

Development and Application of a Teacher Assessment for Learning (AfL) Literacy Tool

Author:

Alonzo, Dennis

Publication Date:

2016

DOI:

<https://doi.org/10.26190/unsworks/18809>

License:

<https://creativecommons.org/licenses/by-nc-nd/3.0/au/>

Link to license to see what you are allowed to do with this resource.

Downloaded from <http://hdl.handle.net/1959.4/55663> in <https://unsworks.unsw.edu.au> on 2024-04-26

Development and Application of a Teacher Assessment *for* Learning (AfL) Literacy Tool

Dennis Antonio Alonzo

A thesis in fulfilment of the requirements for the degree of
Doctor of Philosophy

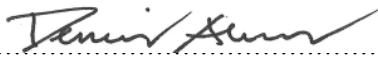


School of Education
Arts and Social Sciences

2016

ORIGINALITY STATEMENT

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed 

19 March 2016
Date

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: ALONZO

First name: DENNIS

Other name/s:

Abbreviation for degree as given in the University calendar: PhD

School: Education

Faculty: Arts and Social Sciences

Title: Development and Application of Teacher Assessment *for* Learning (AfL) Literacy Tool

Abstract

This research presents an empirically driven-tool for teacher assessment for learning (AfL) literacy in response to the weaknesses of existing tests, rating scales and other forms of assessments used to evaluate teacher assessment literacy. This tool is drawn from the philosophical framework of AfL and assessment innovation and change, using a probabilistic model of competence to link latent trait theory to Vygotsky's zone of proximal development. Theoretical and empirical approaches were combined to develop a teacher AfL literacy tool with criteria and explicit standards. Exploratory, confirmatory factor and second-order factor analyses, as well as exploratory structural equation modelling were used to establish the dimensionality of teacher AfL literacy. In addition, the generalised partial credit model of Rasch model was used to explore the characteristics of the tool at the item level, and latent profile analysis was employed to explore the classes of teachers that can be identified based on their AfL literacy profile.

Results show that there are six dimensions at the first-order factor level that can describe teacher roles as *assessors, pedagogy experts, student partners, motivators, teacher learners, and stakeholder partners*. A single general factor emerged at the second-order factor level, which was labelled as *teachers as AfL literate professionals*. At the item level, the tool met the measurement requirements in terms of reliability, dimensionality, difficulty and item discrimination. Latent profile analysis extracted five classes of teachers based on their AfL practices, a result which established the typology of teachers' stages of AfL development.

The findings highlight new conceptualisations of teacher AfL literacy, including new ways of categorising and describing teachers' AfL practices, understanding the interdependence of assessment skills, highlighting the most important assessment skills of teachers, and developing a typology of stages of teacher AfL literacy development. Furthermore, the robust empirical evidence supports the utilisation of the tool for teachers' self-assessment, for collegial evaluation of teaching practices, and for school heads' monitoring of teachers' AfL literacy development to support teachers to develop high levels of confidence and skills in making highly contextualised, consistent, fair and trustworthy assessment decisions to more effectively support student learning.

3

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).



19 March 2016

.....
Signature

.....
Witness

.....
Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY

Date of completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

PUBLICATION AND CONFERENCE PRESENTATIONS

- Alonzo, D., & Davison, C. (2012, November). *Evaluating EAL/D teacher assessment literacy: Towards a standards-based framework*. Association for Language Testing and Assessment of Australia and New Zealand. Sydney, Australia.
- Alonzo, D., Davison, C., & Ahmad, K. (2013, March). *A competency-based teacher assessment literacy: Development and application*. Assessment in Language Education. 48th Regional Language Centre. Southeast Asian Ministers Education Organization. Singapore.
- Alonzo, D., & Davison, C. (2013, March). *Evaluating teacher assessment literacy: Towards a standards-based framework*. American Association for Applied Linguistics. Texas, USA.
- Alonzo, D., & Davison, C. (2013, December). *Designing an EAL teacher assessment literacy framework: Linking results to professional development*. Australian Education Research Association. Adelaide, Australia.
- Alonzo, D., & Davison, C. (2013, December). *An application of latent trait theory in evaluating teacher assessment literacy and addressing professional needs*. Australian Education Research Association. Adelaide, Australia.
- Alonzo, D., & Davison, C. (2014, March). *An application of latent trait theory in evaluating language teacher assessment literacy and addressing professional needs*. American Association for Applied Linguistics. Portland, USA.
- Alonzo, D., Davison, C., Ahmad, K., & Aliuddin, D. (2014, May). *Enhancing assessment leadership of school heads*. International Association for Educational Assessment (IEAE) Conference. Singapore.
- Alonzo, D., & Davison, C. (2014, August). *Designing an EAL teacher assessment literacy framework: Linking results to professional development (Symposium)*. 17th World Congress of the International Applied Linguistics Association Conference. Brisbane, Australia.
- Alonzo, D., Mirriahi, N., & Davison, C. (2014, September). *Putting standards into the standards-based assessment practices*. British Educational Research Association. Institute of London, London.
- Alonzo, D., & Davison, C. (2014, September). *A teacher assessment for learning (AfL) competency framework: Development and application*. British Educational Research Association. Institute of London, London.
- Alonzo, D., & Davison, C. (2014, September - October). *Developing EAL teachers' self-reflective practice (Workshop)*. Australian Council of TESOL Associations, Melbourne.
- Alonzo, D., Davison, C., & Davies, G. (2014, September – October). *Development of an assessment for learning (AfL)-infused unit of work (Workshop)*. Australian Council of TESOL Associations, Melbourne.

Alonzo, D., Lee, J., & Davison, C. (2015, April). *A teacher assessment for learning (AfL) competency framework: Development and application*. American Educational Research Association. Chicago, USA.

Alonzo, D., & Davison, C. (in press). Developing a framework for evaluating assessment literacy in Brunei. In C. Davison & K. Hill, *Teacher assessment literacy in second and foreign language education*. Springer.

INVITED SPEAKER/WORKSHOP RELATED TO THIS RESEARCH

- Alonzo, D. (2013, October). *The TEAL Project: Building EAL assessment for learning communities*. International Competition and Assessment for Schools. Educational Assessment Australia. Sydney, Australia.
- Alonzo, D. (2013, October). *Developing teacher self-reflective practice*. Enhancing Language and literacy assessment for EAL/D Learners. TESOL Seminar 4. New South Wales Education & Communities. Sydney, Australia.
- Alonzo, D. (2014, October). *The need for assessment leadership in building a community of assessment literate science teachers*. Assessment in Schools Conference. Educational Assessment Australia. Sydney, Australia.

ACKNOWLEDGMENTS

I started my PhD with a very high level of motivation and enthusiasm. As I progressed in my candidature, the overwhelming pressure and demands of the task threw all of these into the air. I thank God for surrounding me with people who, when I lost my sense of balance, pushed me back on track.

I owe the deepest gratitude to my supervisors Prof. Chris Davison, Dr. Jihyun Lee and Prof. John Bennett. Chris, your trust and confidence in me made me think I was fantastic, even when the reality was, I was in constant crisis. I would like to put on record your support for me including helping me settle in Sydney, providing opportunities for work and research engagement, widening my professional networks, and sourcing funds for attending conferences. Your warmth and your kindness gave me tremendous support when I needed it most and your unmatched expertise is reflected in this work. Jihyun, your ability to see my potential rather than acknowledging my limitations sustained me in the intensely painful yet highly rewarding process of navigating the Mplus software. Your suggestions and professional advice about statistical analyses were fundamentally helpful. John, your advice to stop and focus when I was going to include everything I thought was needed to complete this research made me realise that my PhD was just the beginning of a real academic life. I always valued your support, encouragement, and inspiration.

I would like to thank the School of Education staff who gave advice and encouragement, particularly to Dr. Jae Yup Jared Jung, for answering my emails on issues related to my initial data analysis; and to Prof. Colin Evers for his encouragement and inspiration.

It was an honour to receive an Australia Development Awards Scholarship from the Commonwealth of Australia, without which, this thesis would have not been

realised. Also, I would like to acknowledge the support of the Philippine government, particularly the University of Southeastern Philippines.

I would like to give credit to my team in Optimise International (B) Sdn Bhd: Mr. Brian Galbraith for liaising with the Ministry of Education of Brunei to provide funds for this research and access to the teachers and school heads; and Kalthom Ahmad for your tremendous help in developing and translating the assessment tool into a Malay version.

I was particularly grateful to the Ministry of Education of Brunei for supporting this study through the Deputy Permanent Secretary Haji Suhaila Bin Haji Abd Karim. Likewise, deep appreciation goes to Dina Aliuddin, Hamid Hussein, Baharin Ibrahim, Johnson Lim Chee Siong for organising the logistics needed for the data gathering.

I would also like to express my deep gratitude and love to my wife, Celeste for allowing me to live like I was single and childless just to complete this thesis, and to my children, Adriel and Driego, for reminding me that there are more important things in life than research. The hugs and kisses you gave when Daddy was falling apart put every piece back together.

To my father, who missed seeing the finish of this project, you are terribly missed! My mother, sisters, and in-laws, nieces and nephews, my deepest appreciation for your support, prayers and words of encouragement.

My family down under: Roland and Jenny, Henry and Olivia, Giovanni and Nowena, Jun and Neclette, and Kuya Geng and Ate Joy, your support was phenomenal.

Lastly, I would like to acknowledge my friends: Jason, Esmond, Kyle, Fredd, and Harold for their words of encouragement.

ABSTRACT

This research presents an empirically driven-tool for teacher assessment for learning (AfL) literacy in response to the weaknesses of existing tests, rating scales and other forms of assessments used to evaluate teacher assessment literacy. This tool is drawn from the philosophical framework of AfL and assessment innovation and change, using a probabilistic model of competence to link latent trait theory to Vygotsky's zone of proximal development. Theoretical and empirical approaches were combined to develop a teacher AfL literacy tool with criteria and explicit standards. Exploratory, confirmatory factor and second-order factor analyses, as well as exploratory structural equation modelling were used to establish the dimensionality of teacher AfL literacy. In addition, the generalised partial credit model of Rasch model was used to explore the characteristics of the tool at the item level, and latent profile analysis was employed to explore the classes of teachers that can be identified based on their AfL literacy profile.

Results show that there are six dimensions at the first-order factor level that can describe teacher roles as *assessors*, *pedagogy experts*, *student partners*, *motivators*, *teacher learners*, and *stakeholder partners*. A single general factor emerged at the second-order factor level, which was labelled as *teachers as AfL literate professionals*. At the item level, the tool met the measurement requirements in terms of reliability, dimensionality, difficulty and item discrimination. Latent profile analysis extracted five classes of teachers based on their AfL practices, a result which established the typology of teachers' stages of AfL development.

The findings highlight new conceptualisations of teacher AfL literacy, including new ways of categorising and describing teachers' AfL practices, understanding the interdependence of assessment skills, highlighting the most important assessment skills of teachers, and developing a typology of stages of teacher AfL literacy development.

Furthermore, the robust empirical evidence supports the utilisation of the tool for teachers' self-assessment, for collegial evaluation of teaching practices, and for school heads' monitoring of teachers' A/fL literacy development to support teachers to develop high levels of confidence and skills in making highly contextualised, consistent, fair and trustworthy assessment decisions to more effectively support student learning.

TABLE OF CONTENTS

ORIGINALITY STATEMENT	ii
PUBLICATION AND CONFERENCE PRESENTATIONS	iv
INVITED SPEAKER/WORKSHOP RELATED TO THIS RESEARCH	vi
ACKNOWLEDGMENTS	vii
ABSTRACT	ix
TABLE OF CONTENTS	xi
LIST OF TABLES	xvii
LIST OF FIGURES	xx
LIST OF ACRONYMS	xxi
CHAPTER 1 – INTRODUCTION	1
1.1. Introduction.....	1
1.2. Background of the Study.....	1
1.3. Rationale for the Study	4
1.4. Context of the Study.....	7
1.5. Aims of the Study	11
1.6. Research Questions	12
1.7. Summary and Structure of the Thesis	14
CHAPTER 2 - TEACHER ASSESSMENT PRACTICES AND DEVELOPMENT OF AfL LITERACY	15
2.1. Introduction.....	15
2.2. Empirical Evidence and Theoretical Support for Teacher Assessment Practices.....	15
2.2.1. Overview.....	15
2.2.2. Empirical Evidence for Teacher Assessment Practices	15
2.2.3. Theoretical Support for Teachers' AfL Practices	19
2.3. The Development of the Assessment <i>for</i> Learning Concept.....	30
2.3.1. Overview	30
2.3.2. Wave 1: Adoption, Dichotomisation and Formalisation	31
2.3.3. Wave 2: Complexification and Exploration	34
2.3.4. Wave 3: Period of Realisation and Reconciliation	42
2.4. Definition and Development of Teacher AfL Literacy	49
2.4.1. Overview	49
2.4.2. The Changing Conception of Assessment Literacy	50
2.4.3. The Existing Dimensions of Teacher Assessment Literacy	52
2.4.4. Teacher Assessment Literacy in the Context of AfL.....	56
2.4.5. Key Characteristics of Teacher AfL Literacy	58

2.4.6.	Key Principles of Teacher AfL Literacy Professional Development Programs ..	61
2.5.	Theoretical and Conceptual Framework of my Study	68
2.5.1.	Overview	68
2.5.2.	Assessing Student Learning	70
2.5.3.	Using Assessment Information to Plan a Lesson	71
2.5.4.	Engaging Students in Learning and Teaching Processes	73
2.5.5.	Using Assessment to Develop Student Motivation	77
2.5.6.	Using Student Data to Identify Professional Needs	78
2.5.7.	Informing Stakeholders of Assessment Practices	79
2.9.	Summary	80
CHAPTER 3 – THE EVALUATION OF TEACHER AfL LITERACY: THE USE OF FRAMEWORKS AND TOOLS		82
3.1.	Introduction	82
3.2.	Existing Frameworks for Mapping Teacher Assessment Practices	83
3.2.1.	Overview	83
3.2.2.	Descriptive Frameworks and Models	83
3.2.3.	Prescriptive Frameworks and Models	85
3.2.4.	A Research-Based Model	87
3.3.	Teacher Assessment Literacy in Teacher Professional Standards	88
3.3.1.	Overview	88
3.3.2.	Teacher Assessment Literacy in Teacher Professional Standards	88
3.3.3.	Career Stages of Teacher Development	92
3.4.	Existing Teacher Assessment Literacy Evaluation Tools	94
3.4.1.	Overview	94
3.4.2.	Competing Teacher Assessment Literacy Evaluation Tools	94
3.4.3.	The Need for a Robust Methodology in Tool Development	99
3.4.4.	The Need for Criteria and Standards for AfL Literacy	100
3.5.	Approaches in Tool Development	104
3.5.1.	Overview	104
3.6.2.	Theoretical Approach to Scale Development	104
3.5.2.	Issues Associated with the Use of an Empirical Approach	106
3.5.3.	The Rasch Model	114
3.6.	Factor Analyses	118
3.11.1.	Overview	118
3.11.2.	Exploratory Factor Analysis	119
3.11.3.	Confirmatory Factor Analysis	121

3.11.4. Exploratory Structural Equation Modelling	124
3.12. Latent Profile Analysis.....	125
3.13. The Trustworthiness of Teacher Assessment Tools	127
3.14. Summary	132
CHAPTER 4 - METHODOLOGY	133
4.1. Introduction.....	133
4.2. Research Design	133
4.3. Research Context	136
4.4. Participants in the Main Study and Sample Size.....	137
4.5. The Development of the Teacher AfL Literacy Tool.....	139
4.5.1. Overview.....	139
4.5.2. Identification of Indicators.....	140
4.5.3. Establishment of Standards	141
4.5.4. Expert Validation	143
4.6. Pilot Testing	144
4.7. Factor Analysis.....	147
4.7.1. Overview.....	147
4.7.2. Exploratory Factor Analysis.....	150
4.7.3. Confirmatory Factor Analysis	153
4.7.4. Higher Order Factor Analysis	154
4.7.5. Exploratory Structural Equation Modelling	154
4.8. Latent Profile Analysis	154
4.9. Item Characteristics of the Teacher AfL Literacy Assessment Tool	155
4.10. Validation of Results	156
4.11. Summary	157
CHAPTER 5 - RESULTS OF THE PRELIMINARY STEPS AND THE PILOT STUDY ...	158
5.1. Introduction.....	158
5.2. Identification of Indicators and Performance Criteria	158
5.3. Expert Validation	159
5.4. Pilot Testing	163
5.4.1. Overview	163
5.4.2. Reliability Index	163
5.4.3. Item Fit.....	164
5.4.4. Category Disordering.....	166
5.4.5. Item Discrimination.....	167
5.4.6. Difficulty Index	168

5.5. Insights for the Main Study.....	170
5.6. Summary.....	171
CHAPTER 6 - RESULTS OF THE MAIN STUDY	172
6.1. Introduction.....	172
6.2. Population Sample.....	172
6.3. Results of Exploratory Factor Analysis	174
6.3.1. Overview	174
6.3.2. Model Test Statistics	174
6.3.3. Convergent and Discriminant Validities	183
6.4. Results of Confirmatory Analysis of the 6-Factor Model.....	185
6.4.1. Overview	185
6.4.2. Model Test Statistics and Fit Indexes of the Proposed Model.....	185
6.4.3. Convergent and Discriminant Validities of the Proposed Model	192
6.5. Second-order Factor Analysis.....	194
6.5.1. Overview.....	194
6.5.2. Model Fit Statistics and Fit Indexes	194
6.5.3. Convergent and Discriminant Validity.....	196
6.6. Exploratory Structural Equation Modelling	201
6.6.1. Overview	201
6.6.2. Model Fit Statistics and Fit Indexes.....	201
6.6.3. Convergent and Divergent validity	204
6.7. Comparison of CFA, Second-Order CFA and ESEM Results.....	207
6.8. The Final Teacher AfL Literacy Tool.....	210
6.9. Latent Profile Analysis.....	223
6.9.1. Overview.....	223
6.9.2. Fit Indices of Latent Profiles	223
6.9.3. Composition of Each Profile	226
6.9.4. Labels of Latent Profiles	226
6.9.5. Mean Scores Differences between Teacher Classes.....	228
6.10. Item Characteristics of the Teacher AfL Literacy Tool	232
6.10.1. Overview	232
6.10.2. The Reliability Index.....	232
6.10.3. Item Fit.....	233
6.10.4. Item Discrimination	239
6.10.5. Difficulty Index	241
6.10.6. Hierarchy of Teacher AfL Skills	245

6.11. Summary.....	249
CHAPTER 7 - CROSS-COUNTRY VALIDATION	250
7.1. Introduction.....	250
7.2. Population Sample.....	250
7.3. Teachers' Self-Assessment	251
7.4. Factor Analyses.....	251
7.4.1. Overview	251
7.4.2. CFA Model	251
7.4.3. Second-Order Factor.....	257
7.4.4. Exploratory Structural Equation Modelling	264
7.5. Latent Profile Analysis	269
7.6. Summary.....	271
CHAPTER 8 - DISCUSSION.....	272
8.1. Introduction	272
8.2. The Teacher AfL Literacy Tool	272
8.2.1. Overview.....	272
8.2.2. A New Way of Conceptualising Teacher AfL Literacy	273
8.2.3. Expanding the Existing Theoretical Models of Teacher Assessment Literacy ...	278
8.2.4. The Teacher AfL Literacy Tool as the Only Tool Aligned with AfL Principles.	284
8.2.5. The General Factor of Teacher AfL Literacy	288
8.3. The Criteria in Teacher AfL Literacy Tool and Classroom Practices.....	291
8.3.1. Overview.....	291
8.3.2. The Criteria and Standards as a Way to Operationalise AfL Principles	291
8.3.3. The Potential Use of the Criteria and Standards of the Teacher AfL Literacy Tool in Teacher Standards	295
8.4. The Measurement Properties of the Teacher AfL Literacy Tool	301
8.4.1. Overview.....	301
8.4.2. Internal and External Consistency of the Tool	302
8.4.3. The Teacher AfL Skills as a Continuum of Practice	303
8.4.4. Consistency of the Factor-Level and Measurement Properties of the Tool between Similar Contexts	304
8.5. Grouping of Teachers based on AfL Profiles	306
8.5.1. Overview.....	306
8.5.2. An Emerging Typology of Teacher AfL Literacy Development.....	306
8.5.3. The Interdependency of Teacher AfL Skills	308
8.5.4. Characteristics of High Performing Teachers	309

8.5.5. Teachers' Actual Level of A/fL Literacy.....	310
8.6. The Trustworthiness of the Tool	311
8.6.1. Overview	311
8.6.2. Reliability and Validity: Practical Perspectives.....	312
8.6.3. The Nature of Teachers' A/fL Literacy Assessment Tool.....	313
8.6.4. The Components of Teacher A/fL Literacy Assessment.....	317
8.6.5. The Social Consequences of Using the Teacher A/fL Literacy Tool	325
8.6. Summary	325
CHAPTER 9 - SUMMARY AND CONCLUSION	326
9.1. Introduction.....	326
9.2. Summary of Findings.....	326
9.3. Limitations of the Research	330
9.4. Implications of the Research.....	332
9.4.1. Implications for Theory.....	332
9.4.2. Implications for Practice	334
9.5. Areas for Future Investigation.....	336
9.6. Summary	338
REFERENCES.....	339
APPENDIXES.....	389
Appendix A: Support Letter from the Ministry of Education of Brunei.....	389
Appendix B: Ethics Approval from HREA Panel B of UNSW Australia.....	390
Appendix C. Participant Information Sheet	391
Appendix D: Certification from Optimise International (B) Sdn Bhd (OIB)	392
Appendix E: Workshop Schedule	393
Appendix F: Syntax for Exploratory Factor Analysis	395
Appendix G: Syntax for Confirmatory Factor Analysis	396
Appendix H: Syntax for Second-order Factor Analysis.....	397
Appendix I: Syntax for Exploratory Structural Equation Modelling	398
Appendix J: Syntax for Latent Profile Analysis	399

LIST OF TABLES

Table	Title	Page
2.1	Comparison of the Three Macro Processes in Feedback, FA and AfL	37
2.2	Comparison of Dimensions from Existing Teacher Assessment Literacy	55
3.1	Teacher Assessment Literacy used in some Teacher Professional Standards	94
3.2	Teacher Career Stages in the Existing Literature	96
4.1	Sample Size used in Various Stages of the Research	140
4.2	Recommended Sample Size for Factor Analysis	141
4.3	The Generic Descriptions of Literacy Levels Used in Establishing	145
5.1	The Item-Content Validity Indexes of Teacher AfL Skills in the Tool	164
5.2	Item Parameter Estimates of the Rasch Model for 72 Items of Teacher AfL Literacy (N=197)	168
6.1	Demographic Profile of Participants in the Main Study	176
6.2	Goodness of Fit Statistics of Models for Teacher AfL Literacy Derived from Exploratory Factor Analysis with Weighted Least Squares Means and Variance Adjusted Estimation for 72 Items (N=325)	179
6.3	Eigenvalues of the 10 Factors	180
6.4	Summary of Exploratory Factor Analysis Results for Teacher AfL Literacy Measure Using Weighted Least Squares Means and Variance Adjusted Estimation for 72 Items (N=325)	181
6.5	Summary of Final EFA Results for Teacher AfL Literacy Measure Using WLSMV Estimation for 52 Items (N=325)	185
6.6	Correlations amongst the Six Factors of Teacher AfL literacy Derived from Exploratory Factor Analysis using WLSMV Estimation for 52 Items (N=325)	187
6.7	Evaluation of Fit Indexes of the 6-Factor Confirmatory Model (N=345)	190

6.8	Unstandardised and Standardised Loadings for the 6-Factor Confirmatory Model of Teacher AfL Literacy (N=354)	193
6.9	Correlations amongst the Six Factors of the Confirmatory Model of Teacher AfL Literacy (N=354)	196
6.10	Evaluation of Fit Indexes of the Second-Order Confirmatory Factor Model (N=345)	198
6.11	Unstandardised and Standardised Loadings for the 6-Factor Second-order Confirmatory Model of Teacher AfL Literacy (N=354)	200
6.12	Correlations amongst the Six Factors of the Second-Order Confirmatory Model of Teacher AfL Literacy (N=354)	202
6.13	Evaluation of Fit Indexes of the 6-Factor Exploratory Structural Equation Model (N=345)	206
6.14	Unstandardised and Standardised Loadings for the 6-Factor Exploratory Structural Equation Model of Teacher AfL Literacy (N=354)	208
6.15	Correlations amongst the Six Factors of the Exploratory Structural Equation Model of Teacher AfL Literacy (N=354)	209
6.16	Comparison of CFA, Hierarchical and ESEM Results of Teacher AfL Literacy Model	211
6.17	Model Fit Indexes of Tested Latent Profile Models of Teacher AfL Literacy (N=354)	227
6.18	One-Way Analysis of Variance of Mean Scores of Five Classes of Teachers in Six Factors (N=354)	233
6.19	Post-hoc Comparison Test of the Means of Five Classes of Teachers for Six Factors using Tukey HSD (N=354)	235
6.20	Item Parameter Estimates and Item Fit of the 42 Teacher AfL Skills	238
6.21	Item Parameter Estimates and Fit of the Standards	239
6.22	Discrimination Indexes of the Final Items in Tool	243
7.1	Evaluation of Fit Indexes of the 6-Factor Confirmatory Model (N=383)	256
7.2	Unstandardised and Standardised Loadings for 6-Factor Confirmatory Model of Teacher AfL Literacy (N=383)	258

7.3	Correlations amongst the Six Factors of the Confirmatory Model of Teacher AfL Literacy (N=383 Philippine Teachers)	260
7.4	Evaluation of Fit Indexes of the Second-order Confirmatory Factor Model (N=383)	262
7.5	Unstandardised and Standardised Loadings for 6-Factor Second-order Confirmatory Model of Teacher AfL Literacy (N=383 Philippine Teachers)	264
7.6	Correlations amongst the Six Factors of the Second-order Confirmatory Model of Teacher AfL Literacy (N=383 Philippine Teachers)	267
7.7	Evaluation of Fit Indexes of the 6-Factor Exploratory Structural Equation Model (N=345 Philippine Teachers)	269
7.8	Unstandardised and Standardised Loadings for 6-Factor Exploratory Structural Equation Model of Teacher AfL Literacy (N=383)	271
7.9	Correlations amongst the Six Factors of the Exploratory Structural Equation Model of Teacher AfL Literacy (N=383 Philippine Teachers)	272
7.10	Model Fit Indexes of Tested Latent Profile Models of Teacher AfL Literacy (N=383 Philippine Teachers)	273
8.1	A Summary of Existing Models of Teacher Assessment Practices	284
8.2	Comparison of the Dimensions of Teacher Assessment Literacy	290
8.3	The Alignment of Criteria of Teacher AfL Literacy Tool to AfL Principles	297
8.4	Comparison between the Dimensions of the Teacher AfL Literacy Tool and the Selected Professional Teacher Standards	301

LIST OF FIGURES

Figure	Title	Page
2.1	The theoretical framework for the study showing the six theoretical dimensions of teacher AfL literacy	69
4.1	Summary of the methods of the study	138
5.1	Example of an item that shows category disordering	170
5.2	Map of latent distributions of teacher AfL practices based on difficulty level, X =1.3 sample	172
6.1	The 6-factor model of teacher AfL literacy	194
6.2	The 6-factor model with second-order general factor of teacher AfL literacy with the standardised factor loadings of variables to their corresponding factor	203
6.3	The latent profiles of teachers based on their AfL literacy	229
6.4	Map of latent distributions of teacher AfL practices based on difficulty level, where X represents 4.8 numbers of teachers	246
6.5	Map of latent distributions of teacher AfL standards based on difficulty level, where X represents 4.8 numbers of teachers	248
6.6	Hierarchical order of items based on difficulty level	250
6.7	The distribution of standards showing the arbitrary cut-off logit scores	252
7.1	The 6-factor model of teacher AfL literacy using data from Philippine Teachers	259
7.2	The 6-factor model with second-order general factor of teacher AfL literacy using data from Philippine teachers	266
7.3	The latent profiles of teachers based on their AfL literacy	275
8.1	Revised theoretical framework of teacher AfL literacy	281
8.2	The components of teacher's AfL literacy assessment	322

LIST OF ACRONYMS

AERA	<i>American Educational Research Association</i>
AIC	<i>Akaike Information Criteria</i>
AITSL	<i>Australian Institute for Teaching and School Leadership</i>
AfL	<i>Assessment for Learning</i>
ALI	<i>Assessment Literacy Inventory</i>
ANOVA	<i>Analysis of Variance</i>
ARC	<i>Assessment Resource Centre</i>
ARG	<i>Assessment Reform Group</i>
BCATs	<i>Brunei Common Assessment Tasks</i>
BIC	<i>Bayesian Information Criteria</i>
BOSTES	<i>Board of Studies Teaching and Educational Standards</i>
BLRT	<i>Bootstrap Likelihood Ratio Test</i>
CFA	<i>Confirmatory Factor Analysis</i>
CFI	<i>Comparative Fit Index</i>
CI	<i>Confidence Interval</i>
CoP	<i>Community of Practice</i>
CTT	<i>Classical Test Theory</i>
CVI	<i>Content Validity Index</i>
DepEd	<i>Department of Education</i>
DWLS	<i>Diagonally Weighted Least Squares</i>
EAL/D	<i>English as an Additional Language or Dialect</i>
EFA	<i>Exploratory Factor Analysis</i>
ESEM	<i>Exploratory Structural Equation Modelling</i>
FA	<i>Formative Assessment</i>
HSD	<i>Honest Significant Test</i>
ICC	<i>Item Characteristics Curve</i>
IRT	<i>Item Response Theory</i>
LCA	<i>Latent Class Analysis</i>
LPA	<i>Latent Profile Analysis</i>
LMR	<i>Lo-Mendell-Rubin</i>
ML	<i>Maximum Likelihood</i>
MoE	<i>Ministry of Education</i>
NAPLAN	<i>National Assessment Program for Literacy and Numeracy</i>
NFI	<i>Normative Fit Index</i>
NNFI	<i>Non-Normative Fit Index</i>
OIB	<i>Optimise International Brunei</i>
PCA	<i>Principal Component Analysis</i>
PD	<i>Professional Development</i>
RMSEA	<i>Root Mean Square Error Approximation</i>
SA	<i>Summative Assessment</i>
SBA	<i>School-Based Assessment</i>

SBAfL	<i>School-Based Assessment for Learning</i>
SPA	<i>Student Progress Assessment</i>
SPE	<i>Student Progress Examination</i>
SRMR	<i>Standardised Root Mean Square Residual</i>
TLI	<i>Tucker-Lewis Index</i>
UK	<i>United Kingdom</i>
ULS	<i>Unweighted Least Squares</i>
UNSW	<i>University of New South Wales</i>
USA	<i>United States of America</i>
VLMR	<i>Young-Lo-Mendell-Rubin</i>
WLS	<i>Weighted Least Squares</i>
WLSMV	<i>Weighted Least Squares Means and Variance</i>
WRMR	<i>Weighted Root Mean Square Residual</i>
ZPD	<i>Zone of Proximal Development</i>

CHAPTER 1 – INTRODUCTION

1.1. Introduction

This thesis presents a theoretically and empirically driven-tool for teacher assessment for learning (A/fL) literacy in response to the weaknesses of existing tests, rating scales and other forms of assessment used to evaluate teacher assessment literacy. This tool is drawn from the philosophical context of A/fL, assessment innovation and change, latent growth and probabilistic model of competence. I combined both theoretical and empirical approaches to develop the teacher A/fL literacy tool with detailed criteria and five levels of standards. A series of factor analyses, with a subsequent application of generalised partial credit analysis of the Rasch model, was used to establish the dimensions and the item level characteristics of teacher A/fL literacy.

In this chapter, I describe the background, rationale, context, and purpose of the study, as well as introduce the research questions and scope of the research.

1.2. Background of the Study

Around the world, educational systems are trying to identify and describe the teaching competencies that ensure high quality learning and teaching. This is particularly evident with the movement to develop and apply teacher professional standards in most, if not all educational systems. Such professional standards are believed to enhance teaching performance by providing greater support and specificity for teacher roles and responsibilities.

Currently available teacher professional standards (e.g., the Australian Institute for Teaching and School Leadership (AITSL) Standards, the National Competency-Based Teacher Standards for Philippines, the United Kingdom Teacher Standards, and Teaching Standards in the Asia-Pacific Region) and teacher performance assessment

tools are generally broad, covering all learning and teaching functions. Although it is relatively easy to agree upon the importance of a teacher being competent in the classroom (Ferguson & Brown, 2000), the specific attributes of effective teaching are yet to be fully described. This is primarily because teacher competency is such a broad construct (Spanierman et al., 2010; Vogt & Rogalla, 2009), including a range of dimensions such as classroom management (Emmer & Stough, 2001), pedagogical and content knowledge (Loughran, Mulhall, & Berry, 2004), assessment competence (Hattie, 2003; Stiggins & Duke, 2008), student-teacher rapport (Goddard, Hoy, & Hoy, 2000; Hoy & Miskel, 2001; Tschannen-Moran & Hoy, 2001), and other important dimensions. Hence, some teacher professional standards are too generic and others are too specific to capture the particular knowledge and skills required to teach effectively (Samson & Collins, 2012).

The current emphasis on teacher roles in improving student learning and performance is rooted in the evidence that student achievement is significantly associated with the competence of teacher (Ferguson & Brown, 2000; Sanders & Rivers, 1996). Research shows that high performing teachers create a learning environment where students learn best. The contribution of teachers to student learning explains roughly 30% of the variance observed in student performance (Hattie, 2003). This means that teacher competence is the next most important determinant of student achievement after the students themselves, who account for 50% of the variance in their learning. However, not everything that teachers do in the classroom brings about the desired learning improvement. Hattie (2008) emphasises that about 95% of what teachers do in the classroom influences student achievement, but not all of these teacher activities positively impact student learning.

In order to optimise student learning and achievement, it is vital to focus efforts on improving the specific components of teaching competencies that positively impact learning. In addition, if schools and the whole system are committed to supporting the key teacher functions that help students to learn effectively, the focus of performance evaluation and professional development for continuous improvement should be on those teacher functions that are known to most positively impact the improvement of student learning and achievement. Amongst these various dimensions of teaching competency, assessment experts and researchers have long argued for the importance of teachers' ability to use assessment and student assessment information to improve student learning (e.g., Black & Wiliam, 1999; Hattie, 2008; Popham, 2009; Stiggins, 2005). This teacher assessment ability is broadly represented by a philosophical framework referred to as assessment for learning (AfL).

The AfL paradigm calls for a shift in teachers' assessment practices from recording and evaluating student achievement to helping and encouraging students' active engagement in learning through assessment (Black & Wiliam, 1999; Hattie, 2008; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Some of the core AfL strategies include detailed elaboration of learning outcomes, success criteria and performance standards (Gray & Tall, 1994; Nicol & Macfarlane-Dick, 2006); assessment tasks specifically designed for learning (Davison, 2007; Thompson & Wiliam, 2007); effective use of feedback (Hattie & Timperley, 2007; Sadler, 1989b); and the use of self and peer assessment in the classroom teaching (Klenowski, 1995; McDonald & Boud, 2003; Price, O'Donovan, & Rust, 2004; Taras, 2003). In a meta-analysis conducted by Black and Wiliam in 1998 using 250 studies, they found that students whose teachers used formative assessment strategies significantly improved their performance. A decade later, a larger scale meta-analysis was conducted by Hattie

(2008) exploring the effects of more than 100 educational interventions on learning. From this study, teacher assessment practices emerged as the interventions with the highest potential to enhance student learning.

The work of Black and William (1998) and Hattie (2008), along with other studies (Black, McCormick, James, & Pedder, 2006; Keppell & Carless, 2006; Marshall & Drummond, 2006; Munns & Woodward, 2006), provides strong empirical support for the claim that teacher assessment practices associated with AfL can be grouped together into one single construct. Combining the effect sizes of all such teacher assessment practices would make teacher AfL literacy the most influential single factor in improving student learning. This empirical support for the effectiveness of teacher AfL literacy to improve student learning is further discussed in Chapter 2.

1.3. Rationale for the Study

As indicated above, teacher AfL literacy is a critical component of teaching effectiveness. To ensure that teachers and students take advantage of the positive effects of assessment, Stiggins (1999a), Popham (2009), and Davison and Michell (2014) argue that AfL literacy must be a key component of teachers' professional knowledge, skills, and professional development. There is a need to ensure that teachers develop a high level of expertise in the use of AfL because the levels of their AfL literacy affect their confidence in using a range of assessment strategies, both teacher-developed and system-imposed. Similarly, teachers' levels of AfL literacy enable them to effectively use assessment information to make critical decisions related to learning and teaching to further support student learning.

Researchers and educators argue that to support teachers to develop a high level of AfL literacy, an assessment literacy program needs to be congruent with the

philosophy of AfL itself (Davison, 2013). It must begin with measuring individual teachers' current levels of assessment literacy, followed by identifying their specific training and support needs. To perform these processes, a teacher AfL literacy tool needs to be developed that can be used both as a performance evaluation tool and a professional development tool.

There have been numerous attempts to develop and use tests, rating scales and other forms of assessments to evaluate teacher assessment literacy, but none of these have used the principles of AfL in their development and application. To adhere to the principles of AfL, just as in classroom assessment practices, the assessment of teacher AfL literacy should be based on established criteria and standards of practice, which should be clearly communicated to individual teachers. The need for such criteria and standards was first expressed by Bailey and Brown in 1995, and later on emphasised by Inbar-Lourie in 2008. The latter argues that “such initiative will greatly facilitate the meaningful construction of ... assessment courses and make a major contribution to the field at large” (p. 398).

The problem with existing teacher assessment tools is that they are not aligned to the principles of AfL, to be discussed more fully in Chapter 3. Furthermore, the construct of teacher AfL literacy has not yet been clearly defined and supported by empirical evidence. Although teacher assessment literacy has been extensively studied at the item level, the evidence to support the dimensionality of teacher AfL literacy is yet to be substantiated. The ‘fuzzy’ dimensions of teacher AfL literacy are illustrated in the original four areas of teacher assessment practices advocated by Black and Wiliam in their seminal paper in 1998, that is, questioning, feedback, sharing criteria, and self-assessment. These areas were revised in 2003 by Black, Harrison, Lee, Marshall and Wiliam, retaining questioning and feedback, and adding peer assessment along with

self-assessment and the formative use of summative tests. This revision was informed by the results of their project in transforming teacher formative practices. However, many researchers and teachers are not convinced that these four revised areas of teacher assessment practices capture the entirety of the AfL construct. In an attempt to establish the dimensions of this construct, some authors used theoretical support (e.g. Stiggins, 2009), others adopted conceptual frameworks from existing teacher standards (e.g. Mertler & Campbell, 2005), and a few used factor analysis (e.g. Newfields, 2006; Fulcher, 2012). Although the latter authors used an empirical approach, their analysis was limited only to exploratory factor analysis, which is just a starting analysis for a more sophisticated empirical approach. Hence, the results of teacher assessment using these existing assessment tools are not as trustworthy as they need to be from the psychometric point of view.

Establishing the dimensions of teacher assessment literacy is critical from both a psychometric perspective and for the practical applications of any assessment tool. In principle, if the assessment tool does not meet necessary psychometric standards of validity, reliability, item fit, discrimination index, and difficulty index, the data gathered would be problematic and the subsequent interpretations of the assessment results would be compromised. Hailaya, Alagumalai, and Ben (2014) illustrate the need for a review and further testing of existing teacher assessment literacy tools using more advanced psychometric methodologies and a thorough factor analytic approach to establish the trustworthiness of these tools.

Due to misalignment with AfL principles and a lack of trustworthiness of existing assessment literacy tools (to be elaborated in Chapter 3) there is a need to further explore the construct of teacher AfL literacy both at the item and dimension levels. This will enable the development of a teacher AfL literacy tool that is

underpinned by a strong theoretical rationale and robust empirical evidence. The use of rigid methodologies, including the application of the Rasch analysis to analyse the tool at the item level and the use of exploratory and confirmatory factor analyses to establish the dimensions of the construct, will ensure that the tool developed meets the psychometric and factor analytic requirements. Hence, the overall approach adopted in this study needs to give confidence to the users of the teacher AfL literacy tool, particularly teachers in their self-assessment and school leaders in their periodic monitoring of teacher performance, that meaningful information about an individual teachers' level of AfL literacy can be provided across the stages of their professional development. Furthermore, the same should be able to be used to identify individual teachers' AfL literacy needs in order to develop a needs-based professional development program to further support and advance their assessment literacy.

1.4. Context of the Study

The development of a teacher AfL literacy tool originated as a part of a larger project for the Brunei Ministry of Education. In 2010, the Brunei government initiated a nationwide education reform to change the assessment culture from testing student knowledge through rote learning to a more learner-centred assessment approach. The concept AfL was actively incorporated into Brunei's existing School-Based Assessment (SBA), renamed School-Based Assessment *for* Learning (SBAfL). Under the old SBA program, teachers had already been given limited responsibility to develop assessment tasks and assess student learning to complement national tests developed and administered by the national examination board. However, during the early implementation of the program from 2008 to 2010, there was no fundamental change in the assessment practices of teachers. In fact, the SBA continued to rely on testing with teachers administering more

tests than before (SBAfL, 2011). As most of the assessments were written, teachers devoted most of their time preparing students to take the test. Consequently, rote learning was seen as the fundamental skills acquired by students. Also, the over emphasis on high-stakes summative assessment led to the over-assessment of students where all homework and all class work was marked for summative purposes. Teachers spent most of their time assessing and marking student performance without using the assessment information to assist students in their learning.

This practice was a result of teachers' misunderstanding of SBA, believing that more frequent testing generated higher learning outcomes. To address this issue, SBAfL was introduced by the lead consultant, Prof. Chris Davison (Davison, 2013), to change the philosophical and cultural beliefs underpinning the assessment culture of Brunei. Under this reform, the concept of SBA was retained but integrated with the philosophies and principles of AfL. SBAfL advocates key principles of AfL focused on empowering students to take active roles in their learning. Similarly, teachers are trained to ensure that assessment is fundamentally integrated into learning and teaching activities.

Under the SBAfL reform, teachers were provided with support to design and use a variety of assessment tasks, accommodating various student ability levels and the diverse needs of students. Teachers were trained in how to use assessment data to inform their key decisions related to learning and teaching with the aim of enhancing student learning. The introduction of SBAfL aimed to benefit students, teachers and the nation in general.

The benefits claimed for students were drawn directly from the existing literature about the principles and practices of assessment for learning (SBAfL Guidebook, 2011, p. 14):

1. Criteria are shared with students, so they know what they need to do to succeed in each assessment task.
2. SBAfL assumes everyone can be successful and students know that their teachers are helping everyone in the class to learn and achieve as well as they can.
3. The nature and timing of the task is part of teaching and learning so students are able to perform to their potential.
4. The teachers and peers give high quality feedback all the time, so students know where they are, where their weaknesses are, and what they need to do to achieve success.
5. The teacher coaches the students to self-assess so that they can be drivers of their own learning.
6. There is a variety of assessment tasks, so students develop various skills which cannot be readily assessed in a one-off external written exam.

These benefits for students were meant to be facilitated by teachers' use of the principles of SBAfL, drawing heavily from the principles and practices of AfL. It was claimed that by adhering to SBAfL policy, teachers:

1. can ensure quality and effective teaching and learning.
2. can provide achievable set targets for their students.
3. are able to obtain a clear picture of students' progress and thus administer appropriate remedial and intervention programs.
4. gain useful data for them to reflect on their own teaching methodology.
5. can provide engaging activities and varied tasks that are built into the process of teaching and learning.

(SBAfL Guidebook, 2011, pp. 14-15).

