the teacher and student levels, the hierarchical linear modeling (HLM) was employed. The HLM 6.08 (Raudenbush, Bryk, & Congdon, 2009) was used in running the HLM analysis. To support the aims of the study, triangulation of data was used and included analysis of qualitative responses from selected elementary and high schoolteacher participants to help enrich the interpretation of the quantitative results on assessment literacy. Interview questions were employed to gather the qualitative data. These qualitative data were analysed by using the common themes and by employing SPSS Text Analysis software. All analyses underpinned by a mixed-methods design were carried out following the conceptual framework presented in Chapter 2.

12.3 Summary of the Findings

This section presents the summary of the key findings that were drawn from the analysed data. The findings served to answer the research questions presented in Chapters 1 and 9, and as the contributions of

this study to the Philippine education system and broadly to the assessment literature. These findings are presented below according to the variables and the specific research questions.

12.3.1 Assessment literacy

- RQ1: What is the level of assessment literacy of the elementary and secondary school teachers?

The assessment literacy levels of the elementary and secondary school teachers in the province of Tawi-Tawi, Philippines were relatively low. In terms of the specific standards, both groups of teacher respondents were all below average. Of the standards tested, both groups performed highest on Standard 1 (Choosing assessment methods appropriate for instructional decisions) and lowest on Standard 2 (Developing assessment methods appropriate for instructional decisions). The results suggest that while Tawi-Tawi teachers, to a certain extent, possessed knowledge in selecting assessment methods as illustrated by their highest performance in Standard 1, they nonetheless lacked knowledge in developing them as indicated by their lowest performance in Standard 2. These findings are supported by the interview results. Teachers reported that they choose assessment methods and tools that are valid and reliable. However, when asked about their views on valid and reliable assessment forms, some teachers provided responses that were not in accordance with the concepts of validity and reliability. Moreover, some of them appeared to be unfamiliar with the methods of establishing these two important qualities of any measuring instrument. Specifically, some teacher respondents associated validity and reliability with the test scores or with passing the test, and with their own operational definitions.

12.3.2 Assessment practices

- RQ2: What are the assessment practices of the elementary and secondary school teachers?

The elementary and secondary school teachers generally indicated that they frequently practise assessment with respect to purpose, design, and communication. Of these keys to quality classroom assessment, their foremost consideration was ‘purpose’ in employing assessment, though they also

consider using appropriate assessment design and communicating assessment results frequently. In other words, when doing assessment activity, they often consider the purpose of doing it, the procedure in choosing and applying the relevant assessment methods/tools, and the proper communication of assessment results. Besides, most teacher respondents reported that they commonly use multiple choice and completion types of test as their assessment tool. Considering these results, the elementary and secondary school teachers in the province of Tawi-Tawi, Philippines generally appeared to practise useful assessment strategies by giving attention to its purpose, its design, and its results.

12.3.3 Teaching practices

- RQ3: What are the teaching practices of the elementary and secondary school teachers?

The elementary and secondary school teachers generally appeared to practise a mix of direct transmission and alternative approaches in more than half of their lessons. Their dominant teaching practices were on 'structuring activities', although they also practise 'student-oriented activities' and 'enhanced activities' in more than half of their lessons. Of the three specific teaching activities, the 'enhanced activities' was the least used. These results pointed out that, although both tested methods were used in more than half of the lessons, the elementary and high school teachers in the province of Tawi-Tawi were more inclined to use the direct transmission method as indicated by 'structuring activities' as their main practice. In other words, instructional activities were mostly prepared and structured by teachers.

12.3.4 Perceptions of assessment

- RQ4: What are the perceptions of the elementary and secondary school students on assessment?

The students' (Grade 6 pupils, Second Year and Fourth Year high school students) perceptions of assessment appeared to be more positive as indicated by their high mean scores. Particularly, the student respondents exhibited positive perceptions of test and assignment as indicated by their average mean scores. However, between these two types of assessment activities/tools, they appeared to have more

positive perceptions towards test. In other words, student respondents view test as a preferential assessment tool. This is expected as the education system in the Philippines considers test as one of the major assessment tools and as the students were more familiar with this assessment mode.

12.3.5 Attitude towards assessment

- RQ5: What is the attitude of the elementary and secondary school students towards assessment?

The student respondents generally exhibited positive attitude towards assessment as indicated by their high mean score. In other words, students consider assessment as contributory to their academic achievement, success in school, and to their education in general.

12.3.6 Academic achievement

- RQ6: What is the level of academic achievement of Grade 6 and Second Year high school students?

The overall level of academic achievement of the Grade 6 and Second Year high school students was below average as indicated by their obtained NAT mean score. This implies that the concerned students obtained low performances in the core areas tested in the NAT namely, Filipino (Philippine national language), Mathematics, English, Science, and HEKASI (Heograpiya, Kasaysayan, and Sibika or Geography, History, and Civics).

12.3.7 General aptitude

- RQ7: What is the level of general aptitude of Fourth Year high school students?

The level of general aptitude of Fourth Year high school students was also below average as indicated by their NCAE mean score. The result implies that their aptitude was low in the core tested areas of NCAE namely, Filipino, Mathematics, English, Science, and Araling Panlipunan (Social Studies).

12.3.8 Significant mean differences

- RQ8: Is there any significant difference on the levels of elementary and secondary school teachers' assessment literacy, assessment practices, and teaching practices in terms of gender, age range, academic qualification, years of teaching experience, school level, and school type?

The t-test/ANOVA results indicated the following:

- Male teachers had more knowledge than their female counterpart in Standard 4 (Using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement);
- Teachers whose age was below 25 years had higher assessment literacy than those whose age was within 40 to 49 years in Standard 2;
- Those with postgraduate qualifications had better assessment literacy than those with bachelor degree in Standard 3 (Administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods), Standard 4, assessment communication, and in terms of the overall assessment literacy;
- High school teachers appeared to possess higher assessment literacy than elementary school teachers in Standard 4, Standard 6 (Communicating assessment results to students, parents, other lay audiences, and other educators), Standard 7 (Recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information), and in their overall assessment literacy;
- Teachers who had 1-5 years of teaching experience appeared to have more knowledge than those who had 11-15 years and 21-25 years of teaching experience in Standard 2. Teachers who had 6-10 years of teaching experience also exhibited higher assessment literacy than those with 1-5 years of teaching experience in Standard 5 (Developing valid pupil grading procedures, which use pupil assessments). Moreover, teachers who had 6-10 years of teaching experience appeared to be more literate than those with 16-20 years of teaching experience in Standard 7. Furthermore, teachers with more than 30

years of teaching experience appeared to practise 'student-oriented activities' more than those with 6-10 years of teaching experience.

- Those from private school appeared to have higher assessment literacy than teachers from public school in Standard 2 (Developing assessment methods appropriate for instructional decisions) and in terms of their overall assessment literacy. However, public school teachers appeared to be more structured in their teaching activities than their counterpart in the private school.

12.3.9 Relationships among tested factors

12.3.9.1 Teacher-level factors

- RQ9.1.1: What is the influence of gender, age range, academic qualification, years of teaching experience, and school type on teacher assessment literacy, assessment practices, and teaching practices?

  The structural equation modeling (SEM)/path analysis results indicated the following:

- Higher academic qualification positively influenced teachers' assessment literacy, including those that are specific to Standard 3, Standard 4, and Standard 6, and teachers' assessment practices, including those that are related to assessment purpose and communication;

- Teaching in private school had a positive impact on teachers' overall assessment literacy and on specific literacy pertaining to Standards 2, 5, and 6; however, it had a negative influence on assessment practices concerning purpose and on teaching practices concerning structuring activities;

- Teachers' young age had a positive impact on assessment practices, including those related to assessment purpose and design, and on teaching practices pertaining to structuring and enhanced activities;

- Longer teaching service/experience as determined by the number of years positively influenced teaching practices, including those related to structuring and student-oriented activities; and

- Male teachers had a positive influence on assessment literacy pertaining to Standard 4, and

on teaching practices involving structuring activities;

- RQ9.1.2: What is the influence of teachers' assessment literacy on their assessment and teaching practices?

Teachers' assessment literacy had a negative impact on their teaching practices. However, teachers' assessment literacy on Standard 5 positively influenced their assessment practices involving assessment purpose and design and their assessment literacy on Standard 7 negatively affected their teaching practices involving enhanced activities;

- RQ9.1.3: What is the influence of teachers' assessment practices on their teaching practices?

Teachers' assessment practices had a positive effect on their teaching practices; specifically, teachers' assessment practices concerning assessment purpose positively influenced their teaching practices involving structuring and student-oriented activities, and their assessment practices concerning assessment communication positively influenced all their teaching activities (structuring, student-oriented, and enhanced activities).

12.3.9.2 Student-level factors

- RQ9.2.1: What is the influence of gender on student perceptions of assessment, student attitude towards assessment, academic achievement, and aptitude?

Grade 6 and Second Year high school female students had a positive impact on attitude towards assessment and academic achievement. In addition, Fourth Year high school female students had an impact on attitude towards assessment;

- RQ9.2.2: What is the influence of students' perceptions of assessment on their attitude towards assessment?

  The students' perceptions of assessment generally appeared to positively influence their attitude towards assessment.

- RQ9.2.3: What is the impact of Grade 6 and Second Year high school students' perceptions of assessment and attitude towards assessment on their academic achievement?

The assessment perceptions of Grade 6 and Second Year high school students had a positive influence on their attitude towards assessment and academic achievement. Besides, Grade 6 and Second Year high school students' perceptions of test positively affected their attitude towards assessment. Moreover, Grade 6 and Second Year high school students' perceptions of assignment positively affected their attitude towards assessment and academic achievement.

- RQ9.2.4: What is the impact of Fourth Year high school students' perceptions of assessment and attitude towards assessment on their aptitude?

The Fourth Year high school students' perceptions of assessment positively influenced their attitude towards assessment and their attitude towards assessment positively affected their aptitude. Specifically, their perceptions of test and assignment positively affected their attitude towards assessment, and their perceptions of assignment had a positive impact on their aptitude. Additionally, their attitude towards assessment had a positive influence on their aptitude.

12.3.9.3Effect of teacher assessment literacy on academic achievement and aptitude through the mediating variables at the teacher and student levels

- RQ9.1.4: What is the influence of teacher assessment literacy on student academic achievement and aptitude through assessment practices, teaching practices, student perceptions of assessment, and student attitude towards assessment?

The HLM analysis results indicated that teachers who were 60 years old and above, female students, student perceptions of assessment, student attitude towards assessment, and being in the public school all had direct effects on the academic achievement of Grade 6 and Second Year high school students. Being in the public school also influenced the effects of these students' gender, assessment perceptions, and assessment attitude on their academic achievement. Moreover, teachers with high

academic qualification and student attitude towards assessment positively influenced the aptitude of Fourth Year high school students. Conversely, teachers who were below 25 years old, assessment perceptions, and teachers with 16 to 20 years of teaching experience had a negative effect on Fourth Year high school students' aptitude. Furthermore, teachers' academic qualification had a negative effect on students' attitude towards assessment.

The HLM analysis results on the interaction of school type on the assessment perception indicated that Grade 6 and Second Year high school students in the private school had lesser positive perception than those in the public school. On the interaction of school type on the assessment attitude of Grade 6 and Second Year High School students, the result indicated that those from the public school tended to have more positive attitude towards assessment than their counterpart in the private school. Besides, the result on the interaction of school type on gender implied that female students had more or less the same level of achievement in both school types. However, male students in the public school tended to obtain higher achievement than male students in the private school. On the interaction of academic qualification on Fourth Year High School students' assessment attitude, the result implied that teachers with higher academic qualification (postgraduate units and degrees) tended to positively influence students' attitude towards assessment and aptitude.

12.4 Theoretical Implications

The issue of assessment literacy among teachers has appeared in the literature. From this available literature, it has been stressed that assessment literacy is one of the essential attributes that classroom teachers need to possess. This emphasis arises from the view that teachers who are assessment literate are in a better position to carry out good teaching and to promote greater learning. As a result, assessment standards such as those implemented in the U.S. and Australia have been developed as a guide to help boost teachers' capability in the area of assessment and to help ensure the needed competency in this domain. A number of educational researchers have also conducted studies on

assessment literacy using the standards to provide evidence of teachers' knowledge and skills in student assessment and to identify relevant areas for possible intervention or improvement.

However, despite the continuing emphasis on the importance of assessment literacy on teacher competence, there are still shortcomings that can be cited. First, the research studies on assessment literacy are still insufficient. There have been many studies conducted on assessment preparation of teachers and on teachers' use of assessment but only few have focused directly on teachers' actual knowledge/skills in the area of student assessment (Plake & Impara, 1997). Besides, most of the studies conducted on assessment literacy have been undertaken in the United States and perhaps few Western countries. Research of this kind has not been widespread in countries in the Asia-Pacific region, including the Philippines. Second, studies on assessment literacy have been limited to the investigation of teachers' knowledge/skills on specific areas as described in the assessment standards. There has been no attempt to examine the possible effect of other factors such as demographic variables and other teacher characteristics on assessment literacy. Moreover, despite the perceived relationships of assessment literacy with other education variables such as those related to teaching and learning, an attempt to link it with these variables has not been carried out or has not been widespread, if any. And third, the absence of research and information on teachers' assessment literacy in a number of countries, including the Philippines, has led to inattention to assessment literacy studies and to unprioritised assessment intervention/reform. These gaps formed part of the rationale to conceptualise and administer this study.