From the time the reform was introduced into the school system, there was a great deal of interest from both the government and local schools to learn more about AfL principles and practices. To fully guide teachers in their implementation of SBAfL, large scale professional development (PD) programs were implemented. The PD was designed to develop the skills of teachers in sharing and developing learning outcomes and success criteria, designing assessment tasks, eliciting and giving feedback, and encouraging student self and peer assessment. However, the implementation was challenged by the lack of shared understanding of AfL amongst teachers and across the various management levels of the education bureaucracy. Another challenge encountered was the absence of a tool that could evaluate the level of assessment literacy of teachers and could identify their professional development needs.

The stimulus to develop a teacher AfL literacy tool came from the desire to describe more clearly what characterised effective assessment, hence establishing a common understanding amongst teachers, teacher educators, school leaders, researchers, professional organisations, and Ministry of Education officials. Specifically, the teacher AfL literacy tool needed to be used by teachers for self-reflection to identify their current and developing levels of abilities, and to support them for their on-going professional development. Also, the tool needed to be used by teachers and school leaders to engage in a peer-review of teaching performance to identify the professional support needed by teachers. It is in this context that the Ministry of Education of Brunei, through the consulting team led by University of New South Wales Australia researchers and Optimise International (B) Sdn Bhd, funded my access to the research site and supported the participation of teachers, school leaders, and other key stakeholders for this study.

1.5. Aims of the Study

As argued earlier, whilst several large studies have documented effective practices in A/fL (e.g., James, & Pedder, 2006; Black & Wiliam, 1999; Hattie, 2008; Keppell & Carless, 2006; Klenowski, 1995; Marshall & Drummond, 2006; McDonald & Boud, 2003; Munns & Woodward, 2006), and there have been numerous attempts to develop and use tests, rating scales and other forms of assessments to describe various aspects of teacher's assessment literacy (e.g. Davison, 2004, 2007; Mertler & Campbell, 2005; Newfields, 2006; Thomson, 2012), there is no existing framework derived from an empirical approach. Also, there is no assessment tool that takes into account both A/fL and measurement principles in its development, and that has involved teachers and key education stakeholders in the process.

To bridge the gap, I worked with teachers, assessment practitioners, school leaders, and key education management people to develop a scale, framed as rubrics with five levels of increasing difficulty of assessment performance per assessment skill. In doing so, I drew on the work of Griffin (2007), who made a strong connection between Glaser's (1963,1981) and Rasch's (1960, 1980) latent trait theory and Vygotsky's (1978, 1986) zone of proximal development, and incorporated the works of Andrich (1978) and Masters (1982) in extending the probabilistic model of competence to polytomously scored items. Specifically, my research study aimed to:

1. develop a theoretically and empirically-driven tool for describing and understanding teacher A/fL literacy, to address the psychometric and measurement properties of the scale and establish the dimensionality of teacher A/fL literacy profile;
2. explore the latent profiles of teachers based on their profiles of A/fL literacy, to establish empirical evidence about the groupings of teachers

with similar assessment profiles. The results can be used to establish a typology of teacher AfL literacy; and

3. determine the extent of the level of generalisability of the tool to other contexts, to provide evidence that the dimensions and the levels of performance of teacher AfL literacy are consistent across different contexts.

However, it must be emphasised that even if the measurement properties of the teacher AfL literacy tool prove to be invariant across different contexts, the assessment process and teacher AfL literacy remain largely context-dependent. There may be different teacher AfL criteria and standards (e.g. interpretation and use of large scale assessment data, measurement skills, use of technology in assessment) that need to be added to the tool to account for the system-imposed assessment practices of diverse teacher cohorts from different educational systems.

1.6. Research Questions

To achieve the aims of this study, I drew upon the principles of AfL and used a robust methodology to develop the teacher AfL literacy tool. This research sought answers to the following questions:

1. What are the emerging AfL practices of teachers? Specifically,
 - a. What are the indicators of teachers' effective AfL practices?
 - b. What are the standards of performance for each stage of professional growth?
2. Are there distinct factors/subdomains of teacher AfL literacy?
3. What are the item characteristics of the final teachers AfL literacy tool?
4. What are the latent classes of teachers based on their AfL literacy?

5. To what extent can the tool be applied and used in another context?

Question 1 identifies the elements or indicators of teacher AfL literacy, which form the criteria in the tool. As the teacher AfL literacy tool takes the form of rubrics to ensure its alignment to the principles of AfL, each level comprises a description of the performance that characterises the teacher AfL tasks.

Question 2 uses a factor analytic approach (described in Chapter 3) to establish the dimensionality of teacher AfL literacy, whilst Question 3 uses the generalised partial credit model of the Rasch model. The answers to these questions provide empirical evidence at the factor level and item level respectively. The findings are critical to support the measurement requirements in terms of reliability and validity of the tool, which would give confidence to its users.

Question 4 uses the results of latent profile analysis to discover if there are distinct classifications of teachers based on their AfL latent profile. The answer to this question will inform the design and delivery of the professional development program, in that if teachers can be grouped based on common abilities in their assessment practices, specific professional development programs can be designed to address their particular assessment literacy needs.

Question 5 uses another set of data from teachers in another country to see the extent to which the teacher AfL literacy tool could be applied to another related context. This was assessed with new data, using the same analyses in Questions 2 and 4, although no comparisons could be made between the teachers from the two countries. This was due to the limited access to the data source owned by Brunei government, and to protect the privacy and integrity of teachers.

1.7. Summary and Structure of the Thesis

In this chapter, I have provided the background, rationale and context of the study. I have also explained the reasons for developing the teacher AfL literacy tool and its potential application to support teachers both in their classroom practices and in their professional development. Furthermore, I have defined the focus and scope of the study, including the research questions being addressed.

From this chapter, it can be seen that there is a need to develop a teacher AfL literacy tool to address the weaknesses of existing teacher assessment literacy tools, and to further explore the dimensionality of the construct of teacher AfL literacy. The results of the study, particularly the teacher AfL literacy tool, is deemed to further support the implementation of the SBAfL policy by providing a tool that can be used both as a measurement instrument and a professional development tool.

There are eight remaining chapters in this paper. Two chapters are devoted for the literature review: Chapter 2 argues for the pivotal role of teacher assessment practices in improving student learning and presents the theoretical framework for the study, and Chapter 3 presents a literature review of existing teacher assessment literacy framework and tools and the best practices in scale development, which is needed to inform the methodology of my study. In Chapter 4, the methodology for the study is presented including the research design and the analyses used. Chapters 5, 6 and 7 present the results of the pilot study, main study, and cross country validation respectively. The findings of the study are discussed in Chapter 8, with an emphasis on the contributions of the study to the literature. The concluding chapter (Chapter 9) summarises the major findings of the study and its conclusions, and presents the implications for theory, research, policy and practice.

CHAPTER 2 - TEACHER ASSESSMENT PRACTICES AND DEVELOPMENT OF AfL LITERACY

2.1. Introduction

In this chapter I present a review of the literature to support the argument that the underlying construct, teacher AfL literacy, has a pivotal role in improving learning and teaching. This chapter begins with a review of the empirical evidence for the influence of teacher assessment practices on student learning, followed by a discussion of the theoretical support for teacher AfL practices to ensure that this construct is aligned with widely-accepted theories of learning. Next, I trace the development of the AfL concept from its emergence up to the present to clarify its conceptualisation. Then I define teacher AfL literacy construct based on prior definitions, conceptualisations and theorisation, incorporating the elements of best practices in assessment. Finally, I critique existing dimensions of teacher assessment literacy, on which the theoretical framework of my study is built.

2.2. Empirical Evidence and Theoretical Support for Teacher Assessment Practices

2.2.1. Overview

The role of teacher assessment practices in learning and teaching needs to be supported both empirically and theoretically to provide strong evidence that this component of teacher competency defines teaching effectiveness. The strength of the empirical evidence and the alignment to theories of learning for this construct is discussed below.

2.2.2. Empirical Evidence for Teacher Assessment Practices

As indicated in Chapter 1, the significant effects of teacher assessment practices on

various aspects of learning and teaching are strongly supported by research evidence. A range of research studies highlights the significant contributions of teacher assessment practices to the improvement in learning and teaching. Amongst these studies, the work of Black & Wiliam (1999) and Hattie (2008) provides strong empirical evidence that teacher assessment practices are the most influential interventions in improving student learning. Both studies have achieved an unmatched level of status and credibility, with more recent studies (Beaumont, O'Doherty & Shannon, 2011; Birenbaum, Kimron, & Shilton, 2011; Clements et al., 2011; Hendry, Bromberger, & Armstrong, (2011); Love, 2013; Ruiz-Primo, 2011) validating the results of the earlier studies and supporting the role of teachers' assessment practices in improving student learning. In both studies, the authors compared a large number of educational interventions and used effect size measures to establish and interpret the impact of these interventions on student achievement, with Black and Wiliam (1999) finding a range of effect sizes from $d=0.4$ to $d=0.7$ whilst Hattie (2008) reported a range between $d=0.41$ to $d=1.44$.

Mathematically, the effect size of an intervention is obtained by dividing the difference between the mean scores of students who have received the intervention and the mean scores of those who have not received the intervention by the square root of the average of the squared standard deviations (Cohen, 1988). To interpret the magnitude of effect size, Hattie (2008) developed a scale where the effects of interventions on student achievement are seen as a continuum ranging from reverse effects, developmental effects, teacher effects, to a zone of desired effects. He called this continuum the "barometers of influence", with an effect size of $d=0.4$ upwards as the zone of desired effects where interventions at or above this cut-off point are believed to advance the learning of students far more than could normally occur in a year of teaching. An intervention with an effect size of $d=1.0$ can advance the learning of

students by 2 to 3 years or improve learning by 50%, or in the case of experiments, those students who received the intervention could exceed the performance of those who did not receive the treatment by 84% (Hattie, 2008). Hattie interprets effect size through the use of average percentile gain. Since an effect size is equivalent to a point in normal distribution, it can be equated on the same scale with the percentile gain. To highlight the desired effect of an intervention, $d = 0.40$ would mean that students who received this treatment would put them an additional 16 percentile points from their original standing which means that they can perform better (by 66%) than those students who did not received the intervention. In general, for every increase in effect size, there is a corresponding increase in the percentile gain. Hence, the higher the effect size observed in an intervention, the higher the gain in student achievement.

Using the concept of effect size, the results of Hattie's (2008) work show that teacher assessment practices are amongst the top interventions: self-report grades being the highest with an effect size of $d = 1.44$, formative evaluation being the 3rd ranked ($d=0.90$), teacher clarity (e.g. explicitly stating and communicating learning outcomes, using criteria and standards, and informing students of the assessment approach) ranked 8th ($d=0.75$), reciprocal teaching (actively involving students in learning and teaching activities by letting students to take the role of a teacher and lead their classmates in summarizing, questioning, clarifying and predicting) being the 9th ranked ($d=0.74$), and eliciting and giving feedback being 10th ($d=0.73$). Other teacher assessment skills with effect sizes greater than the threshold value of 0.4 include using prior achievement in teaching ($d=0.67$), effective teacher-student relationships ($d=0.72$), and engaging students in self-questioning ($d=0.64$). Based on these results, key assessment strategies associated with teacher AfL literacy are arguably the most important components of effective learning and teaching. Also, highlighted in this study is the benefit of teacher

engagement in professional development, which is amongst those interventions with a higher effect size ($d=0.62$). Thus, it can be argued that teacher use of assessment in the classroom can be fully maximized if they continuously engage in professional development related to assessment. This argument supports the need for my study as it provides a tool for teacher AfL literacy, as will be discussed in the latter section of this chapter.

However, despite the prominence and status of the work of Black and Wiliam (1998) and Hattie (2008), both studies have not escaped criticism. For example, Black and Wiliam's study has been challenged for its claimed lack of theoretical pedagogic context (Perrenoud, 1998), concerns about the reliability of some included studies and discrepancies in results (Smith & Gorard, 2005), methodological limitations (Dunn & Mulvenon, 2009), and inconsistencies in the use of terminologies (Taras, 2007). Similarly, Hattie's study has been criticised for its exclusion of some factors that might have changed the results of his meta-analysis, his supposedly unrealistic conceptions of teaching and effective teachers (Terhart, 2011), problems with the dependency of effect size to sample size, social, background and context effects not being considered when discussing the results, concerns about the inclusion of some low quality studies and some not being assessed for validity, achievement defined only by quantitative measures, inapplicability of some studies to the real classroom (Snook, O'Neill, O'Neill, & Openshaw, 2009), and inappropriateness of inferences drawn from the results to inform policy due to the weakness of methodology used when studies included in the meta-analysis have different social and cultural contexts (Evers & Mason, 2011). However, despite these issues, the work of both researchers has been used to justify the introduction of AfL as a key component of government assessment policy (e.g. Brunei Ministry of Education, 2008, 2011; Queensland Department of

Education and the Arts, 2005; Hong Kong Curriculum Development Council, 2001; Learning and Teaching Scotland, 2006; Singapore Ministry of Education, 2008).

As shown above, a range of evidence demonstrates the effectiveness of this construct of teacher assessment literacy in enhancing learning and increasing achievement. However, to further support the use of this construct in learning and teaching, I look at the alignment between this empirical support and the theoretical support for teacher AfL practices.

2.2.3. Theoretical Support for Teachers' AfL Practices

Although teachers' AfL practices can be viewed through a behaviourist lens [with a focus on outcomes (Nuthal, 1999)], through a humanist lens [with focuses on learning progress, prior learning (Friedman and MacDonald (2006), and the role of motivation (Martin & Dowson, 2009)], or through a constructivist lens [with a focus on learning tasks as assessment tasks (Keppel and Carless, 2006), learning autonomy and the active construction of knowledge (Marshall and Drummond, 2006)], the emphases of behaviourism and constructivism in the interactions of thinking processes and learning and teaching are confined to individual learners, ignoring the role of social and cultural processes in the learning process. Recent research in AfL suggests that AfL can only be fully understood and implemented by teachers when they recognise the social, political and economic contribution and impact of their assessment practices and how these factors shape their assessment practices (Gipps, 1999). Through this sociocultural lens, teachers should see assessment as a tool that can be used to create a learning community (Lave & Wenger, 1991) of students and teachers, which then shapes teacher classroom practices and student learning behaviour. Also, from this perspective, learning is defined as more than just individual cognitive processes and in addition, it takes into

account the social processes and cultural settings that influence learning. Research in AfL (Black, McCormick, James, & Pedder, 2006; Keppell & Carless, 2006; Marshall & Drummond, 2006; Munns & Woodward, 2006) highlights the importance of social interactions, cultural contexts, and the belief systems of both students and teachers and how these factors shape students' identity (Cowie, 2005), and the nature of power and control in the classroom (Black & Wiliam, 2006). Although one of the aims of using AfL in the classroom is to guide students to become self-regulated and independent learners, the process of self-regulation is critically dependent on interactions with their teachers and their peers to activate and support their learning.

The conceptualisation of teacher AfL literacy from a sociocultural perspective is seen in the roles and beliefs of teachers required to activate student learning. The seven key points explained below summarise the conceptualisation of teacher AfL literacy from a sociocultural perspective.

1. *Teachers' beliefs about learning in an AfL context.*

Effective teaching in AfL context starts with teachers being clear about learning outcomes (Hattie, 2008), able to determine the current knowledge and skills of students and help them achieve the desired learning outcomes. This is consistent with the belief in a sociocultural perspective that teachers need to view learning as a process of moving students from novice to expert status whilst supporting them to constantly negotiate their identity in the classroom (James, 2006). The belief of teachers that learning occurs in a social environment and is influenced by the cultural setting of the classroom and the school in general enables them to develop socially and culturally dependent learning and teaching activities that facilitate interaction. This belief of teachers

about learning and teaching in an AfL context is a critical factor in helping students to engage in their learning. As Willis (2009) argues:

In order for teachers to engage students in AfL practices and create a community of practice that enables all students to develop an identity as an autonomous learner, teachers need to work beyond the curriculum, pedagogy and assessment of a purely behaviourist paradigm (p. 4).

One important aspect of the role of teachers in AfL from a sociocultural perspective is to help students acquire autonomy in their learning. This cannot be achieved through cognitive processes alone, but rather must be facilitated by social processes in the classroom. Through students' interactions in the classroom, they develop and shape a community of practice, which eventually shapes their identity as autonomous learners (James, 2006). Willis (2009), drawing from the definition of Lave and Wenger (1991) defines learner autonomy in the context of sociocultural perspective as the, "socially constructed identity of a self-monitoring student who participates in culturally accepted ways within a community of practice" (p. 3). The challenge for teachers is how to provide opportunities for student interactions in the class to support the negotiation and formation of student identity. To facilitate these interactions, teachers using AfL principles need to view students not as mere passive receivers of external knowledge but rather as socially active individuals who see knowledge as a dynamic output of interactions (Gipps, 1999).

2. *Teachers' beliefs about assessment in an AfL context.*

Consistent with a sociocultural perspective, in AfL assessment strategies, tasks and processes are seen as tools to help students identify both their strengths and weaknesses and the same assessment processes can be used to further

enhance their learning (Gipps, 1999). Teachers need to have a strong belief that assessment is only useful if it helps students to learn more effectively (Brookhart, 2003; Kane, 2006; Nichols, 2005; Sadler, 2012). To ensure that assessment is aligned with this kind of sociocultural perspective, teachers need to develop tasks that are authentic, match the level of student ability, and facilitate interactions amongst students and between teacher and students (James, 2006).

The work of Keppel and Carless (2006) demonstrate the use of assessment tasks to fulfil both learning and measurement functions. Using a learning-oriented assessment framework, students were highly involved in the assessment process through clear communication during the process of assessment, providing students with the opportunity for collaboration to complete the assessment tasks and requiring them to assess and provide feedback to their peers' work. It was found the learning experiences of students in completing assessment tasks, through active collaboration, interaction and negotiation of feedback, facilitated active construction of knowledge. In a more detailed study by Marshall and Drummond (2006), apart from explicitly using assessment tasks as learning tasks, they required teachers to consider the scope of the tasks and to ensure that learning processes undertaken by students promoted learning autonomy. They showed that explicitly sharing learning outcomes and criteria is not enough to develop student learning autonomy but rather it requires that students are highly involved in the development of criteria. Also, teachers' use of model or exemplar materials to demonstrate what is meant by quality performance could challenge students to produce the same high level performance, but in a variety of ways. The use of exemplars gives students the

idea that quality is not a fixed and finite concept. Students then use these criteria to actively engage in self-assessment, allowing them to develop the capacity to identify any misconceptions, capitalise on their strengths, and monitor their own progress. These processes support their development of learning autonomy. Also, teacher-student interactions, whilst students are progressing in their work, particularly the use of questioning, direct students to engage in deep thinking. Furthermore, the feedback provided by teachers has a significant role in the construction of knowledge and teachers need to guide students to use feedback to establish personal learning goals. The feedback received by students both promotes student motivation (Dweck, 2007) and provides scaffolding for students as they progress in their learning. The timely feedback received by students enables them to affirm their perceptions of their learning experiences and identify the next steps to undertake to move towards higher level of learning.

Another important assessment approach in AfL that embodies the principles of a socio-cultural theory of learning is the use of differentiated assessment to give opportunities for students to best demonstrate their learning. The teachers' effort in broadening their assessment approach helps students to offer alternative evidence of their learning (Linn, 1992), which avoids the conception of assessment as a 'one size fits all' activity. The use of differentiated assessment tasks, apart from ensuring fairness and accessibility, also develops student motivation and engagement (Koshy, 2013). According to Reeves (2013) there are three overlapping concepts that are essential to initiate and sustain the motivation and engagement of students. These are student choice, power, and competence. Reeves argues that:

Empowered students who exercise choice, doing what they want to do, may be temporarily engaged. But if they never become competent, they will become frustrated and distracted. Conversely, competent students who master a skill but never have the opportunity to enjoy a degree of choice or exercise power over the content and nature of their assessments, will dully go through the motions but never achieve a high level of engagement (para. 4).

The interplay of choice, power and competence are embodied in AfL practices.

As teachers adhere to AfL principles in their practices, apart from using differentiated assessment to provide choices to students, they develop assessment tasks that progress in terms of complexity to further challenge and move students to a higher level of performance.

3. *Teachers' view of student roles in an AfL context.*

The move of the Assessment reform Group (ARG) (2002) to use the term 'assessment for learning' instead of 'formative assessment' emphasised the key role of students in assessment as primary participants, and not just passive recipients of assessment processes and outcomes. This definition highlights the socio-cultural belief of teachers about students' role in learning as active collaborators and co-participants in constructing their knowledge and skills within the appropriate context of learning (Lave & Wenger, 1991). This notion is supported by several studies in AfL that demonstrate that students who assume active roles in all aspects of learning, teaching and assessment have higher achievement levels (Dochy, Segers, & Sluijsmans, 1999; Kuh & Ewell 2010; McDonald & Boud, 2003). Apart from the positive effect of student engagement on their achievement level, the social environment and processes created by teachers and students promote self-esteem, encourage engagement, and increase motivation. The active involvement of students in the learning

process and the ability of the teachers to develop culture of trust and respect inside the classroom are important practices of AfL, which consequently increases student motivation.

The nature and quality of assessment tasks and the overall assessment approach used by teachers also impact on the emotional aspect of student learning. This was demonstrated in a study by Cowie (2005), in which students perceived a teacher action inside the classroom as a threat or as a motivator depending on how teachers developed and demonstrated trust and respect inside the classroom. Students perceived trust and respect as reciprocal activities. They trusted and respected teachers who respected the way they wanted to learn and who trusted that they could learn on their own. According to Cowie, one consequence of this high level of trust between teacher and students is the likelihood that students will act on the feedback given by the teacher. As argued by Sadler (1989), the effective use of feedback lies in its power to motivate students to set personalised goals to improve their learning.

Another strategy in AfL that is valuable in promoting student engagement and interaction is the use of peer-assessment. The work of Marshall and Drummond (2006) highlights the role of students as active assessors of their peers' output and as an important resource who can provide feedback to each other. The use of peer-assessment initiates the engagement process between students, whilst dialogue between students and teachers to reinforce the peer-feedback further enhances engagement and interaction. The nature of a learning environment where every student is valued and seen as active contributors to knowledge construction empowers students "to contribute in a positive and productive way both inside and outside the school" (Clark, 2011, p. 9).

4. *Teacher and student relationships in an AfL context.*

The interpersonal relationships between teachers and students affects motivation and achievement (Martin & Dowson, 2009). As argued by Willis (2009), the nature of the teacher-student relationship in the classroom determines the level of participation of students in both learning and assessment. Teachers' classroom assessment practices and the power and control between teachers and students can promote student autonomy (Pedder, 2006), depending on how teachers establish relationships in the classroom to develop students' active participation. In a social learning environment, the multiple identities of teachers and students are recognised. The nature of learning is seen as a product of social interaction, which raises the issue of power and control (Pryor & Crossouard, 2005). One could ask, if students are considered to be independent learners, are they the ones controlling the social environment of learning? What power do teachers have in the social environment? These issues of power and control can be addressed if negotiation between students and teachers is open. These identity issues are important considerations in the meta-contextual reflection of students and teachers as they engage in assessment processes (Pryor & Crossouard, 2005). The identities formed by students and teachers enable them to relate to their social environment and they use these identities when they are faced with a new context for learning.

One indicator of a good teacher-student relationship is that success and failure in learning are a shared responsibility of teachers and students (Marshall & Drummond, 2006). This shared responsibility reinforces both the roles of teachers as activators of learning (Hattie, 2008) and the role of students as independent learners. Both teachers and students constantly monitor learning

progress and identify learning needs and they collaboratively work together toward meeting the desired learning outcomes. If failures occur, teachers and students work together to change the context of learning or use new learning and assessment tools to find better ways to achieve the desired outcomes (James, 2006).

5. *Teachers' roles as assessors in an AfL context.*

In AfL, teachers are seen as the more able individuals who provide assistance to students. This is consistent with the ideas of Lave and Wenger (1991) that teachers, together with students, monitor student progress and provide assistance to develop students' learning autonomy. One way to support students in their learning is for teachers to identify individual students' zone of proximal development (ZPD), which is defined as the difference between what the students can independently do and what they can do with assistance (Vygotsky, 1978). The ZPD has a wide application in AfL practices as it not only provides information to teachers about the current level of ability of individual students, but it also enables teachers to identify appropriate learning interventions and support to address individual learning needs of students. Also, the ZPD enables teachers to develop a differentiated assessment approach to further empower students in the assessment process. Ash and Levitt (2003) argue that working with the construct of the ZPD facilitates the transformation of teachers' practices and enhances professional development. As teachers assess the work or performance of students, they develop further understanding of the assessment task and define quality work using the success criteria. This is further enhanced when they work collaboratively with other teachers as their collaboration initiates the development of shared understanding of the quality of

performance. In addition, as teachers assess the performance of students, they can identify if there is a mismatch between their expectations and student output. The identified mismatch can facilitate teachers' reflective thinking to evaluate their practices and other components of learning and teaching activities, and use the results of their reflection to adjust their teaching, success criteria, and their expectations to appropriately meet the needs of students. This regular reflection and adaptation of teachers enhances their skills in the use of assessment information to inform learning and teaching.

6. *Teachers as part of a community of practice.*

In a socio-cultural paradigm, teachers' involvement in a community of practice (CoP) helps them put knowledge into practice (Schlager & Fusco, 2004), validates their ideas and beliefs, gains access to and evaluates the application of new information from colleagues (Rhodes & Beneike, 2002), and co-constructs new knowledge and skills together (Flagg & Ayling, 2011). According to Harrison (2005), teachers' engagement in discussions related to their practices not only deepens their understanding of their profession, but serves as agents of change for others to establish shared common goals in enhancing their capabilities. The effects of these teacher interactions to student learning are shown by the study of Goddard, Goddard & Tschannen-Moran (2007), in which they have found that the level of teachers' collaboration influences student achievement. This is because teachers who are actively engaged in discussions around curriculum, instruction, and assessment are using context-based learning and teaching practices.

7. *The role of cultural setting and policy in the AfL context.*

Both at the system level and in a wider context, cultural, political and economic factors affect the way AfL is being implemented (Davison, 2013). Inside the classroom, the different backgrounds of the students need to be considered by teachers in order to develop and implement responsive assessment and curriculum practices (Klenowski et al., 2010). Apart from the classroom context, teachers' AfL practices are also influenced by the school culture and the education department's policy. Clark (2011), drawing from the work of Damon and Phelps (1989) on equality and mutuality as components of collaborative learning, argues that a socio-cultural classroom is supported by mutuality, particularly from policy makers. This mutuality is evident in the support of government for AfL reforms. The countries that have adopted and implemented AfL reforms acknowledge the importance of creating and implementing policies that are coherent with AfL principles. Davison (2013), leading large-scale assessment reforms in Hong Kong, Brunei, and Singapore, attributed the success or failure of these initiatives to a shared understanding of AfL principles and practices across schools and all levels of bureaucracy. Such political support strengthens the implementation of AfL and reshapes practices that are inconsistent with the principles of AfL. Similarly, the economic aspect of AfL implementation is equally important as it is costly and needs strong financial support from higher authorities. Another important layer that adds to the complexities of the context-dependency of AfL practices is addressing the expectations and beliefs of the community, parents, and carers (Davison, 2013; Hutchinson & Young, 2011).

These seven key points demonstrate that teacher AfL practices are strongly aligned with a socio-cultural theory of learning. In this perspective, teachers use assessment as a tool to support students in their learning, give greater responsibility to students in terms of monitoring their learning, and develop a learning environment in which students are actively engaged in making decisions related to their learning. In addition, teachers establish a community of practice to support each other in their on-going AfL literacy development. Lastly, teachers use AfL skills that are appropriate within the social, cultural and economic context of their particular educational system.

This theoretical support for teacher assessment practices highlights the elements that constitute a teacher AfL literacy. However, in order to provide greater clarity around what constitutes teacher AfL literacy, the construct of assessment for learning needs to be clearly defined based on existing conceptualisations as supported by empirical evidence. In the next section, I outline the development of the construct of assessment for learning, focusing on current understanding of sound assessment practices.

2.3. The Development of the Assessment *for* Learning Concept

2.3.1. Overview

The debates over the types and purposes of assessment divide many assessment practitioners, researchers and teachers into different camps. There are different perspectives on the key elements of assessment (Baroudi, 2007), its nature and process (Bennett, 2011), the distinction between summative and formative assessment (Sadler, 1989a), and the appropriateness of the use of assessment types (William, 2011). However, the concept of AfL has probably had the greatest influence on debates over

assessment in the last three decades.

Although AfL was first used by H. Black (1986) in a published book chapter, the paper presented by James (1992) on assessment for learning at an American Educational Research Association (AERA) conference and the book published by Sutton (1995) with the title ‘Assessment for Learning’ provided the major impetus for the use of the concept and name in the educational field. The most cited definition of AfL is the one from the Assessment Reform Group (2002) which states that “assessment for learning (AfL) is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go, and how best to get there” (p. 1). Although most educators agree on this definition of AfL, the conceptualisations and theorisation of formative assessment and its relationship to AfL remain problematic. To address these definitional and conceptual issues surrounding AfL, this review of the literature of the AfL concept is divided into three subsections that correspond to three different waves of understanding of AfL concepts and practices. Although these three periods are chronological, they are not distinct but overlap as each successive period attempted to account for the weaknesses of its predecessor.

2.3.2. Wave 1: Adoption, Dichotomisation and Formalisation

Although formative and summative assessments have always been an implicit part of educational practice, it was in 1967 that Scriven made the distinction between formative and summative evaluation for program implementation. The former term was coined to describe the original concept of looking for data that could be used to modify the program during its entire implementation, whilst the latter term was used to describe the collection of data to judge the worth of the program.

In 1971, the concept was adopted by Bloom, Hastings and Madaus and applied in the context of student learning by replacing the word ‘evaluation’ with ‘assessment’. They describe the distinguishing features of summative assessment (SA) versus formative assessment (FA), including the purpose and timing of assessments and the interpretability of data gathered by test items (Bloom, Hastings, & Madaus, 1971). Since then, both terms have been used operationally to distinguish the types and functions of assessment. In general, those assessments being conducted by teachers regularly or daily to gather data aimed at improving learning were collectively referred to as FA whilst those assessments conducted to determine whether learning had occurred over a period of teaching were classified as SA. This distinction led to the dominant belief that SA is mutually exclusive from FA.

However, despite the introduction of FA, the pressure of external examinations for accountability purposes continued to institutionalise SA as the dominant paradigm, especially in the United States of America (USA) as well as in the traditional examination-driven cultures of the non-English speaking world. The effort of schools to perform well in local, state, national and even international tests has reinforced the prominence of summative assessment practices even in the classroom. This is evident in the countless numbers of books on summative assessment and testing published during this period with only a handful on formative assessment. According to Sadler’s (1989) inventory, from the time the word ‘assessment’ was used as a key concept, only the works of Rowntree (1977), Bloom, Madaus, and Hastings (1981), H. Black and Dockrell (1984) and Chater (1984) mentioned feedback and formative assessment. The dominance of SA in the USA has continued to persist even though there have been widespread calls for assessment reform (Stiggins, 2005).

During this period, the term assessment for learning had not been coined yet and formative assessment was given only superficial attention. Crooks (1988) published a review of the literature at the time which highlights that evaluation of student learning was rarely used to help students in their learning. He was one of the first assessment theorists who emphasised that feedback on the performance of the students must be communicated to induce further learning, a concept later strengthened and formalised by Sadler (1989).

The conception of and distinction between FA and SA in the USA spread to other countries, including Australia, which during this period, was not very reliant on high-stakes tests, hence judgments about student learning involved a relatively large degree of teacher-based assessment, even for formal secondary graduation requirements. In Queensland Sadler (1989) formalised the distinction between FA and SA, attributing it to the work of Ramaprasad (1983), who puts feedback as the central feature of FA. Sadler argued that FA differs from SA in terms of purpose and effect (like Bloom et al., 1971) but excluded timing as a distinguishing feature. He further posited that “many of the principles appropriate to summative assessments are not necessarily transferable to formative assessment, the latter requires a distinctive conceptualization and technology” (Sadler, 1989, p. 120). The work of Sadler further enhanced the dichotomy between FA and SA and further divided teacher assessment practices into either FA or SA.

However, reviewing Sadler’s conception of feedback as the central feature of FA, (after the collection and analysis of evidence of learning), it can be argued that SA can be used to provide feedback as well. In other words, conceptually, SA can be part of FA as the information gathered in SA about individual students’ current level of learning can and should be able to be used by both students and teachers to develop

learning goals and approaches to meet the required standards. This point was later highlighted by Davison (2007) and Taras (2009), but was not acknowledged by Sadler. He maintained his view that SA and FA were mutually exclusive.

2.3.3. Wave 2: Complexification and Exploration

The second wave of complexification and exploration of the key concepts underpinning A/fL was seen when the United Kingdom (UK) implemented an assessment reform in 1999 focused on the use of formative assessment, based on the work of Black & Wiliam (1998). Their meta-analysis of over 250 studies on the effects of formative assessment in learning and teaching was commissioned by the Assessment Reform Group to explore the utilization of FA to improve student learning. Shortly after the publication of their seminal paper, which showed the powerful effect of formative assessment in improving student learning and the benefits it had for low performing students, various educational systems adopted formative assessment as a centrepiece of their educational reforms. However, even before Black and Wiliam's work was commissioned, the value of formative assessment in the United Kingdom had been acknowledged. In fact, from 1970s until the release of Black and Wiliam's ground-breaking paper in 1998, although SA was dominating educational systems, there had been numerous studies documenting the effects of assessment in supporting learning. In 1982, FA had started to gain considerable attention with the recommendation of the Inquiry into the Teaching of Mathematics in Schools in the UK that "assessment should be diagnostic and supportive and teaching should be based on the scheme of work which is appraised and revised regularly" (Cockcroft, 1982, p. 243). In 1987, the UK government required the integration of national assessment and teachers' assessment to report individual achievement of students. This stirred national debates because of the difficulty in

integrating the results of FA with SA. It was emphasised by the National Curriculum Task Group on Assessment and Teaching (1998), the agency commissioned to develop a format in reporting the results of FA and SA as one, that the results of FA could be combined together to be used for summative purposes but the results of SA could not be used for formative purposes. This further confused teachers and did not encourage any interplay between FA and SA.

During this period, the definitions of FA and AfL started to be formalised. In the report of their study, Black and Wiliam (1998) offered a definition of FA:

We use the general term assessment to refer to all those activities undertaken by the teachers - and by their students in assessing themselves - that provide information to be used as feedback to modify teaching and learning activities. Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching to meet the student needs (Black & Wiliam, 1998, p. 140).

This definition of FA is consistent with the work of Sadler where feedback is considered to be the main feature of FA. It is evident from this definition that results of FA have to be used by teachers to modify their teaching to address learning needs of students. However, attributing to FA the worth of assessment activities to inform teaching and learning is quite simplistic as there are other factors that contribute to the effectiveness of utilising assessment information to improve student learning.

The ARG encapsulated the basic problem with the term FA:

The term “formative” itself is open to a variety of interpretations and often means no more than that assessment is carried out frequently and is planned at the same time as teaching. Such assessment does not necessarily have all the characteristics just identified as helping learning. It may be formative in helping the teacher to identify areas where more explanation or practice is needed. But for pupils, the marks or remarks on their work may tell them about their success or failure but not about how to make progress towards further learning (ARG, 1999, p. 7).

For ARG, the worth of assessment in improving student learning is guided by the key characteristics of AfL, of which FA is just a part. They defined AfL as “the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there” (ARG, 2002, p.2). With this definition, the ARG (2002) formulated 10 principles of AfL based on the research evidence of effective teacher assessment practices. According to ARG, assessment:

- Is part of effective planning
- Focuses on how student learn
- Is central to classroom practice
- Is a key professional skill
- Is sensitive and constructive
- Fosters motivation
- Promotes understanding of goals and criteria
- Help learners how to improve
- Develops the capacity for self-assessment
- Recognises all educational achievement

The ARG’s definition of AfL and its characteristics broadened Black and Wiliam’s pedagogical definition of FA. It emphasised the responsibility of students for their learning, and the need to ensure they are highly involved in assessment processes and in monitoring their learning so they had substantial information (including feedback from teachers and the results of self and peer assessments) to identify what they needed to do in order to attain the desired learning outcomes. The three processes described in the definition offered by ARG are consistent with the core premises identified by Ramaprasad (1983) and adopted by Sadler (1989) to allow feedback to function (see Table 2.1).

Table 2.1.

Comparison of the Three Macro Processes in Feedback, FA and AfL

ARG (1999, p. 2) Learners and their teachers to decide	Sadler (1989, p. 121). Learners have to	Ramaprasad (1983, p. 6). Conditions required for feedback to function include the availability of
where the learners are in their learning	compare the actuals (or current) level of performance with the standard	data on the actual data level of the same parameter
where they need to go	possess a concept of the standard (or goal, or reference) being aimed for	data on the reference level of the focal system parameters
how best to get there	engage in appropriate action which leads to some closure of the gap	mechanism [for] comparing the two data to generate information about the gap between the two levels

The work of ARG in promoting the concept of AfL received significant attention in many educational systems, including England, Wales, Hong Kong, Singapore and most states of Australia, all of which took AfL up as a basis for their assessment reform. However, the attention received by AfL and the continuous promotion of Black and Wiliam and Black et al. of FA created some confusion amongst teachers, researchers, and even amongst themselves. In 2003 Black et al. extended their definition of FA but used it synonymously with AfL and still emphasised that the main feature of the extended definition is the utilisation of the results of assessment as feedback to improve student learning.

Assessment for Learning is any assessment for which the first priority in its design and practice is to serve the purpose of promoting pupils' learning. It thus differs from assessment designed primarily to serve the purposes of accountability, or of ranking, or of certifying competence. An assessment activity can help learning if it provides information to be used as feedback, by the teachers, and by their pupils in assessing themselves and each other, to modify the teaching and learning activities in which they are engaged. Such assessment becomes 'formative assessment' when the evidence is actually used to adapt the teaching

work to meet learning needs (Black, Harrison, Lee, Marshall, & Wiliam, 2003, pp. 2-3).

It is interesting to note that Black and Wiliam acknowledge the participation of students in assessment processes, particularly self and peer assessments but still use the term FA to describe the practice of using assessment data to adjust teaching practices to meet student learning needs. This revised definition of FA by Black et al. was criticised by Taras (2009) [although Taras herself earlier had used the term interchangeably (see Taras, 2005)], as it contradicted Black and Wiliam's earlier definition as there was no mention of student responsibilities in assessment. She commented, "FA is part of teaching methodology and has more to do with teachers than with learners, thus seeming to contradict the initial definitions" (p. 61). Even Stiggins (2002) was at one point tempted to use A/fL as synonymous with FA but quickly clarified that these two terms were not the same. Stiggins emphasised that the processes in FA of frequent assessment to provide information to teachers to adjust learning is just part of A/fL. According to him, in the process of learning, FA becomes A/fL if students become actively involved in the assessment processes and become controllers of their own learning.

Not only did Black and Wiliam and Black et al. change their definitions of FA and A/fL several times, they also widened the dichotomy between FA and SA. Their exclusion of the effects of SA on students' learning in their research tended to make teachers and other researchers think that SA must function very differently from FA. Shortly after publication of their 1998 paper, Biggs (1998) critiqued the premise of the study which focused only on FA without accounting for the interplay of SA. Biggs' critique focused on the effects of SA on learning which he called 'backwash' (Biggs, 1996), arguing that it can negatively affect learning and its effect can be much stronger

than the effects of feedback. Biggs wanted to see Black and Wiliam deal with the broader context in which learning occurs. Biggs (1996) argued that:

There is a powerful interaction between FA and SA that could be incorporated in an overall synthesis, so that both backwash (from SA) and feedback (from FA) are conceptualised within the same framework. Then indeed one might have a powerful enhancement of learning, using such a synthesis to engineer backwash from SA so that the effects were *positive*, the backwash from SA supporting the feedback from FA (p. 106).

Biggs's argument highlights the interconnectedness of FA and SA, that is, the fact that it is impossible to fully isolate one from the other, as they both support each other. Despite the wide range of evidence and criticisms of the weaknesses (both theoretically and practically) of their conception of FA, Black and Wiliam and Black et al. continued using the term FA instead of AfL to collectively describe all assessment practices that improve learning and teaching¹.

In 2009, Black and Wiliam made a slight revision of their definition of FA, which they claimed to be consistent with the definition of AfL by the ARG:

Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited (p. 109).

Not only did Black and Wiliam revise their definition of FA, they also added more elements or key characteristics to FA. They incorporated many of the elements of assessment practices that had been proven by research to improve teaching and learning.

¹ In 2007, Stiggins who made reference to and cited Black and Wiliam's (1999) work on FA, commented that he had stop using the term FA because through the years of its development, the term has lost its meaning (Bennett, 2011).

For example, the critique by Biggs (1998) of their article regarding the exclusion of the effect of SA in learning pressed them to incorporate SA as part of formative assessment (Black & Wiliam, 2009), adjusting their earlier position that SA cannot be used for formative purposes. Black and Wiliam, together with their colleagues, Harrison, Lee, and Marshall, revised the four areas suggested in their original paper, *Inside the Black Box*, to include the formative use of summative tests (Black et al., 2003). They argued that the result of SA shows evidence of students' achievement and, therefore, can be used to elicit and give feedback. Also, they emphasised the value of SA, "these can also communicate to learners what is and is not valued in a particular discipline, thus communicating criteria for success" (Black & Wiliam, 2009, p. 8). Another revision of Black and Wiliam's conception of FA is their adoption of Perrenoud's (1998) argument that feedback is not the main issue, but rather it is just one of the many elements that affect learning. In 2009, Black and Wiliam linked their analysis of feedback to other learning theories 'so that the concept of formative interaction may be enriched and contextualised in the light of relevant theories' (p. 6) including the self-regulated learning model (Boekaerts & Corno, 2005; Boekaerts, Maes, & Karoly, 2005) which was linked to the role of feedback.

The complexification and exploration of the use of FA and A/L and the exclusion of SA from FA had driven researchers and teachers to find empirical and theoretical evidence to resolve the issue. This triggered much research and many arguments revolving around the use of FA and SA to improve student learning. As more and more research findings were made available and more theorisations of FA, SA and A/L developed, teachers, researchers, policy makers and other stakeholders started to understand what was fundamental to assessment practices that promote student learning. However, this realisation did not stop other researchers (i.e. Black & Wiliam, 2009)

from continuing to widen the dichotomy between FA and SA, and hence, seemingly creating more confusion and resistance (Bennett, 2009).

During this time, a large number of studies explored the effects of FA and SA on student learning. Most of the studies were focused on investigating the effects of SA on student learning in an attempt to find strong evidence to support the superior role of SA over FA in learning and teaching. However, as researchers gathered evidence to show that SA raises standards, what was also highlighted was its negative effects in limiting learning (Harlen & Deakin Crick, 2003), narrowing the curriculum (Johnston & McClune, 2000; Reay & William, 1999) and compromising teaching (Gordon & Reese, 1997; Leonard & Davey, 2001; Pollard, Triggs, Broadfoot, Mcness, & Osborn, 2002; L. Shepard, 2000).

Another area of exploration during this period was the interplay between FA and SA. Researchers came to the conclusion that SA could be used to enhance learning and teaching in various ways. One of the findings was that if the content, format and design of SA reflect the development of skills, then students' preparation could increase their motivation as they try to understand the links between different concepts (Shepard, 2006). Also, a synergy between SA and FA exists (Harlen, 2005) as evidence of student learning can be used for any purposes. This is exemplified by the use of portfolios for high-stakes assessments in Queensland where the evidence of student learning collected is used to provide regular feedback to students. Then, at the end of the specific period of teaching, the documented evidence of student learning is summarised to describe the overall performance of the students. To support the interplay of FA and SA, Taras (2009) argues that feedback always relies on SA as all assessment processes start with a summative assessment. According to her, the effectiveness of FA is being constrained by resistance to acknowledging the interplay between FA and SA.