From the shortcomings/gaps identified above, the general contributions of this study are fourfold. First, this study used the established assessment standards and revealed findings on assessment literacy of in-service teachers who came from a different context. As such, it provides additional information to the available literature on assessment literacy and helps highlight this issue. In addition, it provides support to the previous studies while it presents new information on teachers' assessment literacy. For instance, the study confirms the previous finding that teachers possess low assessment literacy while it reveals that in-service teachers were strong in Standard 1 and weak in Standard 2, a new finding concerning the specific

assessment standards. Second, the study expanded the focus by not only examining the teachers' assessment literacy on assessment principles as expressed or delineated by the assessment standards but also the relationships of this attribute in general, and specific standards in particular, with relevant variables. As mentioned earlier, experts have stressed that assessment literacy facilitates high quality assessment and good teaching. This implies that assessment literacy supports and exerts influence on assessment and teaching practices, which may further impact on student learning and other student attributes. This is in agreement with available theories such as the 3-P Model as cited in Chapter 2, in which assessment literacy can be viewed as part of the so-called 'presage factors' that are expected to influence the 'process' and eventually the 'product' in educational setting. To capture experts' assertion and the implied relationships, this study included a number of variables, which were believed to interact with assessment literacy. These include the demographic factors/teacher characteristics such as gender, age range, academic qualification, years of experience, and school type and other education variables such as assessment practices and its sub-constructs, teaching practices and its sub-factors, students' perceptions of assessment and its sub-variables, student attitude towards assessment, academic achievement and aptitude. From the results, some findings did not confirm the relationships as hypothesised in this study. However, there are also findings that indicate interactions between assessment literacy and other variables. For instance, school type and teachers' academic qualification were found to exert a positive effect on teachers' assessment literacy. Also, teachers' assessment literacy in Standard 5 (Developing valid pupil grading procedures which use pupil assessments) was found to positively influence assessment practices involving purpose, which further impacts on teaching practices concerning structuring and student-oriented activities. Thus, some relationships can be traced. However, these findings cannot be considered conclusive because of the possible influence of other factors that were not covered in the study. Hence, further research involving these variables and their assumed relationships is warranted. While not being conclusive, this study helps set the context and provide information that can be the basis for developing and testing new propositions in the relevant studies in the future. Third, the study attempted to examine

constructs such as assessment perceptions and assessment attitude that have not been covered in the previous studies or perhaps that have been least investigated, if any. On assessment perceptions, few studies are available but test and assignment as covered in this study were not part of the investigation. Moreover, studies on attitude towards assessment have not been available at the time of this study. Hence, this study provides new information on factors associated with assessment that can be made part of the bases for improving student learning. However, while these factors were deemed important and related to the issue of assessment literacy, relevant findings are still inconclusive and cannot be generalized to other contexts. As new findings, they are still subject to confirmation or refutation by other studies. Thus, further research on these factors is also warranted. Nevertheless, this helps provide information for the development and testing of new framework in future research. And finally, this study helps highlight the issue and provides new information/empirical evidence on basic education teachers' assessment literacy that can be one of the bases in formulating and excogitating relevant policies and programs and in launching assessment reform in the Philippines, including Tawi-Tawi.

This study might be the first one to investigate the link/relationships of assessment literacy with other factors. In the Philippines and in the province of Tawi-Tawi, this study is the first to be conducted among in-service teachers in the elementary and secondary levels.

12.5 Methodological Implications

This research study involved questions (see Chapters 1,9, and 10) that sought to examine assessment literacy and other relevant factors, including the possible relationships that exist among them. To investigate these factors and their relationships, theoretical framework (see Chapter 2) was developed from related information in the available literature. The theoretical framework guided the analyses of the tested factors and their relationships.

To gather quantitative data for the involved variables, survey instruments/scales were used. These instruments/scales were validated and calibrated to obtain reliable data for subsequent analysis. Initially, the

instruments/scales were subjected to professional/expert validation (content validity). After which, construct validity was established through Rasch Model (Rasch, 1960) and CFA using ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007) and LISREL 8.80 (Jöreskog & Sörbom, 2006), respectively. The data from the validated/calibrated instruments/scales were analysed following a particular method and employing statistical techniques and software.

In the analysis of teacher and student data, the embedded mixed-methods design (Creswell, 2008) was employed. Under this design, both quantitative and qualitative methods were used. However, the quantitative method was a dominant approach as the data collected for this study were mostly in the form of numbers/scales. The qualitative method was a supporting approach providing data that support the interpretation of the quantitative data. Prior to quantitative analysis, the data, which were in the form of raw scores, were transformed into measures to achieve uniformity for more valid interpretation of the results. The weighted likelihood estimation (WLE) technique (Warm, 1989) was employed to carry out the score transformation. Transformed scores using the WLE method were further converted to W scale (developed by Woodcock and Dahl in 1971). Further conversion of WLE scores into *W* scale was done to eliminate the negative values and the decimal values, and for convenient interpretation of the analysis results. To transform raw scores into WLE, ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007) was used. To further transform WLE scores to *W* scores, Microsoft Excel was employed. Moreover, listwise deletion, one of the case methods, was used to handle the missing data. This was employed, as missing data were very minimal in this study and to ensure that analyses were conducted with the same number of cases (Kline, 2011). Listwise deletion was carried out using LISREL 8.80.

The quantitative analysis utilised frequency, mean, standard deviation, and percentage to describe the data. To determine significant differences in the means of variables, t-test of dependent samples and one-way ANOVA were used. SPSS 16.0 (SPSS, Inc. 2007) was employed to run the descriptive analysis, t-test, and ANOVA. As indicated earlier, relationships among the tested factors were also examined in this study. In treating the relationships, factors were first grouped into teacher and student levels. This was to

properly determine the directional influence at each of these levels. To analyse the relationships at each level, structural equation modeling (SEM), or specifically path analysis – a single-level procedure, was carried out using LISREL 8.80. The existence of the two levels (teacher and student levels) was indicative of the hierarchical or nesting structure of the data. As such, analysis of the relationships and interactions between teacher-level variables and student-level variables, and the influence of all variables from the two levels on the outcome variables required proper technique. Thus, hierarchical linear modeling (HLM), a multilevel technique, was employed for the directional relations among the tested variables from the two levels. To run HLM, HLM 6.08 software was used.

To support and enrich the interpretation of the quantitative results, qualitative data from selected teacher and student participants were collected through semi-structured interview. Analysis of interview responses was undertaken by identifying the common themes (thematic analysis). This was carried out using the SPSS Text Analysis software.

With the employed procedures and techniques briefly described above, this study provides a number of relevant implications. In any research study, the objectives as reflected in the research questions and the study's theoretical framework are initially advanced. These should determine the selection of method and statistical techniques. In other words, the kind of method and techniques to be used should be dictated by the aims and theoretical propositions of the study and not otherwise. Moreover, in the case of survey research, it is important to subject the questionnaires/instruments/scales to rigorous validation process to secure dependable data and to achieve desirable degree of objectivity. This requires the use of appropriate techniques. The Rasch model as used in this study has been articulated as a useful psychometric technique when gauging any measuring instrument. The Rasch's special properties of item and person independence and unidimensionality, and its characteristic of being mathematically sound provide the strength and ensure any possible objectivity in deciding whether any instrument/scale possesses measurement capacity. Hence, the use of the Rasch model is promising, especially in the context of the Philippines and Tawi-Tawi where the Rasch model is not widely employed and where

educational research in the form of survey is part of the common practice. It is admitted that in the real world, perfectly reliable data and perfect objectivity can hardly be achieved. However, efforts should be taken to ensure that data are as reliable and objective as possible so that interpretations and findings drawn from data analysis results are meaningful. This is to avoid what Kline (2011, p. 6) describes as “garbage in, garbage out”. Furthermore, in the selection of statistical techniques, it is essential to consider their relevance, strengths and weaknesses in treating the data. Analysing and running the data using statistical techniques and software will always provide output. But whether the output is appropriate for the objectives of the study is something else.

Finally, in the educational context, so much of information needs to be unpacked. The quantitative data is by no means detailed, though high level of objectivity can be achieved. Thus, the use of qualitative data for more information and deeper interpretation about educational phenomena should be meaningful. Moreover, there is a web of educational variables operating at different levels. As such, educational data are nested in nature. It is important to capture this characteristic of educational data to be able to untangle the web of relationships among educational factors. Hence, the use of appropriate techniques such as SEM and HLM in the analysis of this kind of data should be useful. The current developments in multilevel Rasch models, moderation effects, and mediation effects are even more promising in understanding the complex educational phenomena. In the Philippines and in the province of Tawi-Tawi, where the use of mixed-methods design, the Rasch model, SEM and HLM are not really widespread, local educational researchers should find these techniques more advantageous.

12.6 Implications for Policy, Teacher Education Curriculum, Teacher Professional Development, and Assessment Reform and Research

Based on the results of the quantitative and qualitative analyses in this study, findings pertaining to the tested variables were drawn. These findings are believed to have implications on the educational policy, curriculum, development programs, educational reform, and educational research in the area of

assessment, although the limitations of this study should also be considered in viewing these implications or any proposed recommendations.

From the findings, the elementary and secondary school teachers appeared to be less literate in the area of student assessment. This indicates that basic education teachers in the province of Tawi-Tawi still need to acquire assessment expertise to be more competent in classroom assessment. This implies that there is a need to review DepEd policies at the local, regional, and national levels of education system to find out whether student assessment has been made part of the focus and priority. Otherwise, relevant assessment policies need to be formulated and implemented to facilitate upgrading of teachers' competency in the area of assessment. Assessment development programs likewise need to be reviewed and strengthened. More relevant in-service trainings for teachers should be offered and teachers, especially those who completed under the old pre-service teacher education curriculum, should be enjoined to undergo the trainings to boost their assessment competence. Other forms of professional development such as short-term courses and pursuit of higher degrees should also be made available to teachers to help upgrade their capabilities, especially that academic qualification was found to impact on teachers' assessment literacy. Perhaps, similar policies and programs should also be applied to school administrators and other involved personnel to make them competent in devising and implementing assessment programs at the school level. Assessment reform that includes the development and implementation of standalone assessment standards for Filipino teachers is another possible measure that should be launched. This is especially needed in view of the recently adopted K-12 program that seeks to introduce new assessment requirements. A number of local educators have put forward this reform and this study supports their proposition. Moreover, there is a need to revisit teacher education curriculum at the undergraduate and graduate levels and enhance the assessment component. The experience of this researcher as a tutor in the Curriculum and Assessment of Learning course provided the observations that students lack the understanding on key aspects of assessment. Yet, these students are expected to be a facilitator and an assessor of learning when they join the teaching force. Thus, it is important that the assessment component

should be strengthened at all levels of teacher education. Again, some local researchers have made this part of their recommendations and this study provides empirical support. In addition, academic degree that specialises in educational assessment may be offered at the undergraduate and graduate levels. The Commission on Higher Education (CHED) should encourage the offering of this program and should strictly require all teacher education institutions to offer all assessment subjects as prescribed. Furthermore, the Philippine Regulation Commission (PRC) should develop and increase assessment questions in the Licensure Examination for Teachers that reflect the required assessment standards. And lastly, research studies on assessment literacy/educational assessment should be encouraged and supported, and be used as basis in developing and/or strengthening assessment programs/reforms and in providing training for school administrators and teachers.

It has been revealed in this study (see Chapter 9) that teachers also appeared to employ direct instruction method more than the alternative approach in their classroom teaching. Perhaps, this was due to their familiarity with the lecture, one of the observed common teaching methods in the context of Tawi-Tawi. In this instance, teachers in the province of Tawi-Tawi still need to be trained on the alternative approach to make them ready for its use and for the implementation of the current basic education curriculum, which prescribes the use of constructivism approach (see Chapter 1). In other words, more professional development programs should be conducted, especially on aspects where teachers are less prepared and for those who come from the remote areas, to upgrade their professional capabilities. Supportive policies, coherent teacher education programs, relevant reform and research are among the key areas that warrant review. At the student level, this study provides findings pertaining to the direct influence of assessment perceptions and attitude on students' academic achievement and aptitude. This suggests that assessment perceptions and attitude are characteristics that need to be developed among basic education students to help improve their learning and aptitude. These can be developed through teachers' classroom activities, including those pertaining to assessment. Thus, teachers need competence in teaching and assessment.

12.7 Limitations of the Study and Implications for Further Research

It is acknowledged that a perfect research can hardly be achieved. Needless to say, this study has limitations. As this study is perhaps the first one to link assessment literacy with other variables, the findings concerning the directional relationships are far from being conclusive. This is especially so as some proposed relationships did not come out in this study. As such, the hypothesised relations remain at the level of hypotheses and are therefore subject for further research for adequacy. Besides, factors tested in this study are not meant to be the only factors interacting with assessment literacy. There are other variables that can affect and be affected by assessment literacy that can be covered in future research. In the educational context, there are complex webs of factors and relationships that can hardly be covered in a single study. Furthermore, as this study employed purposive sampling in the selection of teacher and student participants, the findings cannot be generalised to the whole populations of teachers and students in Tawi-Tawi and in the Philippines. Teachers and students were purposively chosen from the three targeted-classes to which NAT and NCAE, the outcome variables tested in this study, were administered. Thus, generalisation is limited to these samples. Lastly, a longitudinal study could have preferably been chosen to capture the better picture of the variables and their relationships and to offer stronger findings. However, due to time constraint and limited resources, the researcher only managed to carry out a cross-sectional study.

Data collection had also posed some challenges. Tawi-Tawi is an archipelagic province composed of many islands where the schools are spread across. To gather the needed data, the researcher had to travel from one island to another through commercial motor launch and chartered motorised boats and had to walk from one village to another to reach the schools. The irregular schedule of commercial motor launch, occasional unavailability of chartered boats, bad weather and peace and order conditions were the difficulties that delayed the collection of data.

Due to the limitations/problems cited above, suggestions are therefore advanced for consideration in

future similar studies:

- Samples from all schools and from more classes representative of the target population are needed. This means that proper sampling method such as multistage random sampling should be employed, taking into account the hierarchical nature of the data;
- This study utlised and modified some instruments that were developed in other countries. Although the instruments were validated and found to have acceptable measurement properties, the development of new relevant instruments that are more appropriate for Filipino and specifically Tawi-Tawi teachers and students is also suggested to obtain more meaningful results;
- Administration of the survey instruments/scales should be made consistent (i.e. distribution and collection, and time allotted for completing the instruments/scales) as much as possible throughout the duration of the data collection. This will reduce the additional facets or biases that need to be considered in data analysis;
- Longitudinal study is strongly suggested considering its advantages described briefly above and as mentioned in the earlier chapter;
- Actual observations or video study of teachers' assessment and teaching practices are likewise suggested to cross-check teachers' self-rated/self-reported responses on these variables and to obtain better interpretation of the relationships of these factors with assessment literacy, should similar research be conducted in the future;
- Interview questions need to be revised to elicit more information about the tested variables and to provide in-depth interpretation of the quantitative findings should further research in the same area be undertaken; and
- Mixed-methods, SEM, and HLM be used in future educational research to draw meaningful findings.

Data analysis was also part of the difficulties in completing this study. As an educational research, this study was to examine the data that were multilevel in nature. As such, appropriate analysis techniques were needed in order to obtain meaningful results. However, even when the appropriate techniques were

available, they were not widely known. Nevertheless, these challenges were fairly managed by acquiring and reading the available information.

By addressing the limitations/problems of this study and/or following suggestions provided above, more meaningful results can be obtained from future research undertakings in assessment literacy.

12.8 Concluding Remarks

This study had the aims of examining teachers' assessment literacy and its possible relationships with other variables, especially its influence on the outcome variables. The study generally attempted to add information to the available literature by providing more findings on the assessment literacy of in-service teachers and on its link with other education variables as implied in the literature. It also attempted to specifically highlight the issue of assessment literacy among in-service teachers in the Philippines, particularly in the province of Tawi-Tawi, in view of the experts' assertion that it is one of the essential attributes that classroom teachers need to possess. From the investigation of all the variables involved, new findings concerning in-service teachers' assessment literacy and its relationships with other factors emerged. However, the results provided no clear direct or indirect relationship between assessment literacy and outcome variables. Nevertheless, it is believed that this study has its contributions.

In terms of the contribution to assessment literature, this study is deemed successful in providing additional findings on assessment literacy of in-service teachers from a different context. In fact, it is the first study to provide empirical evidence on the assessment literacy of Filipino teachers from the rural area, as no study of this kind has been conducted in the Philippines and in Tawi-Tawi. In addition, this study is maybe the first to also provide evidence on the relationship of assessment literacy with relevant education variables. While this finding is far from being conclusive and warrants further investigations, this study provides initial data for other educational researchers to confirm or refute, and to develop new framework to advance the study of assessment literacy.

Another contribution that this study provides is related to its methodological approaches to address

the objectives or the research questions. The use of mixed-methods design allowed elicitation of more information and deeper interpretation of some analysis results. Moreover, the use of single-level (SEM) and multilevel (HLM) analysis techniques provided the strength in data handling and analysis, and in the validity of the results because the issues associated with the ordinary statistical techniques (i.e., the loss of information, erroneous estimations, etc.) were addressed. The use of these methods is considered beneficial in educational research, especially that educational phenomena are complex for which appropriate procedures are needed to help obtain proper inferences.

In conclusion, it is believed that this study has provided additional knowledge that helps advance the understanding of assessment literacy, its role in fostering student learning, and its paramount importance in education, training and practice. This study has likewise provided findings based on empirical evidence that could help guide future development efforts in education, especially in the Philippines and the province of Tawi-Tawi. Assessment literacy and practice have key roles to play in the improvement of quality education in any country. In the Philippines, there is an urgent need for a national study in teacher assessment literacy that should be given attention by the government. This study could be replicated at the national level to identify specific needs of teachers to improve their assessment literacy and practice through professional development programs. Consequentially, it is strongly recommended that the Philippine Government, through its Department of Education, employ measures to make teachers' assessment literacy one of its priority elements in pre-service teacher education and training.

References

Abell, S. K. & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.). *The Professional Knowledge Base of Science Teaching* (pp. 205-221). Dordrecht: Springer Science+Business Media B.V.

Airasian, P. W. (1994). *Classroom Assessment (2nd Ed.).* New York: McGraw-Hill.

Alagumalai, S. & Ben, F. (2006). External Assessment: Review of Literature and Current Practices. DECS Commission Report. School of Education, University of Adelaide.

Alagumalai, S. & Curtis, D. D. (2005). Classical Test Theory (CTT). In S. Alagumalai, D. D. Curtis, & N. Hungi. (Eds.). *Applied Rasch Measurement: A Book of Exemplars* (pp. 1-14). Dordrecht, The Netherlands: Springer.

Alagumalai, S., Curtis, D. D. & Hungi, N. (2005). *Applied Rasch Measurement: A Book of Exemplars*. The Netherlands: Springer.

Allison, P. D. (2002). *Missing Data.* Thousand Oaks, CA: SAGE.

Alwin, D. F. & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, *40*(1), 37-47.

American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), and National Education Association (NEA) (1990). *Standards for teacher competence in educational assessment of students*. Retrieved from http://www.unl.edu.buros/article3.html

Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi: 10.1007/BF02293814

Andrich, D. (1995). Further remarks on non-dichotomization of graded responses. *Psychometrika, 60*(1), 37-46.

Applefield, J. M., Huber, R., & Moallem, M. (n.d.). Constructivism in theory and practice: Toward a better understanding. Retrieved from http://peopleuncw.edu/huber/constructivism.pdf

Arbuckle, J. L. (2007). *AMOS (Version 16.0.1) [CFA and SEM analysis program]*. Spring House, PA: Amos Development Corporation.

Asaad, A. S. & Hailaya, W. M. (2001). *Statistics as applied to education and other related fields*. Manila, Philippines: REX Book Store, Inc.

Asaad, A. S. & Hailaya, W. M. (2004). *Measurement and evaluation: Concepts and principles*. Manila, Philippines: REX Book Store, Inc.

Assessment Reform Group (2002). Testing, motivation and learning. Shaftesbury, Cambridge: University of Cambridge Faculty of Education.

Atkins, D. (2010). Overview for analyzing multilevel data using the HLM software. Retrieved from http://depts.washington.edu/cshrb/wordpress/wp-content/uploads/2013/04/Applied-Longitudinal-DataAnalysis-with-HLM-Statistical-Code-HLM-overview.pdf

Baker, S. L. (2006). Dummy variables (to represent categories) and time series. Retrieved from http://hspm.sph.sc.edu/Courses/J716/pdf/7166%20Dummy%20Variables%20and%20Time%20Series.pdf

Baker, F. (2001).*The Basics of Item Response Theory* (2nd ed.). Retrieved from *www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED458219*

Ballada, C. A. (2013). Developing standards for assessment competencies of Filipino teachers. *The Assessment Handbook, 10*, 9-23.

Balagtas, M. U., Dacanay, A. G., Dizon, M. A., & Duque, R. E. (2010). Literacy level on educational assessment of students in a premier teacher education institution: Basis for a capacity building program. *The Assessment Handbook, 4*(1), 1-19.

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182.

Ben, F. (2010). Students' uptake of physics (Unpublished doctoral dissertation). University of Adelaide, Adelaide SA, Australia.

Ben, F., Hailaya, W. M., & Alagumalai, S. (2012). *Validation of the Technical and Further Education-South Australia (TAFE-SA) Assessment of Basic Skills Instrument* (TAFE-SA Commission Report). Adelaide, Australia: TAFE-SA.

Beretvas, N. (2004). Using hierarchical linear modeling for literacy research under no child left behind. *Reading Research Quarterly*, *39*(1), 95-99.

Biggs, J. B. (1993). From theory to practice: A cognitive systems approach. Higher Education Research and Development, 12(1), 73-85. doi: 10.1080/0729436930120107

Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice, 5*(1), 7-74. doi:10.1080/0969595980050102

Black, P. & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80(*2), 139-148.

Black, P. (2001). Dreams, strategies, and systems: Portraits of assessment past, present and future. *Assessment in Education*, 8(1), 66-85.

Black, P. S. (2004). The subversive influence of formative assessment. In Alagumalai, S., Thompson, M., Gibbons, J. A., & Dutney, A. (Eds.). *The Seeker* (pp. 77-92). Adelaide: Flinders University Institute of International Education.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Bock, R. D. (1983). The Discrete Bayesian. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement. A festschrift for Frederick M. Lord* (pp. 103-115). New Jersey: Lawrence Earlbaum.

Bollen, K. A. & Long, J. S. (1993). *Testing Structural Equation Models.* Newbury Park: SAGE Publications.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). NY: Taylor & Francis Group, LLC.

Braun, H., Jenkins, F., & Grigg, W. (2006). *Comparing Private Schools and Public Schools Using Hierarchical Linear Modeling* (NCES 2006-461). U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office.

Brookhart, S. M. (1999). *The art and science of classroom assessment: The missing part of pedagogy*. Washington, D. C.: Eric Clearinghouse on Higher Education (ED432938). Retrieved from http://chiron.valdosta.edu/whuitt/files/artsciassess.html

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: SAGE Publications, Inc.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Byrne, B. M. (2001). Structural equation modeling with AMOS: Basic concepts, applications, and Programming. New Jersey: Lawrence Erlbaum Associates.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York, NY: Taylor & Francis Group, LLC.

Campbell, J., Kyriakides, L., Muijs, D. & Robinson, W. (2004). *Assessing teacher effectiveness: Developing a differentiated model*. London: RoutledgeFalmer.

Caoli-Rodriguez, R. B. (2007). *The Philippines country case study: Country profile prepared for the education for all global monitoring report 2008 education for all by 2015: Will we make it?* UNESCO. Retrieved from unesdoc.unesco.org/images/0015/001555/155516e.pdf

Cavanagh, R. F., Waldrip, B. G., Romanoski, J., Dorman, J., & Fisher, D. (2005). Measuring student perceptions of classroom assessment. In *annual meeting of the Australian Association for Research in Education, Parramatta, Australia*.

Cavanagh, R. F. & Romanoski, J. T. (2006). Rating scale instruments and measurement. *Learning Environ Res, 9*, 273-289. doi: 10.1007/s10984-006-9011-y

Center for Assessment and Evaluation of Student Learning (2004). Making sense of test scores. Retrieved from http://www.caesl.org.

Chatterji, M. (2003). *Designing and testing tools for educational assessment*. U.S.A.: Pearson Education, Inc.

CHED Memorandum Order (CMO) No. 11 (1999). *Revised policies and standards for undergraduate teacher education curriculum*. Retrieved from http://www.ched.gov.ph/chedwww/index.php/eng/Information/CHED-Memorandum-Orders/2004 CHED-Memorandum-Orders

CHED Memorandum Order (CMO) No. 30 (2004). *Revised policies and standards for undergraduate teacher education curriculum*. Retrieved from http://www.ched.gov.ph/chedwww/index.php/eng/Information/CHED-Memorandum-Orders/2004-CHED-Memorandum-Orders

Chen, S-K., Hou, L. & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a-posteriori estimation in CAT using partial credit model. *Educational and Psychological Measurement, 58*(4), 569-595.

Churchill, R., Ferguson, P., Godinho, S., Johnson, N. F., Keddie, A., Letts, W., Mackay, J., McGill, M., Moss, J., Nagel, M. C., Nicholson, P., and Vick, M. (2011). *Teaching: Making a difference*. Milton Qld: John Wiley & Sons Australia, Ltd.

Ciarleglio, M. M. & Makuch, R. W. (2007). Hierarchical linear modeling: An overview. *Child Abuse & Neglect, 31*, 91-98.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye. *Handbook of Classroom Assessment: Learning, Achievement, and Adjustment* (pp. 2-29). California: Academic Press.

Clark, C. M. & Peterson, P. L. (1984). Teachers' Thought Processes – Occasional Paper No. 72. East Lansing, Michigan: The Institute for Research on Teaching, Michigan State University.

Commission on Higher Education (CHED) (2010). *Information on Higher Education System*. Retrieved from http://www.ched.gov.ph/chedwww/index.php/eng/Information

Council of Ministers of Education, Canada (2010). *Pan-Canadian assessment program: Teacher questionnaire*. Retrieved from http://www.cmec.ca/docs/pcap/pcap2010/pcap-teacher-questionnaire.pdf

Council of Ministers of Education, Canada (2007). *Pan-Canadian Assessment Program: PCAP-13 reading, mathematics, and science assessment teacher questionnaire*. Retrieved from http://www.cmec.ca/docs/pcap/pcap2007/TeacherQuestionnaire_en.pdf

Country Reports on Local Government Systems: Philippines (2002). Retrieved from http://www.unescap.org/huset/lgstudy/newcountrypaper/Philippines/Philippines.pdf

Creswell, J. W. (2008). *Educational research: Planning, conducting and evaluating quantitative and qualitative research*. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal*, *5*(2), 125-144.

Curtis, D. D. & Boman, P. (2007). X-ray your data with Rasch. *International Education Journal*, *8*(2), 249-259.

Darmawan, I. G. N. (2003). *Implementation of Information Technology in Local Government in Bali, Indonesia.* Adelaide, South Australia: Shannon Research Press.

Darmawan, I. G. N. & Keeves, J. P. (2009). Using multilevel analysis. In C. R. Aldous, I. G. N. Darmawan, & J. P. Keeves (Eds.), *Change Over Time in Learning Numeracy and Literacy in Rural and Remote Schools* (pp. 48-60). South Australia: Shannon Research Press.

de Castell, S., Luke, A. & MacLennan, D. (1981). On defining literacy. *Canadian Journal of Education*, *6*(3), 7-18.

de Guzman (2003). The dynamics of educational reforms in the Philippine basic and higher education sectors. Asia Pacific Education Review, 4(1), 39-50.

de Leeuw, J. (1992). Series editor's introduction to hierarchical linear models. In A. S Bryk & S. W. Raudenbush, *Hierarchical Linear Models: Applications and Data Analysis Methods* (pp. xiii-xvi). Newbury Park, CA: SAGE Publications, Inc.

Department of Education (DepEd) Fact Sheet (2009). Basic Education Statistics. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/Factsheet2009%20Sept%2022.pdf.

DepEd (2006). National Competency-Based Teacher Standards (NCBTS): A professional development guide for Filipino teachers. Retrieved from http://prime.deped.gov.ph/wp-content/uploads/downloads/2011/09/22June_POPULAR-VERSION-FINAL.pdf

DepEd-NETRC (2013). *National Achievement Test: Assessing learning gains at the end of school year* (Brochure). Pasig City, Philippines.

DepEd-NETRC (2013). *National Career Assessment Examination: Providing information through test results for self--assessment, career awareness and career guidance* (Brochure). Pasig City, Philippines.