2.3.4. Wave 3: Period of Realisation and Reconciliation

The realisation that the results of summative assessment can also be used to improve teaching and learning provided the impetus for the third phase of AfL development. This is the period when SA and FA were seen as mutually supporting each other's purposes. The work on AfL is more influential in this period than Black and Wiliam's (1998) and Black et al's. (2003) conceptions of FA. Education departments, ministries, educators and teachers in the USA, Hong Kong, Singapore, Australia, New Zealand, and Brunei moved forward from the previous conceptions of FA, building on the definition of AfL given by ARG, but incorporating specific elements that have been proven by research to greatly contribute to improving learning. For example, the role and value of SA (Biggs, 1998; Stiggins, 2005; Bennett, 2009), the effect of self-assessment in self-efficacy and intrinsic motivation (Schunk, 1996), and the role of the teachers as activators rather than facilitators of learning (Hattie, 2008) were highlighted during this period.

The distinction made between formative and summative assessments is now seen as invaluable because these types of assessments do not function in isolation, as Biggs (1998) had further explained, especially "when the summative assessment is defining the parameters for the formative assessment, it does not seem helpful to confine a review of the one with the attempted exclusion of the other" (p. 108). As illustrated by Bennett (2009), assessments which are conducted to document student learning can function also formatively and those assessments conducted by teachers regularly to modify their teaching can also be used for summative purposes. It is for this reason that Bennett tried to resolve the definitional issue of FA but he did not acknowledge the use of the term AfL. His key point was "calling formative assessment by another name may only exacerbate rather than resolve, the definitional issue"

(Bennett, 2009, p. 7). This is not, however, the main argument with the definitional issue because these two terms are now clearly being used in different ways. Assessment for learning is now perceived as the overarching umbrella of which FA is just a part.

The current re-thinking of assessment practices, due to weakness in the ability of high-stakes tests to improve learning, triggered the return of interest in FA in the USA (Stiggins, 2005). The result of standardised summative assessment conducted once a year does not help teachers gain significant insights about the learning of the students as SA collectively measures the overall achievement in all standards. Consequently, the lack of evidence of students' mastery in each standard makes it impossible for teachers to identify specific interventions to help individual students perform better.

Furthermore, the timing of SA does not provide sufficient opportunity for teachers to use the data to enhance student learning. These realisations of teachers and researchers have slowly changed the way SA is conducted in the USA. SA is now conducted more frequently and the results are used to inform instruction. Although this approach is helpful in adjusting teaching to help students learn better, the nature of the multiple-choice tests used in SA and the responsibility given to teachers "fall short of tapping the immense potential of formative thinking" (Stiggins, 2005, p. 327). This has led to the adoption of an AfL approach, which Stiggins defines in contrast to two major elements of traditional formative thinking. First, he said, AfL includes not just FA that informs teaching, but also informs students about their own learning. Students are empowered to take active roles in their learning by being trained to use assessment information to make changes in the way they learn and on how they approach the achievement of learning outcomes. Second, whilst AfL incorporates frequent testing of student achievement of learning outcomes, it changes the focus of frequent assessment. Regular formal or informal assessments explore the learning steps to achieve the learning

outcomes, “it tells users if and when students are attaining the foundations of knowledge, the reasoning, the performance skills, and the product development capabilities that underpin the mastery of essential standards” (p. 328). In other words, A/fL provides a whole picture of students’ progression toward the achievement of learning outcomes.

In a high-stakes test environment, A/fL has been seen to operate effectively in improving learning without the negative effects of assessment whilst maintaining accountability (Popham, 2011). This is best exemplified by the practice of Queensland with their state-wide Core Skills Test of student assessment administered in the final years of secondary school. The Queensland Studies Authority collects and analyses assessment data and disseminates the results to the schools, including information on how students responded to the test. The information provided is used by teachers to adjust learning and teaching activities. This practice supports the idea that for a successful implementation of A/fL, assessment data should be used at all levels of authority (i.e., monitoring and evaluation of teaching practices, professional development, policy) to influence classroom practices. What is highlighted in this practice, is the critical role of state authorities in providing support to teachers in their assessment literacy (Klenowski, 2011). Another example of the participation of state authorities in assessment is the case of the Board of Studies Teaching and Educational Standards (BOSTES) of New South Wales in Australia where they created the Assessment Resource Centre (ARC) website aimed at supporting and enhancing professional practice in assessment and reporting student achievement (BOSTES, n.d.).

The current thinking about A/fL has helped to reconcile the different functions of SA and FA, and to see them as a continuum (Davison, 2007) rather than a dichotomy. Other core concepts have also shifted during this current period, including the role of

the teacher. The role of the teachers in student learning, in the 1980s and 1990s widely conceptualised as a facilitator, has shifted to being seen as an activator of learning. Stiggins (2005) argued that when AfL is used effectively, it “always triggers an optimistic response to assessment results from within the learners” (p. 328). The processes identified by Stiggins that help teachers trigger learning were highlighted as a result of Hattie’s (2008) meta-analysis. Hattie compares the effect size of processes that describe teachers as facilitators and those processes that describe teachers as activators of learning. The average effect size ($d=.60$) of reciprocal teaching, feedback, teaching students self-verbalization, meta-cognition strategies, direct instruction, mastery learning, setting challenging goals, and use of behavioural organizers (which are considered as approaches in activating learning) is significantly higher than the average effect size ($d=.17$) of simulations and gaming, inquiry based teaching, smaller class sizes, individualised instruction, problem-based learning, different teaching strategies for boys and girls, web-based learning, whole language reading and inductive teaching (which are considered as approaches in facilitating learning).

There are several key points that underpin the current re-thinking of AfL. The philosophy underlying assessment processes in AfL is empowerment. Collectively, AfL activities should be an enabling mechanism to develop the ability of students to take responsibility for their learning. This makes AfL a truly student-centred approach. In this approach to assessment, students are guided to identify their strengths and weaknesses, to understand success criteria and standards, to assess their own and their peers’ work, to become responsive to the feedback they receive and to set goals to achieve the desired learning outcomes. In other words, teachers are more concerned about the processes of assessment and learning, not just the procedure, and work to ensure that students are actively participating in learning and teaching activities.

As part of this review of AfL the role and nature of teacher professional development specifically related to assessment has also been highlighted. As activators of learning and instructional designers, teachers have multiple, complex and interrelated tasks that need regular moderation. To ensure that AfL is effectively implemented, teachers have a two-fold task: (1) to develop their AfL skills and (2) to ensure that students will respond positively to the activation to develop their skills required in learning. If teachers are focused only in the enhancement of their own skills, then assessment for learning is a failure. Prior knowledge and subsequent training for teachers are important aspect of AfL. Whilst students are learning, it is expected also that teachers should do the same. AfL then requires teachers to continue engaging in professional development (formal or informal) to address their professional needs.

Around this time, AfL also started to gain prominence in Asian countries, which have a long history of summative assessment dominant practices. After Hong Kong, countries like Singapore and Brunei were quick to adopt the concept. In Hong Kong school based assessment (SBA) for English language education is rooted in the principles of AfL (Davison, 2007, 2013). Apart from public examinations, teachers design assessment tasks and conduct formal and informal classroom assessment whilst engaging students in self and peer assessment. AfL in this context is used to gather a holistic view of student achievement because traditional testing cannot provide information about the more important aspects of learning. The most significant contribution of this initiative has been the adoption of a more inclusive model of AfL where all assessments, either formative or summative, are conceptualised under the umbrella of AfL and are conducted for the purpose of helping students in their learning (Kennedy, Sang, Wai-ming, & Fok, 2006).

In Singapore, the principles of A/fL were incorporated into their traditional practice (Davison, 2013). Much of the effort of the government goes into professional development where the focus is on enhancing teacher assessment literacy in using A/fL in the classroom. With Singapore's recent assessment reform, Tan (2011) argues that the effect of assessment on student learning should also be considered. In other words, useful and meaningful assessment practices enable and support the development of students' lifelong learning skills, with A/fL having a proven benefit in the acquisition of such skills (Hanrahan & Isaacs, 2001).

In Brunei, the principles and practices of A/fL have been incorporated into their school-based assessment reform, which they now call school-based assessment for learning (SBA/fL), described in Chapter 1. The SBA/fL is uniquely contextualised to meet the objectives of the Brunei assessment reform. It was originally designed for students in Year 7 and 8 to prepare them in their transition to upper secondary school. SBA/fL is a dominant practice from Term 1 of Year 7 to Term 3 of Year 8. The use of A/fL incorporates the typology of assessment practices developed by Davison (2007). In Terms 1 and 2 of Year 7, students are provided with assessment activities that are designed to scaffold their acquisition of skills in taking responsibility for their learning. The process starts with informal assessment activities that help students prepare for their engagement in assessment starting in Term 3 of Year 7. It eventually progresses into a more formal assessment activities, which are criteria-based that address the needs of the students, whilst adhering to external standards. As students enter Term 3 in Year 7, they will engage with a prescribed set of assessment tasks over 18 months called the Brunei Common Assessment Tasks (BCATs), which comprise a variety of authentic assessment tasks. These assessments have both formative and summative functions. Whilst the combined result of SBA/fL comprises 30% of their final grade, the BCATs

are used by teachers to engage students in self and peer assessments and to provide meaningful feedback to assist them in their learning. In 2012, SBAfL was extended to all Brunei primary schools.

Another equally important development in the reconceptualization of AfL has been the realisation that the responsibility for student learning does not just lie only with teachers and students alone. A concerted effort and common understanding by all stakeholders is needed to ensure that government policies, curriculum design and parents' expectations do not constrain the effective implementation of AfL (Davison, 2013; Bennett, 2011). AfL operates in a wider educational context, hence as argued by Davison (2013), policy makers also play an important role in AfL. The policies and guidelines they create may make or break AfL implementation. Teachers, students and all other stakeholders involved in the educational process, particularly policy makers, need to have common aims, beliefs, and understandings to enable the successful implementation of AfL (Buhagiar, 2007; Hayward & Hedge, 2005; Hayward, Priestley, & Young, 2004). Thus, the definition of AfL should not only focus on the roles of teachers but it should also consider system issues referred to by Bennett (2011). For example, in a high accountability educational culture, the implementation of AfL is impossible even if teachers are experts in all domains of AfL. For the successful implementation of assessment reform, all components of the education system should function coherently.

In summary, the theorisation of AfL has gone through a long process of debate, complexification, exploration, confusion and convergence. The dichotomy of FA and SA once divided teachers, researchers and even policy makers. As each camp sought to establish dominance over the other, significant insights from various research studies on FA and SA, conducted separately or in combination have emerged. After resisting the

dichotomisation of SA and FA, reconciliation has started to take place. Much attention in recent developments in assessment reform has focused on AfL, rather than looking at FA and SA separately.

The current conceptualisation of AfL outlined above is used to define the construct of teacher AfL practices used in this study. As argued, teacher assessment practices are seen as a continuum, rather than a dichotomy. Also, assessment is seen as operating in a wider perspective where every stakeholder has a responsibility in ensuring student learning. .

2.4. Definition and Development of Teacher AfL Literacy

2.4.1. Overview

As indicated in the previous section, teacher AfL knowledge and skills have changed over time and can be characterised as a highly situated construct, which means that it must be operationally defined for a particular educational system. In this section, I define and characterise teacher assessment literacy in the context of AfL. I begin by looking at the diverse views of researchers and practitioners on what knowledge and skills constitute teacher AfL literacy. From the shortcomings of the existing views, I propose a more holistic approach to conceptualising it, outline the key characteristics and principles of assessment literacy, and I identify key issues around in its implementation.

From a synthesis of more than 800 meta-analyses studies, Hattie commented that ‘the remarkable feature of the evidence is that the biggest effects on student learning occur when teachers become learners of their own teaching, and when students become their own teachers’ (Hattie, 2008, p. 22). In other words, improving

achievement means ensuring that students take full responsibility in their learning and they are able to self-assess and to self-monitor their progress. However, these abilities of students are not innate and need to be developed by teachers. Thus, the way to develop self-regulated students lies through the ability of the teachers to provide opportunities for students to self-assess and to self-monitor their learning progress.

Hattie (2008) summarises the value of teacher assessment literacy in improving student learning:

The act of teaching requires deliberate interventions to ensure that there is cognitive change in the student: thus the key ingredients are awareness of the learning intention, knowing when a student is successful in attaining those intentions, having sufficient understanding of the student's understanding as he or she comes to the task, and knowing enough about the content to provide meaningful and challenging experience in some sort of progressive development. It involves an experienced teacher who knows a range of learning strategies to provide the student when they seem not to understand, to provide direction and re-direction in terms of the content being understood and thus maximize the power of feedback, and having the skill to 'get out the way' when learning is progressing towards the success criteria (p. 23).

2.4.2. The Changing Conception of Assessment Literacy

The views of several authors form the theoretical background of my discussion of teacher assessment literacy. From the early 1960s, there has been a continuous stream of publications, including both books and training materials, on assessment. The continuous publication of assessment resources underpins the whole idea of enhancing teacher assessment literacy. Although the rationale for enhancing teachers' assessment literacy has been constant over time, the content and components of sound assessment practices have changed. Davies (2008) identifies two factors that shaped these changes: the growing professionalization of assessment; and the demand for knowledge and skills in assessment. The focus of earlier assessment literacy was on the measurement

principles, but later shifted to the assessment knowledge and skills that teachers needed in the classroom. This is evident in the views of Fullan (2002) who equates assessment literacy with teachers and principals' capacity to use student achievement data to increase learning, and to inform policy makers on the uses and misuses of achievement data. Apart from this, Webb (2002) highlights the importance of using assessment data to improve teaching and effectiveness of the educational programs to help students' learn. Over time, these professional and technical emphases have evolved to include a more holistic view of the concept of assessment literacy. Stiggins (2005) describes the characteristics of assessment literate teachers as teachers who "know the difference between sound and unsound assessment. They are not intimidated by the sometimes mysterious and always daunting technical world of assessment" (p. 240). These teachers are confident in their capacity to undertake the necessary preparation and planning. They determine the object of their assessment, the purpose of doing it, the best way to assess the construct of interest, the best way to generate exemplary performance of students, the misrepresentation of assessment, and the negative effects of inaccurate assessment. This view of Stiggins is consistent with the view of Popham (2011) that teacher assessment literacy involves teachers' "understanding of the fundamental assessment concepts and procedures deemed likely to influence educational decisions" (p. 265). Popham's emphasis is on teacher ability to use assessment to inform decisions related to student learning and effective teaching.

In the recent years, the emphases of assessment literacy have broadened to include critical views of testing and its social consequences (McNamara & Roever, 2006) and the social roles of assessment (Inbar-Lourie, 2008), including the roles of teachers in providing assessment information to stakeholders (Tylor, 2009). Also, assessment literacy is no longer confined to teachers but there is now a growing

emphasis on the assessment literacy of other stakeholders (Davison, 2013; Taylor, 2009). This trend has been shaped by pressures at the system level where the effectiveness of teacher assessment practices are constrained by the external pressures due to the inconsistencies of understanding and expectations of other stakeholders. This call for stakeholder assessment literacy was first evident in Popham's (2009) view that assessment literacy is directly linked to the responsibilities of people, which means that different stakeholders have different assessment literacy needs. Teacher assessment literacy is a critical factor for addressing the assessment needs of, and building the assessment literacy of, stakeholders (Taylor, 2009).

The various views on teacher assessment literacy are partly anchored to the principles of AfL. In the next section, I look at this construct from the perspective of AfL.

2.4.3. The Existing Dimensions of Teacher Assessment Literacy

Apart from studies that explore the individual teacher assessment skills, there are other studies that use a range of dimensions to characterise teacher assessment literacy. Table 2.2 below presents a summary of some of the dimensions used to describe teacher assessment practices. Amongst these dimensions, three were developed using a theoretical approach and two used principal component analysis to establish the dimensions.

The seven competencies used by Stiggins (1999b) to describe teacher classroom practices cover the dimensions described in the Standards for Teacher Competence in the Educational Assessment of Students (American Federation of Teachers, 1990) which was used by Mertler and Campbell (2005) to develop their Assessment Literacy Inventory. Mertler and Campbell presented a comparison between the dimensions they

used and the ones described by Stiggins and made this remark: “while there is some debate as about the extent to which *The Standards* adequately address those competencies which research shows that teachers need to possess, (shows) that there is a great deal of overlap in the original 1990 Standards and the competencies listed by Stiggins (1999b)” (p. 7). However, Stiggins critiqued the standards used for teacher competence in the Educational Assessment of Students as not being able to account for the critical roles of teachers in meeting the demands of their assessment responsibilities in the classroom. He maintains that the list of competencies he originally proposed is more substantial and these competencies are supported by research evidence about their role in improving student learning. However, although the competencies described by Stiggins are comprehensive and can be supported theoretically, no study has been conducted to gather empirical evidence to demonstrate whether these competencies reflect the actual dimensions of teacher assessment literacy, particularly in the context of assessment for learning.

Another study that attempted to explore the underlying dimensions of teacher assessment literacy is the work of Newfields (2006). The methodology he used was somewhat problematic because, just like Stiggins and Mertler and Campbell, he only used a theoretical approach to establish the four dimensions. Although he employed a series of expert validations and pilot testing, there was no empirical evidence provided to support the existence of four dimensions in the data. Another issue of his work is its strong focus on assessment knowledge and measurement principles, which makes the individual teacher assessment skills skewed towards summative assessment practices. Hence, even if these dimensions are supported by empirical evidence, they will be specific only for teacher summative assessment practices, and not for the broader framework of teacher AfL literacy.

The dimensions of assessment literacy that emerged from an empirical approach are the works of DeLuca and Klinger (2010) and Fulcher (2012), which provided preliminary empirical evidence about the existence of 12 and four factors respectively. The dimensions described by DeLuca and Klinger account for teacher assessment literacy related to practice, theory and philosophy. However, some of these dimensions could be collapsed to form more unified dimensions. For example, the dimensions related to summative assessment (i.e. test design and marking, provincially mandated assessment practices, technical knowledge of assessment practice, philosophies of large-scale assessment) can be clustered into one dimension. Another problem with these dimensions is the dichotomisation of assessment practices into formative and summative assessment that is inconsistent with current thinking in the A/L paradigm.

The dimensions established by Fulcher encompass the dimensions described by Newfields but one dimension emerged that relates to classroom assessment practices. Although the inclusion of classroom assessment practices as one dimension highlights the possible integration of other teacher assessment practices that are more psychometric by nature, the explicit naming of this dimension as teacher classroom practices poses an issue about the implied exclusivity of classroom assessment in relation to other teacher assessment activities, particularly those that relate to summative assessment.

Table 2.2

Comparison of Dimensions from Existing Teacher Assessment Literacy

Stiggins (1999)	McMillan (2001)	Mertler & Campbell (2005)*	Newfields (2006)	DeLuca & Klinger (2010)	Fulcher (2012)
1. Connecting assessment to clear purposes 2. Applying proper assessment methods 3. Avoiding bias in assessment 4. Developing quality assessment exercises and scoring criteria and sampling appropriately 5. Using assessment as instructional intervention 6. Clarifying achievement expectations 7. Communicating effectively about student achievement	1. Factors in determining grades 2. Types of assessment 3. Cognitive levels of assessment	1. Choosing assessment methods 2. Developing assessment methods 3. Administering, scoring and interpreting 4. Developing valid pupil grading procedures 5. Recognizing unethical, illegal, and otherwise inappropriate assessment methods 6. Using assessment results when making decisions 7. Communicating assessment results	1. Terminology 2. Procedures 3. Test interpretation 4. Assessment Ethics	1. Test design and marking 2. Provincially mandated assessment practices 3. Technical knowledge of summative assessment practice 4. Assessment for learning practices 5. Types of assessment 6. Assessment of learning theory 7. Theoretical principles in AoL and AfL 8. Assessment item formats 9. Statistical techniques for assessment 10. Philosophies of classroom assessment 11. Philosophies of large-scale assessment 12. Reliability issues 13. Validity issues 14. Rationale for assessment decisions and practices	1. Test design and development 2. Large-scale standardised testing 3. Classroom testing and washback 4. Validity and reliability

*Based on *Standards for Teacher Competence in the Educational Assessment of Students (1990)*

Although the studies described above that explored the dimensions of teacher assessment literacy have obvious limitations in terms of the assessment knowledge and skills included and their clustering to form factors, the use of these dimensions to describe and characterise teacher assessment practices is helpful in understanding teacher assessment skills. Also, it suggests that teacher assessment literacy consists of dimensions where similar teacher assessment practices cluster together and form distinct factors. However, as can be seen from Table 2.2, there is no agreement regarding the interpretations of the dimensions of teacher assessment literacy. Hence, the dimensions of teacher assessment literacy construct for AfL need considerable work. Taking into consideration the limitations of these existing studies, a more robust methodology is needed so that the use of such dimensions to describe teacher AfL literacy is supported by empirical evidence. Also, an empirical approach will determine which of the dimensions of teacher assessment literacy are contributing the most to measuring the actual level of teacher AfL literacy. This is so that any teacher AfL literacy tool that can be developed from the analyses will contain only dimensions, and their corresponding indicators represented by teacher assessment skills, that can explain the variance observed in the improvement of student learning.

In the next section, I propose a theoretical framework for teacher AfL literacy based on the existing literature, which can be used for this study

2.4.4. Teacher Assessment Literacy in the Context of AfL

Defining assessment literacy from the point of view of AfL is important in attempting to ensure a consistent interpretation, as the literature presents different views and conceptions of what knowledge and skills constitute teacher assessment literacy.

Although the term ‘assessment literacy’ was originally used to describe the components

of assessment practices that teachers need to know in order to carry out their functions in terms of assessing student learning (Inbar-Lourie, 2008; Stiggins, 1997; Stoyanoff & Chapelle, 2005), some of the literature equates the term with teacher understanding of measurement principles. In this view, the emphasis in teacher assessment literacy is on preparing students for high-stake tests, with teachers needing to understand and apply key measurement principles in their classroom practices such as reliability, validity, item difficulty, item discrimination, and interpretation of test results. Even a recent definition of assessment literacy by Fulcher (2012), which was expanded to account for diverse contexts, principles and practices, still leans towards describing the competencies of teachers for large-scale standardised testing. Consequently, the emphasis of teacher professional development is focused on test design and analysis, which should only be one element of the key components of teacher AfL literacy. The view of teacher AfL literacy as simply ensuring the reliability and validity of assessment is too narrow to account for a teacher's daily assessment practices as they try to make sense of how students learn through engaging in various decision-making activities, and hence, to adjust their teaching whilst ensuring that student self-regulation is being developed. Thus, in order for assessment literacy to be relevant to classroom activities, it must be viewed across different dimensions to develop a more holistic definition.

Putting together these different views on assessment literacy, I have developed a more comprehensive definition to describe the core content and design of a teacher assessment literacy program consistent with the principles of AfL. I have expanded the definition of assessment literacy by Webb (2002) which emphasises that assessment literacy should focus on the utilisation of assessment information to monitor students' progress in achieving learning outcomes and to improve teaching and effectiveness of the educational programs to help students learn. According to Webb, a clear indicator of

an assessment literate individual is his/her capacity to use assessment data to increase the motivation of all stakeholders to carry out their roles in education. What was not emphasised by Webb is the underlying purpose for enhancing teacher assessment literacy. From this point of view and incorporating the principles of AfL, in this study, I define this as:

Teacher assessment for learning literacy accounts for knowledge and skills in making highly contextualised, fair, consistent and trustworthy assessment decisions to inform learning and teaching to effectively support both students and teachers' professional learning. The aim of teachers is on building students and other stakeholders' capabilities and confidence to take an active role in assessment, learning and teaching activities to enable and provide the needed support for more effective learning.

In summary, this version of teacher AfL literacy sees teachers' skills in using assessment as the central unifying process of all learning and teaching activities. The daily decisions made by teachers to improve learning are informed by various assessment information about student learning characteristics, progress and needs, as well as by information from other stakeholders such as parents and school leaders. To fully understand the definition of teacher AfL literacy proposed above, I identify its key characteristics, elaborated in the following section.

2.4.5. Key Characteristics of Teacher AfL Literacy

Taking into account all the issues discussed so far, the definition of teacher AfL literacy for this study needs to embody a number of key characteristics. Firstly, AfL assessment literate teachers translate AfL principles and knowledge into practice to ensure student learning. These principles and knowledge are manifested by teachers' skills in using assessment and analysing assessment information to make important decisions in their

learning and teaching context. There are six general clusters of knowledge and skills that comprise assessment literacy. First is the ability of the teachers to develop and use a wide range of assessment strategies to gather robust assessment data (Stiggins, 2005). Second is the ability of the teacher to reflect on the assessment data – both student achievement and student characteristics – and use this information to develop lesson plans and implement instruction. Webb (2002) offers a definition of assessment literacy that encompasses this ability of the teachers. He defines assessment literacy “as the knowledge of means for assessing what students know and can do, how to interpret the results from these assessments, and how to apply these results to improve student learning and program effectiveness” (p. 1). Third are the skills of the teachers to actively engage students in the assessment, learning, and teaching activities. Fourth is the teacher’s use of assessment to enhance and sustain student motivation. As argued by McMillan (2003) teachers need to assess students’ motivation to better understand them and to make contextualised assessment decisions. The teacher needs to know if the assessment task is enhancing students’ engagement in learning and if the students are exerting effort in trying to accomplish the task. Fifth is the ability of the teachers to reflect on their practices and assessment data to identify their professional needs both in assessment knowledge and in curriculum-content knowledge (Timperley, 2008, 2011). Sixth is the ability of the teachers to establish strong partnerships with stakeholders by providing them their assessment information needs. Clear understanding of the principles of AfL will help make the implementation of AfL be more effective. For example, policy makers can review assessment policies to ensure that external factors are helping teachers make assessment information more relevant (McMillan, 2003). These six clusters are further described in the theoretical framework in the last section of this chapter.

Secondly, any teacher AfL literacy model must emphasize the responsibility of teachers in working with students, colleagues, parents, and community. As mentioned above, assessment literate teachers provide assessment information to various stakeholder groups (Fullan, 2002; Taylor, 2009). Apart from the ability of teachers to communicate assessment results to students, it is also necessary that they provide assessment information to different stakeholder groups tailored to each groups' information needs (Davison, 2013). The dialogue between the teacher and stakeholders ensures that stakeholders gain deeper understanding of assessment practices, which consequently draws positive support from them.

Thirdly, the construct of teacher AfL literacy stresses the trustworthiness and social consequences of teachers' assessment practices. McNamara and Roever (2006) cite some misuse of language testing, which is also relevant to other disciplines, like sorting out and labelling students into various social classes instead of using assessment to improve student learning. Similarly, they cite the controlling effects of centralised standards and testing in teaching. In the definition proposed above, I emphasise enhancing teacher competence in ensuring that their assessment practices are fair and trustworthy from the point of view of all key stakeholders.

Fourthly, a teacher AfL literacy program should not be confined to just formal training but consider the socio-cultural aspects of teacher learning where teachers and school leaders "form a knowledge community by discussing, critiquing and questioning fundamental issues relevant to their context" (Inbar-lourie, 2008, p. 389). Formal training may build the foundation of teachers' assessment knowledge but studies have consistently shown that teachers forming a community of learners can better improve their assessment literacy. In the study by Dekker and Feijs (2005), they found that the most significant source of support for teachers in continuing their AfL practices after

initial training period was the development of learning groups within the school, where teachers were provided with the environment where they could discuss related assessment issues with peers through professional meetings or through informal contacts.

Finally, the definition of teacher AfL literacy and its characteristics discussed above must address the inconsistencies both in understanding and the actual practices amongst teachers of what constitutes sound assessment practices (Davison & Leung, 2009). Ideally, teachers should have common understanding of their assessment practices. However, in reality teachers have different views and needs as influenced by a range of factors. Davison (2004) has emphasised that in order to resolve conflicts in teachers' needs in assessment literacy, school systems should prioritise the enhancement of teachers' assessment skills. This is consistent with the findings of Hayward and Hedge (2005) that in the primary sector in Scotland, staff development related to AfL is one of the most common areas of concern with teachers arguing that intensive training is needed if they are to implement AfL reform effectively.

These key characteristics of teacher AfL literacy capture the elements of the construct and how it is used in this study. To support teachers in their AfL development, there is a need to look at the key principles of professional development programs aimed at enhancing their AfL literacy. These key principles will guide teacher learning and how they can be strongly supported in their practices.

2.4.6. Key Principles of Teacher AfL Literacy Professional Development Programs

Given that AfL effectively supports learning better than any other intervention (Black and Wiliam, 1998; Hattie, 2008), this principle needs to be applied to the contexts

where teachers are the learners as well, that is, teacher AfL literacy programs should be consistent with the key principles of AfL. However, current practice in teacher AfL literacy professional development does not seem to follow this core principle. Although most assessment literacy programs are clear in terms of what skills are being developed, teacher learning appears to be made up of several small but fragmented components.

Davison (2013) emphasises that for the successful implementation of AfL, program implementers should use the principles of AfL to develop, implement, monitor and evaluate the assessment literacy program. From this perspective, it is necessary that teachers should be considered as learners - just like the students they teach that is, they need support to become independent and self-regulated teacher learners. To achieve this, teacher assessment literacy programs should be organised around a number of the key principles of AfL. Davison (2013) has expanded on seven core AfL principles and contextualised them to a teacher AfL literacy program, as follows:

1. Consistent with a view of effective student learning in AfL, a teacher assessment literacy program should consider where individual teachers are in terms of their AfL competence, where they need to go, and how best to get there (Davison, 2013). Any assessment literacy programs aimed at supporting teachers to enhance their assessment competence should begin with evaluating their current level of performance and identifying their training and support needs.
2. AfL should be embedded in curriculum and assessment institutionally and pedagogically. Teacher AfL literacy programs should start by setting and sharing appropriate learning outcomes, success criteria, and performance standards. These learning outcomes are the teachers' assessment knowledge and skills, the success criteria are the indicators of these knowledge and skills, and the

performance standards are the quality of assessment practices that will be used to monitor teacher AfL development. However, in present practice, there are no clear criteria and standards to evaluate teacher AfL literacy, except for the generic criteria and standards embedded in some teacher professional standards. Specific criteria and standards are critical for teacher AfL literacy as they can give direction to the assessment literacy program with regards to the training design, the learning of the teachers, and how their learning is to be assessed and their progress be monitored. Torrie and Van Buren (2008) stress the importance of knowing the current level of individual teachers' assessment literacy and recommend the development of rubrics to measure individual teacher's level of ability in relation to the assessment standards. Thus, before any AfL assessment literacy or AfL reform movement should be undertaken, the learning outcomes, criteria, and performance standards should be clearly established.

3. Feedback should be used extensively to provide useful information to individual teachers. As proposed by Davison (2013), one characteristic of a successful assessment reform is "constructive qualitative feedback which helps stakeholders (these include teachers) to recognize the next steps needed for reform and how to take them" (p. 265). The effectiveness of feedback on teachers' performance is supported by studies such as Jensen and Reichl (2011) in Australia, who found that if teachers receive feedback related to their performance, their effectiveness could rise by up to 30 per cent. Just as in the classroom, teacher AfL literacy learning needs to utilize feedback in order to discuss with teachers their performance in relation to the learning outcomes, criteria, and standards. Essentially, at any stage of assessment literacy, clear criteria and standards are needed. Feedback should point out the strengths and

weaknesses of individual teachers and enable them to set goals to move to the next level of performance.

4. Teacher A/L literacy programs should also develop the self and peer assessment capability of teachers (Davison, 2013). Teachers should be encouraged to regularly reflect on their practices to assess how effective they are and how well they are progressing in using A/L to improve their professional learning. Similar to student self and peer assessment, teachers need to have criteria and standards readily available to guide them in their reflective practice. However, as indicated earlier, there are no available criteria and standards that teachers can use to determine their level of competence. Self and peer assessment is only possible if there is an assessment tool that can be used to guide them to assess their performance. However, even if there were readily available criteria and standards, the process of self and peer assessment is not an automatic response because it requires sound critical and reflective thinking skills, which enable teachers to ‘gather, interpret, evaluate, and select information for the purpose of making informed choices’ (Bruning, Schraw, & Ronning, 1999, p. 201). The power of critical thinking is most beneficial when used by teachers to self-assess and improve their performance (Facione, 2011). The interplay of critical thinking, criteria and standards in conducting self- assessment is exemplified by Lipman (1995), who emphasised that critical thinking is ‘skilful, responsible thinking that facilitates good judgment because it relies upon criteria, is self-correcting and is sensitive to context’ (p. 38). This implies that successful judgment by teachers of their performance requires them to think about their abilities in relation to pre-established and acceptable criteria and standards.

Their thought processes is guided by these criteria, which enable them to reflect and set personalised goals to improve their practices.

5. AfL literacy should provide teachers with continuing opportunities to engage in further education. Contrary to the common practices of most formal training, professional development, AfL programs are most effective if embedded in teachers' everyday classroom activities. The formal training should provide not only theoretical and conceptual understanding, but also establish a community of learners sharing, reflecting, learning and modifying their assessment practices in a regular basis. Black, Harrison, Lee, Marshall, and Wiliam (2003) emphasise that professional development for teachers to adopt and adapt AfL should be framed in such a way that teachers will be fully engaged in a range of activities where they are treated as learners themselves rather than simply telling them how to use assessment and assessment information. In other words, teachers should undergo authentic learning that fosters inquiry, experimentation, collaboration, and reflection (James et al., 2007). One important skill for teachers to develop in AfL literacy is their reflective skills to think about new ideas, try out these innovations, and monitor and evaluate the results. This will create a community of teachers who are learning together, improving and innovating their practices, and sharing expertise.

The importance of continuous sharing and reflecting on their practices by teachers and their peers goes far beyond acquiring explicit knowledge. The community of learners they create gives them opportunities to share and acquire tacit knowledge, which cannot be transferred so easily through formal training and conferences. Superficially, it may seem easy to create such a learning environment, but there are a number of critical factors that influence its

effectiveness. Amongst these are trust, early involvement, due diligence (Foos, Schum, & Rothenberg, 2006), personal interest and shared values (Dhanaraj, Lyles, Steensma, & Tihanyi, 2001), intrinsic motivation (Osterloh & Frey, 2000) and fit to the organization (Ambrosini & Billsberry, 2007). It is, therefore, imperative that systems identify and adopt the philosophical changes required for effective assessment AfL literacy. Systemic changes should foster trust, develop and communicate shared values, support intrinsic motivation, and find ways for individuals to fit into the school system. The latter requires not only helping teachers to change their assessment practices, but developing personal attributes, which are necessary pre-requisites for AfL literacy.

6. Assessment literacy is not only necessary for teachers but for all other stakeholders, including administrators, students and parents (Davison, 2013). The linkage of assessment literacy to key responsibilities (Popham, 2009) defines its true nature. People with different stakes in education have different needs and so require levels of assessment literacy. Davison (2013), as the lead consultant in assessment reform in Hong Kong, Singapore, and Brunei, found that the most important factor contributing to failure of assessment reform is misconceptions amongst policy makers about what AfL really is. Due to lack of understanding of what it takes to implement assessment reform, policy makers may think that simply changing the assessment practices of teachers will make the assessment reform successful. This, however, is not the case because teachers are not autonomous, nor working in isolation. What is required is the establishment of strong AfL culture at all levels of stakeholders and across the system (Davison & Leung, 2009). Furthermore, Davison argues that issues in AfL implementation should be used to develop an assessment literacy program

for policy makers. At the highest level, the nature of AfL should be clearly understood so as to facilitate the legislation of some pre-requisites needed to institutionalise AfL implementation. Once policy makers have a clear understanding of AfL, then they can also institutionalise the support needed by teachers.

7. An effective AfL literacy program should recognise the diversity of teachers, who, just like students, have individual needs, diverse learning characteristics and different classroom contexts in which they operate. Hence, AfL literacy programs should use the concept of differentiated instruction and adopt various strategies that suit teachers' needs. Above all, program implementers should have a strong belief that all stakeholders can improve their assessment literacy (Davison, 2013).

In summary, the effectiveness of a teacher AfL literacy program starts with a clear understanding of the basic principles of AfL by stakeholders, re-engineering the educational culture and re-aligning educational practices to AfL principles to provide teachers with an environment that models AfL culture, providing the necessary support services to teachers, thus enabling teachers to actualize their learning. Similarly, understanding teacher AfL literacy requires looking at various ways to conceptualise this construct. Some authors have only considered this construct at the item level, whilst others have tried to establish a range of dimensions to offer a broader conceptualisation. To clarify these competing dimensions, in the next section, I present the various dimensions used and/or established by other authors and highlight their strengths and their weaknesses.

2.5. Theoretical and Conceptual Framework of my Study

2.5.1. Overview

Teacher assessment practices are widely labelled as identifying learning outcomes, choosing and using appropriate assessment methods, developing assessment tasks, establishing criteria and standards, ensuring the reliability and validity of assessment, communicating assessment results to students, and using results of assessment for teaching and learning (J. Chappuis, 2009; Popham, 2009; Stiggins, 1999a; Stiggins, Arter, Chappuis, & Chappuis, 2007). These assessment practices are often grouped into factors such as teacher competency in assessment design, administration, analysis and interpretation, use of assessment information to make decisions, communicating assessment results, ensuring appropriateness and ethical assessment practices (Mertler & Campbell (2005); Newfields (2006); and Fulcher (2012), but the summary of my literature review has led to six broad categories of teacher A/fL literacy. Figure 2.1 below summarises these six categories of teacher A/fL literacy.

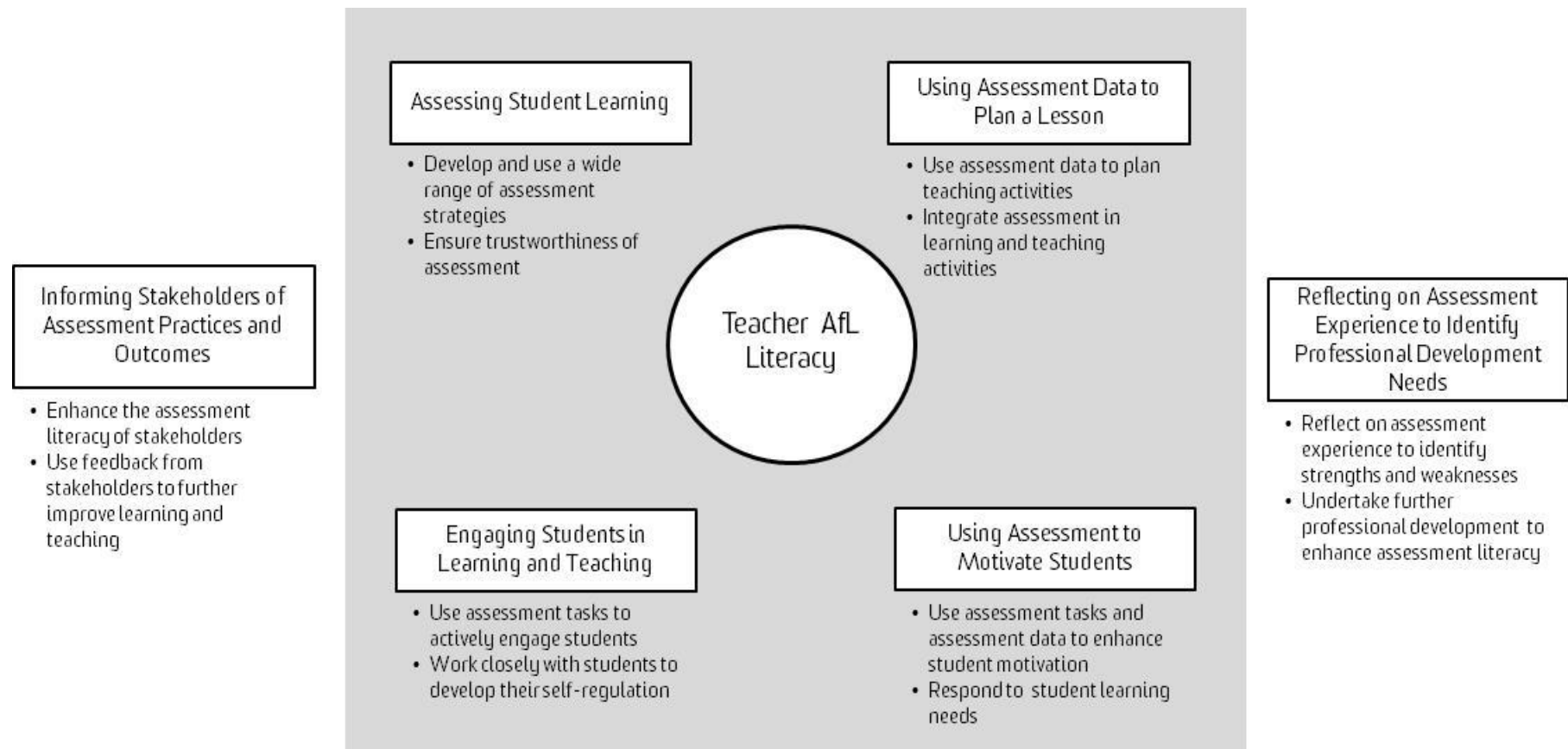


Figure 2.1. The theoretical framework for the study showing the six theoretical dimensions of teacher AfL literacy.

2.5.2. *Assessing Student Learning*

Without classroom assessment, it is hard for teachers to measure and monitor students' learning achievement and progress. Apart from knowing what individual students have learned, teachers need to gather information related to their learning characteristics and needs. This dimension covers the skills of teachers in developing a wide range of assessment tasks to assess student learning. This dimension is clearly described in the first two stages of the guidelines for teacher-based assessment developed by Davison (2008), that is, planning assessments and collecting information about student learning. Teachers need to ensure that assessment strategies are embedded in learning and teaching activities so that students will see assessment as part of their learning activities rather than an isolated activity that occurs at the end of the teaching period. This highlights the principles of AfL that assessment "should be part of effective planning" (ARG, 2000, p. 2).

In addition, this dimension requires teacher judgment in selecting appropriate assessment methods, for which the typology developed by Davison (2007) of teachers' AfL practices is useful. A range of assessment strategies including in-class contingent formative assessment-while-teaching, more planned integrated formative assessment, more formal assessment modelled on summative assessment but used for formative purposes, and prescribed summative assessments are suggested. The more effective assessment methods are determined by looking at how best the learning outcomes can be measured. For example, the use of formative assessment (interview, questioning, checklist) can provide teachers with evidence of learning and achievement (Stiggins & Chappuis, 2005) and to determine what needs to be done to improve learning and teaching (Black et al., 2003). Also, teachers can use it to engage students in the learning

and teaching activities and assessment (Black & Wiliam, 2009a). In a similar manner, teachers can use summative assessment both to determine the final learning outcomes and to inform future teaching and learning activities (Biggs, 1998). To effectively select and use appropriate assessment strategies, issues of trustworthiness (low reliability of formative assessment and limited scope of summative assessment) are critical to the quality of information gathered. Hence, across the range of assessment activities, teachers need to ensure the reliability and validity of assessment which include the application of measurement principles even in formative assessment (Bennett, 2011), the quality (Herman, Osmundson, Ayala, Schneider, & Timms, 2006), sufficiency (Smith, 2003), and various sources (Moss, 2003) of assessment information and their social consequences (McNamara & Roever, 2006).

In general, this dimension requires teachers' judgment in developing and using assessment strategies that help them and their students to monitor and evaluate student learning (Struyven, Dochy, Janssens, Schlthou, & Gielen, 2006). Careful planning by teachers is needed to ensure the alignment of assessment approach with learning outcomes to enhance the trustworthiness of assessment and help students to see the link between learning and assessment (Guskey, 2003). Using appropriate assessment strategies and using clear criteria and standards will increase teacher clarity, which Hattie (2008) has been shown to have a high effect on student achievement. In addition, McMillan (2000) argues that communicating the overall approach to assessment before lessons begin ensures fairness and ethical practice in assessment.

2.5.3. Using Assessment Information to Plan a Lesson

This dimension covers teachers' roles in reflecting on assessment information to inform one's teaching. It highlights the pedagogical use of assessment data. Student assessment

data are important input for teachers' instructional design and planning (Moss, 2003). These data include students' interests (Hébert, 1993), prior knowledge and experience (Kalyuga, Chandler, & Sweller, 2001; Weinert & Helmke, 1998), learning styles (Dunn, Griggs, Olson, Beasley, & Gorman, 1995), resources and support needed. Certainly, planning of classroom activities requires a high level of teacher understanding of student learning characteristics, levels and maturity of student learning. As asserted by Hattie (2003), expert teachers take into account the unique context of each classroom and all other factors that affect student learning and use them in planning learning and teaching. As teachers understand more about the complexities of the classroom climate, they can set higher expectations for learning, proven to enhance achievement (Hattie, 2008).

This dimension is also underpinned by a differentiated teaching approach to fully meet the individual needs of students (Tomlinson & McTighe, 2006). As teachers reflect on the available assessment information and the desired learning outcomes, they can identify and develop learning and teaching activities whilst taking into account the available resources. In planning learning and teaching activities, teachers also need to consider how to properly embed assessment in their teaching. One potential approach is the use of assessment tasks as learning tasks. This approach was fully demonstrated and supported by the studies of Keppel and Carless (2006) and Marshall and Drummond (2006). The engagement of students in various assessment tasks as learning tasks, which requires them to utilise a range of cognitive processes beyond recalling and recognising information will increase knowledge building and skills development (Mayer, 2002).

Furthermore, this dimension highlights the ability of the teacher to integrate assessment into teaching. Assessment information, if efficiently used, can help teachers focus their effort in planning lessons on the attainment of learning outcomes by

establishing the links and interplay of assessment, teaching and learning (Biggs & Tang, 2007; T. Glasson, 2009). Guskey (2003) argues that this ability of the teacher has many benefits both for teachers to help them effectively support students in their learning and for students to understand the whole process of learning and assessment. McMillan (2000) confirms that the ability to integrate assessment and instruction develops teacher capacity to decide what focus and level of teaching is appropriate for specific group of students.

2.5.4. Engaging Students in Learning and Teaching Processes

This dimension requires teachers to use assessment tasks and information to enhance students' motivation. This dimension considers the emotional investment of students in the classroom, particularly in assessment (Dweck, 2007) because students exhibit a range of emotional responses to assessment activities. According Pekrun et al. (2002), the reactions of students to assessment activities affect their psychological and physical health, which consequently affects their ability to self-regulate including their motivation, use of learning strategies, cognitive functions, and ultimately their academic achievement.

Students are considered as the most important players in their learning. Thus, a teacher needs to actively involve each student in assessment, as well as learning and teaching (ARG, 2002). It has been argued that that to fully realise the potential of students, a teacher needs to shift from merely facilitating to activating learning (Hattie, 2008) and treating students as owners of their own learning (William & Thomson, 2008), practices which can be developed by working closely with students and providing them opportunities to understand and assume some ownership of learning outcomes (Black & William, 1998; Gray & Tall, 1994; Hounsell, McCune, Hounsell, &

Litjens, 2008; Sadler, 1989a) and success criteria (Fredericksen & Collins, 1989; Gray & Tall, 1994; D.J. Nicol & Macfarlane-Dick, 2006; O'Donovan, Price, & Rust, 2008), engaging them in self and peer assessment (Kaufmann & Schunn, 2011; Klenowski, 1995; Lew, Alwis, & Schmidt, 2010; McMahon, 2010;), and providing them with timely and effective feedback related to their strengths and suggestions on how to further improve their learning (Bird & Yucel, 2015; Carless, Salter, Yang, & Lam, 2011; Walker, 2015) .

Developing student self and peer assessment capabilities requires various processes to effectively use this assessment approach. Initially, students need to recognise what is required to complete the tasks. Nicol and Macfarlane-Dick (2006) argue that the assessment task set by teachers should activate cognition, motivation and positive attitudes of students towards learning, which is the start of self-regulation. However, this is not an automatic response and it needs support from teachers. One of the things that teachers need to ensure is student understanding of the assessment.

Transferring tacit knowledge has been valued significantly in learning and teaching. Assessment knowledge, including learning outcomes, assessment tasks, criteria and standards are comprised of both explicit and tacit knowledge. Explicit knowledge can easily be transferred and understood by students. The “real” conception by teachers of assessment task, criteria and standards is however hard to transfer to students by discussion alone. Teachers should make sure that students have the same understanding of they do of the individual standards provided in the criteria, even if it is couched in simpler language.

To ensure this, teachers should find ways for tacit knowledge to be transferred. Rust, Price, and O'Donovan (2003) found that this has a twofold effect to learning: improving learning over time and enhancing transferability within a similar context.

Since tacit knowledge is experience-based (Baumard, 1999; Nonaka, 1991), teachers should provide an avenue for students to observe, imitate, and dialogue with them and the time to practice. Bamber (2015) found students who were engaged in an exercise involving marking exemplar materials before marking their own work developed a high level of marking accuracy and calibration of standards.

Tacit knowledge transfer is important in achieving learning outcomes. The mismatch between students' and teachers' conceptions about learning outcomes, criteria, and standards can lead to poor performance (Hounsell, 1997) and negatively affects how students value feedback from teachers (Norton, 1990). The connection between feedback and how students close their learning gap depends on the shared understanding among teachers and students of learning outcomes, criteria, and standards.

Developing critical and reflective thinking in students is a pre-requisite for students' engagement in self and peer assessment. The self and peer assessment capabilities of students require complex processes to enable them to pass on valid judgment and provide meaningful feedback for themselves and for their peers. The way students evaluate theirs and their peers' work requires a thorough understanding of the learning outcomes, criteria, and standards as their judgment and feedback are based on these. Structuring opportunities for the development of critical and reflective thinking helps ensure that students can undertake self and peer assessment effectively (Gamire & Pearson, 2006). The overall aim of empowering students to regularly engage in self and peer assessment is to develop learners who can self-monitor their learning and can provide significant insights for themselves and to their peers to further improve their performance.

However, setting up a classroom environment for self and peer assessment requires more than student engagement in critical and reflective thinking. The use of self and peer assessment has been widely criticised due to its low validity and reliability (Swanson, Case, & Van der Vleuten, 1991). To address this issue, there is a need for teachers to create a learning environment where apart from understanding the learning outcome, criteria, and standards, teachers can moderate the results of self and peer assessment, establish a review procedure where peer assessment can be discussed in cases where doubts arise and incorporate teacher feedback to confirm peer evaluation (Magin & Helmore, 2001). The teacher feedback is the expert voice that validates peer feedback, thus improving student capacity for peer feedback. The effectiveness of the design of learning environment assists students in giving valid and reliable comments about their peers' work (Luca & McLaoughlin, 2002).

A range of activities have been shown to increase achievement (McDonald & Boud, 2003) by helping students understand success criteria (Boud, Cohen, & Sampson, 1999; Gibbs, 1999) and by using the process to correct mistakes in learning (Taras, 2003), using timely and effective feedback to establish partnerships (Hattie & Timperley, 2007) and dialogues (Angelo & Cross, 1993) with students and to increase their confidence and self-trust in conducting (Nicol & Macfarlane-Dick, 2006) and adjusting (Luca & McLaoughlin, 2002) the results of self and peer assessment (Hanrahan & Isaacs, 2001; Patri, 2002) to make it more accurate. The key to these processes is teacher clarity, which is defined by Hattie (2008) as the ability of teachers to clearly communicate the learning outcomes and the expected level of performance.

2.5.5. *Using Assessment to Develop Student Motivation*

Enabling and sustaining student motivation in learning is a complex process that teachers need to effectively manage (Pintrich, 2003; Pintrich & Zusho, 2002). Both empirical evidence and theoretical argument support placing assessment as the centrepiece of learning and teaching, providing unprecedented potential to enhance student engagement, and hence improve learning outcomes (Wiliam, 2011).

Teacher assessment practices can either be positive or negative (Brookhart & Durkin, 2003). The improper use of assessment (i.e., labelling and stigmatising students, negative feedback, unclear learning outcomes, criteria and standards, and so on) has backwash effects that compromise student motivation and hence, reduce student learning (Biggs, 1996). In contrast, if the focus of assessment is to ensure learning gains, it has a powerful effect on student motivation, and consequently improves learning (Dweck, 1999, 2007).

Williams and Williams (2011) argue that apart from students and teachers themselves, there are three other key teacher-related factors that can affect student motivation. First, the content and methods of teaching need to be carefully selected to ensure that they are relevant to the interests of the students and address their learning styles and needs. Hence, the use of differentiated instruction is necessary. A second factor is the classroom environment, which teachers need to transform into a learning environment where learning processes are explicit and allow students to practice self-regulation (Schunk & Zimmerman, 1994) by giving them a high level of autonomy in the learning process. Finally, assessment can be used as a means of developing a high level of interpersonal relationship between students and teachers, which has been proven to promote motivation and achievement (Martin & Dowson, 2009).

In an innovative classroom, the student control over some aspects of assessment activities, including choice of assessment tasks, has been argued to enhance student engagement as it empowers students to shape their learning goals (Bevitt, 2015). Not only does control over assessment activities affect student motivation but also their choice of assessment tasks. A key concept to address this issue is the differentiation of assessment. This is underpinned by the philosophy that students who find tasks interesting and connected to real-life situations are obviously likely to engage with those tasks (William, 2011). Another goal of using assessment to sustain motivation is to develop self-regulation to ensure that students are actively monitoring (Sadler, 1989a) and controlling their thinking processes, learning behaviour and motivation per se (Pintrich & Zusho, 2002).

2.5.6. Using Student Data to Identify Professional Needs

Apart from teacher self-assessment, student assessment data, including student work, , are rich sources of information for identifying teacher professional needs (Darling-Hammond, 1998). Using this information clearly identifies the needed expertise to respond to the learning needs of the students, and eventually develop teacher adaptive expertise, which enables teachers to become flexible in responding to student needs (Timperley, 2011).

The ability of the teachers to reflect on their needs based on student assessment information and their assessment experience is highlighted by Timperley et al. (2008) in their teacher inquiry and knowledge-building cycle. They stress that after collecting information about the current level of students' learning and the knowledge and skills needed by students to achieve their learning outcomes, teachers need to identify what knowledge and skills they need to have to assist students in their learning. This on-

going and sustained approach to professional learning is amongst the most valued strategy by teachers (Koh, 2001). The role of professional development in building assessment literacy has also been recognised by various experts. For example, Popham (2009) argues that the necessity of continuing the delivery of assessment literacy programs is fundamental to improving teaching practices. Similarly, the report of Kahl, Hofman and Bryant (2013) for the Council of Accreditation of Educator Preparation recommends the mastery and continuous building of assessment literacy.

The key to using these assessment data is for teachers to engage in reflective practice where they become learners from all the outcomes of the classroom processes. Black et al. (2003) emphasise that professional development for teachers to adopt and adapt AfL should be framed in a way that teachers will be fully engaged in a range of activities where they treat themselves as learners rather than telling them how to use the principles of AfL. The outcome of self-reflection is the consciousness to think of new innovations, try these innovations and see the results. In other words, teachers should undergo authentic learning that fosters inquiry, experimentation, collaboration, and reflection (James et al., 2007).

2.5.7. Informing Stakeholders of Assessment Practices

This dimension covers the ability of the teacher to work closely with stakeholders to improve their assessment literacy to ensure that their expectations and beliefs are consistent with the teacher's AfL practices. This is described by Davison (2013) as a critical component for AfL reform within and across school system. In every assessment reform, there should be a shared understanding and a common belief system amongst all stakeholders from the school level, including students, parents and schools heads, and across all levels of the education bureaucracy. Indeed, the existence of this factor

reinforces the work of Taylor (2009) who emphasises that part of a teacher's assessment literacy is to develop the assessment literacy of other stakeholders. The continuing dialogue and interactions of teachers with school leaders, parents and the community related to student learning and assessment practices in general, provide avenues for building shared understanding amongst them.

Ensuring that stakeholders have understood the rationale and theoretical support for teacher AfL practices ensures that external pressures from parents and the community support, rather than undermine, teacher AfL practices. Also, as part of this dimension, teachers need to consider that each group of stakeholders has different AfL literacy needs because they have different assessment knowledge, assumptions and responsibilities (Popham, 2008). Whilst preparing reports for various stakeholders, teachers need to think about the trustworthiness of the report. All stakeholders need to understand quality of information provided and how the report addresses their information needs in order to build their trust in teachers (Knight, 2002).

These six theoretical dimensions of teacher AfL literacy underpin the development of the tool in the present study. These dimensions are tested using an empirical approach to gather substantial evidence to support or refute their existence.

2.9. Summary

This chapter has defined the construct of teacher AfL literacy being investigated in this study. I have presented a literature review of the empirical and theoretical support for teacher assessment practices, the development of the concept of AfL construct, and the various views related to the content and approaches to teacher assessment literacy. In addition, I have highlighted the dimensions related to the conceptualisation of teacher AfL literacy and the weaknesses of approaches used in their development. This chapter

is the basis for my arguments in the next chapter, particularly for the need to further explore and evaluate the dimensions of teacher AfL literacy using a more robust methodology. Finally, this chapter has concluded with a discussion of the theoretical framework for teacher AfL literacy used in this study.

CHAPTER 3 – THE EVALUATION OF TEACHER AfL LITERACY: THE USE OF FRAMEWORKS AND TOOLS

3.1. Introduction

The various conceptualisations of assessment for learning (AfL) and teacher AfL literacy described in Chapter 2 have resulted in the development of different instruments to describe and evaluate teacher assessment literacy. In this chapter, I review the literature with an emphasis on describing and evaluating existing teacher assessment literacy frameworks, models and tools used to guide and evaluate teacher AfL literacy. I have also included a review of the best practices in measurement tool development, which then informed the methodology of my study.

The aim of this chapter is to justify the need for the development of a new tool to measure teacher AfL literacy using a more robust methodology to support its psychometric requirements, and to establish the dimensions of teacher assessment literacy. To achieve this aim, I critically analyse the weaknesses of the existing frameworks, models and tools and present an argument for the need to develop a new tool that is aligned to the principles of AfL. Then I present a review of current practices and approaches in scale development, including theoretical approaches, the use of the Rasch model in item level analysis, and the use of exploratory structural equation modelling (ESEM) in addition to the conventional use of exploratory and confirmatory factor analyses in establishing the dimensions of the construct. Finally, I discuss the need to use latent profile analysis (LPA) to discover groups of teacher who have the same AfL profiles, as this grouping of teachers have potential implications for the development of teacher AfL literacy programs.

3.2. Existing Frameworks for Mapping Teacher Assessment Practices

3.2.1. Overview

There are a number of frameworks or models for teacher assessment literacy that have been developed and/or recommended for use by several authors, either for mainstream education or specific fields. These models can be categorically identified based on their uses and application. Some are descriptive and are useful guides for teachers' classroom practices (e.g., Cowie & Bell, 1999; Harlen's time dimension, 2007; Brookhart et al., 2006), others are prescriptive and more suitable for the evaluation of teacher assessment literacy for professional development (Harlen's component dimensions, 2007; and one is a specific framework for researching teacher assessment practices (Hill & McNamara, 2012).

3.2.2. Descriptive Frameworks and Models

Descriptive frameworks and models describe the assessment knowledge and skills that teachers need but describing different kinds of practices. For example, the model presented by Cowie and Bell (1999), which was generated from their 2-year study of science classrooms involving 10 teachers of students in Years 7-10, revealed from their 128 classroom observations throughout the research period that teachers were using two types of formative assessment. Teachers engage in either planned or interactive formative assessment. The former happens when teachers use pre-planned assessment tasks to elicit students' performances. Teachers then interpret these performances and act according to the purpose of the assessment. On the other hand, interactive formative assessment is what Davison (2007) calls 'in-class contingent formative assessment', which happens when teachers respond to assessment opportunities during learning interactions. Although the model presented by Cowie and

Bell (1999) is useful to account for two different types of formative assessment, this model is not sufficient to describe all assessment activities that have the potential to further support student learning.

The typology of teacher assessment practices developed by Davison (2007) adds to Cowie and Bell's model by incorporating two types of more formal assessments that teachers might use for classroom assessment, that is, more formal mock or trial assessment modelled on summative assessments but used for formative purposes that are focused to determining the gap between students' current performance and the desired performance; and prescribed summative assessments, but where the results are also used formatively to guide future teaching and learning focused mainly on determining what the students have achieved in relation to the standards and what the students need to do in their future learning. In Davison's model, all four assessment types are placed on a continuum rather than taking formative and summative assessment as a dichotomy of practice.

Another model, which is descriptive by nature, is the time dimension framework of Harlen (2007). This model clearly shows the interaction between formative and summative assessments, hence supporting the inclusion of summative assessment under the umbrella of AfL (as argued in Chapter 2.3). In this model, Harlen argues that in every episode of teaching and learning activity across the entire duration of the course, a cycle of formative assessment occurs. The range of evidence collected is used to help students to achieve the specific lesson goals of the next episode of learning. In addition, this collected evidence can be used to report the overall achievement of the students to satisfy the summative report. This model looks at the longitudinal relationship between formative and summative assessments and puts summative assessment at the end. This shows only the linear relationship between formative and summative assessments where

summative assessment is at the terminal end of the spectrum of assessment activities. However, in reality, summative assessment can be administered at any period of learning and teaching. It has been widely argued that the results gathered in summative assessment can also be used for formative purposes (Davison, 2007; Taras, 2009). In fact, according to Biggs (1999) formative assessment starts with summative assessment where feedback given to students is based on the assessment conducted by teacher, which is summative by nature.

3.2.3. Prescriptive Frameworks and Models

Prescriptive frameworks and models provide more detail about the “what and how” of teacher assessment practices, which can be used by teachers to reflect on their practices and used to provide dialogic feedback between school leaders and teachers. For example, the model of Wiliam & Thompson (2008), which captures both Cowie and Bell and Harlen models described above, use the three key purposes of instructional processes embedded in the definition of AfL by the ARG (2002).

In this model, it shows the roles of teacher, peer and learners (for each of the three key areas) in ensuring the achievement of learning outcomes. The model provides clarity of, and ways to operationalise, the definition of AfL. However, just like the previous models described, it does not account for other factors that affect assessment and student learning. This is similar to the framework developed by Harlen (2007), which specifies seven components, and their corresponding key concepts, and describes the assessment system by which teachers operate. Although this model covers a wide range of assessment knowledge and skills, the interactions of various assessment skills are not established. It appears that teachers’ individual assessment skills operate in isolation.

Other models view teacher assessment practices within the broader context in which assessment occurs. For example, the model proposed by Brookhart, Walsh, and Zientarski (2006), founded on Brookhart's (1997) earlier model of classroom assessment, is fairly comprehensive. It retains the original components of the older model such as the causal relationship amongst teacher assessment practices, student effort, and student learning, but adds the function of student motivation (the desire to do something) as directly affecting the level of students' effort exerted to complete the task. Another feature of this model, as opposed to the earlier version, is the expansion of teacher assessment practices, to what it refers to as the unique 'classroom assessment environment'. Hence, every assessment experience is different from the rest because of the concept of context-dependent assessment experiences. This framework of classroom assessment explains the multifaceted nature of student learning, where several interrelated factors drive student to either engage or withdraw from their learning. These factors could either result from assessment activities or drive the assessment activities per se.

The more comprehensive prescriptive model is the one developed by Davison (2008), adopted in Singapore schools. The model demonstrates a step-by-step process from planning assessment, collecting information, making professional judgement to providing appropriate feedback or advice to help students achieve the learning outcomes. For each stage in Davison's model, there are suggested assessment strategies or activities. This model can be used either to guide teachers' assessment practices or as a tool for professional development. However, the ability of teachers to reflect on their assessment and teaching experiences and identify their professional needs both in terms of content and assessment skills is not included in this model, even though it is widely acknowledged that is the most effective method to increase levels of teacher assessment

literacy. In turn, the increase in teacher assessment literacy brought about by engagement in professional development, either formal or informal, contributes greatly to improving student learning.

3.2.4. A Research-Based Model

One specific model of AfL literacy was intentionally developed to guide researching classroom assessment. The framework proposed by Hill and McNamara (2012), can be used to evaluate existing assessment practices. It covers the various aspects of assessment practices that include planning and actual conduct of assessment, the assessment constructs (enterprises, qualities and standards) available, the theoretical and epistemological bases for teacher assessment practices, and students' beliefs and understanding of assessment processes. The questions provided in the framework are useful guides for eliciting responses to gather detailed information about assessment practices in the classroom. Apart from its use in researching assessment practices, the framework can be used as a checklist for teachers to reflect on their assessment practices. However, the framework lacks the other factors that are described by Brookhart et al. (2006) that affect student engagement in their learning. The model of Hill and McNamara, if used to guide practices and for professional development may make teachers think that assessment can be isolated from other factors affecting student learning.

Overall, although these frameworks and models have limitations in terms of their functions and their scope, they are useful in understanding, describing, or guiding teacher assessment practices. The use of these frameworks is evident in the inclusion of teacher assessment literacy in most teacher professional standards, another important source of information about frameworks for evaluating teacher assessment literacy.

3.3. Teacher Assessment Literacy in Teacher Professional Standards

3.3.1. Overview

Teacher assessment literacy is an integral component of teacher professional standards used by several countries. Some of these standards take the form of dimensions but others used individual teacher assessment skills.

3.3.2. Teacher Assessment Literacy in Teacher Professional Standards

In Table 3.1, some of the existing assessment standards embedded in teacher professional standards are presented. It can be seen that none of the standards is comprehensive enough to include all the important teacher assessment competencies. Some of the assessment competencies identified by Stiggins (1999) are missing in these teacher professional standards. For example, Stiggins explicitly stated that teachers need to clarify achievement expectations with students, but none of the standards in Table 3.1 included this competency. The same is true with the competency related to using assessment as an instructional intervention.

The assessment competencies described in these standards are too generic and fail to operationalize the principles of AfL. For example, in the AITSL and the Philippines teacher standards, the link between using assessment and assessment information to inform learning and teaching activities is not explicitly included, whereas in the UK standards, this is included. Similarly, in the AITSL and the Philippines standards, reporting of student achievement to various stakeholders is stated but is absent in the UK standards.

Brookhart is one of a number of leading researchers to call for a new set of teacher assessment literacy standards. Brookhart (2011) argues that current teacher

assessment standards in the USA, particularly the Standards for Teacher Competence in the Educational Assessment of Students (1990), do not account for emerging conceptualisations of effective teacher assessment practices based on how students can be effectively supported in their learning. She presented a new compendium of teacher assessment knowledge and skills presented below:

1. Teachers should understand learning in the content area they teach.
2. Teachers should be able to articulate clear learning intentions that are congruent with both the content and depth of thinking implied by standards and curriculum goals, in such a way that they are attainable and assessable.
3. Teachers should have a repertoire of strategies for communicating to students what achievement of a learning intention looks like.
4. Teachers should understand the purposes and uses of the range of available assessment options and be skilled in using them.
5. Teachers should have the skills to analyse classroom questions, test items and performance assessment tasks to ascertain the specific knowledge and thinking skills required for students to do them.
6. Teachers should have the skills to provide effective, useful feedback on student work.
7. Teachers should be able to construct scoring schemes that quantify student performance on classroom assessments into useful information for decisions about students, classrooms, schools, and districts. These decisions should lead to improved student learning, growth, or development.
8. Teachers should be able to administer external assessments and interpret their results for decisions about students, classrooms, schools, and districts.
9. Teachers should be able to articulate their interpretations of assessment results and their reasoning about the educational decisions based on assessment results to the educational populations they serve (student and his/her family, class, school, community) .
10. Teachers should be able to help students use assessment information to make sound educational decisions.
11. Teachers should understand and carry out their legal and ethical responsibilities in assessment as they conduct their work (p. 7).

According to Brookhart, these standards were formulated in reference to the existing teacher assessment practices but “with reference to formative assessment and standards-based reform” (p. 6). The 11 standards described captured the philosophy of using assessment to effectively support students in their learning but there was no empirical evidence presented to support the validity of these standards.

Table 3.1

Teacher Assessment Literacy used in some Teacher Professional Standards

Australian Professional Standards for Teachers (AITSL, 2011)	UK Teachers' Standards (Department of Education, 2011)	Philippines National Competency Based Teacher Standards (Department of Education, 2004)	Teaching Standards in the Asia-Pacific Region (DEEWR, 2008)
• Assess student learning • Make consistent and comparable judgment • Interpret student data • Provide feedback to students on their learning • Report on student achievement	• Know and understand how to assess the relevant subject and curriculum areas, including statutory assessment requirements • Make use of formative and summative assessment to secure pupils' progress • Use relevant data to monitor progress, set targets, and plan subsequent lessons • Give pupils regular feedback, both orally and through accurate marking, and encourage pupils to respond to the feedback.	• Develops and uses a variety of appropriate assessment strategies to monitor and evaluate learning • Communicates promptly and clearly to learners, parents, and superiors about the progress of learners • Monitors regularly and provides feedback on learners' understanding of content	• Demonstrate appropriate knowledge of learning activities, programs and assessment • Engages students in learning • Provides feedback to students and assesses learning • Demonstrates flexibility and responsiveness • Reflect on teaching with view to improvement

Some of these teacher standards explicitly describe the levels of teacher performance for each dimension (i.e., AITSL uses graduate, proficient, highly accomplished, and lead teacher; Philippines uses three levels of performance) while others just make generic descriptions for each level. This is important in described teacher assessment literacy so that clear progression or development in assessment knowledge and skills can be captured.

3.3.3. Career Stages of Teacher Development

It has been argued that teachers, through their experiences, move through a series of stages in their professional career. There are several models used to describe these career stages and most of these are described as developmental stages across a continuum of teachers' levels of ability (see Table 3.2). Although the stages described are not specific for teacher A/L development stages, they can still be used in this particular context.

Comparison of these career stages shows some interesting insights. The stages described by Katz, (1972), Burden (1990) and the ones used by AITSL use teacher qualifications and levels of ability to characterise teacher development along the continuum of practice, whilst Fuller and Brown (1975) used the development of teacher identity in the classroom. Also, they characterised the highest career stage as when teachers have acquired the necessary skills to respond appropriately to the needs of students and have become confident in their profession, whilst the AITSL described a higher level of teacher career stage when they take an active leadership role in their respective schools. At this stage, teachers do not only possess a high level of ability but they also initiate and provide expert assistance and training to their colleagues.

Table 3.2

Teacher Career Stages in the Existing Literature

Theory	Stages
Katz's Theory of Teacher Development (Katz, 1972)	1. Survival – early stage where teachers feel inadequate and unprepared 2. Consolidation – teachers begin to apply what they have learned from their early experiences and develop effective routines 3. Renewal – teachers try new pedagogical approaches and materials. 4. Maturity – teachers has acquired the necessary teaching skills and confidence in their profession
Fuller and Brown's Theory of Teacher Development (Fuller & Brown, 1975)	1. Pre-teaching stage – teachers continue to identify themselves with students 2. Concern about Survival – teachers become concern of their perceived-inadequacies of their own ability to teach 3. Limitations of the Teaching Context – teachers are concerned about the demand of teaching. 4. Concern with Pupil Needs – teachers become aware of the learning needs of the students and develop approaches to respond appropriately.
Burden's Theory of Teacher Development (Burden, 1990)	1. Survival Stage – first year of teaching where teachers have limited ability 2. Adjustment Stage – 3rd or 4th year of teaching where teachers greatly improved in their practices 3. Mature Stage - starting from the 5th year of teaching and beyond, teachers have high level of confidence in teaching and have acquired higher level of teaching ability.
National Professional Standards for Teachers (AITSL, 2011)	1. Graduate teachers – completed the qualification requirements 2. Proficient teachers – demonstrate achievement of the specified standards 3. Highly accomplished teachers – engage in independent and collaborative work to improve practice 4. Lead teachers – innovative and provide leadership to school community to advance student learning.

Although the career stages described above are named differently, there is one common aspect, that is, the movement of teachers across levels is associated with the number of years of teaching experience. These stages can be seen as a continuum of expertise that is dependent on years of teaching, where teachers move from the first stage in their first year of teaching and move to the succeeding stages in later years. The association of these stages with years of teaching could be seen as a simplistic articulation of teachers' acquisition of skills. Instead of associating teachers' development with how fast they can increase their levels of practices through professional engagement, it is rather described using the years of teaching engagement.

Despite the weakness described, the stages used provide a useful way of thinking about teaching expertise and implementing teachers' professional development. The professional growth of teachers can be fully supported if their current level of expertise is used as a starting point for professional development. These career stages are evident in some assessment tools that use various levels of teacher skills to evaluate the skills development of teacher.

3.4. Existing Teacher Assessment Literacy Evaluation Tools

3.4.1. Overview

There are a few teacher assessment literacy evaluation tools that have been widely used to measure teacher assessment knowledge and skills. In this subsection, I critique these tools and identify various issues associated with their development and use.

3.4.2. Competing Teacher Assessment Literacy Evaluation Tools

The earliest version of a teacher assessment literacy evaluation tool discussed here is the 3-component model of teacher assessment literacy developed by McMillan (2001). The

tool measures the factors used by teachers in determining grades, types of assessment employed, and the cognitive levels assessed by teachers. However, the individual items in the tool have a strong inclination towards measurement or psychometric concepts, teacher assessment practices heavily criticised due to their inability to significantly improve student teaching (Nichols, Meyers, & Berling, 2009).

The most prominent and widely used tool is the Assessment Literacy Inventory (ALI), developed by Mertler and Campbell (2005) through two-stage pilot testing program to establish the psychometric properties of the instrument. The ALI is a scenario based multiple-choice exam with five scenarios, and each has a corresponding set of seven questions that directly link into the seven standards for teacher competence in the educational assessment of students. The properties of ALI, as reported by the researchers, meet the psychometric requirements for reliability, ($r_{KR20} = .74$) and item difficulty with the 35 items representing low to highly difficult with indexes range from .212 to .992. However, in terms of discrimination, even though the mean discrimination value is .313, the item with the lowest discrimination index has a value of .014, which is below the .20 cut off value. This item, along with other items with low discrimination indexes were retained by the researchers, which raises the issue of the usefulness of these items as they cannot discriminate low from high performing teachers. In other words, if the results of this tool are to be used for professional development, the scores of teachers might not be a true indicator of their current perceived level of ability as some of the items do not have the ability to distinguish low from high performers.

There are two evaluation tools developed for specifically for language teachers. One of these was developed by Newfields (2006), called the 'Assessment Literacy for High School Foreign Language Teachers Inventory'. The strength of this tool lies on its inclusion of various item types including matching exercises, multiple choice, open

response, and short completion tests with a total of 70 items in four content areas of assessment identified by Newfields. However, the development of this tool was based purely on theories and expert validation. The judgment of validation group as to include or exclude the item was used to finalise the items comprising the tool. Although a pilot test was done, no available information is given as to the psychometric properties of the tool. The other tool was developed by Fulcher (2012), described later.

Two other tools were developed with the use of factor analysis. The first one was developed by DeLuca and Klinger (2010) using three domains related to teacher assessment practices, understanding of assessment theory and philosophy. It uses 57-fixed response items with a 5-point Likert Scale. The development process included pilot testing but used only five teachers as respondents. DeLuca and Klinger analysed the dimensionality of the domain with 288 teacher participants generating 14 constructs. One of the problems in the development of this tool is the use of principal component analysis instead of factor analysis, which I argued in Chapter 3.3 is an inappropriate statistical analysis for tool development.

The other tool was developed by Fulcher (2012), with the intent of exploring the assessment training needs of language teachers. It is a survey containing two sets of questions related to assessment and the rest eliciting teacher personal information. The first set of questions use constructed-response items to ask teachers about their assessment needs and resources for assessment literacy. The second set of questions is composed of 23 items in a Likert scale format with 5 response categories utilising unimportant, not very important, fairly important, important, and essential as descriptors. Analysis of data from 278 responses enabled the author to establish both psychometric and factor analytic properties of the instrument. However, the analyses used were very limited as the report at the item level includes only the Cronbach's

alpha, which is equal to .93. This value indicated high reliability, but it was not sufficient to account for all the psychometric requirements, even though it can be argued that the tool can provide a consistent measure of teacher assessment literacy over time. The tool failed to report item difficulty and item discrimination. At the factor level, the results were not conclusive, even though the sample size is reasonable (close to 300 as recommended by Field, 2009) because the analysis was limited only to exploratory factor analysis. A subsequent confirmatory factor analysis (CFA) or an exploratory structural equation modelling (ESEM) is needed to further verify the factor model extracted from exploratory factor analysis (EFA).

In general, there are major issues related to the operationalisation of the underlying construct and its defence, the methodology used in the development of these assessment literacy tools, and the proficiency of these assessment tools to measure the construct. Firstly, the scope of the earlier versions of the tools, which defines the parameter of teacher assessment literacy, is designed to measure general concepts of testing and assessment, knowledge of standardised testing and classroom assessment (see McMillan, 2001; Mertler & Campbell, 2005) whereas the succeeding tools contain a range of items measuring teachers' understanding of assessment terminologies, procedures, test interpretation and ethical practice (see Newfields, 2006; Deluca & Klinger 2010). These assessment tools have a very strong emphasis on the understanding and use of measurement principles, and as pointed out by Newfields (2006), as a consequence of his desire to include assessment abilities of teachers that can be easily measured, the test contains a large number of items about statistics. Most available assessment literacy tools are focused on psychometric and measurement principles, testing teachers' knowledge and application of knowledge rather than evaluating teachers' actual performance in the classroom. However, the strong emphasis

on the measurement capabilities of teachers violates the principles of AfL which holds as a central principle that assessment literacy should focus on assessment practices that have been proven to improve student learning (Brookhart, 2003; Nichols et al., 2009; Sadler, 2012). In cases like the above where assessment literacy tools privilege the use of psychometric principles in the classroom, it gives an implicit message to teachers that these practices should be given more importance than pedagogically-oriented classroom assessment practices.

A second major issue is that most of the tools were not developed using a robust methodology. Although two of the tools used principal component analysis, the process was not sufficient to provide strong empirical evidence as to the existence of the four dimensions of teacher assessment literacy.

A third issues is that the item formats of these existing tools provide very limited information about the actual perceived level of individual teachers' assessment literacy. The multiple-choice items provide only information about the knowledge and judgment of teachers about a particular case or scenario, whilst the Likert scale lacks the necessary descriptions of performance for each level of teacher performance. These item formats do not embrace the principles of AfL where assessing actual performance needs explicit criteria and descriptions of each level of performance.

The first issue discussed above is addressed in this study by clarifying the present conceptualisation of AfL and teacher AfL literacy discussed in Chapter 2.3 and Chapter 2.4 respectively. The two other issues will be addressed by using a more robust methodology in tool development and by incorporating descriptions for each level of performance.

3.4.3. The Need for a Robust Methodology in Tool Development

The inconsistencies in various studies in terms of establishing the dimensions of teacher assessment literacy can be attributed to two factors: competing definitions of the construct and the weaknesses of the methodology used. The first reason has already been addressed in Chapter 2.4 by providing an operational definition of the construct of teacher AfL literacy that is used in this study. The second reason is briefly discussed below.

Hailaya, Alagumalai, and Ben (2014) demonstrate why there is a need to review existing teacher assessment tools, and the need to use a robust methodology in developing any assessment tool. In their study, they revisited the assessment literacy inventory (ALI) developed by Mertler and Campbell (2005) and employed the Rasch model and confirmatory factor analysis (CFA) to establish item-level characteristics and the dimensions of the construct. Their results showed that even though the ALI works well at the item level, and with high separation reliability (0.99), there are some issues associated with it. These issues will be dealt with in turn.

First, there was one item that did not tap into the underlying construct. This item contributed error to the measurement of teacher assessment literacy, and hence the results of individual teachers' assessment do not reflect their "true" level of their assessment literacy.

Second, the results of CFA show more profound inaccuracies in the ALI, the fit indexes of the 7-factor model are low, which indicate a poor fitting model. Hence the dimensions used and described in this tool are not the true dimensions of teacher assessment literacy.

Third, the factor loadings of the latent variables to their corresponding factor are also relatively low ranging from 0.11 to 0.58 only. The highest variance explained is

only 33.64%, which is not an acceptable value to support the association of these variables with their corresponding factor. From these results, the items comprising each of the seven factors are not greatly contributing to the actual measurement of teacher assessment literacy. It can therefore be said that there are other assessment skills that can be used to give the best and appropriate measure of teacher assessment literacy.

Considering the results of Hailaya, Alagumalai, and Ben (2014), it can be argued that those teacher assessment literacy evaluation tools that were developed using a theoretical approach only are not trustworthy due to issues of reliability and validity and specific factor level characteristics. However, teacher assessment literacy tools were developed using empirical approaches also have issues related the appropriateness of the empirical methods used. This is further argued in Chapter 3.8.

3.4.4. The Need for Criteria and Standards for AfL Literacy

Apart from the need for a more robust methodology to provide stronger empirical evidence concerning the dimensionality of teacher AfL literacy, it was argued in the earlier section of Chapter 2 for the importance of using criteria and standards to describe teacher AfL performance.

The teacher assessment literacy evaluation tools described above (Fulcher, 2012; Mertler & Campbell, 2005; Newfields, 2006; Thomson, 2012) have demonstrated how to link the results to professional development programs. Although these assessment literacy tools have been extensively used, and reports have supported their effectiveness, the richness of information gathered by these assessment tools is insufficient to account for the actual performance of teachers. The insufficient quality of assessment information gathered consequently affects the quality of professional development programs in addressing the diverse AfL literacy needs of teachers.

Researchers and educators argue that an assessment literacy program aimed at supporting teachers to enhance their assessment knowledge and skills should begin with measuring their current level of performance and be followed by identifying their specific training and support needs. With the available assessment tools, the scales were developed using levels of extraversion, quality of the products/performance, magnitude of continuum or ordering of entities using different levels of measurement.

The generic descriptions common to all score categories or performance levels do not reflect the true competencies of teachers as each category lacks description as to what the individual teachers can actually do in relation to individual A/fL skills. Using descriptors like ‘often’, ‘always’ and the like, may be interpreted differently by different teachers. According to the review of Kahl, Hofman and Bryant (2013) of criteria used in evaluating teacher assessment literacy, they are mostly superficial and cannot provide sufficient information about the teachers’ mastery of assessment. They contend that there is a need to clarify what assessment skills constitute effective teacher assessment literacy.

To address the issues related to the weaknesses of the existing teacher assessment literacy tool, there is a need to develop a tool that utilises the principles of A/fL, particularly the use of rubrics where descriptions of various levels of performance are provided. The idea of developing teacher standards that account for the principles of A/fL was emphasised by Brookhart (2011) because there are no available teacher assessment standards that provide descriptions of what teachers can actually do in each level of assessment literacy development. These individual descriptions of performance are needed to guide teachers *to determine their current level of assessment literacy* and to identify where *they need to go and what level of assessment literacy they need to acquire* to respond to diverse student learning needs. The gap between teacher current

and desired level of AfL literacy will facilitate teachers, with the help and support of school leaders and the presence of enabling mechanisms, *to identify a professional development strategy that suits their needs to further improve their assessment literacy.*

The call to establish a set of standards for teachers' assessment practices was first expressed by Bailey and Brown in 1995 and was emphasised again by Inbar-Lourie in 2008. The use of rubrics in this scale has intended benefits in assessment literacy. Noting the benefits of using rubrics in improving the performance of the students, it is also hoped for that the application of rubrics in teacher assessment literacy will bring significant improvement. The benefits of rubrics as identified by Wolf and Stevens (2007) are reiterated below in the context of this research:

1. Rubrics make the learning targets clearer for teacher AfL literacy programs. The criteria and standards will provide a common understanding amongst teachers and school leaders of what constitutes teacher AfL literacy at various stages. The criteria and standards can be used to determine individual teachers' current level of assessment literacy and their literacy needs.
2. Rubrics allow the selection of appropriate instructional design and delivery. The key learning targets of teachers identified using the tool will assist school leaders and teachers to identify the best training approach to help teachers learn best.
3. Rubrics accurately track the development and progress of teachers. Using the tool will facilitate an accurate and fair on-going assessment of teachers' AfL literacy. The use of the criteria and standards to measure teachers' AfL literacy will give a true picture of their actual performance.
4. Rubrics facilitate teachers' engagement in self and peer assessment. Consistent with the principles of the construct being measured here, the use of rubrics will

guide teachers in their self-reflective practice and in their engagement with collegial peer review of teaching performance to provide feedback to their colleagues. Hafner and Hafner (2003) stress that when the assessment criteria are available, individuals are able to better critique their own performance.

5. Rubrics give equal access for teachers coming from different backgrounds and contexts. The use of rubrics makes the teacher A/L tasks more explicit and all teachers can better understand the required performance.

The teacher assessment tool designed this way will give a better understanding of the construct being measured. The quality criteria identify the indicators of teacher A/L literacy whilst the standards, arranged in order of difficulty, describe the performance required for each level. The progression of standards, which indicate item difficulty, allows the placement of individual teachers based on their ability level. Thus, teachers can be accurately placed at a particular stage of the continuum of A/L development.

The kind of assessment tool described above embraces the principles of a probabilistic model of competency-based assessment. In this assessment, the Rasch notion of underlying growth continua or latent traits specifies that any construct can be observed through people's performance of tasks (Griffin, 2007). The indicators of competence, as defined by criteria, can be arranged in increasing order of difficulty and task demands. This arrangement has an important implication in learning and teaching because it allows the development of literacy along the continuum of competence. This is consistent with the work of Glaser (1963) where performance and development are also described in terms of the nature and difficulty of the tasks. Glaser emphasised that performance is affected by the diversity of learners and the nature of the task. Thus,

performance should be interpreted using the processes used to complete the tasks. The description of the performance should be directly linked to the criteria, which are viewed as thresholds or difficulty point in the continuum of competence. Drawing together the works of Rasch and Glaser, and the zone of proximal development (ZPD) of Vygotsky, Griffin (2004) used the probabilistic model of competence to demonstrate how assessment information gathered using this approach can inform learning and teaching. Also, the competency-based assessment provides an appropriate framework to provide stakeholders with a more meaningful interpretation of what a person can actually do in the task rather than just providing a numerical score of performance.

3.5. Approaches in Tool Development

3.5.1. Overview

Several authors have suggested various approaches in measurement tool development (F. Brown, 1983; Friedenberg, 1995). Early works used a rational or logical approach, which rely mostly on the competence of the scale developers in constructing items or identifying the indicators of the construct. Another approach includes theoretical methodology where items are generated based on the theoretical support of the construct. Recently, the scale development process has significantly improved due to the advancement in measurement theories and mathematical and statistical computations.

3.6.2. Theoretical Approach to Scale Development

Ideally, scale development has two phases: the preliminary phase and the factor analytic phase. The preliminary phase requires the following steps as suggested by DeVellis (2003): (1) define the construct to be measured, (2) construct the item pool, (3) establish

the appropriate format of the measure, (4) have experts' review the items, (5) pilot test the items, (6) analyse the items, (7) optimise the scale length.

The first step to all scale development research is defining the construct of interest. The definition of the construct must be supported by theories and research. The quality of the items that can be generated is dependent on the conceptual foundation of the construct. The clarity and conciseness of construct definition affects the succeeding phases of scale development.

After the construct has been defined, the next step requires generating item pools, which are indicators of the construct. Generally, for each sub-domain or factor, there should be more than one item to reduce the effect of the item error. This step requires the writing ability of the item developer to construct items that are clear, concise, and within the level of the target respondents. Items which are poorly worded and do not tap the underlying construct are the potential sources of error variance, which consequently reduces the correlation amongst items (Quintana & Minami, 2006).