DepEd Order No. 1 (2003). *Promulgating the implementing rules and regulations (IRR) of Republic Act No. 9155 otherwise known as the Governance of Basic Education Act of 2001*. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/DO%201_1-06-03_00001.pdf

DepEd Order No. 43 (2002). *The 2002 Basic Education Curriculum*. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/DO%2043_08-29-02_00001.pdf

DepEd Order No. 79 (2003). *Assessment and evaluation of learning and reporting of students' progress in public elementary and secondary schools*. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/DO%2082_11-19-03_00001.pdf

DepEd Order No. 04 (2004). *Additional guidelines on the new performance-based grading system*. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/DO%204_2-12-04_00001.pdf

DepEd Order No. 92 (2004). *Assessment for learning: Practices, tools and alternative approaches*. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/DM%2092_2-24-04_00001.pdf

DepEd Order No. 33 (2004). *Implementing guidelines on the performance-based grading system for SY 2004-2005*. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/DO%2033_5-31-04_00001.pdf

DepEd Order No. 5 (2005). *Student assessments at the national and division levels of basic education*. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/DO%20No.%205,%20s.%202005.pdf

DepEd Order No. 32 (2009). *National adoption and implementation of NCBTS-TSNA and IPPD for teachers, and integration of its system operations in the overall program for continuing teacher capacity building*. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/DO%20No.%2032,%20s.%202009.pdf

Department of Education (2010). *Discussion Paper on the Enhanced K+12 Basic Education Program*. Retrieved from http://www.imarksweb.net/book/k+12+basic+education+pdf+philippines/

DepEd Order No. 31, (2012). *Policy Guidelines on the Implementation of Grades 1 to 10 of the K to 12 Basic Education Curriculum (BEC) effective School Year 2012-2013*. Retrieved from http://www.deped.gov.ph/cpanel/uploads/issuanceImg/DO%20No.%2031,%20s.%202012.pdf

DepEd-Tawi-Tawi (2008). Division Report on Enrolment & Attendance. Bongao, Tawi-Tawi.

Diamantopoulos, A. & Siguaw, J. A. (2000). *Introducing LISREL: A guide for the uninitiated.* London: SAGE Publications.

Din, F. S. (2000). Direct instruction in remedial math instructions. *National Forum of Special Education Journal*, *9*E, 3-7.

Dorman, J. P. & Knightley, W. M. (2006). Development and validation of an instrument to assess secondary school students' perceptions of assessment tasks. *Educational Studies, 32*(1), 47-58.

Du Toit, S., du Toit, M., Mels, G., & Cheng, Y. (n.d.). LISREL for Windows: PRELIS user's guide. Retrieved from http://www.ssicentral.com/lisrel/techdocs/IPUG.pdf

Dunn, K. E. & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation*, *14*(7), 1-11.

Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks: Corwin Press.

Ewing, M. T., Salzberger, T., & Sinkovics, R. R. (2005). An alternate approach to assessing cross-cultural measurement equivalence in advertising research. *Journal of Advertising, 34*(1), 17-36.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE.

Freeth, D. & Reeves, S. (2004). Learning to work together: Using the presage, process, product (3P) model to highlight decisions and possibilities. Journal of Interprofessional Care, 18(1), 43-56. doi: 10.1080/13561820310001608221

Garavaglia, S. & Sharma, A. (2004). A smart guide to dummy variables: Four applications and a macro. Retrieved from http://www.ats.ucla.edu/stat/sas/library/nesug98/p046.pdf

Garson, G. D. (n.d.). *Introductory guide to HLM with HLM 7 software*. Retrieved from http://www.sagepub.com/upm-data/47529_ch_3.pdf

Gay, L. R. & Airasian, P. (2003). *Educational research: Competencies for analysis and applications* (7th ed). Upper Saddle River, New Jersey: Pearson Education, Inc.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: The Falmer Press.

Glasman, L. R. & Albarracin, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation. Psychological Bulletin, 132(5), 778-822. doi: 10.1037/0033-2909.132.5.778

Goldstein, H. (2011). *Multilevel statistical models* (4th ed). West Sussex, U.K.: John Wiley & Sons, Ltd.

Gonzales, P., Guzman, J., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., & Williams, T. (2004). Highlights from the Trends in International Mathematics and Science Study (TIMMS) 2003 (NCES 2005-005). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved from http://nces.ed.gov/pubs2005/2005005.pdf

Grossen, B. (1995). The story behind Follow Through. *Effective School Practices*, *15*(1).

Guo, S. (2005). Analyzing grouped data with hierarchical linear modeling. *Children and Youth Services Review*, *27*, 637-652.

Guskey, T. R. (2003). How classroom assessments improve learning. Educational Leadership, 60(5), 6-11.

Hair, J. F. Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and their implications to test development. *Instructional Topics in Educational Measurement* (Module 16, pp. 253-262). Retrieved from http://www.ncme.org/pubs/items/24.pdf

Hardy, M. A. (1993). Regression with Dummy Variables. Iowa: SAGE Publications, Inc.

Hargreaves, D. J. (1997). Student learning and assessment are inextricably linked. European *Journal of Engineering Education*, *22*(4), 401-409.

Harris, K. R. & Graham, S. (1994). Constructivism: Principles, paradigms, and integration. *The Journal of Special Education*, *28*(3), 233-247.

Hoyle, R. H. (1995). The structural equation modeling approach: Basic concepts and fundamental issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 1-15). Thousand Oaks, CA: SAGE Publications, Inc.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

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

Human Development Network (2008). Department of Education: When Reforms Don't Transform, Chapter 2 in Philippine Human Development Report 2008/2009. Retrieved from http://hdn.org.ph/wp-content/uploads/2009/05/chapter-2-department-of-education-when-reforms-dont-transform.pdf

International Association for the Evaluation of Educational Achievement (1999). Third International Mathematics and Science Study – repeat: Science teacher questionnaire main survey. Retrieve from http://timssandpirls.bc.edu/timss1999i/pdf/BM2_TeacherS.pdf

Johnson, B. & Christensen, L. (2004). *Educational research: Quantitative, qualitative, and mixed approaches* (2nd ed.). Boston, MA: Pearson Education, Inc.

Jöreskog, K.G. & Sörbom, D. (1993). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language.* Lincolnwood, IL: Scientific Software International, Inc.

Jöreskog, K.G. & Sörbom, D. (2006). LISREL for Windows (Version 8.80) [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Pscychometrika, 56*(2), 255-278.

Keeves, J. P. & Masters, G. N. (1999). Issues in educational measurement. In G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 268-281). The Netherlands: Pergamon.

Kehoe, J. (1995). Basic Item Analysis for Multiple-Choice Tests. Practical Assessment, Research, & Evaluation, 4(10), 1-4. Retrieved from http://PAREonline.net/getvn.asp?v=4&n=10

Kellaghan, T. & Greany, V. (2001). Using assessment to improve the quality of education. Paris, France: UNESCO - International Institute for Educational Planning.

Kennedy, K. J. (2007, May). Barriers to innovative practice: A socio-cultural framework for understanding assessment practices in Asia. Paper presented at the symposium: "Student Assessment and Its Social and Cultural Contexts: How Teachers Respond to Assessment Reforms". *Redesigning Pedagogy – Culture, Understanding and Practice Conference*, Singapore.

Kennedy, J. K., Chan, J. K. S., Fok, P. K. & Yu, W. M. (2008). Forms of assessment and their potential for enhancing learning: Conceptual and cultural Issues. *Educational Research Policy Practice*, *7*, 197-207.

Kim, J. S. (2005). The effects of a constructivist teaching approach on student academic achievement, self-concept, and learning strategies. *Asia Pacific Education Review*, *6*(1), 7-19.

Kim, J., Kaye, J., & Wright, L. K. (2001). Moderating and mediating effects in causal models. *Issues in Mental Health Nursing*, *22*, 63-75.

Kinder, D. & Carnine, D. (1991). Direct instruction: What it is and what it is becoming. *Journal of Behavioral Education*, *1*(2), 193-213.

Klenowski, V. (2008). The changing demands of assessment policy: Sustaining confidence in teacher assessment. Australia: Queensland University of Technology.

Kline, P. (1994). *As easy guide to factor analysis*. NY: Routledge.

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling.* New York: The Guilford Press.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Kreft, I. G. G., de Leeuw, J. & Kim, K-S. (1990). *Comparing four statistical packages for hierarchical linear regession: GENMOD, HLM, ML2, and VARCL* (CSE Technical Report 311). Los Angeles, CA: UCLA Center for Research on Evaluation, Standards, and Student Testing.

Lapus, J. A. (2008). The Education System Facing the Challenges of the 21st Century: The Republic of the Philippines. Geneva, Switzerland. Retrieved from http://www.ibe.unesco.org/National_Reports/ICE_2008/philippines_NR08.pdf.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist*, *35*(2), 125-141.

Lei, P. & Wu, Q. (2007). Introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practice*, 33-43.

Leighton, J. P., Gokiert, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom- versus large-scale tests: Implications for assessment literacy, *Assessment in Education: Principles, Policy, & Practice, 17*(1), 7-21. doi: 10.1080/09695940903565362

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square, and Standardised Mean? *Rasch Measurement Transactions*, *16*(2), 878.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., &Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, 527-537. doi: 10.1016/j.learninstruct.2008.11.001

Little, R. J. & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research, 26*, 3-33.

Lohmöller, J. B. (1989). Basic principles of model building: specification, estimation, evaluation. In H. Wold (Ed.), *Theoretical Empiricism: A General Rationale for Scientific Model-Building* (pp. 1-26). New York: Paragon House.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* New Jersey: Lawrence Earlbaum.

Luistro, A. A. (2012, March). *The state of basic education: Gaining ground*. Retrieved from http://www.slideshare.net/arangkadaph/state-of-education-in-the-philippines-2012.pdf

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage Publications, Inc.

Ma, X., Ma, L. & Bradley, K. D. (2008). Using multilevel modeling to investigate school effects. In A. A. O'Connel & D. B. McCoach (Eds.), *Multilevel Modeling of Educational Data* (pp. 59-102). Charlotte, NC: Information Age Publishing, Inc.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83-104.

Maclellan, E. (2001). Assessment for learning: The differing perceptions of tutors and students. *Assessment & Evaluation in Higher Education*, *26*(4), 307-318.

Magno, C. (2013). Standards of teacher competence on student assessment in the Philippines. *The Assessment Handbook, vol. 10*, 42-53.

Magliaro, S. G., Lockee, B. B., & Burton, J. K. (2005). Direct instruction revisited: A key model for instructional technology. *Educational Technology Research and Development*, *53*(4), 41-55.

Maligalig, D. S. & Albert, J. G. (2008). Measures for assessing basic education in the Philippines. Paper presented at the 6th Social Science Congress, Quezon City, Philippines. Retrieved from http://dirp3.pids.gov.ph/ris/dps/pidsdps0816.pdf

Maligalig, D. S., Caoli-Rodriguez, R. B., Martinez, A., & Cuevas, S. (2010). Education outcomes in the Philippines. *ADB Economics Working Paper Series*. Mandaluyong City, Philippines: Asian Development Bank. Retrieved from http://www.adb.org/Documents/Working-Papers/2010/Economics-WP199.pdf

Marcoulides, G. A. & Kyriakides, L. (2010). Structural equation modeling techniques. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds). *Methodological Advances in Educational Effectiveness Research* (pp. 277-302). OXI4 4RN, UK: Routledge.

Marsh, C. J. (2008; 2010). *Becoming a teacher: Knowledge, skills and issues* (4th ed; 5th ed.). Frenchs Forest NSW: Pearson Australia.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research, 3*(1), 97-110.

McMillan, J. H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research & Evaluation, 7(11)*. Retrieved from http://pareonline.net/getvn.asp?v=7&n=8

McMillan, J. H. & Workman, D. J. (1998). Classroom assessment & grading practices: A review of the literature. Richmond, VA: Metropolitan Educational Research Consortium.

Mcnair, S., Bhargava, A., Adams, L., Edgerton, S., & Kypros, B. (2003). Teachers speak out on assessment practices. *Early Childhood Education Journal*, *31*(1), 23-31.

Mertler, C. A. (2003, October). Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference? Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Mertler, C. A. (2005). Secondary teachers' assessment literacy: Does classroom experience make a difference? American Secondary Education, 33(2), 76-92.

Mertler, C. A. & Campbell, C. (2005, April). Measuring teachers' knowledge & application of classroom assessment concepts: Development of the Assessment Literacy Inventory. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Mickelson, R. A. (1990). The Attitude-Achievement Paradox Among Black Adolescents. *Sociology of Education*, *63*(1), 44-61.

Mindanao State University Secondary Education Department (2009). *Report on Secondary Schools and Enrolments.* Bongao, Tawi-Tawi.

Miralao, V. A. (2004). The Impact of Social Research on Education Policy & Reform in the Philippines. *International Social Science Journal*, *56*(1), 75-87.

Mislevy, R. J. (1986). Bayes Modal Estimation in item response models. *Psychometrika, 51*(2), 177-195.

Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS*. New York: Springer-Verlag New York, Inc.

Mullens, J. E. & Kasprzyk (1999). *Validating item responses on self-report teacher surveys*. Retrieved from http://www.amstat.org/sections/srms/proceedings/papers/1999_118.pdf

Muthén, L.K. & Muthén, B.O. (2007). *Mplus User''s Guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*(3), 431–462.

Myers, L. S., Gamst, G. & Guarino, A. J. (2006). *Applied Multivariate Research: Design and Interpretation.* London: SAGE.

Naumann, J., Richter, T., Groeben, N., & Christmann, U. (n.d.). From theories of attitude to questionnaire design. A research paper. University of Cologne, Germany.

Neale, M. C., Heath, A. C., Hewitt, J. K., Eaves, L .J. & Fulker, D. W. (1989). Fitting genetic models with LISREL: Typothesis testing. *Behavior Genetics, 19*(1), 37-50.

Nichols, P. D. & Mittelholtz, D. J. (1997). Constructing the concept of aptitude: Implications for the assessment of analogical reasoning. In G. Phye (Ed.). *Handbook of Academic Learning: Construction of Knowledge* (pp. 127-147). CA: Academic Press, Inc.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity, 41*, 673-690.