The next step is to establish the number of response categories for each item. Researchers have diverse views in terms of the best number of response categories that will optimise the psychometric measures of the scale. For example, Muñiz, García-Cueto, and Lozano (2005) recommend the use of seven categories, whereas Kroh (2007) suggests to use an 11 point-scale. Other researchers recommend a range. Lozano, García-Cueto, and Muñiz (2008) support the use of between four to seven categories, whilst Preston and Colman (2000) recommend between seven and 10. However, a recent study of J. Lee and Paek (2014) found that either four, five, and six response categories will generate virtually the same psychometric properties. They found that the use of two or three response categories would substantially decrease the ability correlation, IRT reliability, Cronbach's alpha, convergent validity, divergent validity

and interitem correlations. They recommend that if two or three response categories are used, the scale length and discrimination indexes of all items should be optimised to reduce the impact of the low number of response categories on the psychometric properties of the scale.

The next step requires experts in the field to review the generated items. This step ensures the content validity of the instruments. Feedback from the experts includes item construction, clarity, coverage of the items, grammatical correctness, level of language used, and item duplication.

Although DeVellis (2003) recommends that the next step is the inclusion of validation items, Worthington and Wittaker (2006) limit this step for two reasons. First, they argue that the length of questionnaire should be as short as possible and all items should be related to the construct of interest. This is because the longer the questionnaire the less likely that participants will complete the survey (Converse & Preser, 1986). Second, the addition of unrelated items in the questionnaire may contaminate the responses.

To check for the reliability and the characteristics of the items, pilot testing is done at this stage. The succeeding stages require an empirical approach where items are tested based on their predictive utility for criterion group or homogenous item groupings. However, the analysis is not straightforward as there are competing methodologies.

3.5.2. Issues Associated with the Use of an Empirical Approach

Despite the agreement of test developers and psychometricians on the use of an empirical approach in measurement tool development as the most appropriate method, there are still contentious issues about which mathematical and statistical approaches

provide better empirical evidence. There are four main arguments in the selection and use of statistical tools.

First, the argument between the uses of the Rasch model and a factor analytic approach. Worthington and Whittaker (2006), who discuss recommendations for best practice in scale development, emphasize the use of a factor analytic approach but with no mention of the Rasch model. However, Dwight (1996) showed that factor analysis does not produce linear measurement of the construct and thus, the use of the Rasch model gives more meaningful results to quantify the construct being studied. This argument is supported by the results of the study by Ewing, Salzberger, and Sinkovics, (2005) in comparing both approaches in analysing the characteristics of the scale. They came to the conclusion that a “true score theory is certainly not the most advanced way to tackle measurement problems” (p. 30). Hence, the Rasch model is more appropriate when addressing the measurement requirements of a construct because it uses a logistic model where true scores are converted into logit scores.

In some instances, both approaches are used to evaluate the dimensionality of the construct but the Rasch analysis has a particular limitation in determining the underlying sub-construct. Unlike in the factor analytic approach, where the specific components of the construct are identified by extracting factors, the Rasch analysis depends on an *a priori* theoretical base and the construct is split into separate dimensions and each dimension is analysed separately (Sick, 2011). Hence, a factor analytic approach is more appropriate if the focus of the study is to account for and establish the multiple dimensions of the construct.

The strength of the Rasch model lies on its ability to support dimensionality of the construct with evidence of local dependence (or if it is desired to determine if the items are not related), differential item functioning (Randall & Engelhard, 2010) and

item fit. For these analyses, factor analysis is inappropriate because of the limitation of factor analysis to address spurious evidence (Kreiner & Christensen, 2011). Although the Rasch model and CFA are used to identify items that function differently between or amongst subgroups of individuals (Randall & Engelhard, 2010), the Rasch analysis is more powerful because it capitalises on the principle of specific objectivity and it accounts for the residuals in different groups (Ewing, Salzberger & Sinkovics, 2005).

However, there are studies that demonstrate the successful complementation of these two approaches in assessing the dimensions of a particular construct of interest. For example the work of Lamoureux et al. (2007) and Siegert, Jackson, Tennant and Turner-Stokes (2010) initially used a factor analytic approach and then individually analysed the extracted dimensions using the Rasch model. In contrast, Allison, Baron-Cohen, Stone and Muncer (2015) initially used the Rasch model then proceeded to use factor analysis. In this complementation method, the key is to determine the strengths of each approach and use them appropriately to address a particular condition in the analysis being sought.

Recognising the strengths and appropriate applications of both approaches, in this study, the factor analytic approach is used to establish the dimensions of teacher A/L literacy whereas the Rasch model is used to establish the item-level characteristics of the tool that includes determining the difficulty of the criteria, the discrimination index of each criterion, and the relative hierarchy of the standards. The DIF is not within the scope of this paper due to limited access to data and prior agreement (between the researcher and the Ministry of Education of Brunei) that there should be no comparison by any means. Also, the Rasch analysis is used to analyse the results of pilot testing to ensure that the theoretical ordering of the standards are empirically supported so as to confirm the construct validity of the tool. This was done by

determining the category disordering characteristics of each item. This is further discussed in the later section of this chapter.

Second, the tension between the use of exploratory factor analysis (EFA) or principal component analysis (PCA). Although EFA and PCA are both statistical techniques used to reduce a large number of variables into smaller components or factors, these two analyses are not the same. EFA accounts for the common variances among items whereas PCA reduces the number of items but retains the variances of the original observed variables. There has been an on-going debate surrounding the suitability of EFA over PCA as a variable reduction method, which up to this moment, statisticians and scale developers are split in their support for these two perspectives. In the reviews which recounted the use of PCA and EFA in tool development or understanding the dimensionality of the construct in social science research, there is an agreement that more than half of the published papers used PCA, but most of these studies have been heavily criticised for the inappropriate use of PCA (Conway & Huffcut, 2003; Henson & Roberts, 2006; Park, Dailey & Lemus, 2002; Russell, 2002). The popularity of PCA may be due to its simplicity, lesser demand in terms of statistical computations, and convenience as most statistical software use PCA by default (de Winter & Doddou, 2015). However, recent studies comparing these two analyses support the earlier arguments and empirical evidence that true factor analysis outperforms PCA.

As shown by De Winter and Dudou (2015), Kim (2008) and Widaman (2007) there are several issues related to the use of PCA in scale development. First, the factor loadings extracted by PCA are either higher or lower than in CFA due to its tendency for under- and over-estimation. Second, due to this under- and over-estimation, the factor loadings in PCA weakly correlate to the true factor loading. Third, the correlation

coefficients for cross loadings and model error are higher in PCA than in CFA. Hence, EFA is preferred because of its ability to capture single underlying components with higher correlation value and also, unlike in PCA, factor analysis does not assume orthogonal relationship with factors (Wang, Kammerer, Anderson, Lu, & Feingold, 2009).

Worthington and Wittaker (2006), Slavec and Drnovsek (2012) stress the use of EFA because in this analysis, the underlying factors provide the bases for the measured responses as opposed to PCA. Another feature of EFA which make it preferable over PCA is the ability of EFA to decompose variances into two categories: those accounted for by common factors and unique factors (DeCoster, 1998). From a more technical perspective, the main difference between the two analyses lies in their assumptions about communalities. In PCA, the communalities are assumed initially to be 1, which means that the factors account for the total variance of the variables. This is contrary to the EFA where it assumes error variance (Field, 2009). The communalities are rather estimated through calculating the squared multiple correlations of the variables with other variables (Rietveld & Van Hout, 1993). These calculated communalities are the bases for calculating eigenvalues and the extraction of the factors.

Third, the requirement of running an EFA using a different data set before using the confirmatory factor analysis (CFA) with another data set. There are several studies that used only either EFA or CFA to establish the dimensionality of the construct, but there is a strong argument that CFA precedes EFA (Martens, 2005; Slavec & Drnovsek, 2012; Worthington & Whittaker, 2006). The argument is not about which of the two analyses could provide stronger empirical support to the latent traits of the construct, but more on the utility of EFA before conducting CFA. This argument is evident from a large number of studies published that used only CFA to describe the dimensionality of

the construct or to establish and refine the construct validity of the measurement tools (Losada et al., 2008; Mak & Sockel, 2001; Scott-Parker, Watson, King, & Hyde, 2012).

In practice, EFA is done in the initial stage especially when the construct of interest is not well-explored or there is a weak theoretical support for the dimensionality of the construct, whereas CFA is performed with variables with strong theoretical background. From the name itself of EFA, it is exploratory by nature where all items are allowed to load on all factors based on the assumptions that any variable may be associated with any factor. The aim of EFA is to reveal the underlying factors associated with the construct (Costello & Osborne, 2005; Williams, Brown & Osman, 2010). In contrast, CFA is confirmatory by nature where the factor structure derived through either theory or from the results of EFA, is tested to see if it fits the data well (Jackson, Gillapsy & Purc-Stephenson, 2009). Unlike in EFA, the variables in CFA are made to load onto their specific factors and cross loadings of variables are restricted. In psychometrics, CFA is widely used for the refinement of measurement tools including assessing construct validity, discovering the existence of a second-order factor, evaluating measurement invariance, and establishing relationships of the underlying factors (Brown, 2006).

Fourth, the limitation of confirmatory factor analysis and the emergence of exploratory structural equation modelling (ESEM). The most recent literature in factor analyses argues that the use of CFA to test if the model fits the data well is problematic. Asparouhov and Muthén (2009) cite two arguments that call for the need of a more realistic and accurate factor analytic approach.

First, the process of finding the best fitting model is often exploratory rather than confirmatory by nature. The first-derived model in EFA, and once tested in CFA, undergoes a series of modification and/or tests of competing models to determine the

best fitting model. In essence, what is happening in CFA is still exploratory by nature because the factor structure derived from EFA would not be the best fitting model when tested in CFA. Similarly, those models that are not derived from EFA, but have strong a priori support and are being tested in CFA, are further re-specified to determine the best fitting model.

The second issue is the effect of strict specification of zero cross-loadings, which consequently produces factors that are highly correlated, and thus compromises the divergent validity of the factors. In any construct, especially those that are multi-dimensional, the indicators of each factor are correlated to some extent. In CFA, the model tested required to set zero correlations, and as a consequence “the correlation between factor indicators representing different factors is forced to go through their main factors only, usually leading to overestimated factor correlations and subsequent distorted structural relations” (p. 398). Hence, the results of CFA are not indicative of the true nature of the factor structure of the construct being studied.

Due to these issues in CFA, Asparouhov and Muthén (2009) developed a methodology and used actual data and simulation to demonstrate the suitability of this approach that uses an EFA in structural equation measurement model that is less restrictive in terms of cross-loadings. To further demonstrate the advantage of using ESEM over CFA estimation, several studies were conducted by investigating the previous models derived from a CFA model. Marsh et al. (2009) used a large data set derived from students’ evaluations of university teaching (SETs) over a period of 13 years with 30,444 classes, and analysed it using both CFA and ESEM. The ESEM-derived model has better fit indices (e.g., CFI = .96; TLI = .927; RMSEA = 0.04) than the CFA-derived model (e.g., CFI = .887; TLI = .871; RMSEA = .111). Also, the CFA results showed that the median correlation amongst factors is .72 whereas in ESEM-

derived model, it is .34. These results demonstrated that CFA has a high tendency to produce more highly correlated factors than ESEM. The researchers argue that using the results of CFA has strong implications for interpreting both the nature of the construct and the results obtained from using such instruments. The high factor correlations obtained from CFA could not support the multidimensionality nature of the construct and the discriminant validity of the tool. Similarly, the rating obtained could not provide a valid interpretation of teachers' performance, and hence, the feedback they receive is less useful in improving their performance.

Another study by Marsh et al. (2010) revisited the 60-item NEO-Big Five personality factors, where the results of CFA do not support the a priori structure. They used the ESEM approach to analyse a data set gathered from German schools. Results of their study support the claim that ESEM produces a more accurate estimation. They have shown that an ESEM derived model fits the data well and the factors are less correlated than factors derived from CFA. This good discriminant validity of the factors is due to non-restrictive model specification in ESEM (Asparouhov & Muthén, 2009). These features of ESEM were subsequently explored by Marsh, Lüdtke, Nagengast, Morin, and Von Davier (2013) using two real and one simulated data sets. They were able to demonstrate that in CFA, cross-loadings amongst factors affect the results and consequently, affect the decisions to pick the best fitting model. They arrived at a conclusion, which is consistent to the earlier findings of Asparouhov and Muthén (2009) that using ESEM to find the best fitting model answers the limitations and issues in using CFA.

As shown, the use of factor analysis in tool development is undoubtedly the most advanced in terms of establishing the dimensionality of the construct. However,

the factor analysis cannot provide all the item level characteristics of the tool. Hence, the use of the Rasch model is needed.

3.5.3. *The Rasch Model*

To fully understand the underlying traits in the scale being developed, application of the Rasch model is widely favoured over the classical test theory (CTT). This is because the Rasch model is based on the assumption that the latent construct can be explained by the responses of the test takers. The use of the Rasch model has an advantage over CTT because it offers a better approach to understanding item characteristics, including item fit and multidimensionality and also the estimation of an individual's ability level. The item parameter calibration is sample-free which gives it a stable measure whilst individual's ability estimate is item independent.

Several models of the Rasch model are being used in scale development. The choice of model is based on different assumptions. This review focuses only on the polytomous models since the scale that was developed is polytomously scored or categorically scored. There are three major polytomous Rasch models; partial credit model (Masters, 1982), generalised partial credit model (Muraki, 1992) and graded response model (Samejima, 1995).

The partial credit model and the generalised credit model are the two widely used models for scale development. The similarity between the two models lies on the assumption that each of the adjacent categories are dichotomous by nature where the higher performing individual has greater probability of scoring the higher category than its counterpart. The difference between the two models, which is the basis for selecting which model fits the data, is their assumption with item discrimination. The partial credit model assumes that the discrimination index for all items in the scale is constant,

which is contrary to the assumption of the generalised partial credit model where discrimination indices are not necessarily the same for all items. Since the development of the items in the scale provides no guarantee that all items will have the same discrimination indices, the generalised partial credit model will be used in this study.

The relationship between responses and construct is used to establish the probability of success of choosing the level of skills. In other words, those individuals with higher abilities are expected to place themselves in the higher level of the scale. Those individuals with lower abilities are expected to be at the lower level of response category. The value of the probability of success for each item is provided by the item characteristics curve (ICC).

The ICC of each item provides two meaningful parameter estimates. One is the discrimination index, which determines how well each item can distinguish high performing from lower performing individuals. It is recommended that items with a discrimination index greater than or equal to .4 will be included in the questionnaire. The higher the discrimination index, the better the item in locating individuals in the continuum of competence. This is a critical characteristic of individual items in the questionnaire as it provides meaningful information in identifying the current level of performance of each test-taker, which may indicate the zone of proximal development (Griffin, 2004). This ZPD is a critical point for learning and teaching.

Another estimate is the difficulty index, which is the same as the threshold index for dichotomously scored items. For the polytomous items, the difficulty index is presented as a set of threshold indexes, which explains the difficulty of the interaction between the adjacent categories for each item. This is known as item step and steps within an item. Thus, for example, in a 5-point scale, there are 4 threshold indices generated as a result of the interactions between Categories 1 and 2, 2 and 3, 3 and 4,

and 4 and 5. To highlight the application of difficulty index in scale development and understanding the underlying traits, Tang (1996) offers interpretations of the step parameter relative to the ability of the individuals. In a layman's term, it can be said that:

- If the ability of the individual is equal to the value of the threshold index of the score category say, Categories 2 and 3, then the probability of reaching the score Categories 2 and 3 will be equal.
- If the ability of the individual is less than the value of the threshold index of the score categories, say Categories 2 and 3, then the probability of reaching the score Category 3 will be less than the probability of reaching the score Category 2.
- If the ability of the individual is greater than the value of the threshold index of the score categories, say Categories 2 and 3, then the probability of reaching the score Category 3 will be greater than the probability of reaching the score Category 2.

The ICC obtained, when using a large group, is stable and is not affected by the group of samples. With this feature of the Rasch analysis output, the level of the trait of each respondent can be precisely estimated, which gives empirical evidence if the scale provides estimates of the traits across continuum of competence. This is useful in the parametisation process where all item categories are arranged in increasing complexity of task. The resulting hierarchy can then be used to identify the zone of proximal development for each individual if his/her ability level is known. This is advantageous for identifying the learning needs of the individual.

Although it is often thought that all tests or scales are multidimensional by nature, psychometricians assume unidimensionality. Studies on the unidimensionality of the dichotomously and polytomously scored items by Carlson (1993) and Huynh and Ferrera (1994) show that even with the complexities of the cognitive processes used by the individuals in answering the test, data from these tests can meet the assumption of unidimensionality.

Another important assumption of the Rasch model is the local item independence (LID) of the items, which means that items should be uncorrelated after the person's ability measure is controlled. In the study by Huynh and Ferrera (1994), it was shown that sets of items measuring the same subset of constructs showed local independence. However, those items in a certain subset showed local dependency to some extent. In the strictest sense, construction of set of items tapping to the construct identified violates the assumption of local independence. Since all items are expected to measure the same construct, it is expected that these items will show local dependency.

In test development, the assumption of LID is met when the performance of the examinee in a particular item will not influence his/ her performance on other items in the test. In other words, the response of the individual to each of the items in the test is only influenced by his/ her ability and the item characteristics (Hambleton, 1989). However, the concept of LID is within the premise of the ability level of the individual test taker or test takers. It means that items can be correlated over the total group of examinees with different ability levels.

The presence of the LID in the test affects the item characteristics (Yen, 1993) including the reliability index (Sireci, Thissen, & Wainer, 1991) which cause overestimation up to 10-15%, especially if the item-based method of computing coefficient alpha is used. To illustrate this, supposing Item 1 at a certain ability level is

highly correlated with the total score and is also highly correlated with Item 2, then Item 2 is expected to be highly correlated also with the total score. This causes the overestimation of the reliability index, which consequently leads to misinterpretation of the score. The scores will be interpreted as more highly consistent than they actually do (G. Lee & Frisbie, 1999).

To minimise the existence of LID in the test, it is necessary that independent items be developed, although this is difficult to achieve due to the fact that all items being written for a particular test relate to the same construct. After the pilot testing, analysis of the existence of LID will identify those items which exhibit interdependence. These items should be revised.

Allen and Sudweeks (2001) recommend the use of the result with care when using context-based items. To minimise the influence of LID in the test, mixed scoring accompanied by mixed procedures for estimating reliability should be used. To do this, those items that show LID should be considered as one set and treated as single item, and those that do not show LID should be scored independently. With this approach, the resulting coefficient alpha is more accurate than if all items are scored independently. Another approach to identify items that are highly correlated, and to gather empirical support for grouping them into a single factor is to conduct factor analyses.

3.6.Factor Analyses

3.11.1. Overview

Factor analysis has been extensively used in measurement tool development. This section discusses the reasons for using factor analysis in scale development (Jackson, Gillapsy, & Purc-Stephenson, 2009; Worthington and Wittaker, 2006) alongside the

capability of the Mplus software v7 (Muthén & Muthén, 1998-2012).

3.11.2. *Exploratory Factor Analysis*

In general, the primary objectives of EFA are to establish the number of common factors comprising the construct and determine the strength of the relationship between each factor and each observed measure. In other words, EFA empirically establishes the linear relationship of observed variables to latent factors. This feature of EFA is linked to the construct validity issue in scale development. Whilst factors or constructs are unobservable traits, the indicators (commonly called as variables or items), provide the needed measures to quantify the amount of the factors. The result of the EFA provides evidence on which items are related to a particular factor and which items are simultaneously measuring different factors (if there are) which subsequently provide an empirical basis for grouping items or deleting items in the final version of the questionnaire.

In using EFA, there are important considerations including sample size, factorability of the data, rotation, criteria for factor retention and deletion and theoretical interpretability.

Sample size. This is a necessary requisite in conducting EFA because it directs the interpretation of the results. Several authors cite various views in terms of how many samples are needed to satisfy the sample size requirement: at least 10-15 samples per variable (Field, 2009) and 50 observations and 5 times as many observations as variables (Habing, 2003). Generally, when using EFA, the minimum sample size recommended is 300 for the initial analysis. It is argued that the bigger the sample size (300+), the more stable the factor solution is (Field, 2009). However, the required sample size is influenced by the estimator used. In the study of Beauducel and

Herzberg (2006) comparing maximum likelihood (ML) and weighted least squares means and variance adjusted (WLSMV), WLSMV provides accurate estimation even with lower sample size. This is consistent with the results of the study of Hox, Maas, & Brinkhuis (2010).

Rotation. Factor rotation helps in interpreting factors. From its name itself, it rotates variables loading and hence, improves interpretation. There are two types of rotation being used: the orthogonal rotation (which assumes that there is no correlation between extracted factors) and the oblique rotation (which assumes the other way). The choice of which factor rotation should be used depends primarily on the strength of the theory to support the relationship of the factors. Field (2009) recommends that both rotations should be used, and if the oblique rotation shows that there is no correlation existing amongst factors, then the output of the orthogonal rotation should be used for interpretation. However, the work of Worthington and Wittaker (2006) emphasise that oblique rotation is recommended regardless of the nature of the correlation of the factors based on theories to ensure that factors are uncorrelated. The default rotation in Mplus is geomin, which is an oblique rotation and more suitable for categorical or ordinal data.

Factor retention and item deletion. The criteria for factor retention and item deletion include: factor loadings less than .32 or cross-loadings less than .15 difference from an item's highest factor loading; approximate simple structure; and parallel analysis. Factors with two items must be deleted unless the items are correlated ($r > .70$). Field (2009) and Rietveld and Van Hout (1993) suggest retaining factors with eigenvalues greater than 1. Analysis of scree-plot is also recommended to have a visual inspection of the number of factors. It is also recommended that every time an item is

deleted, it is necessary to rerun the entire process of EFA to see the stability of factor structure, factor correlations, item communalities, factor loadings, or cross-loadings.

Theoretical Interpretability. The statistical bases for factor identification need to be considered in conjunction with theoretical interpretability. This is the qualitative side of EFA where the items that are common for each factor must have common characteristics (Field, 2009). Even if the result of EFA suggests that particular set of items belongs to a certain factor, yet there are items which are inconsistent with the rest of the items, then those inconsistent items must be deleted.

The scale produced in EFA should not be altered in any form (adding/deleting items, deleting items, changing item contents, altering the rating scale) before subjecting it to confirmatory analysis.

3.11.3. Confirmatory Factor Analysis

The most common approach to CFA is the use of structural equation modelling (SEM). After obtaining theoretical factors in EFA, the next step required is to specify the obtained factor solution in the CFA to determine the ability of the factor model to fit in an observed data set. This confirmatory method is the most commonly used for scale development where a single model composed of a set of relationships is specified and tested. The result obtained in EFA is tested if the model works or not. In this approach, there is no attempt to modify or alter the existing model.

There are other methods of structural equation modelling (SEM) that are used. These include the competing models approach, which allows researchers to evaluate alternative models, which are not based on the results of the EFA but are still theoretically plausible. In this approach, if another model best fits the data, an explanation must be provided to account for the discrepancies observed and, if the

alternative model will be adopted, a new EFA and CFA will be conducted using a different data set. In the study by Worthington and Wittaker (2006), of the 10 studies they analysed, two used the single-model approach where the resulting model in EFA was specified in the CFA, and eight of the studies used the competing models approach.

Another SEM strategy is the use of model development. The advantage of using this is that the result provides insights on re-specifications of the model which is theoretically justifiable. However, in doing this, the re-specified model must be verified using different samples (Field, 2009).

Similar to EFA, CFA requires a sufficient number of samples to obtain more stable factor loadings. Hair, Black, Babin, and Anderson (2010) recommend the following guidelines in determining the appropriate sample size:

- 100 for models with < 5 constructs, each with more than 3 items, high item communalities (> .6)
- 150 for models with 7 or fewer constructs, modest communalities (.5), no constructs with fewer than 3 items
- 300 for models with 7 or fewer constructs, lower communalities (< .45), multiple constructs with fewer than 3 items
- 500 for models with large numbers of constructs, some with lower communalities, multiple constructs with fewer than 3 items

The next step in conducting SEM is to estimate the model. It starts with the determination of a structural model. At this stage, a decision is needed as to whether each parameter in the model is free or fixed. The method of estimations needs to be specified. The most common approach to parameter estimation is the maximum likelihood (ML) in which the most “likely” parameter values to achieve the best model fit are found unless the measures lack multivariate normality. However, for categorical

and ordinal data, there are three primary estimators to choose from. These are the unweighted least squares (ULS), weighted least squares (WLS) and diagonally weighted least squares (DWLS). Other secondary estimators are derived through correcting the estimated means and/or means and variances.

The weighted least squares means and variance adjusted (WLSMV), derived from correction to means and variances of DWLS, performs best in CFA modelling (Brown, 2006). This is a robust estimator that does not assume normally distributed variables. In the study by Beauducel and Herzberg (2006) which compared the performance of ML with the WLSMV using models with a different number of factors and sample size, the WLSMV outperforms the ML in various aspects of factor analysis. They found that the chi-square test ($p=0.05$) of WLSMV with two or three categories has a higher tendency to reject the null hypothesis than the chi-square test of ML. Similarly, the factor loadings are more accurately estimated using WLSMV. Moreover, WLSV does require a larger sample size to obtain a more accurate model. In the analysis of ordinal data, which is the case of my research, it is done by first estimating a polychromic correlation matrix and an asymptomatic covariance matrix and employing these within a weighted least square estimator.

The results of SEM must be evaluated by looking at the overall fit of the model to the data. There are several fit indices being considered aside from the chi-square value which is typically used but is not adequate to test the goodness of fit due to its sensitivity to sample size (Hu & Bentler, 1999). Supplemental fit indices include incremental, absolute and predictive (Kline, 2005).

Incremental fit indices assess how well the estimated model fits relative to some alternative baseline structural model. The most common baseline model is called the null model, which assumes that all observed variables are uncorrelated. Field (2009)

suggests checking the (1) normed fit index (NFI) which is the ratio of the difference between the χ^2 value for the fitted model and a null model divided by the χ^2 value for the null model (desired value is close to 1); (2) Tucker-Lewis index (TLI), which is conceptually similar to NFI, but is actually a comparison of the normed χ^2 values for the null and specified model (value close to 1); (3) Comparative Fit Index (CFI) which is normed so that values range between 0 and 1 and values $> .90$ are usually associated with a well-fitting model. The absolute fit indices measure how well the structural equation model reproduces the data. The Goodness of Fit Index ranges between 0 and 1 (desired values are those close to 1). The predictive fit indices measure how well the model would fit in other samples from the same population.

In case where the results of CFA are problematic, further analysis is needed using the most recent addition to the family of factor analyses, the exploratory structural equation modelling.

3.11.4. Exploratory Structural Equation Modelling

The same conditions (sample size, factor loadings, etc) are needed and fit indexes in CFA are evaluated for exploratory structural equation modelling (ESEM). This analysis is accomplished using both EFA and structural equation modelling by a one-step approach and is “more accurate [than CFA] because it avoids the potential pitfalls due to the challenging EFA to CFA conversion...the ESEM approach avoids this problem by estimating the measurement and structural model parts simultaneously (Marsh et al., 2010, p. 44). Also, ESEM avoids model overestimation because it does not require zero-loading specification because it accounts cross-loadings amongst factors. The use of ESEM was argued extensively in Chapter 3.11.4.

Although ESEM has been shown by several studies to have great advantages over CFA, it is not recommended that the ESEM approach should replace the CFA approach (Marsh et al., 2010). It is still recommended that a preliminary analysis be conducted using EFA, then followed by using the CFA model to compare the results of subsequent analyses (Marsh, Morin, Parker, & Kaur, 2014). If the model fits the data well using the traditional CFA estimation compared to the one derived in ESEM, then the CFA results should be used. However, if the CFA results are problematic then a subsequent analysis using ESEM needs to be carried out. In other words, the ESEM should serve only as a viable option if the CFA model does not fit the data well.

The use of EFA, CFA, and ESEM establishes only the dimensions of the construct being studied, and provide empirical evidence to the factor-level properties of the tool. The resulting tool can be used with utmost confidence that the results obtained reflect the perceived level of assessment literacy of teachers. However, if the results of the teacher self-assessment should be used for developing a professional development program, which is the intention for the development of this tool, investigating the patterns of teacher use of the various AfL skills is necessary to establish particular groupings of teachers who have the same levels of ability and the same AfL literacy needs. These groupings of teachers can be established using latent profile analysis.

3.12. Latent Profile Analysis

Latent profile analysis (LPA) is the application of latent class analysis (LCA) in continuous, ordinal and categorical variables. Both LCA and LPA are subsets of structural equation modelling used to find groups or subtypes of cases in multivariate data (McCutcheon, 1987). This analysis has been widely used to establish a typology of practices, profiles, or characteristics of individuals or phenomena. For example, LPA

has been used to establish college students' achievement goal orientation (Pastor, Barron, Miller, & Davis, 2007); patterns of eating disorders apart from anorexia nervosa or bulimia nervosa (Mitchell et al., 2007); profile structure of coping strategies of minority adolescents (Aldridge & Roesch, 2008); classes of adolescent sedative/anxiolytic misusers (Hall, Howard, & McCabe, 2010); and amongst others.

In my study, the LPA has the potential use of investigating teacher patterns of A/L practices using the factor scores derived from factor analyses. This analysis can reveal the number of groups of teachers that can be identified by their use of the A/L concept. It was used to test if teacher career stages described by previous researchers are true also in the context of teacher A/L literacy.

The use of LPA is advantageous over traditional hierarchical and non-hierarchical traditional clustering techniques. Pastor et al. (2007) demonstrated in their study of re-examining the profile of college students' achievement goal orientation that LPA gives more confidence to the users of the results as it uses a range of information criteria to determine the number of classes or profiles. Also, unlike in cluster analysis, LPA allows the means, variances and covariances to vary across clusters, which provides a superior fit. In addition, LPA does not aim to create a cluster of equal sizes, and hence, the actual composition of the clusters is fully described.

To determine the number and pattern of classes existing in the data the statistical information criteria such as Akaike information criteria (AIC) (Akaike, 1987); Bayesian information criteria (Schwartz, 1978); and sample-size adjusted Bayesian information criteria (BIC) (Sclove, 1987) are used. Although no absolute rule of thumb has been proposed as a model evaluation guideline, LPA models with the lowest values in AIC and BIC indicate a better fit, provided that the model has a good theoretical interpretation. In cases where these three criteria will have inconsistent results, the

sample-size adjusted BIC result will be used to decide because it is superior to the other two criteria (Nylund, Asparouhov, & Muthén, 2007; Yang, 2006).

In addition, the likelihood tests are used to decide the number of profiles present in the data. These include the Vyoung-Lo-Mendell-Rubin (VLMR) likelihood ratio test (Vuong, 1989), the Lo-Mendell-Rubin (LMR) adjusted test (Lo, Mendell, and Rubin (2001), and the bootstrapped parametric likelihood ratio test (BLRT) which uses bootstrap samples to determine the p-value (McLachla, 2000). These likelihood tests tell whether the model with k -class being examined is enough to explain the observed classes in the data or an additional class is needed. A significant p-value in these three tests (VLMR, LMR, and BLRT) indicates that the model with $(K-1)$ classes is rejected and the K model gives the exact number of classes of participants. In cases where the results of these three tests give inconsistent results, the BLRT is usually used for final decision because it is the most consistent and reliable indicator of the latent classes (Nylund et al., 2007).

The advancement in empirical approaches in test or scale development paved the way for a more thorough exploration of the dimensionality and item characteristics of the construct being investigated. In the actual application of any measurement tool for professional development, it is suggested that the assessment process and professional development program should adopt the context of a classroom assessment.

3.13. The Trustworthiness of Teacher Assessment Tools

Reliability and validity are always important issues for discussing the trustworthiness of any assessment tool. The scale development process described above ensures that the tool developed has met both the psychometric and factor analytic requirements. These properties are important considerations in assessment because the failure of test

developers to ensure these allow the measurement error to significantly affect the results (Field, 2009). The high reliability and validity indexes of the assessment tool gives confidence to users of assessment data that the results are consistent across time and context and hence, the inferences drawn are useful.

In this research, these measures of internal and external consistencies of the teacher AfL literacy framework utilise the factor analytic approach and the application of the Rasch model of item response theory, the latter of which has gained prominence in test design and analysis (McNamara & Knoch, 2012). To further enhance the trustworthiness of the teacher AfL tool, the premise of classroom assessment are applied to teachers' assessment using the tool. Several arguments on reliability and validity hold true in this context.

First, there is a growing argument that strict adherence to these measurement principles detaches theories and concepts from practices. For example, Brookhart (2003) argues that authentic assessment does not use much of the traditional concepts of measurement. Rather, in actual assessment and teaching practices, the emphasis shifts from measurement principles to the quality of information considering the context-dependence of assessment, the use of assessment information to teaching, and the formative and summative functions of assessment. This argument is supported by Smith (2003) who called for a reconsideration of reliability to account for the context under which assessment occurs. He argues that the different forms of reliability measures negate the principles of learning. For example, the test-retest or parallel forms of reliabilities, which are founded in the concept of stability of students' performance across the time of testing, are impractical for classroom assessment as each teacher works hard to bring about significant change to the performance of students in the least possible time.

Along the same line of argument, the Coefficient alpha, which is tied to the concept of score variance, is an unreasonable measure if invoked for classroom assessment. In psychometrics, any items that all students get right or wrong do not affect the test variance, hence are considered useless items. However, in classroom assessment, these items are important for teachers to determine which learning objectives are achieved well by students or which learning objectives need further exploration to enhance student learning. Because of the inappropriateness of psychometric measures of reliability, Smith proposes an alternative conceptualization of reliability as a sufficiency of information to make robust decisions about learning and teaching. This conceptualization of reliability is contextualized in the nature of classroom learning and teaching which account for “the multidimensionality of the underlying assessment and does not require a rank order of the students” (p. 31).

Similarly, the traditional view of validity, which has varied from the time of Thorndike (1918) who put forward a criterion-based model of validity, later amended by Messick (1989) who proposed the content-based validity model, and Cronbach and Meehl (1955) who adopted the construct model of validity, requires reconsideration. In this regard, Moss (2003) reconceptualises validity in the context of classroom assessment using an interpretive approach. She challenges the assumptions of traditional validity by reflecting on actual classroom practices. First, contrary to the view that assessment is a distinct episode in learning and teaching, in the actual classroom setting, assessment is an integral part of the learning and teaching processes. Assessment forms the network that binds all other classroom activities to support student learning.

Second, the concept that validity requires the appropriateness of interpretation of student learning based on the assessment results is too limiting or too encompassing. According to Moss, in actual practice, when teachers are concerned about student

progress, there is no need to have a fixed interpretation of student competence. Rather, teachers need to make trustworthy decisions regularly using assessment information to support student learning and to monitor the learning consequences of their decisions. These decisions should be informed by various assessment results, as a single assessment cannot provide sufficient information. Thus, when considering the validity of individual assessment practices, one should look into how each assessment “fits with the other assessment practices, in progression, to support (and illuminate) learning” (Moss, 2003, p. 16).

Third, the consideration that the individual student is the unit of analysis excludes the role of classroom context in assessment and learning. Moss elaborates the argument of Mehan (1998) on using social situation as the unit of analysis and emphasized that learning and teaching decisions should be based on evidence derived from the analysis of the interactions of all the classroom activities.

Fourth, the assumption that interpretations are based on combining all judgments from assessment results will not give much evidence in situations where aggregation is impossible or is undesirable. Moss points out that drawing the right interpretations about student learning is an iterative process which involves repeated measurement of student learning. This requires teachers to test the pieces of evidence by gathering information from various sources until these pieces of evidence can be formed into a coherent picture of student achievement.

Fifth, any assessment practice has a consequence contrary to the widely accepted concept that consequence matters only if the source of a particular consequence can be traced to construct underrepresentation or construct irrelevant variance. This last view was further elaborated by McNamara and Roever (2006) by studying in detail the social consequences of assessment.

The arguments of Moss described above are not new. Shepard (2001) argued nearly fifteen years ago that in a less standardized test, there is a need for a new methods of analysing and interpreting students' responses. Kane (2001) proposed an argument-based approach to validity which he "suggests that the proposed interpretation be specified in terms of a network and inferences and assumptions, that these inferences and assumptions be evaluated using all available evidence, and that the plausible alternate interpretations be considered" (p. 339-340). In a more detailed way, Killen (2003) looked into the processes which help improve the validity of the assessment. He argued that the appropriateness and usefulness of teacher's judgments of student performance are influenced by all classroom activities from developing learning outcomes to providing learning opportunities, which help students to achieve targets to developing and implementing assessment tasks and to interpreting the assessment results. What was emphasised by Killen is the necessity to ensure the coherence of all the learning, assessment and teaching activities. The interplay and the direct focus of all these activities to student learning ensure the trustworthiness of the teachers' decisions on student learning. Clearly, as each learning and assessment practice happens in various classroom contexts, there is a need to extend the concepts of reliability and validity to the context of actual and authentic assessment practice.

To demonstrate the applications of the arguments above, in a high-stake test-dominated system, test developers and researchers spend a huge amount of time and rigorous effort to construct test items and apply psychometric analysis to ensure the reliability of the test and to ensure that valid inference are drawn. In contrast, in a more school- or classroom-based assessment-dominated system, teachers spend a considerable amount of time developing differentiated assessment tasks to gather a wide range of evidences of student learning and using this assessment information to provide

feedback to students to help them improve in their next stage of learning. In both systems, the issue of reliability and validity are considered, the former consider these psychometric principles in a more formal way involving statistical analyses whilst the latter addresses these issues in more practical approaches.

In this research study, apart from the application of measurement principles to the development of the framework, I extended reliability and validity to the actual use of the tool. The use and application of the framework I developed uses and extends the traditional psychometric principles to what is called by Brookhart (2003) as ‘classroometric’ principles.

3.14. Summary

In this chapter, I have presented a critical analysis of the existing teacher assessment literacy frameworks, models and tools and have argued for the need to develop a tool that is underpinned by the principles of A/L using a more robust methodology. Also, I have presented a literature review of various approaches used in measurement tool development to ensure that my research study draws on more advanced empirical approaches as well as an initial theoretical process. Furthermore, I have justified the complementary use of both the Rasch model and the factor analytic approach, the preference for EFA versus PCA, why CFA is preceded by EFA, and the further exploration of the construct using exploratory structural equation modelling. Finally, I have concluded this chapter by arguing that although the development of measurement tool needs to follow the most advanced empirical approach, the concept of reliability and validity needs to be extended to the actual use of the tool to account for the context-dependent nature of assessment.

CHAPTER 4 - METHODOLOGY

4.1. Introduction

This chapter builds from the literature review of the best practices in scale development using both theoretical and empirical approaches as described in Chapter 3. The aim of this chapter is to justify the choice of and describe the methods used in the study. To achieve this aim, I present the research design and sampling method with an emphasis on the scale development procedure using a factor analytic approach, which includes exploratory factor analysis (EFA), confirmatory factor analyses (CFA), second-order CFA, and exploratory structural equation modelling (ESEM) to establish the dimensions of teacher AfL literacy. Similarly, the application of item response theory (IRT) is argued to analyse the results of pilot test, to determine the final item characteristics of the framework, and establish the continuum of teacher AfL literacy. Finally, the use of latent profile analysis (LPA) is discussed to discover the existence of latent profiles of teachers in using AfL.

This chapter first presents the research design and context, the participants, the development of teacher AfL literacy tool, the pilot testing, the factor analyses, the latent profile analysis, parametisation of AfL standards, and finally, the validation of results.

4.2. Research Design

This study used a multi-method research design. In the early phase, a theoretical approach to scale development was used drawing on research evidence, together with the practical insights of principals, assistant principals, AfL coordinators, teachers, and higher officials from the Ministry of Education (MoE) of Brunei. This approach was used as a starting point for identifying the domains, capabilities, indicators and

standards of performance. In the final stage of this research, an empirical approach was used to develop a more reliable assessment instrument and to determine the degree of consistency between the theoretical rationale and the empirical evidence to support the dimensionality of the teacher AfL literacy tool.

Although there is a strong theoretical rationale in describing effective teacher assessment practices, the research design also incorporated an exploratory case study with teachers' actual assessment activities inside the classroom as the unit of analysis. This approach was utilised to explore the nature and characteristics of teacher competencies in AfL. Workshops, interviews and surveys through self-assessment were used to gather data. The final output of this study is a teacher AfL literacy tool, which the dimensions obtained were used to develop the teacher AfL literacy framework. Figure 4.1 provides an overview how this study was conducted.

During the entire process, participants in each stage were recruited through the MoE. The active involvement of these participants in all phases of the research process is rooted in the assumption that involving them in establishing and in deciding what performance and quality criteria are required will help develop their understanding of AfL principles and practices, which will consequently draw a stronger commitment and support for AfL literacy, at the same time, this ensure that the tool developed is as authentic as possible.

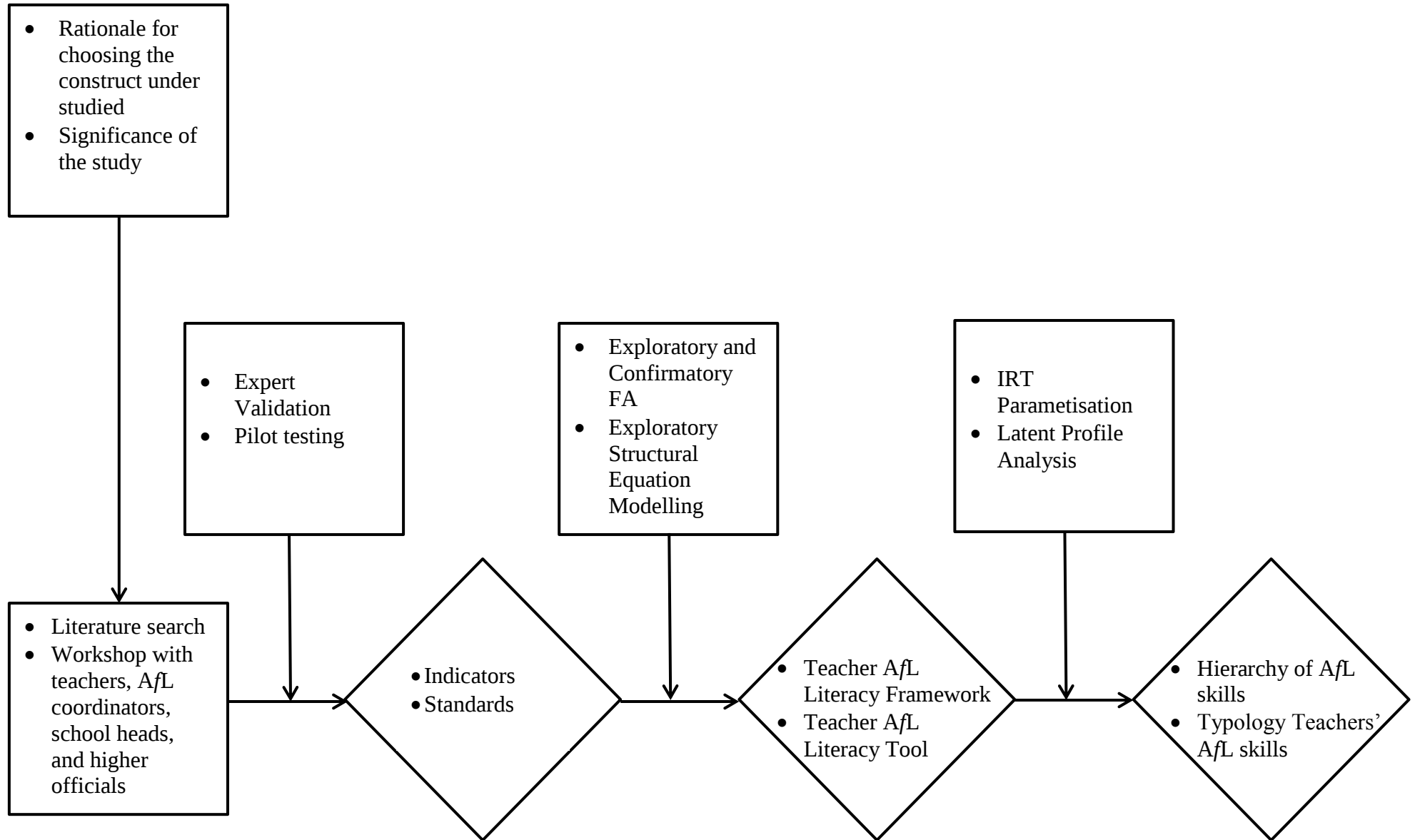


Figure 4.1. Summary of the methods of the study.

4.3. Research Context

The initial phase and the main study were conducted in Brunei Darussalam, a sovereign state located in the Southeast Asia. As mentioned in Chapter 1, the implementation of SBAfL by the Ministry of Education of Brunei required the development of a more effective AfL literacy program, hence, their interest in developing a highly contextualised teacher AfL literacy tool to evaluate the professional learning and development of individual teachers. For this reason, the Ministry of Education of Brunei through their Optimise International (B) Sdn Bhd (OIB) consultants supported this research. The data gathering was conducted from 3 July 2012 to 28 April 2013.

The validation study was conducted in the Philippines involving teachers from the southern part of the country from 10 – 27 January 2013. Data from another country were gathered to explore the applicability and transferability of the framework across the region.

These two countries were selected due to access to teachers and their different levels of AfL implementation. In Brunei, AfL had been extensively adopted with a strong policy that mandated its implementation of AfL and an intensive assessment literacy program had been developed to enhance the competencies of teachers, whereas in the Philippines, AfL had not yet been implemented as part of the educational reform but there was evidence that teachers were using it to some extent. The sole purpose of using two sources of data was to validate the tool to establish its generalizability in terms of use for similar context. Data collected from teachers in each country were not compared in any form as agreed between me and the MoE.

4.4. Participants in the Main Study and Sample Size

In Brunei, teachers who have participated in one or many professional developments in AfL and have been embedding AfL in their practices were recruited through the Ministry of Education (MoE). The MoE distributed and retrieved the recruitment forms. In the Philippines, individual recruitment was done but with permit from the Division's office of the Department of Education (DepEd). The total number of the samples and the date of data collection are presented in Table 4.1.

Table 4.1

Sample Size used in Various Stages of the Research

Date	Phase of Study	Brunei	Philippines
3 - 7 July 2012	Ocular		
3 - 5 Sept 2012	Workshop – identifying best practices	- 30 teachers - 6 principals - 5 assessment coordinators - 4 department heads - 5 MoE officials	
	Validation	- 4 assessment experts - 10 AfL coordinators	
16 - 18 Oct 2012	Pilot Testing	197 teachers	
11 – 13 Nov 2012	EFA	325 teachers	
25 – 27Feb 2013	CFA/ESEM/LPA	354 teachers	
10 - 27 Jan 2013	Validation (CFA/ESEM/LPA)		383 teachers

Different criteria are used in the literature to identify the actual sample size. For example Marsh, Hau, Balla, and Grayson (1998) use the number of items per factor; Fan, Thompson, and Wang (1999) suggest basing the sample size on the estimation method used; West, Finch, and Curran (1995) advocate the use of non-normality of the data; and Velicer and Fava (1998) favour analysing the strength of relationships amongst variables and latent factors. Table 4.2 summarises the recommendations regarding the calculation of sample size by various authors.

Table 4.2

Recommended Sample Size for Factor Analysis

Author	Recommendation
Tabachnick and Fidell (2007)	300
Hair et al. (2010)	>100
Comrey (1973)	100 – poor 200 – fair 300 – good 500 – very good 1000 – excellent
P.M. Bentler and Chou (1987)	5:1 samples to parameters ration 10:1 optimal
Kline (2005)	100-200
Grimm and Yarnold (1995)	5-10 samples per observed variable

Research conducted to find out which sample size is best for factor analysis shows several key findings. MacCallum, Widaman, Zhang, and Hong (1999) found out a data set with high communalities ($>.60$) and with several numbers of items per factor need only a

small sample size. The same is true for a data set with item correlation coefficients greater than .80 (Guadagoli & Velicer, 1998) whilst sample size based on number of items per factor does not affect the result of the factor analysis (Hogarty, Hines, Kromney, Ferron, & Mumford, 2005). In this study, I used the guideline set by Tabachnick and Fidell (2007) to satisfy the relationship between sample size and model complexity.

In this study, although there is empirical evidence that sample size is not an issue in factor analysis if weighted least square means variance adjusted estimation is used (Beauducel & Herzberg, 2006; Hox, Maas, & Brinkhuis, 2010), a relatively large sample size was used to avoid sampling error, so as to ensure the valid interpretation of results.

4.5. The Development of the Teacher AfL Literacy Tool

4.5.1. Overview

The development of the teacher AfL literacy tool followed the model used by Griffin (2000) as explained in Chapter 2. The competency-based performance assessment framework proposed by Griffin has four components, which were adopted in this study:

1. Domains – these are the sub-variables of teacher AfL literacy that describe the job roles of teachers.
2. Capabilities – these are the abilities of teachers on each domain that specify the job functions of teachers.
3. Indicators/criteria – these define what should be done for each capability or the evidence requirements demonstrating that teachers possess the capabilities identified. The indicators include knowledge (cognitive), skills (psychomotor) and attitude and values (affective).

4. Performance – for each indicator/criterion identified teachers will perform differently on a continuum of AfL performance. The range of performance is captured in levels of standards that answers question: “How well is the task performed?” For each criterion, a performance description is explicitly and clearly described. The standards are developed in such a way that the results of individual teachers’ self-assessment can be used to draw inferences about their actual performance.

In the actual process of scale development, Components 3 (criteria) and 4 (performance) are identified first and the capabilities and domains emerged from the results of factor analyses. In most scale development process, the capabilities are not included as only the domains are generated by factor analyses. However, careful examination and categorisation of criteria will give specific clusters, which specify certain capabilities.

4.5.2. Identification of Indicators

As a preliminary step, indicators of teacher AfL literacy were identified from the existing literature. These indicators were used later to verify the results of the workshop.

The actual data gathering started with a three-day workshop with five principals, seven deputy principals, eight assessment coordinators, and 15 teachers. Participants were grouped into five clusters with equal representation of principals and assessment coordinators in each group. On the first day, each group was asked to identify teachers’ best AfL practices. After each group presented their output, a discussion was undertaken to reach at a consensus in terms of the final indicators to include in the assessment tool. The

output of the participants was compared to the indicators derived from the literature review. Teachers and school leaders discussed the discrepancies observed, and a decision was made as to whether include or exclude the inconsistent items from the final set.

The next step required participants to group these indicators based on their commonalities and thematic lines. The judgment was cross-referenced with the relevance and the association of the indicators. From the results, participants came up with six major groupings, which are consistent with the theoretical framework of the study.

4.5.3. Establishment of Standards

On the second and third days, the participants were asked to describe teachers' performance on each indicator of AfL practice at five levels of performance, based on the generic descriptions provided in Table 4.3. The use of five-response categories ensured that the psychometric measures of the teacher AfL literacy tool is optimised (J. Lee & Paek, 2014).

The descriptions are based on the Dreyfus Model of skills acquisition (Dreyfus & Dreyfus, 1980), which is underpinned by the principle that an individual progresses through a continuum of competence starting from novice, beginner, competent, proficient, to expert. This model is based on the sociocultural assumption that moving from novice to expert is grounded in the notion that experience and feedback provides the opportunity to integrate and internalize theoretical knowledge into professional expertise.

Table 4.3

The Generic Descriptions of Literacy Levels Used in Establishing Standards

Performance Levels	Descriptions on Teacher AfL Performance
1	Performs basic tasks required in implementation. Application is limited and generally mechanistic. Implementation 'rules' are followed in a step-by-step way without much, if any, consideration of the context in which optimal teaching and learning occurs. AfL is seen as a somewhat limited and inflexible set of skills applicable in all situations.
2	Demonstrates acceptable performance characterised by personal development of an AfL implementation routine based on recurring situations. In new or novel learning situations, some critical implementation elements are typically omitted.
3	Deliberately plans AfL implementation. The implementation is more effective and organised as a result of clear strategies and mechanisms. New situations and contingencies in the learning process are managed effectively.
4	Implements AfL in a holistic way. All components of AfL are incorporated in a coherent way across the whole process of students' learning that makes sense to both students and parents/carers. Important decisions through assessment and evaluation of the different issues in learning and teaching are made to help develop better strategies and approaches in learning, teaching and assessment.
5	Easily adjusts the implementation of AfL in various situations with implementation relying on the teacher's perception of each learning and teaching situation. At this level the teacher is also providing expert AfL leadership, advice and support to colleagues.

Dreyfus argues that a novice stage is characterised by strict adherence to the rules given, the performance is unlikely to be satisfactory, there is a need for close supervision in

the execution of task, the complexities of tasks are not addressed, and actions are seen as isolated. The next stage, advanced beginner, is achieved when an individual starts to sort out key knowledge required, tasks are completed at acceptable standard, less supervision is required, complex situations are appreciated and actions are seen as a series of steps. Competent performance is demonstrated when an individual starts to use his/her analysis and planning skills and develops an established routine to carry out the same task. The proficiency stage is demonstrated by an individual who can use deliberate decision-making skills and who is able to set rules to formulate plans to achieve the task. Finally, the expert level is demonstrated by the achievement of excellence with relative ease.

4.5.4. Expert Validation

After the results of the workshop were finalised, expert validation of the criteria and standards was sought. The aim of the expert validation was to ensure that the teacher AfL assessment tool contained only items that related to the core AfL literacy construct and that these items were sufficiently dispersed along a continuum of difficulty to enable identification of levels of difficulty in accordance with the intention of the tool to be able to discriminate levels of teacher assessment competence (Lunz, Wright, & Linacre, 1990). This step used AfL experts the extent to which each indicator represented the construct. Thus, this step helped ensure construct validity and gave confidence to the process. Four assessment experts were asked to validate the tool, all of whom have a long track record in research and practical engagement in assessment.

To further support the validity of the assessment tool, 10 AfL coordinators from Brunei who had not participated in the earlier phase of the research were asked to rate the

content validity of individual items using the instrument suggested by Davis (1992). The suggestions made by evaluators were used to revise the teacher AfL literacy tool. Before pilot testing was conducted, 40 teachers were asked to critique the assessment tool to enhance its face validity. Feedback was sought from teachers in relation to the clarity of the criteria and standards and the overall accessibility of the tool.

4.6. Pilot Testing

Pilot testing was conducted to determine the adequacy of the items in the emerging assessment tool; to establish the feasibility of the research methods and sampling design; to identify any logistical problems; and to obtain a snap shot of the possible results of the study (van Teijlingen & Hundley, 2001). Pilot testing is widely utilised in research studies but only a few research papers reports in detail the process and outcome (Lindquist, 1991; Muoio, Wolcott, & Seigel, 1995; van Teijlingen & Hundley, 2001). van Teijlingen and Hundley argue that it is the ethical obligation of researchers to account for all phases of the study, hence the results of any pilot test should be included. Any revision or improvements of the original research protocol as informed by the results of pilot testing should be reported.

The pilot testing gave me a sense of what the large scale testing would look like. The experience gained in pilot testing, however, was not a guarantee that the actual research would follow the same patterns. Even though the pilot testing simulated the main study, the results were not used to make predictions or assumptions about the results of that study.

In the pilot study, the teacher AfL literacy tool developed was administered to 197 Bruneian teachers. The data gathered from the pilot test were analysed using the generalised partial credit model of item response theory (IRT). This was done using Conquest software (Wu & Adams, 2007). The analysis used the following criteria in selecting good items and revising problematic items:

1. Fit Indexes. The fit index of each item indicates whether a particular item fits the item response model well. In other words, it measures the general association or contribution of the item to the measurement of the construct. Generally, items with fit indexes, given by mean fit square value (MNSQ), equal to 1 are said to be perfectly fitted to the model. However, it is rarely that all items in the test will have fit indices equal to 1.0. It is acceptable to include items which have fit indices close to 1.0 and within the range of their respective confidence interval (CI). Items with MNSQ greater than 1.0 or outside the CI need to be revised or deleted.
2. Discrimination indexes. Each item in the test has one discrimination index. This is calculated through correlating the person's score on each item to his/her total score on the questionnaire. In general, the higher the discrimination index, the better the item is able to discriminate individuals based on their level of competence. In this study, only items with a discrimination index of 0.4 and higher were included. This measure supports the distinct performance of individual teachers in any of the five standards. Indicators with high discriminating power can discriminate between different stages and professional learning needs of a very diverse teacher cohort.

3. Category disordering. Ideally, the levels of performance in each criterion are developed based on the demands or difficulty of the performance. Category disordering happens when the scores assigned to each category do not correspond to the level of ability (Linacre, 1999), that is, when high performing individuals choose the lower response category and the low performing individuals choose the high response category. This may happen due to the lack of clarity of the descriptions of the standards so participants interpret the descriptions differently.

Apart from the fit mean square values described above, there are two other measures which indicate category disordering. First, the point-biserial correlation, which is mathematically equivalent to product moment correlation should increase along with the values of the response categories. In the event that the two adjacent categories behave erratically, then category disordering is evident. Second, the average measure value also indicates fit index. Good items have increasing average measure values as the response category increases. If the average measure value for the higher response category is lower than its associated lower categories, then category disordering is said to have occurred. Inspection of the item characteristic curve (ICC) provided a visual cue for category disordering. In general, as the level of ability increases, the probability of success for lower response categories decreases whilst the probability of success for higher response categories increases in an orderly fashion. In other words, the observed ICC of each score category should be approaching its theoretical ICC. If any of the ICCs are flatter than the theoretical ones, then category disordering has occurred.

Aside from these measures, the reliability index was established utilising Cronbach's alpha. This is a measure of internal consistency, which is defined by the average correlation of all items in a test or scale. Cronbach's alpha is the average of all possible split-half estimates. In principle, the split half correlation is calculated by randomly splitting the items into two groups and the correlation of the scores between these groups is then determined. Cronbach's alpha then takes the average of all correlation coefficients from all possible split half-estimates.

4.7. Factor Analysis

4.7.1. Overview

The succeeding analyses used the same estimation, factor rotation and criteria for interpretation and were carried out using Mplus v7 (Muthén & Muthén, 1998-2012).

The weighted least squares means and variance adjusted (WLSMV) estimation was used as it outperforms other methods of estimation when dealing with categorical or ordinal data (Muthén & Muthén, 1998-2012; T.Brown, 2006). This estimator works well if the sample size is 200 or more (Muthén, du Toit, & Spisic, 1997; Flora & Curan, 2004; Rhemtulla, Brosseau, & Savalei, 2012) and it measures more precisely the magnitude of factor loadings than any other estimations (Beauducel & Hersberg, 2006). In this study, the WLSMV is the perfect choice because the data used is ordinal and the sample sizes for each analysis are greater than 300. Also, the geomin oblique rotation was used in all analyses because of the strong theoretical rationale that all items in the tool are correlated as they measure one specific construct. The geomin rotation (Yates, 1987) is advantageous with variable complexity greater than one (McDonald, 2005) and simulation studies of

Asparouhov and Muthen (2009) revealed that it is the most useful to discover factor structure when little is known about the true loading structure and also it can handle simple to complicated loading matrix structures.

From the result of the analysis, the following indexes were available in Mplus results: chi-square statistics, the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardised root mean square residual (SRMR) for EFA and weighted root mean square residual (WRMR) for CFA and exploratory structural equation modelling (ESEM). These indexes used have been proven through simulation and actual use of data to perform reasonably well with categorical and ordinal model estimation (Beaudacel & Herzberg, 2006; Hutchinson & Olmos, 1998; Muthen & Muthen, 1998-2012; Yu & Muthen, 2002)

The result of the chi-square was the initial measure used to evaluate if the pattern of co-variation of the model generated was the same amongst the observed variables. Non-significant chi-square results ($p > .05$) indicate no statistical differences between the original covariance matrix and the covariance matrix of the proposed model, indicating a good model fit to the data. However, chi-square tests are typically not an optimal measure of model fit as they tend to be greatly influenced by sample size (i.e., showing significance with large sample size) and the reconstructed covariance matrix does not usually produce a perfect fit to the original covariance matrix (Browne & Cudeck, 1993). Hence, chi square tests are not powerful enough to conclude that the observed statistical significance is due to the nature of the model fit because significant values might be due to the sample size. This is the reason why other fit indexes are used to evaluate the fit of the model generated.

Another measure used was the root mean square error of approximation (RMSEA) developed by Steiger and Lind (1980). The RMSEA, which takes into account the complexity of the proposed model to explain the covariance between the variables, is relatively insensitive to sample size compared to chi-square statistics. This is a parsimony-corrected index that includes a built-in penalty in the calculation of the square root of the average of the covariance residuals. To interpret the RMSEA values, $RMSEA \leq .05$ means a perfect fit of the proposed model to the observed data; values between .05 to .08 indicate a reasonable fit; and values higher than .08 indicates a poor fit with corresponding 90% confidence interval (Hooper, Coghlan, & Mullen, 2008).

In addition, the CFI developed by Bentler (1990), was used to further evaluate the fit of the model. It is an incremental fit index that compares the proposed model with some alternative model in which the variables are assumed to be uncorrelated, and unlike the chi-square, CFI is not affected by sample size. This index measures the difference between the observed and predicted covariance or correlation matrices. The interpretation of the CFI value is guided by the following thresholds: $CFI \geq .95$ indicates a perfect model, $CFI \geq 0.90$ indicates good fit whilst $CFI < .90$ indicates a poor fitting model (Hu & Bentler, 1999; Tabachnick & Fidell, 2007).

To ascertain that underestimation or overestimation was not a factor in the proposed model, the TLI or Non-normed fit index (NNFI) was considered (Tucker & Lewis, 1973). This index is the dividend of the difference between the relative chi-square vales of the proposed and independence model divided by the relative chi-square value of the independence model subtracted by 1. The relative chi-square values are the dividends of the models' corresponding chi-square values divided by their respective degrees of freedom.

Values greater than .90 are indicative of a good model whilst $TLI \geq .95$ indicates a perfect model (Hu & Bentler, 1999).

Lastly, the SRMR was considered for EFA, which is the measure of the square root of the discrepancy between the sample covariance matrix and the model covariance matrix (Hooper, Coughlan & Mullen, 2008). Alternatively, for CFA, the WRMR is used, which is a relatively new fit index that is more superior measure for categorical data (Muthen & Muthen, 1998-2012; Yu, 2002). It measures the weighted average difference between the sample and estimated population variances and covariances (Yu, 2002). Both SRMR and WRMR should be less than 1.0 for a good model (Hancock & Mueller, 2006; Yu & Muthen, 2002).

4.7.2. Exploratory Factor Analysis

The revised questionnaire after pilot testing was administered to another group of 325 teachers. The names of the teachers were checked by the personnel from the Ministry of Education (MoE) of Brunei to ensure that nobody from the group had participated in the pilot testing or in the earlier phase of the study. This was done to avoid contamination of the data.

The 325 teachers were gathered by the MoE, through OIB, to a lecture theatre. Before the teachers engaged in self- assessment using the teacher AfL literacy tool, the criteria and standards were discussed to ensure their understanding of AfL practices. They were then asked to reflect on their own classroom practices using the teacher AfL literacy tool and were asked to mark the standard for each indicator that best described their current performance.

The responses were coded as 1 for A, 2 for B, 3 for C, 4 for D, 5 for E, and 9 for missing responses. Data cleaning was performed to check for outliers before EFA was conducted. The factors extracted by EFA using the syntax in Appendix F were validated using a more subjective criterion, the theoretical interpretability. This takes careful and thorough analysis of the items comprising each factor to make sure that there is consistency and similarities across these items in each factor. In cases where there are items that are incoherent with other items, these items must be removed. This is an important step in factor analysis because a “meaningful application of factor analysis is unthinkable without theory” (Pedhazur & Schmelkin, 1991, p. 124). Coupled with this interpretation is naming the factors based on the similarities of the items. As noted by Henson and Roberts (2006), ‘the meaningfulness of latent factors is ultimately dependent on researcher definition’ (p. 396). The description provided for each of the factors will define the usefulness of the variables.

In a more traditional view of exploratory factor analysis (EFA) there are measures that are critical for EFA. First, the correlations between each item given by the correlation matrix should be determined. This is the preliminary step to ensure that there is no multicollinearity observed in the data set. Items that are significantly highly correlated to each other are said to measure the same variable and if these are observed in the data set, either one of the items should be removed. Second, the suitability of the data for factor analysis should be established. The first test to consider for factorability is the Kaiser-Meyer Olkin (KMO) Measure of Sampling Adequacy which gives the relationship of partial correlations to the sum of squared correlations. The preferred value is greater than or equal to 0.50 which signifies the presence of factors. Second, the Bartlett’s Test of Sphericity should be

significant ($p < .05$) (Hair et al., 2010; Tabachnick & Fidell, 2007). This accounts for the probability that there is no correlation observed in the matrix.

Following this, the number of factors will be determined. Multiple extraction methods will be used to determine the number of criteria. These include the Guttman-Kaiser's criteria using eigenvalues equal to 1, the Scree plot test, the cumulative percent of variance extracted and theoretical interpretability. The eigenvalues are an important measure because the amount of variance in the data accounted for by the factors is directly proportional to their values. The relative importance of the factor extracted is given by the large eigenvalues.

Investigation of the Scree plot provides a graphical way of examining how many factors are present by looking at the point of inflexion on the graph. The number of factors must be counted from the point where the curve of the line changes abruptly (Cattell, 1966). Aside from the Scree plot test, several authors suggest a different cut off percentage of the total variance accounted for by the factors identified. Field (2009) and Rietveld and Van Hout (1993) suggest that the factors should account for about 70-80% of the variance whilst Hair et al. (2010) suggest 50-60%.

In a more recent use of EFA, new statistical programs give better estimations. In this study, I used the Mplus v7 software and some of the previously described tests for factorability were not considered. For example, the variance explained was not determined because factor analysis in Mplus attempts to reproduce correlations rather than maximising variance (Jöreskog, 1969; Jöreskog & Sörbom, 1979). In this case, the eigenvalues do not relate to the percent variance of the data (B. O. Muthén, 2000). Hence, the variance explained and eigenvalues become irrelevant in factor analysis.

4.7.3. Confirmatory Factor Analysis

Another round of data gathering was done for confirmatory factor analysis. Three hundred fifty four teachers in Brunei expressed their interest to participate in this stage. The same process was followed as in the EFA procedure above from recruitment of participants up to the actual teacher self-assessment. The six factors derived from EFA were loaded onto CFA model using the new data set (see syntax in Appendix G). The results were analysed using the fit indexes described earlier. Apart from the fit indexes, there were other measures considered in evaluating the CFA model. These include the following:

1. Unstandardized factor loadings. The ratio of the estimates to their standard errors should have absolute values greater than 1.96 to indicate significant at 0.05 level.
2. Standardised factor loading. Although there is no cut-off value for an acceptable standardised factor loading, the rule of thumb is that it should be high enough to indicate that each indicator is highly associated to the other indicators comprising each factor (Nunnally & Bernstein, 1994). The high correlations observed amongst indicators support the convergent validity of the model.
3. Discriminant validity. This measures how distinct each factor is from the other factors. This is given by the correlations amongst factors and as a rule of thumb, the lower the correlations amongst factor, the more acceptable the model.

4.7.4. Higher Order Factor Analysis

This analysis explores if the six factors derived from EFA and further verified using CFA are loading into a higher order factor. The same data set used in CFA was used for this second-order factor analysis using the Syntax in Appendix H. The six factors were loaded onto a second-order general factor denoted by G_f . The same fit indexes and other measures were used to evaluate the model fit.

4.7.5. Exploratory Structural Equation Modelling

Current literature in factor analyses argues that using confirmatory factor analysis is problematic for various reasons, as argued in Chapter 3. The limitations of CFA (e.g., model re-specification, inflated factor correlations, ignoring cross loadings) are addressed by using exploratory structural equation modelling (ESEM). This analysis was used to provide further evidence of the existence of the 6-factor model of teacher A/L literacy. The same data set used in CFA was used for this analysis using the Syntax in Appendix I, and the model was evaluated using the same fit indexes and other measures discussed earlier.

4.8. Latent Profile Analysis

Latent profile analysis (LPA) was performed to identify and empirically test groupings of teachers with similar A/L literacy profile (McCutcheon, 1987). The same data set used in CFA and second-order CFA is used in LPA modelling using the Syntax in Appendix J.

To determine the number and pattern of classes existing in the data, both statistical information criteria and likelihood-based results were used such as Akaike information criteria (AIC) developed by Akaike (1987), Bayesian information criteria (BIC) by

Schwartz (1978), and sample-size adjusted BIC (Schlove, 1987). In cases where results of these criteria were inconsistent in terms of showing which one was the best model, the findings of Yang (2006) and Nylund et al. (2007) were used as they have demonstrated that the sample-size adjusted BIC is superior to the other two criteria. Although no absolute rule of thumb has been proposed as a model evaluation guideline, LPA model with the lowest values in AIC and BIC indicate a better fit, provided that the model has a good theoretical interpretation. Additional model fit indexes used include the Vounag-Lo-Mendell-Rubin likelihood ratio test (VLMR, Vuong, 1989), the Lo-Mendell-Rubin adjusted test (LMR, Lo, Mendell, and Rubin (2001). The bootstrapped parametric likelihood ratio test (BLRT) was also considered, which uses bootstrap samples to determine the *p*-value (McLachlan & Peel, 2000). Nylund et al. (2007) have demonstrated that amongst the likelihood tests, the BLRT is the most consistent and reliable indicator of the latent classes. A widely used practice to determine the number of classes in the LPA is to compare the fit indices between the two adjacent models with *K*-classes and (*K*-1) classes. A significant *p*-value in these three tests (VLMR, LMR, and BLRT) indicates that the model with (*K*-1) classes is rejected and the *K* model gives the exact number of classes of participants.

4.9. Item Characteristics of the Teacher A/fL Literacy Assessment Tool

The data used for CFA was also used for final generalised partial credit analysis to determine the characteristics of the final items comprising the teacher A/fL assessment tool. In addition, the hierarchy of indicators and standards was established using the threshold level of the step parameters given by the item delta. The item map parameters derived from the analysis was converted to hierarchy of teacher A/fL practices by mapping the criteria

based on their placement in the latent map. The resulting hierarchy map provides an overall overview of the items ranging from the easiest to the most difficult items. Similarly, the plot of the step parameters of standards was used to determine the clusters of performance, which relates to the different levels or stages of teacher AfL literacy development. A normal distribution of the items is desired to indicate that the assessment tool developed can measure over a broad range of assessment literacy levels of teachers ranging from the easiest AfL skills to the most difficult one.

The identification of clusters to establish the cut off points for different levels of AfL literacy was based on the two criteria used by Bateman and Griffin (2003). First, the items that cluster together on the same difficulty range were identified; and second, the common theme of the clustered items was determined by looking at their similarities or common thematic lines. Another consideration was that the clusters of items exhibit demand of ability level with their adjacent level. An arbitrary limit for each level of competence was used to form the groups of the items that define novice to expert teachers in terms of their AfL competencies.

4.10. Validation of Results

A validation of the results was conducted for two purposes. First, to gather further evidence of the trustworthiness of the teacher AfL assessment tool developed and second, to determine the applicability and transferability of the teacher AfL literacy framework for different teachers from other contexts.

The same processes of data gathering were followed in the Philippine context except that permission to recruit teachers was done at the Division level. Two hundred

ninety seven teachers participated in the self-assessment using the teacher AfL literacy tool. The data was used to conduct CFA, second-order factor analysis, ESEM and LPA analysis using the same methods in the main study.

4.11. Summary

In this chapter, I have described and justified the methods used in this study. The theoretical approach used both the results of literature review and the initial workshop conducted with principals, deputy principals, assessment coordinators, teachers and higher officials from the Ministry of Education (MoE) of Brunei was followed by a series of validations and a pilot study. The empirical approach used extensive factor analyses, which included second-order factor analysis and exploratory structural equation modelling. Additional analysis included latent profile analysis to establish a typology of teachers' AfL literacy. The generalised partial credit model of the Rasch model was used to determine the item characteristics of the tool and the hierarchy of teacher AfL standards.

CHAPTER 5 - RESULTS OF THE PRELIMINARY STEPS AND THE PILOT STUDY

5.1. Introduction

The aim of this chapter is to describe and discuss the results of the validation and the pilot study. This stage of the research was conducted to determine the sufficiency, reliability and fit of the items in the tool to account for teacher A/fL literacy, to establish the feasibility of the research method, and to get a snapshot of the possible results of the study. To achieve this aim, I describe how the results were used to inform decisions regarding the revisions of the teacher A/fL literacy tool and other critical steps in the main study.

This chapter is organised as follows: firstly, the results relating to the identification of indicators and performance criteria, then the expert validation, and finally the results of the pilot testing which includes the results and discussion of the reliability index, item fit, category disordering, item discrimination and item difficulty.

5.2. Identification of Indicators and Performance Criteria

The participants (30 teachers, six principals/deputy principals, five assessment coordinators, four department heads, five Ministry of Education higher officials) in the initial three-day workshop in Brunei identified the indicators of teacher A/fL literacy and established the five levels of standards for each indicator identified. Initially, 98 indicators were generated. After crosschecking, similar indicators were merged and the total number of items was narrowed down to 72 items to develop the first draft of the teacher A/fL literacy tool. These indicators were finalised by matching to the preliminary indicators identified from the literature. To be consistent with the principles of A/fL, the indicators are

referred to as criteria.

All criteria used five response categories, indicating teacher assessment literacy levels. Levels were named as A,B,C,D, and E to ensure that these labels did not bias the workshop participants in terms of describing the actual teacher performance. These levels of performance were described to provide a clear picture of what teachers can actually do in each level. The use of descriptions rather than a numerical scale enhances the concept of assessment literacy as Taylor (2009) argues that “the concept of assessment literacy could be expanded to describe the level of knowledge, skills and understanding of assessment principles and practice” (p. 24). The final criteria and standards included in the teacher AfL literacy tool are presented in Chapter 6.

5.3. Expert Validation

Apart from typographical errors and sentence construction, the four expert evaluators suggested the following:

1. Descriptions of literacy levels for each item should be double checked to ensure that the levels follow a growth continuum. Each level should follow the guide in establishing the descriptions of standards to make a clear distinction between levels.
 - Level A - teachers do the basic requirement of AfL,
 - Level B - teachers perform at acceptable standards,
 - Level C - teachers use various strategies,
 - Level D - teachers assess and evaluate assessment practices, and
 - Level E- teachers provide expert leadership.

2. Descriptions of performance should be made explicit. Terms like appropriately, considerably, better, and so on should be replaced with more specific descriptors.
3. Items related to the trustworthiness of classroom assessment and using assessment to motivate students to learn should be revised to provide further clarity.
4. Directions related to the use of the assessment tool for self and peer assessment should be included in the final copy.
5. The purpose of the tool should be briefly explained to give teachers the sense why they need to engage in self-assessment. It should relate to the purpose of the study and to their professional development needs.
6. It should be communicated to the teachers that Level 5 describes the roles of lead teachers. It is assumed that teachers who are at this level are regularly engaged in developing a portfolio of effective assessment practices.

All these recommendations were taken into account in the final revision of the teacher AfL literacy tool. After incorporating all the recommendations given by the experts the revised tool was given back to the original participants for comments and suggestions.

To further enhance the content validity of the tool, 10 AfL coordinators in Brunei were asked to rate individual items on the tool using a 4-point scale (*irrelevant, somewhat relevant, quite relevant, highly relevant*) following the recommendation of Lynn (1986) to avoid a neutral point for rating. The content validity indexes (CVI) of individual items were computed as the number of raters who gave a rating of either 3 or 4 divided by the total number of raters. The results are presented in Table 5.1.

Table 5.1

The Content Validity Indexes of Teacher AfL Skills included in the Tool

Indicators of Teacher's AfL Literacy	I-CVI
V1 Develops appropriate assessment strategies	1.0
V2 Identifies appropriate assessment tasks	0.8
V3 Designs assessment tasks	1.0
V4 Uses a range of assessment methods	1.0
V5 Gathers a range of evidence of students' learning	1.0
V6 Uses strategic questioning to explore students' learning	0.8
V7 Uses rubrics to assess students' learning	1.0
V8 Records evidence of student's learning progress and performance	1.0
V9 Considers factors that affect students' performance	0.9
V10 Avoids interference in task completion	0.9
V11 Minimises unintended consequences of assessment	0.7
V12 Engages in social moderation	1.0
V13 Uses standards/criterion referencing in interpreting performance	0.9
V14 Uses assessment information for varied purposes	0.8
V15 Identifies students strengths and weaknesses	0.8
V16 Communicates assessment results to students	1.0
V17 Gives feedback on students' strengths and weaknesses	1.0
V18 Gives feedback related to criteria	1.0
V19 Assists students in using feedback to feed forward	0.9
v20 Guides students in goal achievement	1.0
V21 Provides guidance to close learning gaps	0.9
V22 Ensures that assessment does not harm students	1.0
V23 Ensures that assessment results reflect students' "true" learning	1.0
V24 Maintains confidentiality in dealing with assessment results	1.0
V25 Translates learning standards (curriculum content) to learning outcomes	1.0
V26 Uses different sources of information in developing learning outcomes	1.0
V27 Involves students in the development of learning outcomes	1.0
V28 Identifies appropriate teaching methods	1.0
V29 Considers students' prior knowledge in lesson planning	0.9
V30 Considers students' current level of abilities	0.9
V31 Considers students' interest	0.9
V32 Plans lessons according to students' learning needs	1.0
V33 Tailors lessons to available resources	1.0
V34 Develops teaching and learning resources	1.0
V35 Makes students understand the learning outcomes	0.9
V36 Discusses feedback with students	1.0
V37 Involves students in the development of criteria and standards	1.0
V38 Uses a variety of teaching and learning methodologies	1.0
V39 Explains the criteria and standards	1.0
V40 Develops students' capabilities in self and peer assessment	0.9

V41 Engages students in self-assessment	1.0
V42 Engages students in peer-assessment	1.0
V43 Moderates feedback and results of self and peer assessment	1.0
V44 Makes informed decisions in adapting teaching	1.0
V45 Embeds assessment in teaching	1.0
V46 Uses flexible teaching activities	0.9
V47 Respects individual learners	0.9
V48 Develops an environment of trust	1.0
V49 Ensures openness in the class	1.0
V50 Builds students' interest to learn	1.0
V51 Demonstrates belief in the ability of every student to improve	1.0
V52 Sets high expectations of achievement	0.8
V53 Affirms students' good performance	1.0
V54 Clarifies students misconceptions	1.0
V55 Reinforces positive learning attitude of students	0.8
V56 Seeks out new and relevant assessment knowledge	0.8
V57 Forms a community of teacher learners	0.9
V58 Participates in professional development related to assessment	1.0
V59 Engages in self-assessment/ reflection	0.9
V60 Engages in peer-review of teaching performance	0.9
V61 Seeks feedback from colleagues/experts about their performance	1.0
V62 Identifies subject-content knowledge needs	0.9
V63 Searches new and relevant subject-content information	0.9
V64 Undertakes further education/ training	1.0
V65 Informs parents/ guardians about students' learning	0.8
V66 Collaborates with family to establish family support activities	1.0
V67 Informs community of the assessment practices	1.0
V68 Reports to community about students' performance	1.0
V69 Reports the overall performance of students for accountability	1.0
V70 Identifies key assessment and teaching issues for review	1.0
V71 Analysis and uses information from stakeholders	1.0
V72 Incorporates findings into the lesson plan	0.9
Scale-CVI/Ave	0.95

It can be seen from Table 5.1 that all identified indicators of teacher's AfL literacy have a content validity index (CVI) above the cut-off score of 0.78 (Lynn, 1986). The scale-content validity index was also calculated based on the arguments and recommendations of Polit and Beck (2006) that the scale-CVI is best calculated by determining the average of the I-CVI, which can then be interpreted as the average item

quality. From the table, the scale-CVI/Ave is 0.95, which is higher than the threshold value of 0.9. With these results, all 72 items were included in the succeeding stages of research.

5.4. Pilot Testing

5.4.1. Overview

The results of the self-assessment of 197 teachers in Brunei were analysed to establish the reliability index, item fit, item difficulty, and construct validity and to determine if category disordering existed for any of the indicators. The results of the pilot testing related to the measurement properties of the tool are presented below.

5.4.2. Reliability Index

In the pilot study, the teacher AfL literacy tool had a Cronbach's alpha coefficient of .98. The reliability index was exceptionally high which accounts for the internal consistency of the assessment tool (George & Mallery, 2005), and this suggested that each item in the tool had a strong AfL component (i.e., there was a strong internal consistency in the scale). Also, it indicated that the tool could give consistent measures of individual teachers' AfL literacy at any stage of teacher AfL literacy development. However, this result did not give enough confidence that all 72 items were indeed the indicators of teacher AfL literacy. The high reliability index might have been due to the large number of items included (as the number of items directly influences the reliability index) and the presence of redundant items. Further analysis in the main study was required to ensure that the internal consistency of the tool was not just due to the number of and the presence of redundant items.

5.4.3. Item Fit

The item fit determines if the items are unidimensional to some extent or are part of the construct of teacher AfL literacy. The results in Table 5.2 show that there were only 2 items whose weighted fit mean square values were equal to 1, which indicated perfect fit with the construct. Furthermore, 57 items had a weighted fit mean square values either below or greater than 1 but were within the confidence interval with corresponding t-statistics within the range of -2 to +2 values. These results provided an initial indication that these 58 items were truly measuring the teacher AfL literacy. The remaining 14 items (highlighted) had weighted mean square values outside their corresponding confidence interval and their t-statistics were outside the range of acceptable values. These items were problematic which could either be because they were poorly constructed or because they were measuring different construct.

Table 5.2

*Item Parameter Estimates of the Rasch Model for 72 Items of Teacher AfL Literacy**(N=197)*

VARIABLES				UNWEIGHTED FIT			WEIGHTED FIT		
-----				-----			-----		
item		ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
1	1	-0.31	0.07	1.25	(0.79, 1.21)	2.2	1.24	(0.79, 1.21)	2.1
2	2	-0.29	0.07	1.31	(0.79, 1.21)	2.7	1.30	(0.78, 1.22)	2.5
3	3	0.09	0.07	1.33	(0.79, 1.21)	2.8	1.30	(0.79, 1.21)	2.6
4	4	-0.44	0.07	1.23	(0.79, 1.21)	2.0	1.23	(0.78, 1.22)	2.0
5	5	-0.35	0.07	1.11	(0.79, 1.21)	1.0	1.10	(0.78, 1.22)	0.9
6	6	-0.21	0.07	1.35	(0.79, 1.21)	2.9	1.28	(0.78, 1.22)	2.3
7	7	-0.02	0.07	1.35	(0.79, 1.21)	3.0	1.35	(0.78, 1.22)	2.9
8	8	0.04	0.07	1.02	(0.79, 1.21)	0.2	1.01	(0.78, 1.22)	0.1
9	9	0.54	0.07	1.49	(0.79, 1.21)	4.0	1.35	(0.81, 1.19)	3.2
10	10	0.50	0.07	1.06	(0.79, 1.21)	0.6	1.12	(0.80, 1.20)	1.2
11	11	-0.06	0.07	1.10	(0.79, 1.21)	0.9	1.14	(0.80, 1.20)	1.4
12	12	-0.41	0.07	1.28	(0.79, 1.21)	2.4	1.26	(0.80, 1.20)	2.3
13	13	-0.09	0.07	1.20	(0.79, 1.21)	1.8	1.21	(0.80, 1.20)	2.0
14	14	-0.46	0.07	1.16	(0.79, 1.21)	1.4	1.17	(0.79, 1.21)	1.6
15	15	-0.71	0.07	1.21	(0.79, 1.21)	1.8	1.22	(0.80, 1.20)	2.0
16	16	-0.84	0.07	1.18	(0.79, 1.21)	1.6	1.17	(0.80, 1.20)	1.6
17	17	-0.80	0.07	1.06	(0.79, 1.21)	0.6	1.01	(0.80, 1.20)	0.2
18	18	-0.21	0.07	0.95	(0.79, 1.21)	-0.4	0.96	(0.79, 1.21)	-0.3
19	19	0.09	0.07	1.00	(0.79, 1.21)	0.0	0.99	(0.80, 1.20)	-0.0
20	20	-0.08	0.07	0.99	(0.79, 1.21)	-0.0	1.01	(0.79, 1.21)	0.2
21	21	-0.03	0.07	0.98	(0.79, 1.21)	-0.2	0.98	(0.79, 1.21)	-0.1
22	22	-0.32	0.07	1.50	(0.79, 1.21)	4.0	1.40	(0.80, 1.20)	3.5
23	23	0.06	0.07	1.05	(0.79, 1.21)	0.5	1.08	(0.80, 1.20)	0.8
24	24	-0.07	0.07	1.04	(0.79, 1.21)	0.4	1.05	(0.79, 1.21)	0.5
25	25	-0.26	0.07	1.14	(0.79, 1.21)	1.3	1.15	(0.79, 1.21)	1.3
26	26	-0.49	0.07	0.90	(0.79, 1.21)	-0.9	0.91	(0.79, 1.21)	-0.8
27	27	0.13	0.07	1.10	(0.79, 1.21)	0.9	1.11	(0.80, 1.20)	1.1
28	28	-0.43	0.07	1.08	(0.79, 1.21)	0.7	1.09	(0.80, 1.20)	0.8
29	29	-0.22	0.07	0.93	(0.79, 1.21)	-0.7	0.91	(0.79, 1.21)	-0.8
30	30	0.04	0.07	0.86	(0.79, 1.21)	-1.4	0.88	(0.80, 1.20)	-1.2
31	31	0.29	0.07	1.00	(0.79, 1.21)	0.1	0.96	(0.80, 1.20)	-0.3
32	32	0.31	0.08	0.93	(0.79, 1.21)	-0.6	0.91	(0.79, 1.21)	-0.9
33	33	0.05	0.07	1.35	(0.79, 1.21)	3.0	1.35	(0.80, 1.20)	3.1
34	34	0.31	0.07	0.83	(0.79, 1.21)	-1.6	0.80	(0.80, 1.20)	-2.1
35	35	0.32	0.07	0.99	(0.79, 1.21)	-0.1	1.01	(0.80, 1.20)	0.1
36	36	0.12	0.07	0.94	(0.79, 1.21)	-0.5	0.96	(0.80, 1.20)	-0.4
37	37	-0.51	0.08	0.97	(0.79, 1.21)	-0.2	0.96	(0.76, 1.24)	-0.3
38	38	0.35	0.07	0.94	(0.79, 1.21)	-0.6	0.92	(0.80, 1.20)	-0.8
39	39	0.29	0.07	0.82	(0.79, 1.21)	-1.7	0.82	(0.79, 1.21)	-1.7
40	40	0.04	0.07	0.93	(0.79, 1.21)	-0.6	0.95	(0.80, 1.20)	-0.5
41	41	0.18	0.08	0.99	(0.79, 1.21)	-0.1	1.00	(0.76, 1.24)	0.0
42	42	0.62	0.07	1.08	(0.79, 1.21)	0.7	1.05	(0.81, 1.19)	0.6
43	43	0.88	0.07	0.94	(0.79, 1.21)	-0.5	0.95	(0.80, 1.20)	-0.5
44	44	-0.29	0.08	0.85	(0.79, 1.21)	-1.5	0.85	(0.78, 1.22)	-1.4
45	45	-0.58	0.07	1.06	(0.79, 1.21)	0.6	1.03	(0.80, 1.20)	0.4
46	46	-0.73	0.07	1.00	(0.79, 1.21)	0.1	1.01	(0.77, 1.23)	0.1
47	47	-0.49	0.07	1.04	(0.79, 1.21)	0.4	1.09	(0.81, 1.19)	1.0
48	48	-0.24	0.07	0.92	(0.79, 1.21)	-0.7	0.92	(0.78, 1.22)	-0.7
49	49	-0.59	0.08	0.87	(0.79, 1.21)	-1.2	0.88	(0.79, 1.21)	-1.2
50	50	-0.24	0.07	0.83	(0.79, 1.21)	-1.6	0.84	(0.80, 1.20)	-1.6
51	51	0.04	0.07	0.85	(0.79, 1.21)	-1.4	0.85	(0.80, 1.20)	-1.5
52	52	-0.14	0.07	0.84	(0.79, 1.21)	-1.5	0.85	(0.80, 1.20)	-1.6
53	53	-0.31	0.08	0.81	(0.79, 1.21)	-1.8	0.81	(0.79, 1.21)	-1.8
54	54	0.70	0.07	0.96	(0.79, 1.21)	-0.4	0.98	(0.79, 1.21)	-0.2

55	55	0.25	0.07	1.09 (0.79, 1.21)	0.8	1.08 (0.80, 1.20)	0.8
56	56	0.20	0.07	1.07 (0.79, 1.21)	0.7	1.00 (0.80, 1.20)	-0.0
57	57	0.02	0.07	0.90 (0.79, 1.21)	-0.9	0.91 (0.80, 1.20)	-0.9
58	58	-0.07	0.07	0.95 (0.79, 1.21)	-0.5	0.95 (0.79, 1.21)	-0.5
59	59	-0.20	0.07	1.01 (0.79, 1.21)	0.2	1.05 (0.77, 1.23)	0.4
60	60	0.43	0.07	0.94 (0.79, 1.21)	-0.5	0.97 (0.80, 1.20)	-0.3
61	61	0.01	0.07	1.01 (0.79, 1.21)	0.2	1.03 (0.79, 1.21)	0.3
62	62	-0.84	0.07	1.44 (0.79, 1.21)	3.6	1.42 (0.80, 1.20)	3.7
63	63	0.04	0.07	0.82 (0.79, 1.21)	-1.7	0.85 (0.80, 1.20)	-1.6
64	64	0.39	0.07	0.88 (0.79, 1.21)	-1.1	0.92 (0.80, 1.20)	-0.8
65	65	0.93	0.07	1.08 (0.79, 1.21)	0.7	1.09 (0.79, 1.21)	0.8
66	66	0.94	0.07	1.19 (0.79, 1.21)	1.7	1.19 (0.80, 1.20)	1.8
67	67	0.67	0.07	0.96 (0.79, 1.21)	-0.3	0.99 (0.79, 1.21)	-0.0
68	68	0.75	0.07	1.06 (0.79, 1.21)	0.6	1.08 (0.79, 1.21)	0.8
69	69	1.17	0.07	1.08 (0.79, 1.21)	0.8	1.11 (0.80, 1.20)	1.1
70	70	0.53*	0.59	1.03 (0.79, 1.21)	0.3	1.04 (0.80, 1.20)	0.4
71	71	-0.07	0.07	0.96 (0.79, 1.21)	-0.5	0.96 (0.79, 1.21)	-0.6
72	72	-0.58	0.07	1.02 (0.79, 1.21)	0.6	1.04 (0.80, 1.20)	0.5

$\chi^2 = 2477.62$, $df = 69$, $p = 0$; $MNSQ$ = weighted mean square; DI – confidence interval; t – t-statistics

5.4.4. Category Disorder

To support any decisions about the 14 problematic items, further investigation was undertaken. The point-biserial correlation coefficients for all response categories for every item were analysed. Item with response category, which has a lower point-biserial correlation coefficient than the preceding lower response category showed category disordering. This happens when high performing teachers rated themselves lower than low performing teachers for that particular item or criterion. This is a case where the performance described in lower category is harder than that required in the higher response category, or the description provided is unclear, hence, results in variable interpretations (See Figure 5.1 for example). Investigation of the point-biserial correlations of the 14 items showing misfit showed category disordering. The category disordering was observed between Levels D and E. This might be due to the descriptions of Level Es about teachers providing professional advice or support to their colleagues, which can be interpreted by other participants as their competence level for these 14 criteria if they provide any support to their colleagues even if they are not meeting Level Ds.