OECD (2005). *Teachers matter: Attracting, developing and retaining effective teachers*. Paris, France: OECD Publishing.

OECD (2009a). *Creating effective teaching & learning environments: First results from Teaching & Learning International Survey (TALIS)*. Paris, France: OECD Publishing.

OECD (2009b). *Education today: The OECD perspective*. Paris, France: OECD Publishing.

OECD (2010). *Teaching and learning international survey (TALIS) 2008 technical report*. Paris, France: OECD Publishing.

Ornstein, A. (1973). *Accountability for teachers and school administrators*. California: Fearon Publishers/Lear Singler, Inc.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. Practical Assessment, Research, and Evaluation, 7(1), 1-3. Retrieved from http://ericae.net/pare/getvn.asp?v=7&n=1

Parsian, N. & Dunning, T. (2009). Developing and validating a questionnaire to measure spirituality: A psychometric process. *Global Journal of Health Science, 1*(1), 1-10.

Patrician, P. A. (2002). Focus on research methods: Multiple imputation for missing data. *Research in Nursing & Health, 25*, 76-84.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds) (2001). *Knowing what students know: The science and design of educational assessment* (pdf version). Retrieved from http://www.nap.edu/catalog/10019.html

Phye, G. D. (1997). Classroom assessment: A multidimensional perspective. In G. D. Phye (Ed.). *Handbook of Classroom Assessment: Learning, Achievement, and Adjustment* (pp. 33-51). California: Academic Press.

Pickens, J. (2005). Attitudes and perceptions. In N. Borkowski (Ed.). *Organizational Behavior in Health Care* (pp. 43-76). Sudburry, MA: Jones & Bartlett Publishers, Inc.

Plake, B. S. (1993). Teacher Assessment Literacy: Teachers' Competencies in the Educational Assessment of Students. *Mid-Western Educational Researcher*, 6(1), 21-27.

Plake, B. S. & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment?. In G. D. Phye (Ed.). *Handbook of Classroom Assessment: Learning, Achievement, and Adjustment* (pp. 53-67). California: Academic Press.

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

Polissar, L. & Diehr, P. (1982). Regression analysis in health services research: The use of dummy variables. *Medical Care, 20*(9), 959-966.

Polit, D. F. & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*, 489-497. doi: 10.1002/nur.20147

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental?. *Theory Into Practice*, 48, 4-11.

Pratt, D. D., Collins, J. B., & Selinger, S. J. (2001). *Development and use of the Teaching Perspective Inventory*. Retrieved from https://facultycommons.macewan.ca/wp-content/uploads/TPI-online-resource.pdf

Probst, T. M. (2003). Development and Validation of the Job Security Index and the Job Security Satisfaction Scale: A classical Test Theory and IRT Approach. *Journal of Occupational and Organizational Psychology, 76*(4), 451–467. DOI: 10.1348/096317903322591587

Quilter, S. M. & Gallini, J. K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator*, *36*(2), 115-131.

Raudenbush, S. & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*(1), 1-17.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. (2nd ed.). London: Sage Publications, Inc.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F. & Congdon, R. T. (2004). *HLM 6: Hierarchical and Nonlinear Modeling*. Chicago: Scientific Software International.

Raudenbush, S. W., Bryk, A. S. & Congdon, R. T. (2009). HLM for Windows (Version 6.08) [Computer software]. Chicago, IL: Scientific Software International, Inc.

Raykov, T. & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Republic Act No. 10533 (2012). *Enhanced Basic Education Act of 2012*. Retrieved from http://www.senate.gov.ph/lisdata/1417511918!.pdf

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Amalyzing sentence reading times with hierarchical linear models, Discourse Processes, 41(3), 221-250. doi 10.1207/s15326950dp4103_1

Rintaningrum, R., Wilkinson, C. & Keeves, J. P. (2009). The use of path analysis with latent variables. In C. R. Aldous (Ed.). *The Learning of Numeracy and Literacy in South Australian Primary Schools* (pp. 46 – 58). South Australia: Shannon Research Press.

Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modeling. *Learning Disabilities: A Contmeporary Journal*, *2*(1), 30-38.

Rohaan, E. J., Taconis, R. & Jochems, W. M. G. (2010). Reviewing the Relations Between Teachers' Knowledge and Pupils' Attitudes in the Field of Technology Education. *International Journal of Technology and Design Education*, 20, 15-26.

Rookes, P. & Willson, J. (2000). Perception: Theory, development and organisation. London: Routledge.

Rose, B. M., Holmbeck, G. N., Coakley, R. M., & Franks, E. A. (2004). Mediator and moderator effects in developmental and behavioral pediatric research. *Development and Behavioral Pediatrics*, *25*(1), 58-67.

Rosenshine, B. V. (1986). Synthesis of research on explicit teaching. *Educational Leadership*, *43*(7), 60-69.

Rowe, K. (2006). Effective teaching practices for students with and without learning difficulties: Issues and implications surrounding key findings and recommendations from the National Inquiry into the Teaching of Literacy. *Australian Journal of Learning Disabilities*, *11*(3), 99-115.

Rowntree, D. (1987). *Assessing students: How shall we know them?* London: Kogan Page Ltd.

Schafer, W. D. (1993). Assessment literacy for teachers. *Theory into Practice*, *32*(2), 118-126.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A. & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review**.** *The Journal of Educational Research*, *99*(6), 323-338. doi: 10.3200/JOER.99.6.323-338

Schulz, W. (2004). Scaling Procedures for Likert-type Items on Students' Concepts, Attitudes, and Actions. In W. Schulz & H. Sibberns (Eds.). *IEA Civic Education Study Technical Report* (pp. 93-126). The Netherlands: The International Association for the Evaluation of Educational Achievement.

Schumacker, R. E. (2004). Rasch measurement: The dichotomous model. In E. V. Smith, Jr. and R. M. Smith (Eds). *Introduction to Rasch measurement* (pp. 226-257). Maple Grove, MN: JAM Press.

Schumacker, R. E. & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York, NY: Taylor & Francis Group, LLC.

Scientific Software International (SSI) (n.d.). LISREL for Windows: A brief overview. Retrieved from http://www.ssicentral.com/lisrel/index.html

Scientific Software International (n.d.). *Multilevel structural equation modeling*. Retrieved from http://www.ssicentral.com/lisrel/techdocs/Session12.pdf

SEAMEO-RIHED (2011). Philippines' Higher Education System. Retrieved 3 July 2012 from http://www.rihed.seameo.org/mambo/index.php?option=com_content&task=view&id=34&Itemid=41

Senate Economic Planning Office (SEPO) (2011). K to 12: The key to quality education? The SEPO Policy Brief. Retrieved 3 July 2012 from http://www.senate.gov.ph/publications/PB%202011-02%20-%20K%20to%2012%20The%20Key%20to%20Quality.pdf

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.

Shute, V. J. & Becker, B. J. (Eds)(2010). *Innovative Assessment for the 21st Century*. New York: Springer Science + Business Media, LLC.

Skrivanek, S. (2009). The use of dummy variables in regression analysis. Retrieved from http://www.moresteam.com/whitepapers/download/dummy-variables.pdf

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith, Jr. & R. M. Smith (Eds). *Introduction to Rasch Measurement: Theory, Models, and Application* (pp. 73 – 92). Mapple Grove, MN: JAM Press.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. BMC Medical Research Methodology, 8(33).doi: 10.1186/1471-2288-8-33. Retrieved from http://www.biomedcentral.com/1471-2288/8/33

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE Publications Ltd.

SPSS Inc. (2007a). SPSS for Windows (Version 16.0) [Statistical Analysis Program]. Chicago: SPSS Inc.

SPSS Inc. (2007b). SPSS Text Analysis for Windows (Version 16.0) [Statistical Analysis Program]. Chicago: SPSS Inc.

Stevens, J. P. (2009). *Applied multivariate staistics for the social sciences*. New York, NY: Taylor & Francis Group, LLC.

Stiggins, R. J. (1991a). Assessment literacy. *The Phi Delta Kappan*, *72*(7), 534-539.

Stiggins, R. J. (1991b). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education*, *4*(4), 263-273.

Stiggins, R. J. (1999a). Assessment, student confidence, and school success. *Phi Delta Kappan*, *81*(3), 191-198.

Stiggins, R. J. (1999b). Are You Assessment Literate?. *High School Journal*, 6(5), 20-23.

Stiggins, R. J. (2002). *Assessment Crisis: The Absence of Assessment FOR Learning*. Retrieved from http://www.pdkintl.org/kappan/k0206sti.htm

Stiggins, R. J. (2012). Classroom assessment competence. Retrieved from http://images.pearsonassessments..com/images/NES_Publication/2012_04Stiggins.pdf

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

Stiggins, R. J., Conklin, N. F. & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues & Practice*, National Institute for Education. Retrieved from http://www3.interscience.wiley.com.proxy.library.adelaide.edu.au/cgibin/fulltext/119499296/PDFSTRT

Stiggins, R. J., Arter, J. A., Chappuis, J., &Chappuis, S. (2007). *Classroom assessment: Doing it right – Using it well*. Upper Saddle River, NJ: Pearson Education, Inc.

Syjuco, A. B. (n. d.). The Philippine Technical Vocational Education and Training (TVET) System. Retrieved from http://www.tesda.gov.ph/uploads/file/Phil%20TVET%20system%20-%20syjuco.pdf

Struyven, K., Dochy, F. & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education, 30*(4), 325-341.

Tawi-Tawi Geography (2010). Retrieved from http://www.servinghistory.com/topics/Tawi-Tawi::sub::Geography

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal*, *31*(2), 231-262.

Teddlie, C. & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences.* CA: SAGE Publications, Inc.

Tennant, A. Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis & Rheumatism (Arthritis Care & Research), 57(*8), 1358-1362. doi: 10.1002/art.23108

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Tinsley, H. E. A. & Dawis, R. V. (1975). An investigation of the Rasch Simple Logistic Model: Sample free item and test calibration. *Educational and Psychological Measurement, 35,* 325-339. doi: 10.1177/001316447503500211

The 1987 Constitution of the Republic of the Philippines. Retrieved from http://www1.umn.edu/humanrts/research/Philippines/PHILIPPINE%20CONSTITUTION.pdf

UNESCO-IBE (2011). World data on education. Retrieved from http://www.ibe.unesco.org/en/services/online-materials/world-data-on-education/seventh-edition-2010-11.html

Van Alphen, A., Halfens, R. Hasman, A., & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing, 20,* 196-201.

Volante, L. & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education*, *30*(3), 749-770.

Waldrip, B. G., Fisher, D. L., & Dorman, J. P. (2008). Students' perceptions of assessment process: Questionnaire development and validation. *Sustainable Coomunities and Sustainable Environements: Beyond Cultural Boundaries* (pp. 561-568), 16-19 January 2008, Curtin University of Technology, Perth WA, Australia.

Walter, W. (1999). Defining literacy and its consequences in the developing world. *International Journal and Lifelong Education*, *18*(1), 31-48.

Warm, T. A. (1989). Weighted likelihood estimation of ability in Item Response Theory. *Psychometrika, 54*(3), 427-450.

Wilkins, J. L. M. (2008). The relationship among elementary teachers' content knowledge, attitudes, beliefs, and practices. *J Math Teacher Educ, 11*, 139-164.doi: 10.1007/s10857-007-9068-2

Watkins, D. & Hattie, J. (1990). Individual and contextual differences in the approaches to learning of Australian secondary school students. Educational Psychology: An International Journal of Experimental Educational Psychology, 10(4), 333-342. doi: 10.1080/0144341900100404

White, B. (2011). *Mapping your thesis: The comprehensive manual of theory and techniques for masters and doctoral research*. Victoria, Australia: ACER Press.

Woltman, H., Feldstain, A., MacKay, C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52-69.

Woodcock, R. W. (1999). What can Rasch-based scores convey about a person‟s test performance? In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement: What Every Psychologist and Educator Should Know* (pp. 105-127). New Jersey: Lawrence Erlbaum.

Wright, B. D. & Linacre, J. M. (1989). The differences between scores and measures. *Rasch Measurement Transactions, 3*(3), 1-4.

Wright, B. D. & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith, Jr. and R. M. Smith (Eds). *Introduction to Rasch measurement* (pp. 1-24). Maple Grove, MN: JAM Press.

Wright, B. D. & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, Delaware: Wide Range, Inc.

Wu, M. L. & Adams, R. J. (2007). *Applying the Rasch Model to psycho-social measurement: A practical approach.* Melbourne: Educational Measurement Solutions. Retrieved from www.edmeasurement.com.au

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ConQuest Version 2.0 [Generalised Item Response Modeling Software]. Camberwell, Victoria: ACER Press.

Zhang, Z. & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. Applied Measurement in Education, 16(4), 323-342. doi: 10.1207/S15324818AME1604_4

Zeidner, M. (1987). Essay versus multiple-choice type of classroom exams: The student's perspective. *Journal of Educational Research, 80*(6), 352-358.

Appendices

Appendix A
Permission/Approval Documents

Ethics Clearance from the University of Adelaide, Page 1

THE UNIVERSITY OF ADELAIDE
AUSTRALIA

RESEARCH BRANCH
RESEARCH ETHICS AND COMPLIANCE UNIT

SABINE SCHREIBER
SECRETARY
HUMAN RESEARCH ETHICS COMMITTEE
THE UNIVERSITY OF ADELAIDE
SA 5005
AUSTRALIA
TELEPHONE +61 8 8303 6028
FACSIMILE +61 8 8303 7325
email: sabine.schreiber@adelaide.edu.au
CRICOS Provider Number 00123M

16 September 2010

Associate Professor S Alagumalai
School of Education

Dear Associate Professor Alagumalai

PROJECT NO: H-159-2010 *Teachers' assessment literacy and student outcomes in the division of Tawi-Tawi, Philippines*

I write to advise you that I have approved the above project on behalf of the the Human Research Ethics Committee. Please refer to the enclosed endorsement sheet for further details and conditions that may be applicable to this approval.