Item 2

item:2 (2)
Cases for this item 168 Discrimination 0.48
Item Threshold(s): -2.64 -1.54 0.49 2.52 Weighted MNSQ 1.30
Item Delta(s): -2.27 -1.81 0.53 2.39

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
1	0.00	14	8.33	-0.30	-4.03 (.000)	-1.63	1.25
2	1.00	29	17.26	-0.28	-3.74 (.000)	-1.08	0.97
3	2.00	78	46.43	0.05	0.65 (.517)	-0.41	0.79
4	3.00	39	23.21	0.31	4.15 (.000)	0.15	1.19
5	4.00	8	4.76	0.16	2.06 (.041)	0.29	1.43

=====

The pt-bis correlation of category 4 is higher than category 5.

Figure 5.1. Example of an item that shows category disordering.

5.4.5. Item Discrimination

Another measure considered was the item discrimination index to define the power of the item to distinguish high performing from low performing teachers in terms of their AfL literacy. In general, the higher the discrimination index the better the item. This is a critical measurement property of each item in the tool, which would enable the measurement of the actual literacy level of teachers across all criteria, and hence a needs-based professional development program could be developed. From the results of the analysis, 59 items had high discriminating indexes ranging from .48 to .73 whilst the 13 problematic items had discrimination indexes ranging from .12 to .38. The low discrimination indexes of the 13 problematic items were due to the fact that they exhibited category disordering between adjacent levels of performance.

The Level E descriptions of these 14 criteria were revised, following the descriptions of performance and using more explicit wording to clarify the performance required. The revised 14 items were given back to assessment experts for a second round of

validation to make sure that the descriptions of performance are clear and are not prone to any misinterpretations.

5.4.6. Difficulty Index

The last measure taken into consideration was the difficulty index, which was translated into polytomous scale as item thresholds. Figure 5.2 shows that the items were distributed across the continuum of difficulty and approximated a normal curve. The standards were distributed from -4.56 to 4.29 logits, which is a sufficient range for item difficulty (Aldrich & Styles, 2002), indicated a well-developed tool (Hendricks, Fyfe, Styles, Skinner, & Merriman, 2012). Similarly, standards A for all criteria, and the rest of the standards B, C, D, and E, clustered in a particular logit range limits (Bateman & Griffin, 2003), providing evidence that levels A, B, C, D, and E standards required a particular set of skills.

These results validated the work of the experts in ensuring that items were properly distributed along the continuum of difficulty, which is central to the application of the tool for professional development. The spread of the items along the continuum supports the main purpose for developing this tool, that is, to measure individual teachers' varying level of A/L literacy.

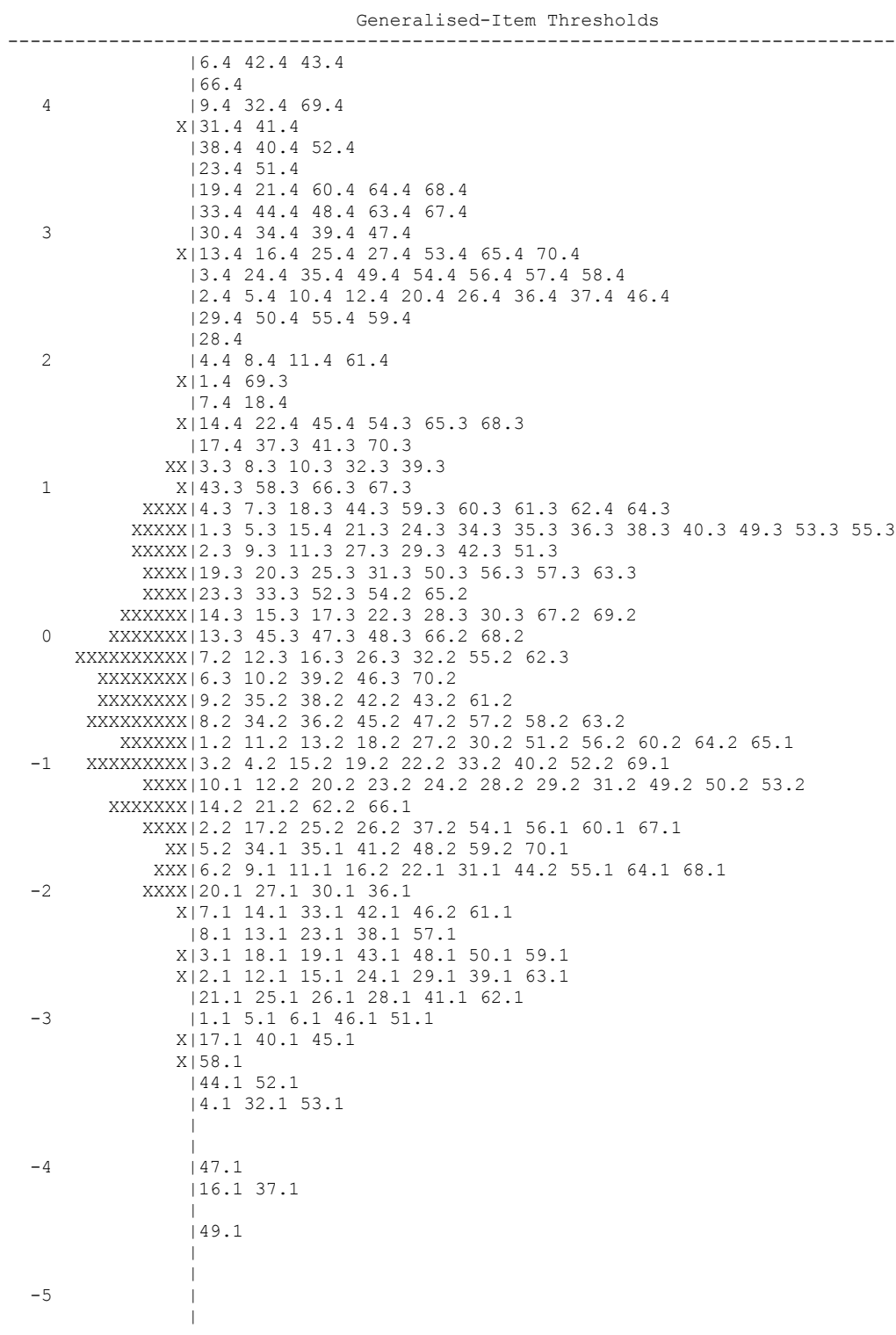


Figure 5.2. Map of latent distributions of teacher A/L practices based on difficulty level, $X = 1.3$ sample

5.5. Insights for the Main Study

Drawing from my experience in conducting and from the results of the pilot testing, there are some significant insights that were considered in the main study. There were no logistical issues as the Ministry of Education of Brunei through the Optimise International Brunei had provided a high level of support in all aspects of the research project. However, in terms of using the tool, the following emerged as issues:

1. There are some criteria and standards that draw several understanding and interpretations. Hence, in the main study, all individual criteria and standards should be explained to ensure common understanding amongst teachers.
2. Levels D and E are prone to confusions. During the data gathering for the main study, it should be emphasised that Level E accounts for teachers' active role in providing professional support and advice to colleagues but it is necessary that teachers in this level are doing or have accomplished Level D performance.
3. Verbal instruction on how to do self-assessment using the tool should be communicated well. The recommendation of the expert validators on the inclusion of the purpose of the self-assessment and directions on how to use the tool should be included in the final teacher A/L literacy tool.

5.6. Summary

This chapter has presented the results of the pilot study. It can be seen that the teacher AfL literacy tool met the measurement requirements of a scale, except for the 13 items that were problematic, which were subsequently revised. These initial results indicated that the tool could potentially be used to measure teacher AfL literacy and could provide robust information about individual teacher's level of AfL literacy.

CHAPTER 6 - RESULTS OF THE MAIN STUDY

6.1. Introduction

The results of the main study are presented in this chapter. These include the results of exploratory factor analysis (EFA), confirmatory factor analysis (CFA), second-order factor analysis, and exploratory structural equation modelling (ESEM), which inform the decisions related to determining the dimensionality of teacher AfL literacy. In addition, I include the results of latent profile analysis (LPA), which establishes a typology of teachers' assessment profile. Lastly, I present the results of the generalised partial credit modelling of item response theory to establish the item characteristics of the teacher AfL literacy scale to explore the hierarchy of teacher AfL performance.

6.2. Population Sample

The demographic profile of the teachers who participated in this study is shown in Table 6.1. Limited information was allowed to be documented to protect the privacy and interest of the teachers in Brunei (*a prior agreement between me and the Ministry of Education of Brunei was made in relation to the limited access of data*). As most schools in Brunei, both primary and secondary, are government owned, most of the participants came from public schools. However, looking at the percentage of participants from private and public schools, there were equal representations from each sector.

Table 6.1

Demographic Profile of Participants in the Main Study

Analysis	Participants								Total (N=8066)	Percent
	Public School				Private School					
	Primary (N=3008)		Secondary (N=3564)		Primary (N=1030)		Secondary (N=464)			
	Count	Percent	Count	Percent	Count	Percent	Count	Percent		
EFA	170	5.65%	127	3.56%	10	0.97%	18	3.88%	325	4.03%
CFA/ESEM	156	5.19%	157	4.41%	15	1.46%	26	5.60%	354	4.39%
Total	326	10.84%	284	7.97%	25	2.43%	44	9.48%	679	8.42%

6.3. Results of Exploratory Factor Analysis

6.3.1. Overview

This section, along with the succeeding factors analyses, answers the second research question, that is, whether there were distinct factors/subdomains of teacher's AfL literacy. The answers to this question are presented first before the answers to the first research question because this section provides empirical evidence for the reliability and validity of the tool derived from both practical insights and theory. In all succeeding factor analyses, the chi-square statistic along with its degrees of freedom and *p*-value, the root mean square error of approximation (RMSEA) and its associated confidence interval (CI), the comparative fit index (CFI), and Tucker-Lewis index (TLI) were used to evaluate the model fit due to their less sensitivity to sample size, model misspecification and parameter estimates (Hooper, Coughlan, & Mullen, 2008).

6.3.2. Model Test Statistics

Results of exploratory factor analysis of the data derived from the self-assessment of 325 Bruneian teachers using the teaches AfL literacy tool using geomin oblique rotation (Asparouhov & Muthen, 2009) and weighted least squares means and variance adjusted (WLSMV) estimator extracted up to a 10-factor solution. The root mean criterion and theoretical interpretability were considered in deciding which factor model fits the data well. As can be seen in Table 6.2, except for the chi-square statistics, which were significant for all factor solutions, all fit statistics of all factor models met the requirements for goodness of fit. Although significant chi-square values indicate unfit models, Kline

(2005, 2010) suggests to further evaluate these values against their respective degrees of freedom where a 3:1 proportion would still indicate a good fit. In all 10 models extracted, the ratios of chi-square values and degrees of freedom met this requirement. Hence, all models extracted were possible models to explain the factor structure of teacher Afl literacy.

In terms of deciding which model fitted the data well, a closer investigation of the fit statistics revealed that the 6-factor model had the lowest RMSEA (0.026), highest CFI (0.966) and TLI (0.965). These values were good indications that this model was the best-fit. Moreover, the 6-factor model is further supported by the results of the SRMR which had the lowest value (0.030). This result conforms to the recommendations of Diamantopoulus and Siguaw (2000) that SRMR value less than .05 indicates a well-fitting model.

Table 6.2

Goodness of Fit Statistics of Models for Teacher AfL Literacy Derived from Exploratory Factor Analysis with Weighted Least Squares

Means and Variance Adjusted Estimation for 72 Items (N=325)

Factor	Chi-square			X^2/df	RMSEA	CFI	TLI	SRMR
	X^2	df	p -value					
1	4025.736	(2484)	< 0.001	1.621	0.043	0.903	0.900	0.063
2	3365.167	(2413)	< 0.001	1.395	0.034	0.940	0.937	0.051
3	3062.578	(2343)	< 0.001	1.307	0.030	0.955	0.951	0.045
4	2873.530	(2274)	< 0.001	1.264	0.028	0.962	0.958	0.042
5	2747.818	(2206)	< 0.001	1.246	0.027	0.966	0.961	0.040
6	2524.217	(2073)	< 0.001	1.218	0.026	0.972	0.965	0.030
7	2457.214	(2002)	< 0.001	1.227	0.032	0.942	0.940	0.520
8	2345.125	(1985)	< 0.001	1.181	0.040	0.958	0.934	0.048
9	2245.324	(1924)	< 0.001	1.167	0.038	0.943	0.936	0.052
10	2198.457	(1845)	< 0.001	1.192	0.036	0.948	0.928	0.054
11	No convergence							
12	No convergence							

X^2 = Chi-square; df = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI – Tucker Lewis index; SRMR = Standardised Root Mean Square Residual.

Investigation of the eigenvalues, which is the variance of the factor, further supported the 6-factor model. Results showed that six factors had values greater than 1 (see Table 6.3). Since this study used Mplus software, the percentage variance is not reported because the EFA in Mplus factor analysis attempts to reproduce correlations rather than maximising variance (Jöreskog, 1969; Jöreskog & Sörbom, 1979). In this case, the eigenvalues did not relate to the percentage variance of the data (B. O. Muthén, 2000).

Table 6.3

Eigenvalues of the 10 Factors

Factor	Eigenvalues
1	27.381
2	3.203
3	2.306
4	1.855
5	1.536
6	1.449
7	0.987
8	0.953
9	0.926
10	0.894

To support the claim that the 6-factor model was the best fitting model derived from EFA, the factor loadings of all models were examined. Except for some cross-loading items as shown in Table 6.4, the factor solution of the 6-factor model was best supported by the theoretical framework of this study.

Table 6.4

Summary of Exploratory Factor Analysis Results for Teacher AfL Literacy Measure Using Weighted Least Squares Means and Variance Adjusted Estimation for 72 Items (N=325)

Indicators	Rotated Factor Loading					
V1 Develops appropriate assessment strategies	0.449*	-0.014	0.027	0.001	0.017	0.294*
V2 Identifies appropriate assessment tasks	0.249*	0.044	0.050	0.015	-0.012	0.257
V3 Designs assessment tasks	0.466*	0.024	-0.083	-0.029	0.197*	0.079
V4 Uses a range of assessment methods	0.484*	0.083	0.093	-0.029	0.028	0.266*
V5 Gathers a range of evidence of students' learning	0.576*	0.112	-0.059	0.023	0.050	0.116
V6 Uses strategic questioning to explore students' learning	0.234*	0.167	-0.095	0.104	-0.051	0.281*
V7 Uses rubrics to assess students' learning	0.462*	0.129	0.034	0.017	0.062	-0.008
V8 Records evidence of student's learning progress and performance	0.592*	-0.038	0.178	0.018	0.063	-0.033
V9 Considers factors that affect students' performance	0.497*	0.032	0.208	0.033	0.060	-0.139
V10 Avoids interference in task completion	0.588*	-0.299*	0.424*	0.009	0.073	-0.128
V11 Minimises unintended consequences of assessment	0.332*	0.036	0.380*	0.015	-0.058	-0.091
V12 Engages in social moderation	0.322*	0.265*	0.077	0.094	0.042	0.021
V13 Uses standards/criterion referencing in interpreting performance	0.308*	0.024	0.346*	0.024	0.173*	-0.099
V14 Uses assessment information for varied purposes	0.243*	-0.008	0.207*	0.056	-0.101	0.084
V15 Identifies students strengths and weaknesses	0.129	0.192	0.221*	0.211*	-0.114	-0.002
V16 Communicates assessment results to students	0.283	0.067	0.391*	0.036	-0.064	0.046
V17 Gives feedback on students' strengths and weaknesses	0.251*	0.082	0.340*	0.128	0.040	0.023
V18 Gives feedback related to criteria	0.079	0.202*	0.450*	0.099	-0.005	-0.007
V19 Assists students in using feedback to feed forward	0.074	0.184	0.425*	-0.086	0.104	0.010
v20 Guides students in goal achievement	0.099	0.399*	0.354*	-0.076	0.087	-0.078
V21 Provides guidance to close learning gaps	0.036	0.424*	0.083	-0.001	0.231*	-0.099
V22 Ensures that assessment does not harm students	0.071	0.383*	-0.024	0.209*	0.151	-0.259*
V23 Ensures that assessment results reflect students' "true" learning	0.049	0.467*	0.039	0.015	0.132	0.039
V24 Maintains confidentiality in dealing with assessment results	-0.002	0.380*	0.127	0.008	0.174*	-0.130
V25 Translates learning standards (curriculum content) to learning outcomes	0.176*	0.441*	0.170	0.012	0.005	0.110
V26 Uses different sources of information in developing learning outcomes	0.023	0.544*	0.026	0.130	0.092	-0.054
V27 Involves students in the development of learning outcomes	0.089	0.279*	0.302*	0.001	0.084	0.087
V28 Identifies appropriate teaching methods	-0.038	0.643*	0.184	0.050	-0.026	-0.112
V29 Considers students' prior knowledge in lesson planning	0.135	0.571*	0.043	-0.038	0.064	0.238
V30 Considers students' current level of abilities	-0.065	0.626*	0.131	0.012	0.040	0.135
V31 Considers students' interest	0.091	0.467*	0.138	0.149	0.006	0.069
V32 Plans lessons according to students' learning needs	0.082	0.354*	0.327*	0.029	-0.013	0.077
V33 Tailors lessons to available resources	0.106	0.543*	0.127	0.087	-0.158*	0.159
V34 Develops teaching and learning resources	-0.011	0.763*	-0.088	0.227*	-0.046	-0.069
V35 Makes students understand the learning outcomes	0.063	0.276	0.244*	0.118	-0.016	0.133
V36 Discusses feedback with students	0.143	0.163	0.358*	0.197*	-0.015	0.085
V37 Involves students in the development of criteria and standards	0.091	0.017	0.592*	-0.011	0.064	-0.008
V38 Uses a variety of teaching and learning methodologies	0.023	0.279	0.076	0.541*	-0.052	-0.203*

V39 Explains the criteria and standards	0.043	0.101	0.527*	-0.001	0.166	0.012
V40 Develops students' capabilities in self and peer assessment	-0.156*	0.180	0.436*	0.100	0.250*	-0.090
V41 Engages students in self-assessment	-0.006	0.024	0.422*	0.175	0.180	0.024
V42 Engages students in peer-assessment	-0.102	0.133	0.469*	0.218*	0.058	0.012
V43 Moderates feedback and results of self and peer assessment	-0.071	0.193	0.506*	-0.070	0.237*	0.068
V44 Makes informed decisions in adapting teaching	0.016	0.211*	0.030	0.445*	0.145	0.073
V45 Embeds assessment in teaching	0.101	0.124	0.112	0.259*	0.279*	0.020
V46 Uses flexible teaching activities	0.000	0.221*	-0.033	0.448*	0.139	0.026
V47 Respects individual learners	0.143*	0.176	-0.062	0.442*	0.042	-0.047
V48 Develops an environment of trust	0.064	-0.121	0.111	0.738*	0.000	-0.022
V49 Ensures openness in the class	-0.025	0.017	0.079	0.525*	0.035	0.154
V50 Builds students' interest to learn	0.150*	0.073	0.125	0.572*	-0.093	-0.013
V51 Demonstrates belief in the ability of every student to improve	-0.015	0.100	0.039	0.585*	0.170*	-0.089
V52 Sets high expectations of achievement	0.018	0.180	0.129	0.269*	0.181*	0.120
V53 Affirms students' good performance	-0.005	0.190	-0.081	0.396*	0.195	0.245*
V54 Clarifies students misconceptions	-0.030	0.215	0.108	0.314*	0.157	0.146
V55 Reinforces positive learning attitude of students	-0.006	0.279*	0.054	0.329*	0.162	0.177
V56 Seeks out new and relevant assessment knowledge	0.122*	-0.008	0.031	0.250	0.290*	-0.030
V57 Forms a community of teacher learners	-0.018	0.003	0.005	0.316*	0.469*	0.066
V58 Participates in professional development related to assessment	0.131*	-0.076	-0.078	0.267*	0.531*	0.037
V59 Engages in self-assessment/ reflection	0.096	-0.053	-0.001	0.213*	0.457*	0.317*
V60 Engages in peer-review of teaching performance	0.010	0.067	0.079	0.055	0.507*	0.326*
V61 Seeks feedback from colleagues/experts about their performance	-0.033	0.171	0.056	0.129	0.370*	0.329*
V62 Identifies subject-content knowledge needs	0.128*	0.114	0.111	0.162	0.391*	-0.028
V63 Searches new and relevant subject-content information	-0.020	-0.019	0.101	0.312*	0.407*	0.095
V64 Undertakes further education/ training	-0.017	0.054	0.054	0.137	0.579*	-0.083
V65 Informs parents/ guardians about students' learning	0.014	-0.060	0.326*	0.261*	0.262*	0.414*
V66 Collaborates with family to establish family support activities	-0.032	-0.083	0.456*	0.038	0.0420	0.481*
V67 Informs community of the assessment practices	0.013	-0.008	0.001	-0.032	-0.062	0.699*
V68 Reports to community about students' performance	0.053	-0.055	0.051	0.026	0.120	0.649*
V69 Reports the overall performance of students for accountability	0.171*	0.076	0.060	-0.139*	-0.040	0.694*
V70 Identifies key assessment and teaching issues for review	0.009	0.356*	-0.046	0.092	-0.003	0.694*
V71 Analysis and uses information from stakeholders	0.064	0.049	0.075	-0.093	-0.075	0.701*
V72 Incorporates findings into the lesson plan	0.150*	0.360*	-0.214*	-0.039	0.005	0.460*

As can be seen in Table 6.4, there were items that were cross loadings to more than one factor, and also there were items that did not significantly load to any factor. The communality values of these items were also low, and hence, these items were subsequently deleted. Also, those items with factor loadings below .32 were deleted. This decision was based on the recommendation of Tabachnick and Fidell (2007) that for a sample size of at least 300, it would need a factor loading greater than .32 to be considered statically meaningful. The higher factor loading, which measures regression coefficients, guarantees higher direct effects of the factor to its indicators. Analysis was re-run every time an item was deleted. The final factor solution of the 6-factor model, where $\leq .32$ factor loadings were suppressed, is presented in Table 6.5.

The final EFA extracted a 6-factor model with 42 items loading significantly, and with a good subscale interpretation. The fit statistics of the 6-factor model had slightly improved compared to the previous EFA analysis, with RMSEA of 0.024, CFI of 0.976, TLI of 0.970, and SRMR of 0.030. All these values suggested that the model met the requirements for goodness of fit, and thus, it can be concluded that the 6-factor model fit the data well.

The factors extracted are described based on the latent construct of the teacher A/L skills comprising each factor:

- Factor 1 is composed of five items representing teacher assessment skills in assessing student.
- Factor 2 is composed of eight items involving teacher assessment skills in using assessment data to inform learning and teaching activities.

- Factor 3 is composed of 10 items representing teacher assessment skills in using assessment and assessment data to engage students in learning and teaching.
- Factor 4 is composed of nine items representing teacher assessment skills in using assessment to enhance student motivation.
- Factor 5 is composed of six items representing teacher assessment skills in identifying and responding to their assessment literacy needs.
- Factor 6 is composed of four items representing teacher assessment skills in working closely with parents, principals and the community.

Table 6.5

Summary of Final EFA Results for Teacher AfL Literacy Measure Using WLSMV Estimation for 52 Items (N=325)

	1	2	3	4	5	6
v3 Designs assessment tasks	0.649*					
V7 Uses rubrics to assess students' learning	0.596*					
V9 Considers factors that affect students' performance	0.455*					
V10 Avoids interference in task completion	0.654*					
V12 Engages in social moderation	0.688*					
V25 Translates learning standards to learning outcomes		0.411*				
V28 Identifies appropriate teaching methods		0.605*				
V29 Considers students' prior knowledge in lesson planning		0.700*				
V30 Considers students' current level of abilities		0.585*				
V31 Considers students' interest		0.651*				
V32 Plans lessons according to students' learning needs		0.519*				
V33 Tailors lessons to available resources		0.592*				
V34 Develops teaching and learning resources		0.849*				
V17 Gives feedback on students' strengths and weaknesses			0.590*			
V19 Assists students in using feedback to feed forward			0.431*			
V27 Involves students in the development of learning outcomes			0.382*			
V35 Makes students understand the learning outcomes			0.665*			
V37 Involves students in the development of criteria and standards			0.592*			
V39 Explains the criteria and standards			0.501*			
V40 Develops students' capabilities in self and peer assessment			0.593*			
V41 Engages students in self-assessment			0.722*			
V42 Engages students in peer-assessment			0.449*			
V43 Moderates feedback and results of self and peer assessment			0.728*			
V46 Uses flexible teaching activities				0.792*		
V47 Conducts assessment with consideration of student background				0.409*		
V48 Develops an environment of trust				0.632*		
V49 Ensures openness in the class				0.592*		
V50 Builds students' interest to learn				0.442*		
V51 Demonstrates belief in the ability of every student to improve				0.546*		
V53 Affirms students' good performance				0.519*		
V54 Clarifies students misconceptions				0.663*		
V55 Reinforces positive learning attitude of students				0.730*		
V58 Participates in professional development related to assessment					0.452*	
V59 Engages in self- assessment/ reflection					0.431*	
V60 Engages in peer-review of teaching performance					0.400*	
V62 Identifies subject- content knowledge needs					0.585*	
V63 Searches new and relevant subject-content information					0.425	
V64 Undertakes further education/ training					0.487	
V66 Collaborates with family to establish support activities						0.515*
V67 Informs community of the assessment practices						0.408*
V68 Reports to community about students' performance						0.633*
V70 Identifies key assessment and teaching issues for review						0.469*

Note: factor loadings $\leq .32$ were suppressed.

6.3.3. Convergent and Discriminant Validities

The measures of construct validity were also considered to establish the strength of the 6-factor model of teacher AfL literacy. Results of the convergent validity showed that the items comprising each factor were related to each other as shown in Table 6.5 where the factor loadings of all items in each factor ranged from .382 to .792. Although there is no agreed cut off score for factor loading values, the sample size ($n=325$) used in this study was relatively large, and hence, smaller factor loadings would be sufficient to account the convergence of the indicators of AfL literacy (MacCallum et al., 1999). The factor loadings demonstrated by the corresponding items of each six factors are highly correlated, and thus support convergent validity.

In terms of discriminant validity or the measure how factors are distinct from each other, investigation of the correlations amongst factors shows that the values range from .021 to .525 (see Table 6.6) with the highest correlation observed between Factors 2 and 4, followed by Factors 4 and 1, and Factors 1 and 2, yet the correlation coefficients between these factors were not substantial enough for these factors to be similar. Hence, the results of discriminant validity showed each factor was relatively distinct from the other factors. It was thought that the correlations observed might be due to the fact that all variables in this study were all measuring a single underlying construct of teacher AfL literacy, which was theoretically interconnected. This assumption is explored further in the subsequent analyses.

Table 6.6

Correlations amongst the Six Factors of Teacher AfL literacy Derived from Exploratory Factor Analysis using WLSMV Estimation for 52 Items (N=325)

	f1	f2	f3	f4	f5	f6
f1 Teachers as Assessors	1.000					
f2 Teachers as Pedagogy Experts	0.328	1.000				
f3 Teachers as Student Partners	0.275	0.247	1.000			
f4 Teachers as Motivators	0.407	0.525	0.146	1.000		
f5 Teachers as Teacher Learners	0.133	0.198	0.216	0.289	1.000	
f6 Teachers as Stakeholder Partners	0.071	0.025	0.021	0.035	0.161	1.000

6.4. Results of Confirmatory Analysis of the 6-Factor Model

6.4.1. Overview

A new data set derived from the self-assessment of 314 teachers in Brunei was used to test the 6-factor model with 42 items derived from EFA with latent factor means and variances set to 0 and 1, respectively. The intercepts, factor loadings, and residual variances were estimated using WLSMV estimator. The results of CFA are presented below.

6.4.2. Model Test Statistics and Fit Indexes of the Proposed Model

The test statistics and fit indexes of the tested model are presented in Table 6.7. The first aspect of the results considered in evaluating the model fit was the model fit statistics, which measure whether the covariance matrixes of the EFA-derived model and of the sample are close enough. Results showed that the chi-square statistics ($\chi^2 = 507.117$, $df=340$, $p<0.001$) was significant, which suggested that the model did not fit the data well. However, this result did not disqualify the model as chi-square tests are typically not an optimal measure of model fit as they tend to be greatly influenced by sample size (i.e., showing significance with large sample size) and the reconstructed covariance matrix does not usually produce a perfect fit to the original covariance matrix (Browne & Cudeck, 1993).

It has been argued by Barrett (2007) that ignoring the chi-square statistics is a lax method in determining model fit in CFA, and Kline (2005, 2010) suggests that the results of chi-square statistics should be reported as they can be used as preliminary evidence to further investigate the fit of the model. Also, Kline further suggests to evaluate the ratio between chi-square and degrees of freedom, which should not be more than 3:1 to indicate

an acceptable model, even if the chi-square value is significant. In this analysis, the values extracted from the data gave a ratio of less than 2 ($X^2/df = 1.500$), which indicated a perfect fit. Hence, the 6-factor model was supported by these results. However, to firmly establish the existence of the 6-factor model, more fit indexes were examined.

The approximate fit indexes show how well the parameter estimates of the CFA solution given by factor loadings, factor correlations, and error covariances, are able to reproduce the relationships that are observed in the data set. These indexes are continuous measures, hence they cannot be used to make dichotomous decisions whether to reject or accept the tested model (Kline, 2005). They can only be used to indicate the relative fit of the tested model. Results showed that the fit statistics of the 6-factor model were all satisfactory within the threshold values as presented in Table 6.7. For example, the RMSEA, which takes into account the complexity of the proposed model to explain the covariance between the variables, had a value equal to 0.031, which was below the 0.50 cut off value to indicate a good fit (Steiger & Lind, 1980). Similarly, the CFI (0.972) and TLI (0.970) were close to the threshold value of 1.0, which also indicated a perfect model. Finally, the WRMR value (0.974) was below 1.0, which indicated a good fitting model. Hence, these indexes confirmed the presence the 6-factor model of teacher A/L literacy.

Table 6.7

Evaluation of Fit Indexes of the 6-Factor Confirmatory Model (N=345)

Indexes	Perfect Fit	Good Fit	Sample Statistics	Decision	Reference
χ^2/df	≤ 2	≤ 3	1.500	Perfect fit	Kline (2005, 2010)
RMSEA	$\leq .05$	$\leq .08$	0.031	Perfect fit	Stegier & Lind (1980); Hooper, Coghlan & Mullen (2008)
CFI	$\geq .95$	$\geq .90$	0.972	Perfect fit	Hu & Bentler (1999) Tabachnick & Fidell (2007)
TLI	$\geq .95$	$\geq .90$	0.970	Perfect fit	Hu & Bentler (1999)
WRMR		≤ 1.0	0.974	Good fit	Hancock & Mueller (2006); Yu & Muthen (2002)

Note: $\chi^2 = 1267.792$; $df = 845$.

Examination of the unstandardised and unstandardised estimates along with their standard errors further supported the six-factor model (see Table 6.8). From the results, the absolute values of all ratios between standards estimates and their corresponding standards errors were all greater than or equal to 1.96, which denoted that all unconstrained loading estimates were significant at 0.05 level. The factor structure of the 6-factor model is presented in Figure 6.1.

Results of the CFA clearly demonstrated that a 6-factor structure defines the dimensions of teacher assessment literacy. The six factors were named as follows:

- Factor 1 was labelled as ***teachers as assessors***. It represents teacher assessment literacy to develop assessment strategy, use assessment tasks and measure student learning ($\alpha = 0.902$).
- Factor 2 covers teacher AfL literacy to use assessment and assessment data to inform teaching and learning activities ($\alpha = 0.935$). I labelled it as ***teachers as pedagogy experts***.
- Factor 3 is characterised by teachers working closely with their students and engaging them in the assessment and learning process ($\alpha = 0.935$). This factor was labelled ***teachers as student partners***.
- Factor 4 is about ***teachers as motivators***. It indicates a dimension of teacher AfL literacy in using assessment and assessment data to respond to individual students' learning needs ($\alpha = 0.930$).
- Factor 5 is composed of items that require teachers to reflect on their assessment experience coupled with the use of assessment data to identify

and respond to their professional development needs (teachers as learners; $\alpha = 0.901$), labelled this dimension as *teachers as teacher learners*.

- Factor 6 was labelled as *teachers as stakeholder partners*. This factor pertains to teacher A/L literacy in working with stakeholders to respond and enhance their assessment literacy ($\alpha = 0.899$).

This 6-factor solution can provided consistent measures of teacher AfL literacy both at the construct and factor levels. The individual Cronbach's alpha coefficients of the six factors were high and even Factors 6, 1 and 5 with a few numbers of items. Hence, the high reliability indexes of the six factors gave confidence that the tool will be trustworthy.

Table 6.8

Unstandardised and Standardised Loadings for the 6-Factor Confirmatory Model of Teacher AfL Literacy (N=354)

Factors	Variables	Unstandardized		Standardized		R ²
		Estimate	S.E.	Estimate	S.E.	
Assessor	V3 Designs assessment tasks	1.000	0.000	0.770	0.033	0.5929
	V7 Uses rubrics to assess students' learning	1.254	0.089	0.753	0.027	0.5670
	V9 Considers factors that affect students' performance	1.235	0.088	0.781	0.034	0.6100
	V10 Avoids interference in task completion	1.045	0.100	0.781	0.034	0.6100
	V12 Engages in social moderation	1.002	0.053	0.770	0.033	0.5929
Pedagogy Expert	V25 Translates learning standards to learning outcomes	1.000	0.000	0.753	0.027	0.5670
	V28 Identifies appropriate teaching methods	1.254	0.089	0.753	0.027	0.5670
	V29 Considers students' prior knowledge in lesson planning	1.240	0.077	0.826	0.024	0.6823
	V30 Considers students' current level of abilities	1.105	0.070	0.736	0.029	0.5417
	V31 Considers students' interest	1.150	0.077	0.766	0.028	0.5868
	V32 Plans lessons according to students' learning needs	1.127	0.073	0.751	0.029	0.5640
	V33 Tailors lessons to available resources	1.120	0.065	0.752	0.026	0.5655
	V34 Develops teaching and learning resources	1.165	0.065	0.783	0.026	0.6131
Student Partner	V17 Gives feedback on students' strengths and weaknesses	1.000	0.000	0.766	0.028	0.5868
	V19 Assists students in using feedback to feed forward	1.063	0.049	0.817	0.026	0.6675
	V27 Involves students in the development of learning outcomes	1.163	0.073	0.782	0.028	0.6115
	V35 Makes students understand the learning outcomes	1.276	0.082	0.766	0.026	0.5868
	V37 Involves students in the development of criteria and standards	1.179	0.085	0.708	0.029	0.5013
	V39 Explains the criteria and standards	1.254	0.091	0.753	0.025	0.5670
	V40 Develops students' capabilities in self and peer assessment	1.290	0.092	0.775	0.026	0.6006
	V41 Engages students in self-assessment	1.256	0.085	0.754	0.026	0.5685
	V42 Engages students in peer-assessment	1.254	0.089	0.753	0.027	0.5670
	V43 Moderates feedback and results of self and peer assessment	1.335	0.093	0.802	0.024	0.6432
Motivator	V46 Uses flexible teaching activities	1.000	0.000	0.783	0.026	0.6131
	V47 Conducts assessment with consideration of student background	1.155	0.066	0.776	0.025	0.6022
	V48 Develops an environment of trust	1.254	0.089	0.753	0.027	0.5670
	V49 Ensures openness in the class	1.061	0.069	0.713	0.033	0.5084
	V50 Builds students' interest to learn	1.155	0.066	0.776	0.025	0.6022
	V51 Demonstrates belief in the ability of every student to improve	1.165	0.065	0.783	0.026	0.6131
	V53 Affirms students' good performance	1.120	0.065	0.752	0.026	0.5655
	V54 Clarifies students' misconceptions	1.163	0.073	0.782	0.028	0.6115
	V55 Reinforces positive learning attitude of students	1.106	0.070	0.743	0.031	0.5520
Teacher Learner	V58 Participates in professional development related to assessment	1.000	0.000	0.754	0.026	0.5658
	V59 Engages in self- assessment/ reflection	1.229	0.082	0.777	0.028	0.6037
	V60 Engages in peer-review of teaching performance	1.124	0.079	0.711	0.033	0.5055
	V62 Identifies subject- content knowledge needs	1.235	0.088	0.781	0.034	0.6100
	V63 Searches new and relevant subject-content information	1.137	0.080	0.719	0.032	0.5170
	V64 Undertakes further education/ training	1.235	0.088	0.781	0.034	0.6100
Stakeholder Partner	V66 Collaborates with family to establish support activities	1.000	0.000	0.768	0.028	0.5898
	V67 Informs community of the assessment practices	1.002	0.053	0.770	0.033	0.5929
	V68 Reports to community about students' performance	1.063	0.049	0.817	0.026	0.6675
	V70 Identifies key assessment and teaching issues for review	1.022	0.052	0.786	0.031	0.6178

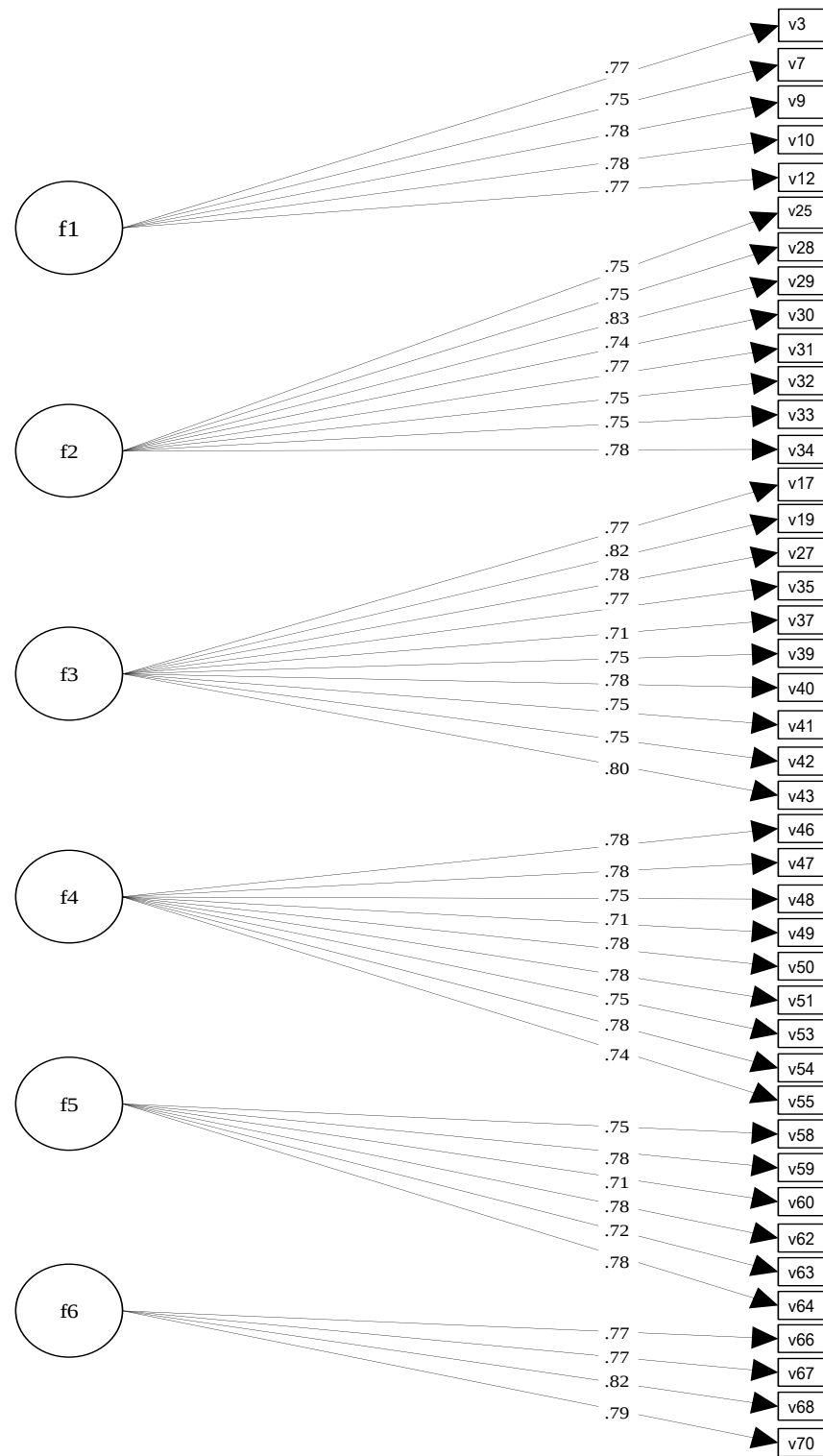


Figure 6.1. The 6-factor model of teacher AfL literacy. The numbers shown are the standardised factor loadings of variables to their corresponding factor.