Approval is current for one year. The expiry date for this project is: 30 September 2011

Where possible, participants taking part in the study should be given a copy of the Information Sheet and the signed Consent Form to retain.

Please note that any changes to the project which might affect its continued ethical acceptability will invalidate the project's approval. In such cases an amended protocol must be submitted to the Committee for further approval. It is a condition of approval that you immediately report anything which might warrant review of ethical approval including (a) serious or unexpected adverse effects on participants (b) proposed changes in the protocol; and (c) unforeseen events that might affect continued ethical acceptability of the project. It is also a condition of approval that you inform the Committee, giving reasons, if the project is discontinued before the expected date of completion.

A reporting form is available from the Committee's website. This may be used to renew ethical approval or report on project status including completion.

Yours sincerely,

Professor Garrett Cullity
Convenor
Human Research Ethics Committee

Ethics Clearance from the University of Adelaide, Page 2

THE UNIVERSITY OF ADELAIDE
AUSTRALIA

RESEARCH BRANCH
RESEARCH ETHICS AND COMPLIANCE UNIT

SABINE SCHREIBER
SECRETARY
HUMAN RESEARCH ETHICS COMMITTEE
THE UNIVERSITY OF ADELAIDE
SA 5005
AUSTRALIA
TELEPHONE +61 8 8303 6028
FACSIMILE +61 8 8303 7325
email: sabine.schreiber@adelaide.edu.au
CRICOS Provider Number 00123M

Applicant: Associate Professor S Alagumalai

Department: School of Education

Project Title: *Teachers' assessment literacy and student outcomes in the division of Tawi-Tawi, Philippines*

THE UNIVERSITY OF ADELAIDE HUMAN RESEARCH ETHICS COMMITTEE

Project No: **H-159-2010** RM No: 0000010643

APPROVED for the period until: 30 September 2011

It is noted that this study will be conducted by Wilham M Hailaya, PhD student.

Refer also to the accompanying letter setting out requirements applying to approval.

Professor Garrett Cullity
Convenor
Human Research Ethics Committee

Date: 15 SEP 2010

Permission from the Philippine Department of Education-National Office, Page 1

REPUBLIKA NG PILIPINAS
REPUBLIC OF THE PHILIPPINES
KAGAWARAN NG EDUKASYON
DEPARTMENT OF EDUCATION
DepED Complex, Meralco Ave., Pasig City, Philippines

gapan ng Pangalawang Kalihim
Office of the Undersecretary
Regional Operations

Direct line: 633-7203
Fax: 631-8492
Email address: rdrivera@deped.gov.ph

16 August 2010

THE UNIVERSITY OF ADELAIDE
Level 8, 10 Pulteney St. Adelaide SA 5005

Attention : **PROFESSOR TANIA ASPLAND**
Head, School of Education

Gentlemen:

This is in response to the letter-request of **MR. WILHAM M. HAILAYA**, dated 30 July 2010, which was received on 3 August 2010.

While we are hereby granting him permission to collect data, conduct surveys and interviews as well as assess the National Achievement Test scores of elementary and secondary students belonging to the Division of Tawi-Tawi, Philippines in connection with his research entitled *"Teachers' Assessment Literacy and Student Outcomes in the Division of Tawi-Tawi, Philippines"* in compliance with the requirements for his degree of Doctor of Philosophy in Education; it is imperative that prior to the commencement of the study, he should coordinate with the Regional Director (ARMM) of the Department of Education.

Further, it is our understanding that this permission is being granted only for the above-mentioned purpose and shall immediately cease upon the completion of the research.

Finally, it is expected that a report of the results of this study shall likewise be submitted to this Office.

Very truly yours,

RIZALINO D. RIVERA
Undersecretary

Permission from the Philippine Department of Education-Regional Office, Page 2

بسم الله الرحمن الرحيم
Republic of the Philippines
Autonomous Region in Muslim Mindanao
DEPARTMENT OF EDUCATION
ARMM Complex, Cotabato City

September 6, 2010

Dr. Kiram Irilis
Schools Division Superintendent
Division of Tawi-Tawi

Dear Dr. Irilis:

This has reference to the letter of Mr. Wilham Hailaya, a faculty member of the MSU at Tawi-Tawi who is currently pursuing degree of Doctor of Philosophy (PhD) at the University of Adelaide under the Australian Leadership Awards Scholarship as favorably indorsed by Usec. Rizalino R. Rivera, requesting any assistance that our office can extend to Mr. Hailaya in the conduct of his study (see attached).

Believing in the benefits that DepEd-ARMM, specifically the Division of Tawi-Tawi may gain from the findings, this office hereby approves and indorses the aforementioned request of the subject.

As such, utmost courtesy and assistance is hereby enjoined.

For your information and guidance.

ATTY. BARATUCAL L. CAUDANG
Regional Secretary

Permission from the Philippine Department of Education-Divisional Office, Page 3

Republic of the Philippines
Autonomous Region in Muslim Mindanao
Department of Education
Division of Tawi-Tawi
Bongao

9 August 2010

MR. WILHAM M. HAILAYA
Phd Candidate
School of Education
Faculty of the Professions
The University of Adelaide
Level 8, 10 Pulteney Street,
Adelaide, South Australia 5005

Dear Mr. Hailaya,

The Department of Education (DepEd)-Tawi-Tawi Division Office hereby grants you the permission to collect/access all the necessary data, including the National Achievement Test(NAT) scores and other examination results of the pupils/students involved in your study.

This letter of permission is issued to you in connection with your proposed PhD research study in the Division of Tawi-Tawi and for whatever legal purpose it may serve you.

Sincerely yours,

DR. KIRAM K. DRILIS, Al-Hadj
Schools Division Superintendent
DepEd-Division of Tawi-Tawi

Permission from Dr. Craig Mertler on the Use of the Assessment Literacy Inventory (ALI)

Wilham Hailaya

From: craig.mertler@gmail.com on behalf of Dr. Craig A. Mertler <cmertler@westga.edu>
Sent: Wednesday, 11 May 2011 9:38 PM
To: Wilham Hailaya
Subject: Re: Request for permission to use Classroom Assessment Literacy Inventory
Attachments: _ALI_v2.pdf; Standards for Teacher Competence.pdf; Mertler_Campbell_AERA2005.pdf

Hello Wilham,

Actually, what I found from the use of the CALI instrument is that it did not have good psychometric qualities. A couple of years later, a colleague and I designed our own similar instrument, called the ALI. I have attached it for your review. I have also attached a paper we presented at AERA in 2005 on the development of this instrument. I have also attached the standards based upon which the ALI was developed.

You have my permission to use the ALI instrument, provided you cite us appropriately. Again, I've attached the following:

- Copy of the ALI (with the item-by-standard list on the last page),
- Copy of the seven standards to which the items are aligned, and
- Copy of our 2005 paper presented at AERA

Let me know if you have questions.

Dr. Craig Mertler

Craig A. Mertler, Ph.D.
Professor and Director
Doctoral Program in School Improvement (http://www.westga.edu/eddsi)
College of Education
University of West Georgia
Phone: 678-839-6096
Email: cmertler@westga.edu
Podcasts: http://web.me.com/mertler/Dr._Mertlers_Podcasts/

Appendix B
Survey Questionnaires
(Teacher & Student Questionnaires)

Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines

Teacher Questionnaire

School of Education
Faculty of the Professions
The University of Adelaide

Teacher Assessment Literacy and Student Outcomes in the Division of Tawi-Tawi, Philippines

(Teacher Questionnaire)

I. Information about this Questionnaire

This questionnaire is addressed to teachers who are handling subjects in Grade 6 (elementary school), Second Year and Fourth Year levels (secondary school). It contains items that ask for general information about the participant and the participant's assessment literacy, assessment practices, and teaching practices. It has been organised into sections (A, B, C, & D) corresponding to the said attributes.

Your responses to this questionnaire are significant in helping describe teachers' assessment literacy and how it relates to student outcomes, thus possibly contributing to the improvement of teaching and learning in the classroom. Hence, it is important that you respond to each item very carefully so that the information provided reflects your situation as accurately as possible. All responses will be combined to make totals and averages in which no individual participant/school can be identified. Your responses and identity will be strictly kept confidential.

II. General Instructions to Teacher Participant:

1. Identify a place and a time in school when you will be able to complete this questionnaire without being interrupted.

2. Please read each item carefully and respond as accurately as you can. Specific instructions in answering the items are given in every section of the questionnaire. If you make a mistake in responding to items that have the given options, simply mark X on your previous choice and check another box corresponding to your new answer. If you make an error in answering questions that require writing of number, words and/or sentences, simply cross out your previous response and write the new answer next to it. **Please don't leave any item unanswered**.

3. The questionnaire needs to be returned to the survey questionnaire administrator at the end of the school day or as soon as it has been completed.

Thank you very much for your time and effort in completing this questionnaire!

A. GENERAL INFORMATION

Instructions: Fill in the box/blank with number/words/sentences that correspond to your answer. For items that have the given options, check the box.

Teacher I.D. Number: ☐

Teacher Name: (Optional) ☐

1. What is your gender?

   ☐ Male
   ☐ Female

2. How old are you?

   ☐ Under 25 years
   ☐ 25 – 29 years
   ☐ 30 – 39 years
   ☐ 40 – 49 years
   ☐ 50 – 59 years
   ☐ 60 years and above

3. What academic qualifications do you have? (Please check all that apply to you).

   ☐ Bachelor degree
   ☐ Master's degree/units
   ☐ Ph.D./Ed.D./units

4. In what grade/year level that you teach most of your subjects? (Please check one box only. If you have equal number of subjects in two levels indentified below, please decide to which level you belong).

   ☐ Grade 6 Elementary
   ☐ Second Year High School
   ☐ Fourth Year High School

5. Name of school where you currently teach: _____
   _____
   _____

6. School Type:

   ☐ Public
   ☐ Private

7. **Including the current year**, how many years of experience do you have as a classroom teacher?

☐ 1 – 5 years	☐ 21 – 25 years
☐ 6 – 10 years	☐ 26 – 30 years
☐ 11 – 15 years	☐ More than 30 years
☐ 16 – 20 years	

B. ASSESSMENT LITERACY

Description of the ALI: *The ALI consists of five scenarios, each followed by seven questions. The items are related to the seven "Standards for Teacher Competence in the Educational Assessment of Students." Some of the items are intended to measure general concepts related to testing and assessment, including the use of assessment activities for assigning student grades and communicating the results of assessments to students and parents; other items are related to knowledge of standardized testing, and the remaining items are related to classroom assessment.*

Directions: Read each scenario followed by each item carefully; select the response you think is the best one by encircling the appropriate letter. Even if you are not sure of your choice, mark the response you believe to be the best.

Scenario #1

Mr. Kalim, a math teacher, questions how well his fourth year high school students are able to apply what they have learned in class to situations encountered in their everyday lives. Although the teacher's manual contains numerous items to test understanding of mathematical concepts, he is not convinced that giving a paper-and-pencil test is the best method for determining what he wants to know.

8. Based on the above scenario, the type of assessment that would ***best*** answer Mr. Kalim's question is called a/an _____.

A. performance assessment	C. extended response assessment
B. authentic assessment	D. standardized test

9. In order to grade his students' knowledge accurately and consistently, Mr. Kalim would be well advised to _____.

A. identify criteria from the unit objectives and create a scoring rubric
B. develop a scoring rubric after getting a feel for what students can do
C. consider student performance on similar types of assignments
D. consult with experienced colleagues about criteria that has been used in the past

10. To get a general impression of how well his students perform in mathematics in comparison to other fourth year high school students, Mr. Kalim administers a standardized math test. This practice is acceptable ***only*** if _____.

A. the reliability of the standardized test does not exceed 0.60
B. the standardized test is administered individually to students
C. the content of the standardized test is well known to students

D. the comparison group is comprised of grade level peers

Note: *Other ALI questions have been excluded from this appendix, as the original instrument is not yet in the public domain. Scenario #1 and first three items are provided following the appendix of Mertler and Campbell's (2005) paper as cited in this thesis.*

C. ASSESSMENT PRACTICES

Instructions: The following items pertain to your classroom assessment practices. Read each item carefully and indicate your response by ticking a box. **Please don't leave any item unanswered**.

	Never (1)	Seldom (2)	Occasionally (3)	Frequently (4)	All the time (5)
11. I use assessment to check the attainment of lesson objectives.	☐	☐	☐	☐	☐
12. I use assessment to establish student learning.	☐	☐	☐	☐	☐
13. I use assessment to increase student learning.	☐	☐	☐	☐	☐
14. I use assessment to develop students' higher order thinking skills.	☐	☐	☐	☐	☐
15. I prepare table of specifications as my guide in constructing test.	☐	☐	☐	☐	☐
16. I construct test that measures attribute/behaviour as stated in my teaching objectives.	☐	☐	☐	☐	☐
17. I use clear directions when giving assessment like tests and projects.	☐	☐	☐	☐	☐
18. I use answer key when marking objective tests like multiple choice, true-false and matching types.	☐	☐	☐	☐	☐
19. I use rubrics when marking other assessment types such as essay test, projects and student demonstration.	☐	☐	☐	☐	☐
20. I use reference table or standard procedure in transmuting scores into grades.	☐	☐	☐	☐	☐

	Never (1)	Seldom (2)	Occasionally (3)	Frequently (4)	All the time (5)
21. I use established procedure in deriving grades from different assessment methods.	☐	☐	☐	☐	☐
22. I interpret assessment results according to the established scale.	☐	☐	☐	☐	☐
23. I use assessment results to plan my instruction.	☐	☐	☐	☐	☐
24. I use assessment results to determine the pace of my instruction.	☐	☐	☐	☐	☐
25. I use assessment results to determine the strategies that suit my student learning needs.	☐	☐	☐	☐	☐
26. I use assessment results to provide feedback to my students.	☐	☐	☐	☐	☐
27. I explain to my students and their parents how grades are derived.	☐	☐	☐	☐	☐
28. I explain to my students and their parents the meaning of assessment results.	☐	☐	☐	☐	☐
29. I explain to my students and their parents the meaning of the national/regional examination results (e.g. average score, percentile rank, etc.).	☐	☐	☐	☐	☐
30. I write comments on student test papers.	☐	☐	☐	☐	☐
31. I write comments on student report card.	☐	☐	☐	☐	☐

D. TEACHING PRACTICES

Instructions: The following items pertain to your teaching practices. Read each item carefully and indicate your response by checking a box. **Please don't leave any item unanswered**.

	Never or hardly ever (1)	In about one-quarter of <lessons> (2)	In about one-half of <lessons> (3)	In about three-quarters of <lessons> (4)	In almost every <lesson> (5)
32. I present new topics to the class in a lecture-style presentation.	☐	☐	☐	☐	☐
33. I explicitly state learning goals.	☐	☐	☐	☐	☐
34. I review with the students the homework they have prepared.	☐	☐	☐	☐	☐
35. Students work in small group to come up with a joint solution to a problem or task.	☐	☐	☐	☐	☐
36. I give different work to students that have difficulties learning the subject matter.	☐	☐	☐	☐	☐
37. I give different work to students that can learn faster.	☐	☐	☐	☐	☐
38. I ask my students to suggest classroom activities including topics.	☐	☐	☐	☐	☐
39. I ask my students to remember every step in a procedure.	☐	☐	☐	☐	☐
40. At the beginning of the lesson, I present a short summary of the previous lesson.	☐	☐	☐	☐	☐
41. I check my students' exercise books.	☐	☐	☐	☐	☐
42. Students work on projects that require at least one week to complete.	☐	☐	☐	☐	☐

	Never or hardly ever	In about one-quarter of <lessons>	In about one-half of <lessons>	In about three-quarters of <lessons>	In almost every <lesson>
	(1)	(2)	(3)	(4)	(5)
43. I work with individual students.	☐	☐	☐	☐	☐
44. Students evaluate and reflect upon their own work.	☐	☐	☐	☐	☐
45. I check, by asking questions, whether or not the subject matter has been understood.	☐	☐	☐	☐	☐
46. Students work in groups based upon their abilities.	☐	☐	☐	☐	☐
47. Students make a product that will be used by someone else.	☐	☐	☐	☐	☐
48. I administer a test or quiz to assess student learning.	☐	☐	☐	☐	☐
49. I ask my students to write an essay in which they are expected to explain their thinking or reasoning at some length.	☐	☐	☐	☐	☐
50. Students work individually with the textbook or worksheets to practice newly taught subject matter.	☐	☐	☐	☐	☐
51. Students hold a debate or argue for a particular point of view which may not be their own.	☐	☐	☐	☐	☐

This is the end of the questionnaire.
Again, thank you very much!

Note: *Original teaching practices items/scale/instrument can be found in the Teaching and Learning International Survey (TALIS) reports (OECD, 2009a; 2010) as cited in this thesis.*

THE UNIVERSITY OF ADELAIDE
AUSTRALIA

'Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines'

(Student Questionnaire)

I. Information about this Questionnaire

This questionnaire is intended for the grade six (elementary school), second year and fourth year (secondary school) students. It contains questions that ask for information about the student and his/her assessment perceptions and attitude towards assessment. It has been divided into sections (A, B, and C) according to the said characteristics.

Your responses to this questionnaire will help improve your teacher's teaching approaches and your own learning in the classroom. Thus, it is important that you respond to each item very carefully so that the information you give will tell about your situation as accurately as possible. Your responses will be combined with the responses of other students in which you and your school will not be individually identifiable. Your responses and the information about you will be strictly kept confidential.

II. General Instructions to Student Participant:

1. Complete this questionnaire in your class. Your class adviser/teacher will help distribute and explain the instructions;

2. Read each item carefully and answer as accurately as you can. Specific instructions in answering the items are given in every section of this questionnaire. If you make a mistake in answering the item, simply mark X on your previous choice and indicate your new response by checking another box. If you make a mistake in writing the information about you, just cross out your response and write the correction next to it. **Please answer all items. If you have any question, ask your class adviser/teacher;** and

3. This questionnaire needs to be returned to your class adviser/teacher as soon as you have completed it.

Note: A class adviser/teacher is requested to communicate the information and the instructions to student participants.

Thank you very much for your time and effort!

A. GENERAL INFORMATION ABOUT YOU

Instructions: Write the information about you in the space provided. For items that have the given choices, check the correct box.

Student I.D. Number: _____

Student Name: (Optional): _____

1 Gender:

- ☐ Boy
- ☐ Girl

2 Grade/Year Level:

- ☐ Grade 6 Elementary
- ☐ Second Year High School
- ☐ Fourth Year High School

3 School Name: _____

B. ASSESSMENT PERCEPTIONS

Instructions: This section is about your perceptions towards test and assignment. Read each item carefully and answer by checking one box only [√] . Please answer all the items.

		Almost Never (1)	Sometimes (2)	Often (3)	Almost Always (4)
1.	Tests in my subject measure what I know.	☐	☐	☐	☐
2.	How I am tested is the same with what I do in class.	☐	☐	☐	☐
3.	I am tested on what the teacher has taught me.	☐	☐	☐	☐
4.	Tests in my subject measure my ability to apply what I learn to real-life situations.	☐	☐	☐	☐
5.	Tests in my subject measure my ability to answer everyday questions.	☐	☐	☐	☐
6.	I am aware how my tests will be marked.	☐	☐	☐	☐
7.	I understand what is needed to successfully complete the test.	☐	☐	☐	☐
8.	I am told in advance when I am being tested.	☐	☐	☐	☐
9.	I am told in advance on what I am being tested.	☐	☐	☐	☐
10.	I understand what my teacher wants in my test.	☐	☐	☐	☐
11.	I have as much chance as any other student at completing the test.	☐	☐	☐	☐
12.	My assignments, including project, are about what I have done in class.	☐	☐	☐	☐
13.	My assignments, including project, are related to what I do outside of school.	☐	☐	☐	☐
14.	I am aware how my assignments will be marked.	☐	☐	☐	☐
15.	I understand what is needed to successfully complete my assignment tasks.	☐	☐	☐	☐

	Almost Never (1)	Sometimes (2)	Often (3)	Almost Always (4)
16. I understand what my teacher wants in my assignments, including project.	☐	☐	☐	☐
17. I have as much chance as any other student at completing my assignments, including project.	☐	☐	☐	☐
18. I complete my assignments, including project, at my own speed.	☐	☐	☐	☐

Note: *Original Students' Perceptions of Assessment Questionnaire can be found in Cavanagh, et al.'s (2005) and Waldrip, et al.'s (2008) papers as cited in this thesis.*

C. ATTITUDE TOWARDS ASSESSMENT

Instructions: This section is about your attitude towards assessment like test, assignment or project. Read each item carefully and answer by checking one box ☐ only. Please answer all the items.

	Strongly Disagree (1)	Disagree (2)	Agree (3)	Strongly Agree (4)
19. Assessment helps me to become successful in my education.	☐	☐	☐	☐
20. If everyone in my school is given an effective assessment, we can gain good education.	☐	☐	☐	☐
21. Assessment in school leads to good academic achievement.	☐	☐	☐	☐
22. I have a chance to be successful if I do well in my tests in school.	☐	☐	☐	☐

This is the end of the Questionnaire. Thank you very much!

Note: *Original attitude items/scale/instrument can be found in Mickelson's (1990) paper as cited in this thesis.*

Appendix C
Information Sheets, and Consent and Complaint Forms

THE UNIVERSITY OF ADELAIDE
AUSTRALIA

School of Education

Level 8, 10 Pulteney Street, University of Adelaide, Adelaide SA 5005; Tel: (+618) 8303 7196, Fax: (+618) 8303 3604

RESEARCH PROJECT INFORMATION SHEET

Dear Colleague,

I am Wilham Hailaya, a faculty member of the Mindanao State University at Tawi-Tawi. I am currently pursuing a degree of Doctor of Philosophy (PhD) in Education specialising in Educational Assessment at the University of Adelaide under the Australian Leadership Awards Scholarship. I am presently conducting a research study leading to the production of thesis on the subject, **Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines**.

The main goal of this research is to investigate the current level of assessment literacy of the elementary and secondary school teachers and its possible link to their assessment practices, teaching practices, and student outcomes namely, student perceptions of assessment, student attitude towards assessment, academic achievement, and aptitude, in the province of Tawi-Tawi. The project is intended to further the research that has been conducted on the topic, but using the Tawi-Tawi/Philippine context. If successful, the results of this study are expected to provide teachers and educational leaders with useful information that can be one of the bases in enhancing teaching and learning, in designing teacher development programs, and in helping improve the quality of basic education in the province of Tawi-Tawi, and possibly in the entire country.

In this research, you are requested to accomplish the survey questionnaire (Teacher Questionnaire) at your free time in school. The questionnaire contains open-ended questions that ask for general information about you and Likert-type questions (questions with the given scale/options/choices) that ask about assessment principles (assessment literacy), assessment practices, and teaching practices. The questionnaire will be collected as soon as it has been completed.

In this study, some teachers will be invited to an interview to obtain their views about assessment methods/tools. The interview will be face-to-face and semi-structured and will be tape-recorded. It will be held for at most 45 minutes at a place and time that are convenient to teachers. If you have been identified for the interview, the researcher will approach you after completing the questionnaire.

In conducting this study, **ethics are strictly observed**. Hence, be assured that any information provided will be treated with strictest confidentiality and neither you nor your school will be individually identifiable in the resulting thesis, report or other publications.

Should you need additional information about this research, please contact me at mobile number +639296819734 or email me at wilham.hailaya@adelaide.edu.au. Alternatively, you can also contact my principal supervisor, Dr. Sivakumar Alagumalai, by telephone on (+618) 8303-5630 in Australia, or email him at sivakumar.alagumalai@adelaide.edu.au.

My sincerest thanks for your participation in this study.

Sincerely yours,

WILHAM M. HAILAYA

THE UNIVERSITY OF ADELAIDE HUMAN RESEARCH ETHICS COMMITTEE

STANDARD CONSENT FORM
FOR PEOPLE WHO ARE PARTICIPANTS IN A RESEARCH PROJECT
(For Teacher Participants in the Province of Tawi-Tawi, Philippines)

1. I, …… *(please print name)*

   consent to take part in the research project entitled:
   "Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines"

2. I acknowledge that I have read the attached Information Sheet entitled:
   Research Project Information Sheet

3. I have had the project, so far as it affects me, fully explained to my satisfaction by the research worker. My consent is given freely.

4. Although I understand that the purpose of this research project is to investigate teachers' assessment literacy, assessment practices, teaching effectiveness, and student outcomes, it has also been explained that my involvement may not be of any benefit to me.

5. I have been given the opportunity to have a member of my family or a friend present while the project was explained to me.

6. I have been informed that, while information gained during the study may be published, I will not be identified and my personal results will not be divulged.

7. When the interview will be held, I understand that it will be audio recorded.

8. I understand that I am free to withdraw from the project at any time and that this will not affect my professional progress, now or in the future.

9. I am aware that I should retain a copy of this Consent Form, when completed, and the attached Information Sheet.

   ……...
   (signature) *(date)*

WITNESS

I have described to …….. *(name of subject)* of

…….. *(name of institution or school)*
the nature of the research to be carried out. In my opinion she/he understood the explanation.
Status in Project**:** …….

Name: ……..….

……...
(signature) *(date)*

THE UNIVERSITY OF ADELAIDE
AUSTRALIA

School of Education

Level 8, 10 Pulteney Street ,University of Adelaide , Adelaide SA 5005; Tel: (+618) 8303 7196, Fax: (+618) 8303 3604

RESEARCH PROJECT INFORMATION SHEET

(For Parents of the Student Participants)

Dear Parent,

I am Wilham Hailaya, a teacher from the Mindanao State University at Tawi-Tawi. I am currently taking a degree of Doctor of Philosophy (PhD) in Education at the University of Adelaide under the Australian Leadership Awards Scholarship. I am presently conducting a research study leading to the production of thesis on the subject, **Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines**.

In this research, your child will be asked to complete the survey questionnaire (Student Questionnaire). The topics covered in the questionnaire and the specific timeframe are mentioned below. Your child will be expected to complete the questionnaire within 25 minutes.

- Student Questionnaire

○ *General Information*	~ 5 minutes
○ Assessment Perceptions	~ 15 minutes
○ Attitude Towards Assessment	~ 5 minutes
Total Time	~ *25 minutes*

The student questionnaire will be distributed by the teacher during break time or right after school in a classroom. The teacher will be requested to collect all questionnaires at the end of the specified time.

The main purpose of this research is to investigate the present level of assessment literacy of the elementary and secondary school teachers and its possible link to student outcomes in the province of Tawi-Tawi. It specifically aims to examine the relationships of teachers' assessment literacy with their assessment practices, teaching practices, and student outcomes namely, assessment perceptions, assessment attitude, academic achievement, and aptitude. The project is intended to further the research that has been conducted on the topic, but using the Tawi-Tawi/Philippine context.

From this project, I hope to develop a model that illustrates the relationships and influence of teachers' assessment literacy on their assessment practices, teaching practices, and student outcomes. If successful, the results of this study would provide teachers and educational leaders with useful information that can be one of the bases in enhancing teaching and learning, in designing teacher development programs, and in helping improve the quality of basic education in the province of Tawi-Tawi, and possibly in the entire country.

In conducting this study, ethics is strictly observed. Hence, be assured that any information provided by your child will be treated with strictest confidentiality and neither your child nor his/her school will be individually identifiable in the resulting thesis, report or other publications. Your child is, of course, entirely free to discontinue his/her participation at any time. Since participation is purely *VOLUNTARY*, non-participation will not affect your child's academic

progress in the school in any way. However, if you allow your child to participate in this project, you will be helping me with my study.

Should you need additional information regarding this research, please contact me by telephone on (+618) 8303-7196, mobile +61433403674, or email at wilham.hailaya@adelaide.edu.au. Should I be unavailable, my Principal Supervisor, Dr. Sivakumar Alagumalai, can also be contacted by telephone on (+618) 8303-5630, or email at sivakumar.alagumalai@adelaide.edu.au.

Please see the attached independent complaints procedure form should you have any complaints about this project.

Thank you for considering this request.

Signed,

Wilham M. Hailaya

THE UNIVERSITY OF ADELAIDE HUMAN RESEARCH ETHICS COMMITTEE

STANDARD CONSENT FORM

For Research to be Undertaken on a Child, and those in Dependant Relationships or Comparable Situations

To be Completed by Parent or Guardian

1. I, ……..…. *(please print name)*

   consent to allow ……... *(please print name)*

   to take part in the research project entitled:

   "Teacher Assessment Literacy and Student Outcomes in the Division of Tawi-Tawi, Philippines"

2. I acknowledge that I have read the attached Information Sheet entitled:

   Research Project Information Sheet

   and have had the project, as far as it affects …… *(name)* fully explained to me by the research worker. My consent is given freely.

   IN ADDITION, I ACKNOWLEDGE THE FOLLOWING ON BEHALF OF …… *(name)*

3. Although I understand that the purpose of this research project is to investigate teachers' assessment literacy, assessment practices, teaching effectiveness, and student outcomes, it has also been explained that my child's involvement may not be of any benefit to me or my child.

4. I have been informed that the information I/he/she provides will be kept confidential. Names will not be disclosed and personal results will not be divulged.

5. In case student interviews will be needed, I understand that they will be audio recorded.

6. In case parent interviews will be needed, I understand that a questionnaire will be sent to my mailing address or e-mail address which I may choose to complete and return to the researcher.

7. I understand that I/my child is free to withdraw from the project at any time and that this will not affect his/her academic progress, now or in the future.

8. I am aware that I should retain a copy of this Consent Form, when completed, and the attached Information Sheet.

……Parent/Guardian ……

(signature and please indicate relationship) *(date)*

----------Lower portion to be returned to Class teacher

CONSENT SLIP

I **agree**/**do not agree** for ……*(name of child)* to participate in this research endeavour **"Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines"**. I understand that my child's participation / non-participation to this project will not affect his/her academic progress, now or in the future.

Name of child: …….. Signature…….. Date……
Name of parent:…… Signature…….. Date……

THE UNIVERSITY OF ADELAIDE
HUMAN RESEARCH ETHICS COMMITTEE

THE UNIVERSITY OF ADELAIDE AUSTRALIA

Level 7, 115 Grenfell Street, The University of Adelaide, SA 5005; Tel: (+618) 8303-7196, Fax (+618) 8303-3700

CONTACTS FOR INFORMATION ON PROJECT AND INDEPENDENT COMPLAINTS PROCEDURE

The Human Research Ethics Committee is obliged to monitor approved research projects. In conjunction with other forms of monitoring it is necessary to provide an independent and confidential reporting mechanism to assure quality assurance of the institutional ethics committee system. This is done by providing research participants with an additional avenue for raising concerns regarding the conduct of any research in which they are involved.

The following study has been reviewed and approved by the University of Adelaide Human Research Ethics Committee:

Project title: **Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines.**

1. If you have questions or problems associated with the practical aspects of your participation in the project, or wish to raise a concern or complaint about the project, then you should consult the project coordinator:

   Name: Wilham M. Hailaya (Researcher)
   Telephone: (+618) 8303- 7196 / +61433403674
   Email: wilham.hailaya@adelaide.edu.au

   Name: Dr. Sivakumar Alagumalai (Principal Supervisor)
   Telephone: (+618) 8303-5630
   Email: sivakumar.alagumalai@adelaide.edu.au

2. If you wish to discuss with an independent person matters related to

   - making a complaint, or
   - raising concerns on the conduct of the project, or
   - the University policy on research involving human participants, or
   - your rights as a participant

   contact the Human Research Ethics Committee's Secretary on phone (+618) 8303-6028.

Appendix D
Letters of Request for Permission to Conduct the Study

THE UNIVERSITY OF ADELAIDE
AUSTRALIA

School of Education

Level 8, 10 Pulteney St., Adelaide SA 5005; Tel: (08) 8303 7196, Fax: (08) 8303 3604

1 November 2010

BR. ARMIN A. LUISTRO, FSC
Secretary, Department of Education (DepEd)
DepEd Complex, Meralco Ave., Pasig City 1600
Philippines

Dear Br. Luistro:

I am Wilham Hailaya, a faculty member of the Mindanao State University at Tawi-Tawi. I am currently pursuing a degree of Doctor of Philosophy (PhD) in Education specialising in Educational Assessment at the University of Adelaide under the Australian Leadership Awards Scholarship. I am presently undertaking a research leading to the production of thesis on the subject, **Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines**.

The main goal of the research is to investigate the current level of assessment literacy of the elementary and secondary school teachers and its possible link to student outcomes in the province of Tawi-Tawi. It specifically aims to examine the influence of teachers' assessment literacy on their assessment practices, teaching practices, and student outcomes namely, perceptions of assessment, attitude towards assessment, academic achievement, and aptitude. The study seeks to contribute to the enrichment of the literature on assessment literacy of teachers. The findings are expected to provide useful information that can be one of the bases in designing teacher development programs and to support the efforts in improving the quality of basic education in the province of Tawi-Tawi, and possibly in the entire country.

In this regard, I would like to seek permission from your office to collect the necessary data using survey questionnaires and interviews from the elementary and secondary schools in the Division of Tawi-Tawi. Samples will come from the grade six elementary teachers and pupils and from the second year and fourth year secondary teachers and students. If granted permission, letters will be sent to the Tawi-Tawi Division Schools' Superintendent, district supervisors, principals, and parents to inform them about the project and to access the lists of school districts, schools, and teacher and student participants. Any information provided will be treated with strictest confidentiality and neither participants nor schools/school districts will be individually identifiable in the resulting thesis, report or other publications. The respondents will, of course, be entirely free to discontinue their participation at any time or to decline to answer particular questions in the study. Since participation is purely voluntary, non-participation will not affect teachers' employment status and students' academic progress in any way.

Once approval has been given at the local level, I will take the responsibilities to obtain the informed consent, to maintain the confidentiality of participant identity, and to ensure that safety precautions are in place. I will also provide the department with a copy of the final report, which can be circulated to interested staff and be made available to educators for future reference.

For any additional information or further question in relation to this research, please contact me by telephone on (+61) 8303-7196, mobile +61433-403-674, or email at wilham.hailaya@adelaide.edu.au. Should I be unavailable, my principal supervisor, Dr. Sivakumar Alagumalai, can also be contacted by telephone at (08) 8303-5630 or email at sivakumar.alagumalai@adelaide.edu.au.

Thank you very much for your attention and assistance.

Sincerely yours,

WILHAM M. HAILAYA

THE UNIVERSITY OF ADELAIDE AUSTRALIA

School of Education

Level 8, 10 Pulteney St., Adelaide SA 5005; Tel: (08) 8303 7196, Fax: (08) 8303 3604

_____
_____
_____
_____
_____

Dear Sir/Madam:

I am **Wilham Hailaya**, a faculty member of the Mindanao State University at Tawi-Tawi. I am currently pursuing a degree of Doctor of Philosophy (PhD) in Education at the University of Adelaide under the Australian Leadership Awards Scholarship. As a requirement for the completion of my Ph.D. program, I am conducting a research study titled, "***Teacher Assessment Literacy and Student Outcomes in the Province of Tawi-Tawi, Philippines***".

The main goal of this research is to investigate the level of assessment literacy of the elementary and secondary school teachers and its possible link to teachers' assessment practices, teaching practices, and student outcomes namely, assessment perceptions, assessment attitude, academic achievement, and aptitude, in the province of Tawi-Tawi. If successful, the results of this study are expected to provide teachers and educational leaders with useful information that can be one of the bases in enhancing teaching and learning, in designing teacher development programs, and in helping improve the quality of basic education in the province of Tawi-Tawi, and possibly in the entire country.

As school(s) under your jurisdiction is (are) included in my research study, I would like to seek permission to conduct surveys to your Grade Six/Second Year/Fourth Year teachers and students. I also would like to seek permission to administer interviews to your selected teachers in the said grade/year levels.

Attached are permissions from the Department of Education (DepEd) Central Office, DepEd-ARMM Regional Office, and DepEd-Tawi-Tawi Division Office for your reference.

Thank you very much.

Very respectfully yours,

WILHAM M. HAILAYA

Appendix E
Multicollinearity Test Results/VIF

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-510.675	296.731		-1.721	.085		
	TSEX_M_1	21.247	3.983	.104	5.335	.000	.747	1.339
	AGE1	.109	7.185	.000	.015	.988	.700	1.429
	AGE2	-35.096	5.851	-.130	-5.998	.000	.604	1.656
	AGE4	-2.678	5.184	-.012	-.517	.606	.551	1.816
	AGE5	11.096	7.305	.037	1.519	.129	.465	2.149
	AGE6	27.666	12.775	.049	2.166	.030	.554	1.806
	ACAD_PG_1	-26.936	5.168	-.095	-5.212	.000	.845	1.183
	SCH_PUB_1	75.892	8.504	.169	8.924	.000	.792	1.262
	EXYEAR1	8.583	5.324	.038	1.612	.107	.508	1.969
	EXYEAR3	8.251	5.616	.032	1.469	.142	.592	1.690
	EXYEAR4	-19.234	7.247	-.063	-2.654	.008	.500	2.002
	EXYEAR5	23.740	7.493	.080	3.168	.002	.439	2.277
	EXYEAR6	29.552	13.595	.048	2.174	.030	.581	1.721
	EXYEAR7	18.967	17.725	.021	1.070	.285	.702	1.425
	SSEX_M_1	-14.434	3.434	-.072	-4.203	.000	.958	1.044
	ALI_WS	1.420	2.311	.078	.615	.539	.017	57.581
	STAND1_WS	-.574	.353	-.055	-1.627	.104	.251	3.981
	STAND2_WS	-.355	.381	-.030	-.930	.352	.265	3.777
	STAND3_WS	1.041	.387	.112	2.688	.007	.162	6.162
	STAND4_WS	-.644	.472	-.057	-1.366	.172	.166	6.041
	STAND5_WS	-.298	.381	-.031	-.782	.434	.175	5.710
	STAND6_WS	.575	.421	.052	1.364	.173	.192	5.204
	STAND7_WS	.054	.395	.006	.136	.892	.161	6.199
	API_WS	-2.812	.881	-.270	-3.192	.001	.040	25.178
	PUR_WS	2.053	.350	.276	5.867	.000	.128	7.816
	DES_WS	.606	.347	.073	1.746	.081	.163	6.149
	COM_WS	.258	.226	.048	1.140	.254	.157	6.380
	TPS_WS	-5.634	1.826	-.344	-3.086	.002	.023	43.884
	STRUCT_WS	1.369	.883	.092	1.550	.121	.080	12.509
	STUDOR_WS	2.350	.603	.195	3.898	.000	.113	8.844
	ENACT_WS	.719	.470	.062	1.528	.127	.174	5.737
	SPAS_WS	8.041	2.312	.497	3.478	.001	.014	71.952
	PTT_WS	-4.891	1.431	-.334	-3.418	.001	.030	33.765
	PTA_WS	-2.158	.985	-.144	-2.190	.029	.065	15.329
	SATS_WS	.559	.161	.070	3.463	.001	.692	1.445

a. Dependent Variable: NAT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-868.280	358.057		-2.425	.015		
	TSEX_M_1	21.378	4.447	.107	4.808	.000	.725	1.380
	AGE1	-66.867	10.628	-.155	-6.292	.000	.585	1.709
	AGE2	52.484	13.199	.091	3.976	.000	.687	1.456
	AGE4	8.196	6.209	.037	1.320	.187	.452	2.213
	AGE5	40.039	9.837	.159	4.070	.000	.233	4.284
	AGE6	12.054	14.695	.029	.820	.412	.290	3.444
	ACAD_PG_1	18.497	5.384	.085	3.436	.001	.584	1.711
	SCH_PUB_1	20.701	6.335	.089	3.268	.001	.481	2.079
	EXYEAR2	45.695	7.792	.183	5.864	.000	.365	2.742
	EXYEAR3	53.183	7.866	.190	6.761	.000	.450	2.222
	EXYEAR4	-23.308	8.735	-.082	-2.668	.008	.380	2.632
	EXYEAR5	4.994	10.061	.018	.496	.620	.264	3.783
	EXYEAR6	48.530	14.560	.117	3.333	.001	.289	3.454
	EXYEAR7	43.107	14.167	.132	3.043	.002	.189	5.287
	SSEX_M_1	-3.862	3.932	-.019	-.982	.326	.946	1.057
	ALI_\W/S	6.352	3.438	.467	1.848	.065	.006	179.005
	STAND1_\W/S	-1.347	.425	-.193	-3.165	.002	.096	10.402
	STAND2_\W/S	-2.654	.579	-.282	-4.582	.000	.094	10.619
	STAND3_\W/S	-2.264	.627	-.233	-3.609	.000	.085	11.712
	STAND4_\W/S	-.770	.656	-.070	-1.173	.241	.100	10.008
	STAND5_\W/S	.901	.568	.091	1.587	.113	.108	9.302
	STAND6_\W/S	.894	.689	.092	1.298	.195	.070	14.194
	STAND7_\W/S	.038	.584	.003	.065	.948	.124	8.084
	API_\W/S	-5.879	1.249	-.591	-4.706	.000	.023	44.186
	PUR_\W/S	2.773	.463	.399	5.986	.000	.081	12.420
	DES_\W/S	1.411	.483	.129	2.922	.004	.182	5.486
	COM_\W/S	1.395	.332	.275	4.205	.000	.083	11.979
	TPS_\W/S	.102	2.678	.006	.038	.970	.014	71.821
	STRUCT_\W/S	2.021	1.175	.131	1.719	.086	.061	16.371
	STUDOR_\W/S	-1.270	.849	-.108	-1.496	.135	.069	14.567
	ENACT_\W/S	.949	.762	.080	1.245	.213	.086	11.586
	SPAS_\W/S	9.682	4.864	.480	1.990	.047	.006	162.839
	PTT_\W/S	-7.665	2.991	-.422	-2.562	.010	.013	75.961
	PTA_\W/S	-2.957	1.947	-.169	-1.518	.129	.029	34.563
	SATS_\W/S	.896	.183	.104	4.884	.000	.793	1.261

a. Dependent Variable: NCAE