6.4.3. Convergent and Discriminant Validities of the Proposed Model

The convergent validity of the 6-factor model was also considered. In simple terms, convergent validity provides evidence that each variable is highly correlated with its corresponding factor. This measure is determined by the factor loadings of the items given by the standardised estimates. These values give the standardised score increase in each variable for every standardised unit increase in its associated factor. An increase in any of the factors would mean a greater magnitude increase in the factor loadings of their associated variables. Although there are different views on what constitute a good factor loading, the most accepted value is equal to or greater than 0.70 (Nunnally & Bernstein, 1994). From the results in Table 6.8, the factor loadings extracted from the data were all substantial ranging from 0.71 to 0.82, all above the threshold value. Similarly, the R^2 values indicated that the variance explained ranged from 50.55% to 66.75%. More than 50% of the variances of all variables was accounted for by their associated factor. Thus, only less than 50% of the residual variances of each indicator were not explained by the latent variable.

Lastly, the discriminant validity of the proposed model was investigated and the results (see Table 6.9) showed that the correlations amongst factors ranged from 0.601 to 0.870, which indicated high correlations between and amongst factors. This means that the specified CFA model did not meet the criterion of discriminant validity. Hence, each factor did not seem to be measuring a particular sub-construct of teacher assessment for learning literacy. Further investigation of the CFA results did not suggest any modification indexes. Hence, the high correlations amongst factors suggested that these six factors were loading

onto a second-order factor. This assumption was further explored in the subsequent analyses.

Table 6.9

Correlations amongst the Six Factors of the Confirmatory Model of Teacher AfL Literacy
(*N*=354)

	f1	f2	f3	f4	f5	f6
f1 Teachers as Assessors	1.000					
f2 Teachers as Pedagogy Experts	0.750	1.000				
f3 Teachers as Student Partners	0.843	0.870	1.000			
f4 Teachers as Motivators	0.695	0.794	0.843	1.000		
f5 Teachers as Teacher Learners	0.641	0.664	0.763	0.774	1.000	
f6 Teachers as Stakeholder Partners	0.601	0.641	0.751	0.739	0.866	1.000

6.5. Second-order Factor Analysis

6.5.1. Overview

A second-order CFA was conducted to account for the high correlations of factors in the CFA model. Due to high factor correlations in CFA model, it was hypothesised that the six factors were loading onto a more generic construct of teacher AfL literacy. The model fit statistics and fit indexes were evaluated to support the existence of the second-order factor.

6.5.2. Model Fit Statistics and Fit Indexes

The model fit shown in Table 6.10 supported the hypothesis that the second-order factor model fitted the data well. Although the chi-square test yielded a value of 1186.026 with 854 degrees of freedom and p -value less than 0.001, yet the ratio between X^2 and df was below 2:1, which indicated a perfect fitting model. Contrary to the results of chi-square statistics, other fit indexes indicated a good fit of the specified model to the observed data. The RMSEA value of 0.047 was below the threshold value of .05, and within the confidence interval (0.040 to 0.053). The CFI (0.962) and TLI (0.960) were all close to 1.0 and the WRMR (0.978) was below 1.0. All these indexes suggested a good fit of the second-order model to the observed data set. The second-order factor was named as teachers as AfL literate professionals to indicate that teachers needed to possess certain level of assessment expertise before they could function well in any of the dimensions of teacher assessment literacy.

The results above were supported by the unstandardized loadings (Table 6.11) where all ratios of the estimates to their standard errors had absolute values greater than 1.96, and hence, all unconstrained loadings estimates were significant at 0.05 level.

Table 6.10

Evaluation of Fit Indexes of the Second-Order Confirmatory Factor Model (N=345)

Indexes	Perfect Fit	Good Fit	Sample Statistics	Decision	Reference
χ^2/df	≤ 2	≤ 3	1.389	Perfect fit	Kline (2005, 2010)
RMSEA	$\leq .05$	$\leq .08$	0.047	Perfect fit	Stegier & Lind (1980); Hooper, Coghlan & Mullen (2008)
CFI	$\geq .95$	$\geq .90$	0.962	Perfect fit	Hu & Bentler (1999) Tabachnick & Fidell (2007)
TLI	$\geq .95$	$\geq .90$	0.960	Perfect fit	Hu & Bentler (1999)
WRMR		≤ 1.0	0.978	Good fit	Hancock & Mueller (2006); Yu & Muthen (2002)

Note: $\chi^2 = 1186.026$; $df = 854$.

6.5.3. Convergent and Discriminant Validity

The factor loadings of the six factors onto second-order factor indicated that the second-order factor model met the requirement for convergent validity (see Table 6.11). The first order factor values ranged from 0.718 to 0.826. Furthermore, the standardized factor loadings of the first-order factors onto the second-order or the general factor were extremely high (ranging from 0.86 to 0.97) across all six factors. It was shown by the second-order CFA that both the first order factor and the second order factor had high convergent validities as presented in Table 6.12. However, the discriminant validity of the first order factors was problematic as they were highly correlated to each other ranging from 0.765 to 0.888. This posed problems as these high correlations suggest that the factors were not distinct from each other, or were not measuring different sub-constructs. Figure 6.2 shows the second-order CFA model where all six factors were made to load to one second-order factor.

The problematic results of the CFA and the second-order factor analysis in terms of the discriminant validity of the teacher AfL literacy were further explored using a more advanced structural equation modelling to find an explanation why there were high correlations observed amongst factors. As CFA and second-order factor analysis did not give any modification index to further modify the model, the exploratory structural equation modelling was used to further explore the factor structure of the tool.

Table 6.11

Unstandardised and Standardised Loadings for the 6-Factor Second-order Confirmatory Model of Teacher AfL Literacy (N=354)

Factors	Variables	Unstandardized		Standardized		R ²
		Estimate	S.E.	Estimate	S.E.	
Assessor	v3 Designs assessment tasks	1.000	0.000	0.768	0.036	0.5898
	V7 Uses rubrics to assess students' learning	1.059	0.095	0.730	0.041	0.5329
	V9 Considers factors that affect students' performance	0.951	0.061	0.768	0.036	0.5895
	V10 Avoids interference in task completion	1.045	0.065	0.790	0.031	0.6241
	V12 Engages in social moderation	0.951	0.061	0.768	0.036	0.5898
Pedagogy Expert	V25 Translates learning standards to learning outcomes	1.000	0.000	0.812	0.034	0.6593
	V28 Identifies appropriate teaching methods	1.089	0.093	0.750	0.044	0.5625
	V29 Considers students' prior knowledge in lesson planning	1.059	0.095	0.730	0.041	0.5329
	V30 Considers students' current level of abilities	1.037	0.060	0.785	0.036	0.6162
	V31 Considers students' interest	1.025	0.060	0.776	0.033	0.6022
	V32 Plans lessons according to students' learning needs	1.025	0.060	0.776	0.033	0.6022
	V33 Tailors lessons to available resources	1.192	0.092	0.822	0.031	0.6757
	V34 Develops teaching and learning resources	1.201	0.097	0.828	0.035	0.6856
Student Partner	V17 Gives feedback on students' strengths and weaknesses	1.000	0.000	0.757	0.037	0.5730
	V19 Assists students in using feedback to feed forward	0.951	0.061	0.768	0.036	0.5898
	V27 Involves students in the development of learning outcomes	1.045	0.065	0.790	0.031	0.6241
	V35 Makes students understand the learning outcomes	1.059	0.095	0.730	0.041	0.5329
	V37 Involves students in the development of criteria and standards	1.026	0.060	0.777	0.033	0.6037
	V39 Explains the criteria and standards	1.025	0.060	0.776	0.033	0.5868
	V40 Develops students' capabilities in self and peer assessment	0.987	0.064	0.747	0.038	0.5580
	V41 Engages students in self-assessment	1.092	0.063	0.826	0.030	0.6823
	V42 Engages students in peer-assessment	1.075	0.076	0.747	0.038	0.5580
	V43 Moderates feedback and results of self and peer assessment	1.037	0.060	0.785	0.036	0.6162
	V46 Uses flexible teaching activities	1.000	0.000	0.716	0.037	0.5127
	V47 Conducts assessment with consideration of student background	1.079	0.068	0.773	0.037	0.5975
	V48 Develops an environment of trust	1.075	0.070	0.770	0.036	0.5929
Motivator	V49 Ensures openness in the class	1.043	0.072	0.747	0.038	0.5580
	V50 Builds students' interest to learn	1.052	0.067	0.754	0.037	0.5685
	V51 Demonstrates belief in the ability of every student to improve	1.052	0.067	0.754	0.037	0.5685
	V53 Affirms students' good performance	1.056	0.074	0.756	0.039	0.5715
	V54 Clarifies students misconceptions	0.951	0.061	0.768	0.036	0.5898
	V55 Reinforces positive learning attitude of students	1.059	0.095	0.730	0.041	0.5329
	V58 Participates in professional development related to assessment	1.000	0.000	0.747	0.038	0.5580
	V59 Engages in self- assessment/ reflection	1.075	0.076	0.747	0.038	0.5580
Teacher Learner	V60 Engages in peer-review of teaching performance	1.186	0.097	0.825	0.042	0.6806
	V62 Identifies subject- content knowledge needs	1.025	0.060	0.776	0.033	0.6022
	V63 Searches new and relevant subject-content information	1.032	0.087	0.718	0.047	0.5155
	V64 Undertakes further education/ training	1.135	0.095	0.789	0.041	0.6225

Stakeholder Partner	V66	Collaborates with family to establish support activities	1.000	0.000	0.808	0.039	0.6529
	V67	Informs community of the assessment practices	1.004	0.062	0.812	0.034	0.6593
	V68	Reports to community about students' performance	0.951	0.061	0.768	0.036	0.5898
	V70	Identifies key assessment and teaching issues for review	1.025	0.058	0.829	0.033	0.6872
G Teacher as an AfL Expert	F1	Teacher as an assessor	1.000	0.000	0.891	0.035	0.7939
	F2	Teacher as a pedagogy expert	1.784	0.303	0.945	0.025	0.8930
	F3	Teacher as a student partner	2.152	0.385	0.972	0.012	0.9448
	F4	Teacher as a motivator	1.894	0.337	0.924	0.018	0.8538
	F5	Teacher as a teacher learner	1.758	0.320	0.864	0.026	0.7464
	F6	Teacher as a stakeholder partner	2.159	0.367	0.913	0.023	0.8336

Table 6.12

Correlations amongst the Six Factors of the Second-order Confirmatory Model of Teacher AfL Literacy (N=354)

	f1	f2	f3	f4	f5	f6
f1 Teachers as Assessors	1.000					
f2 Teachers as Pedagogy Experts	0.808	1.000				
f3 Teachers as Student Partners	0.888	0.860	1.000			
f4 Teachers a Motivators	0.825	0.800	0.879	1.000		
f5 Teachers a Teacher Learner	0.789	0.765	0.840	0.781	1.000	
f6 Teachers as Stakeholder Partners	0.833	0.808	0.888	0.825	0.825	1.000

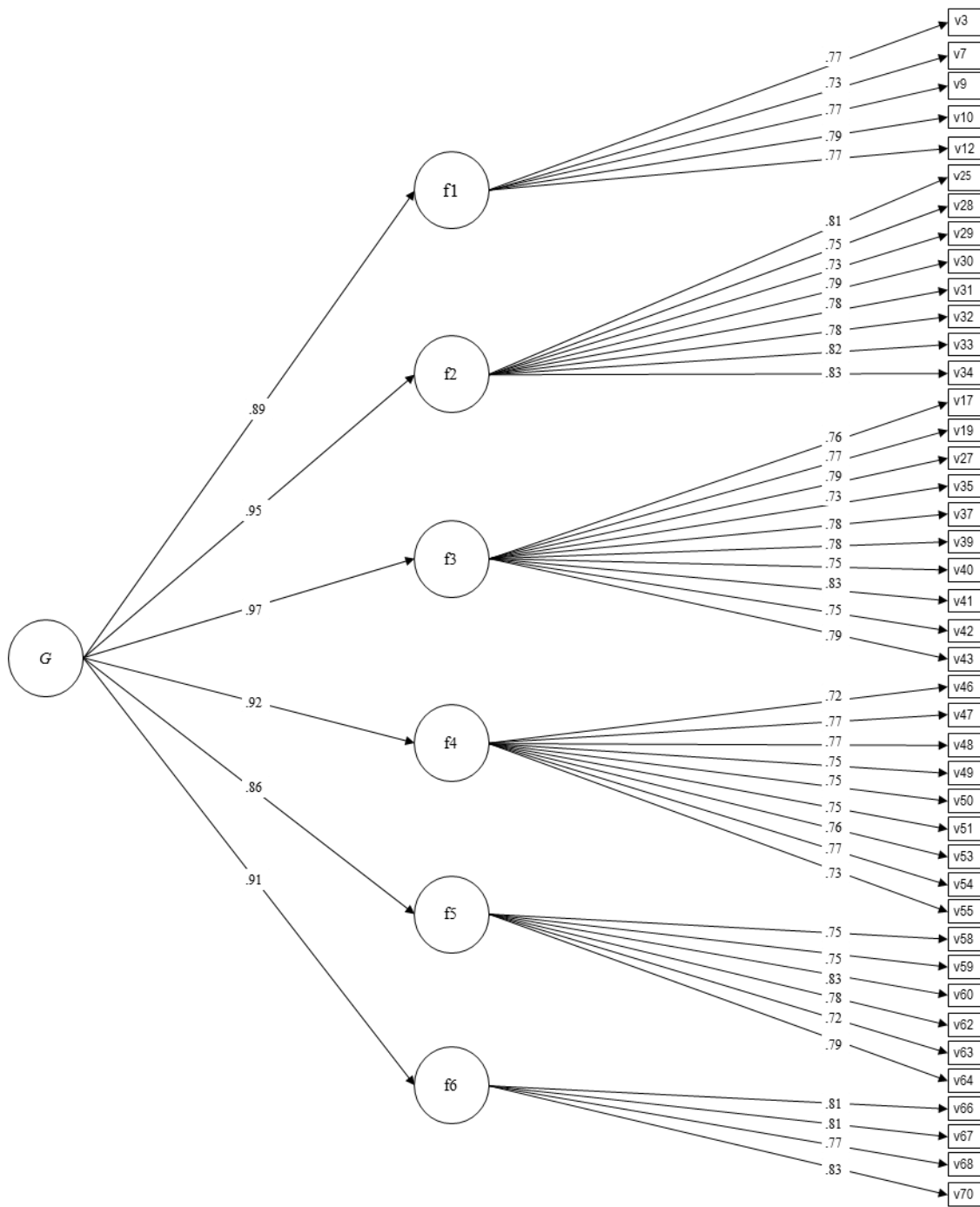


Figure 6.2. The 6-factor model with second-order general factor of teacher AfL literacy with the standardised factor loadings of variables to their corresponding factor.

6.6. Exploratory Structural Equation Modelling

6.6.1. Overview

Although both CFA and second-order CFA models have good fit indexes and convergent validities, their discriminant validities were problematic. The dimensionality of the six-factor model was further tested using exploratory structural equation modelling (ESEM). The use of ESEM overcomes the limitation of CFA which fixes cross-loadings to 0, where in reality there is no single variable that does not cross load to another factor (Marsh et al., 2009). According to Asparouhov and Muthén (2009) the use of ESEM makes the estimation more accurate than in CFA because it accounts for the true nature of the construct as there is no pre-specified zero cross loadings. Hence, ESEM provides better goodness of fit indexes and it models both convergent and discriminant validity more accurately (Marsh et al., 2009). The use of ESEM to reveal the true nature of a particular construct is argued in Chapter 3.

6.6.2. Model Fit Statistics and Fit Indexes

The test statistics and fit indexes of the tested model are presented in Table 6.13. The chi-square value, although of little value in ESEM due to its sensitivity to sample size (Browne & Cudeck, 1993), was still considered and the results ($X^2 = 767.735$, $df=624$, $p<0.001$) showed it was significant. Hence, there was covariance evidence against the 6-factor model. However, the ratio of X^2 and df was almost 1:1, and thus, it was still within the threshold ratio suggested by Kline (2005) to be considered as perfect fit model. To support the 6-factor model in ESEM, the fit indexes were also investigated, showing an excellent

model fit. The RMSEA value (0.026, CI 0.019 - 0.031) was within the 90% confidence interval and the upper bound was less than 0.10, which both indicated an excellent model fit; the CFI (0.990) and TLI (0.987) were all very close to 1, and the WRMR (0.958) was less than 1.0. These fit indexes were all higher than those generated from CFA and hierarchical CFA. Further, the unstandardized loadings were all significant as supported by the absolute values of the ratios between estimates and their corresponding standard errors, which were all greater than 1.96.

Table 6.13

Evaluation of Fit Indexes of the 6-Factor Exploratory Structural Equation Model (N=345)

Indexes	Perfect Fit	Good Fit	Sample Statistics	Decision	Reference
χ^2/df	≤ 2	≤ 3	1.230	Perfect fit	Kline (2005, 2010)
RMSEA	$\leq .05$	$\leq .08$	0.026	Perfect fit	Stegier & Lind (1980); Hooper, Coghlan & Mullen (2008)
CFI	$\geq .95$	$\geq .90$	0.990	Perfect fit	Hu & Bentler (1999) Tabachnick & Fidell (2007)
TLI	$\geq .95$	$\geq .90$	0.987	Perfect fit	Hu & Bentler (1999)
WRMR		≤ 1.0	0.958	Good fit	Hancock & Mueller (2006); Yu & Muthen (2002)

Note: $\chi^2 = 767.735$; $df = 624$.

6.6.3. Convergent and Divergent validity

The factor loadings of variables ranged from 0.685 to 0.788 as shown in Table 6.14. These values showed that the model tested had met the requirement for convergent validity except for some cross loadings observed from -1.66 to 0.381. In other words, the items were highly correlated to their respective factor, and hence it is expected that any standardised unit increase in a factor would result to an increase in a standardised scores in all variables associated to that particular factor. The cross-loadings were expected because the loadings of variables to other factors apart from their associated factor were not constrained to 0.

The discriminant validity, given by the factor correlations, showed that the 6 factors were low to moderately correlated. As can be seen in Table 6.15, the correlation values ranged from 0.101 to 0.481. This means that each factor, although slightly correlated, is measuring a different sub-construct of teacher A/L literacy.

Table 6.14

Unstandardised and Standardised Loadings for the 6-Factor Exploratory Structural Equation Model of Teacher AfL Literacy (N=354)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
v3 Designs assessment tasks	0.790	-0.066	0.263	0.009	0.040	-0.061
V7 Uses rubrics to assess students' learning	0.718	0.088	0.146	0.008	-0.120	0.057
V9 Considers factors that affect students' performance	0.768	0.033	-0.003	0.165	0.012	0.024
V10 Avoids interference in task completion	0.720	-0.159	0.094	-0.104	0.121	0.017
V12 Engages in social moderation	0.709	0.105	0.101	0.076	0.014	-0.194
V25 Translates learning standards to learning outcomes	0.264	0.757	-0.014	-0.143	0.212	-0.020
V28 Identifies appropriate teaching methods	0.143	0.759	0.010	0.089	-0.028	0.100
V29 Considers students' prior knowledge in lesson planning	-0.004	0.788	0.137	0.003	0.126	-0.046
V30 Considers students' current level of abilities	-0.004	0.748	0.015	-0.024	0.112	-0.024
V31 Considers students' interest	0.063	0.728	-0.041	0.247	-0.005	0.105
V32 Plans lessons according to students' learning needs	-0.028	0.723	0.158	0.072	0.010	0.075
V33 Tailors lessons to available resources	0.005	0.770	0.138	0.186	-0.056	0.009
V34 Develops teaching and learning resources	0.206	0.789	0.089	0.129	0.069	-0.124
V17 Gives feedback on students' strengths and weaknesses	0.328	0.052	0.699	0.290	-0.011	0.185
V19 Assists students in using feedback to feed forward	0.381	0.051	0.738	0.062	0.080	0.040
V27 Involves students in the development of learning outcomes	0.315	0.359	0.740	-1.001	-0.014	-0.666
V35 Makes students understand the learning outcomes	0.160	0.310	0.708	0.165	0.003	-0.078
V37 Involves students in the development of criteria & standards	0.004	0.209	0.699	-0.056	-0.012	0.142
V39 Explains the criteria and standards	0.026	0.003	0.714	0.034	0.004	0.221
V40 Develops students' capabilities in self/ peer assessment	0.142	0.004	0.721	0.017	-0.021	0.005
V41 Engages students in self-assessment	0.069	-0.008	0.698	0.027	0.102	0.017
V42 Engages students in peer-assessment	0.171	0.057	0.713	0.030	0.099	0.010
V43 Moderates feedback and results of self and peer	0.024	0.060	0.703	-0.035	0.127	0.029
V46 Uses flexible teaching activities	-0.015	0.083	0.389	0.711	-0.001	-0.033
V47 Conducts assessment with consideration of students' background	-0.047	-0.141	0.273	0.712	-0.028	-0.020
V48 Develops an environment of trust	0.113	0.025	0.027	0.699	-0.022	0.253
V49 Ensures openness in the class	0.028	0.048	-0.020	0.696	0.110	0.182
V50 Builds students' interest to learn	0.033	0.164	0.092	0.685	0.088	0.013
V51 Demonstrates belief in the ability of every student to improve	-0.014	0.138	0.323	0.758	0.105	0.067
V53 Affirms students' good performance	-0.022	0.108	0.272	0.785	0.170	-0.034
V54 Clarifies students misconceptions	0.055	0.089	0.345	0.708	0.227	-0.040
V55 Reinforces positive learning attitude of students	0.089	-0.039	0.359	0.767	0.114	-0.033
V58 Participates in professional development (assessment)	-0.035	-0.062	0.033	0.135	0.712	0.044
V59 Engages in self- assessment/ reflection	0.058	0.048	0.146	-0.034	0.735	-0.078
V60 Engages in peer-review of teaching performance	0.000	-0.009	-0.019	0.256	0.707	-0.115
V62 Identifies subject- content knowledge needs	0.070	0.143	0.013	-0.027	0.730	0.099
V63 Searches new and relevant subject-content information	0.077	0.082	-0.153	0.148	0.699	-0.047
V64 Undertakes further education/ training	-0.024	-0.075	0.004	0.333	0.711	-0.051
V66 Collaborates with family to establish support activities	-0.024	-0.005	0.072	0.101	0.346	0.782
V67 Informs community of the assessment practices	0.048	-0.036	-0.012	-0.014	0.213	0.716
V68 Reports to community about students' performance	0.026	0.007	0.100	0.031	0.297	0.716
V70 Identifies key assessment and teaching issues for review	-0.095	0.206	0.099	-0.024	0.212	0.779

Table 6.15

Correlations amongst the Six Factors of the Exploratory Structural Equation Model of Teacher AfL Literacy (N=354)

	f1	f2	f3	f4	f5	f6
f1 Teachers as Assessors	1.000					
f2 Teachers as Pedagogy Experts	0.356	1.000				
f3 Teachers as Student Partners	0.394	0.223	1.000			
f4 Teachers a Motivators	0.481	0.226	0.355	1.000		
f5 Teachers a Teacher Learner	0.375	0.384	0.372	0.347	1.000	
f6 Teachers as Stakeholder Partners	0.133	0.129	0.138	0.038	0.101	1.000

6.7. Comparison of CFA, Second-Order CFA and ESEM Results

The 6-factor model tested gave slightly different results (i.e., model fit statistics, fit indexes, factor loadings, convergent and discriminant validities) in CFA, second-order CFA and ESEM. Table 6.16 summarises the results of these three analyses, which even given some differences, supported the existence of a 6-factor model of teacher AfL literacy.

From Table 6.16, the chi-square statistics results were all significant but with ratio to *df* less than 3:1. The fit indexes were all within the acceptable values, but the results of ESEM showed better fit in these indexes. The RMSEA values for CFA and hierarchical-CFA were both indicative of a perfect fit model (below 0.05 but greater than 0.03), but the ESEM value showed a superior index (below 0.03). Similarly, the CFI and TLI from ESEM were the closest to 1 compared to those derived from other two analyses. Further, the WRMR is lowest in ESEM.

Further comparison of the results of the analyses showed that the convergent validity is best modelled in second-order CFA, and then CFA than in ESEM. This is due to the fact the CFA has a tendency to overestimate factor loadings (Asparouhov & Muthén, 2009). However, the difference between the ranges of factor loadings was relatively small. In terms of discriminant validity, it was problematic for both CFA and second-order CFA, where all factors were highly correlated, but in ESEM the correlations amongst factors were greatly reduced. The high correlations observed in CFA were the effect of restricting the cross-loadings of variables amongst factors, and hence caused the correlations values amongst variables to be inflated (Asparouhov & Muthén, 2009; Marsh et al., 2009).

Table 6.16

Comparison of CFA, Hierarchical and ESEM Results of Teacher AfL Literacy Model

Analysis	Fit Indexes								Validity Measures	
	χ^2	df	p-value	χ^2/df	RMSEA	CFI	TLI	WRMR	Factor loadings	Factor correlations
CFA	1267.792	845	0.000	1.500	0.031	0.972	0.970	0.974	0.71 - 0.82	0.601-0.870
Second-Order CFA	1186.026	854	0.000	1.389	0.047	0.962	0.960	0.978	0.718 - 0.826	0.86 - 0.97
ESEM	767.735	624	0.000	1.230	0.026	0.990	0.987	0.958	0.685 - 0.788	0.101 - 0.394

The best fit observed in ESEM is consistent with the argument of Asparouhov and Muthén (2009) and Marsh et al. (2009) that ESEM addresses the restrictive property of CFA where cross-loadings are not allowed. From a methodological perspective, the results of this study support the previous results of studies (e.g., Asparouhov & Muthén, 2009; Marsh et al., 2009) that ESEM produces better fit indexes and the factors are less correlated than in CFA.

Overall, whilst the fit indexes and the factor loadings derived from the three analyses were slightly different, they all support the existence of the 6-factor model in the data set. Hence, the claim that the teacher AfL literacy is composed of six factors as initially extracted from the results of exploratory factor analysis is warranted by the results of these three analyses. The EFA, CFA, second-order FA and ESEM reduced the variables of teacher AfL literacy tool from 72 teacher AfL practices to 42 items with six overarching dimensions.

6.8. The Final Teacher AfL Literacy Tool

This section answers the first research question of this study, that is:

What are the emerging AfL practices of teachers? Specifically,

- What are the indicators of teachers' effective AfL practices?
- What are the standards of performance for each stage of professional growth?

Although this was the first question, the empirical evidence to support the final version of teacher AfL literacy tool was derived from the results of the analyses presented in the preceding section. The final version of the teacher AfL literacy tool is presented in the succeeding pages with the directions on how to use the tool for teachers' self-assessment and colleague's peer-assessment are included as suggested by the validators.

The tool was organised using the six domains of teacher AfL literacy with the indicators of teacher AfL practices presented as criteria with their respective five levels of standards. The criteria comprising each factor were further clustered based on thematic analysis to reveal the job roles of the teachers. The purpose of this further clustering is to find an easier nomenclature to describe each factor with sub-domains to provide a better way to describe and understand teacher AfL practices.

A Teacher Assessment for Learning Literacy Tool

Purpose of the Competency Assessment Tool

This assessment instrument is part of a process designed to help teachers improve their use of assessment for learning (AfL) in improving student learning and achievement. As with AfL for students, the assessment instrument can be used for self and peer assessment and for school leader direct assessment.

- *in self-assessment, teachers can identify what they believe to be their current level of AfL competency and so identify the next sets of skills they need to develop to move to the next competency level;*
- *in peer-assessment, colleagues can provide their perception of fellow teachers' current AfL competency levels based on their observation of classroom practices and other information provided by the colleague being assessed. They can also identify opportunities for assisting colleagues seeking further development and for learning from those teachers already well-advanced in AfL competency development and classroom practice*
- *in direct assessment, school principals/heads can use the instrument to assess the current level of individual teachers AfL competence. From the results, they can identify the support required by each teacher including appropriate professional development.*

Importantly, all of these assessment results are brought together within the school for 'moderation' just as in assessment moderation or social standardisation for student work. In this way, in a safe and supportive conversation, each teacher has the opportunity to present their teaching portfolio and other 'evidence' to illustrate their self-rating with a purpose of reaching an agreement of the various assessors on the most accurate 'picture' from which to plan the next step of professional development.

Directions

For self-assessment: Whilst reflecting on your own classroom practices, please read each item carefully then mark the competency level which best describes your current performance. There is no right or wrong answer. Your responses will be used to guide and support you in enhancing your assessment literacy to effectively support student learning.

For peer and principal direct assessment:

Whilst reflecting on your colleague's current AfL competency levels, please take into account any other information or evidence of performance made available to you by your colleague for whom you are doing this assessment. Again, please remember there is no right or wrong answer. As with self-assessment, your responses are needed to guide and support your colleague in their AfL literacy and classroom practices improvement.

Domain 1. Teacher as an ASSESSOR – this domain covers the skills of a teacher to develop a wide range of assessment tasks and to use a comprehensive system of assessing student learning. Teacher needs to evaluate the appropriateness, usefulness, adequacy, trustworthiness and fairness of his/her assessment practices.

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
DEVELOPMENT AND USE OF ASSESSMENT TASKS	Designs assessment tasks	Designs generic assessment tasks	Designs inclusive assessment tasks	Designs differentiated and multi-dimensional assessment tasks that allow students to demonstrate their learning in a variety of ways	Evaluates the suitability of the differentiated assessment tasks developed	Provides training to colleagues in designing differentiated assessment tasks
	Uses rubrics to assess students' learning	Uses valued expertise in assessing student learning	Uses rubrics to assess student learning	Uses rubrics and valued expertise to identify student learning opportunities	Reviews the effectiveness of rubrics against valued expertise in assessing students' performance and identifying learning opportunities	Provides expert assistance to colleagues in effectively using rubrics
ENSURING TRUSTWORTHINESS OF ASSESSMENT	Considers factors that affect students' performance (<i>length of assessment, suitability of task, language, design, readiness of students</i>)	Identifies factors that may affect student performance in the assessment	Determines the readiness of students before participating in the assessment	Develops and uses a checklist to ensure all factors affecting students' readiness to participate in the assessment are minimised or removed	Uses information gathered regarding students' readiness to modify assessment tasks accordingly	Provides professional guidance to other teachers in modifying assessment tasks based on information gathered on students' readiness
	Avoids interference in task completion (<i>e.g. ability to read, write, personality, physical condition or knowledge of irrelevant background information</i>)	Identifies any problems that might hinder students' ability to demonstrate their learning	Uses assessment tasks which do not require irrelevant background knowledge and skills or special talents or abilities to demonstrate learning	Employs several checking mechanisms to identify possible factors, which might hinder students' performance	Evaluates the overall assessment and teaching practices to avoid factors which might hinder students' performance and quickly adjusts the assessment tasks where necessary	Provides training for teachers on how to avoid factors which might interfere with students' performance

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
	Establishes dialogue/ conversation with colleagues to ensure consistent, fair and comparable judgment of students' learning <i>(engages in standardisation)</i>	Contributes to discussions about ensuring consistent, fair and comparable judgment of students' learning	Initiates discussions to ensure consistent, fair and comparable judgment of students' learning	Questions assessment results to confirm the interpretation of students learning to ensure consistent, fair and comparable judgment	Synthesises the results of discussion/ moderation to reach a consensus to ensure consistent, fair and comparable judgment of students' learning	Leads the moderation activity to reach a consensus in decision- making to ensure consistent, fair and comparable judgment

Domain 2. Teacher as a PEDAGOGY EXPERT- this domain covers the teacher's role of reflecting on assessment information to inform one's teaching approach. The teacher determines students' prior knowledge, interests and levels of ability and uses these in planning the lesson and in implementing the curriculum.

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
DEVELOPMENT OF LESSON	Translates learning standards (curriculum content) to learning outcomes	Adopts existing learning standards and outcomes	Uses learning standards (curriculum content) to develop learning outcomes	Analyses learning standards (curriculum content) and other sources of information to develop learning outcomes	Evaluates the coherence of learning outcomes to standards (curriculum content)	Leads colleagues in the development of learning outcomes which are coherent with standards
	Identifies appropriate teaching methods	Uses routine teaching methods	Identifies the best teaching methods to help students maximise their learning	Develops differentiated instruction to meet various student learning needs	Develops and assesses the effectiveness of differentiated instruction in meeting the various learning needs of students	Provides expert assistance to other teachers in developing and using appropriate teaching methods
	Considers students' prior knowledge in lesson planning	Develops lesson plans but with little reference to students' prior knowledge	Assesses students' prior knowledge and uses the information in lesson planning	Develops strategies to effectively integrate students' prior knowledge in lesson planning	Evaluates the effectiveness of strategies developed in integrating students' prior knowledge in lesson planning	Demonstrates to colleagues effective strategies in integrating students' prior knowledge in lesson planning
	Considers students' current level of abilities	Identifies students' current level of abilities in lesson planning	Uses students' current level of abilities to develop lessons within this level	Uses students' current level of abilities as reference in planning lessons including selection of materials, teaching methods and learning activities.	Assesses the impact to learning of incorporating students' current level of abilities in lesson	Demonstrates to colleagues various applications of students' current abilities in lesson planning

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
	Considers students' interest	Refers to students' interests in lesson planning	Modifies teaching methods based on the interests of the students	Demonstrates various ways (modifying teaching methods, content, timing, etc.) to incorporate students' interests in lesson planning/teaching	Evaluates the lesson plan/teaching in terms of the effectiveness of incorporating students' interests	Delivers professional training to teachers on how to incorporate students' interests in lesson planning/teaching
	Plans lessons according to students' learning needs	Plans lessons with reference to students' learning needs	Develops lesson plans that address students' learning needs	Develops differentiated instruction addressing individual needs	Evaluates the effectiveness of differentiated instruction	Trains colleagues in developing differentiated instruction
SELECTION OF LEARNING EXPERIENCE	Tailors lessons to available resources	Develops lesson plans but with little consideration of the resources available	Adjusts some aspects of lesson plans based on available resources	Modifies lesson plans based on available resources to improve the quality of instruction	Evaluates the effectiveness of available resources in teaching and learning	Provides expert assistance to colleagues on how to adjust lesson plans to fit available resources
	Develops teaching and learning resources	Uses available teaching and learning resources	Develops limited teaching and learning resources	Develops teaching and learning resources which are adaptive, relevant and context-based	Assesses the suitability of developed teaching and learning resources	Guides other teachers in developing new learning resources to suit students' needs

DOMAIN 3. Teacher as a STUDENT PARTNER – this domain requires the teacher to use assessment tasks to actively engage students in teaching and learning activities. The focus of the teacher is to work closely with students to provide them with a wide range of opportunities to develop their self-regulation.

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
EMPLOYS STRATEGIES TO ENGAGE STUDENTS IN THE LEARNING PROCESS	Involves students in the development of learning outcomes	Presents the learning outcomes to students	Provides opportunities for students to participate in the development of learning outcomes	Develops a learning environment where students are actively engaged in developing and analysing the learning outcomes	Assesses the impact of students' participation in developing learning outcomes to their learning	Provides expert leadership to colleagues on how to effectively involve students in the development of learning outcomes
	Involves students in the development of criteria and standards (rubrics)	Develops success criteria/rubrics	Provides opportunities for students to critique the success criteria/rubrics	Works with students to develop success criteria/rubrics	Creates an environment where students are actively involved in the development rubrics	Provides professional input and advice for colleagues on how to involve students in the development of rubrics
	Makes students understand the learning outcomes	Shares learning outcomes with students	Discusses and provides work samples to clarify learning outcomes	Employs various strategies to ensure all students have understood the learning outcomes before proceeding to the lesson	Assesses the effectiveness of various strategies in sharing learning outcomes with students	Demonstrates to colleagues how to ensure students understand the learning outcomes
	Explains the success criteria and standards (rubrics)	Discusses with students the success criteria and standards	Provides exemplar materials to clarify the success criteria and standards	Uses various strategies to ensure students have understood the success criteria and standards	Evaluates the effectiveness of exemplar materials and strategies in ensuring students' understanding of criteria and standards	Helps other teachers to effectively use exemplar materials to enhance students' understanding of success criteria and standards
ENGAGES STUDENTS IN ASSESSMENT	Develops students' capabilities in self and peer assessments	Tells students to develop their assessment skills	Provides support to students in developing their reflective thinking and evaluative skills	Uses various strategies for developing students' reflective thinking and evaluative skills	Monitors students' progress in self and peer assessment capabilities and provides opportunities for further growth	Trains colleagues to develop students' self- and peer-assessment abilities

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
	Engages students in self-assessment	Allots time for students to conduct self-assessment	Identifies the exact timing for engaging students in self-assessment	Integrates self-assessment into the teaching and learning process	Sustains students' interest in doing self-assessment even outside the classroom	Provides assistance to colleagues on how to engage students in self-assessment
	Engages students in peer-assessment	Ensures that peer-assessment is part of the teaching and learning processes	Manages the learning experience to ensure all students engage in peer-assessment	Develops and uses strategies to fully engage students in peer-assessment	Determines the effectiveness of strategies used to engage students in peer-assessment	Trains other teachers in using various strategies to engage students in peer-assessment
	Moderates feedback and results of self and peer assessment	Refers feedback to the results of self and peer-assessment	States teacher feedback in a way that confirms results of self- and peer-assessment	Validates any inconsistencies between the results of self- and peer-assessment and teacher feedback	Works closely with students to moderate the results of self and peer assessments and reinforce students' learning by giving timely and appropriate teacher feedback	Expertly trains colleagues on how to moderate feedback and results of self- and peer-assessments
COMMUNICATION OF ASSESSMENT RESULTS	Gives feedback on students' strengths and weaknesses	Tells the students their strengths and weaknesses	Gives suggestions on how to use students' strengths to improve their learning and how to address their weaknesses	Follows through with students on how to use their strengths to improve learning and how to address their weaknesses	Consistently encourages and supports students to using their strengths to improve their learning and address their weaknesses	Shows teachers how to engage students using their strengths to improve their learning and address their weaknesses
	Assists students in using feedback to feed forward	Uses assessment results to tell students what goals to achieve	Works with students to use assessment results to set their learning goals	Creates a learning environment where students are engaged in analysing assessment results to identify their learning needs and to set learning goals	Develops approaches to help students work independently to analyse assessment results to identify their learning needs and to set learning goals	Shows colleagues various ways to work with students to analyse assessment results to identify their learning needs and to set learning goals

DOMAIN 4: Teacher as a MOTIVATOR – this domain requires the teacher to use assessment tasks and assessment information to enhance students' motivation. The aim of the teacher is to provide differentiated instruction and assessment to address individual students' learning needs and motivation.

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
ADAPTATION OF TEACHING AND ASSESSMENT	Conducts assessment with consideration of student background and culture.	Demonstrates recognition of students' worth	Treats students in ways that reflect their worth	Responds appropriately to students' differences	Demonstrates high level of respect to students by modelling and reinforcing respect between teacher and students and amongst students as well.	Demonstrates to colleagues various ways to develop a learning culture where every student is respected
	Uses a variety of teaching and learning methodologies	Uses any teaching and learning methods	Uses culturally relevant and interesting teaching and learning methods	Develops or innovates a range of relevant and interesting teaching and learning methods	Assesses the effectiveness of teaching methods developed and used	Provides support to colleagues on how to develop, innovate and use a range of relevant and interesting teaching and learning methods
DEVELOPMENT OF ENVIRONMENT THAT VALUES LEARNERS	Develops an environment of trust	Builds mutual trust and good rapport with students	Identifies and addresses factors that inhibit students' trust	Provides activities that promote trust between and amongst students and teachers	Evaluates the effectiveness of activities that promote trust amongst teachers and students and improves those activities	Shares with colleagues effective ways to develop an environment of trust
	Ensures openness in the class	Provides opportunities for students to present their views but still maintains high level of authoritativeness	Initiates ways to develop classroom openness	Creates an environment where students are confident in sharing their views, ideas and thoughts and in seeking help	Evaluates classroom activities and identifies those that enhance appropriate openness in the class	Demonstrates to colleagues ways to ensure openness in the class

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
	Uses assessment to build students' interest to learn	Identifies signs of students' becoming disinterested in their learning	Uses assessment to build up students' interest in learning	Employs several assessment activities to sustain students' interest in learning	Evaluates the effectiveness of assessment activities that build up students' interest in learning	Demonstrates to colleagues best practices in building students' interest to learning
	Demonstrates belief in the ability of every student to improve	Advises low performing students to undertake remedial instruction	Works regularly with low performing and with learning needs students to improve their learning	Uses several strategies to help individual students improve their learning	Identifies the best approaches for helping individual and different groups of students improve their learning	Models to colleagues positive attitudes that demonstrate the belief that all students improve
DIRECTING STUDENTS TOWARDS GOAL ATTAINMENT	Affirms students' good performance	Identifies students' good performance and gives praise	Identifies students' good performance and uses praise directed at effort, self-regulation, engagement and learning	Uses praise directed at effort, self-regulation, engagement and learning and other strategies to affirm students' performance	Selects and uses the most appropriate strategies to affirm students' learning	Demonstrates to colleagues various ways to affirm students' good performance
	Clarifies students misconceptions	Identifies students' misconceptions about the lesson	Follows through students' progress in correcting their misconceptions about the lesson	Uses a wide range of strategies to address students' misconceptions about the lesson	Assesses the effectiveness of various strategies in addressing students' misconceptions about the lesson	Helps colleagues to use various strategies in addressing students' misconceptions about the lesson
	Reinforces positive learning attitude of students	Identifies students' positive learning attitudes	Provides activities which enhance students' positive learning attitude	Employs several strategies for reinforcing students' positive attitude to improve learning	Provides information related to the effectiveness of various strategies for reinforcing students' positive attitudes	Models to colleagues various ways for reinforcing positive learning attitudes

DOMAIN 5: Teacher as a TEACHER LEARNER– this domain pertains to the role of the teacher to reflect on the assessment data to identify the effectiveness and appropriateness of his/her classroom practices. From the reflection, the teacher identifies their own strengths and weaknesses both in the AfL skills and in the curriculum-content knowledge and undertakes further training/education.

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
ENGAGEMENT IN PROFESSIONAL DEVELOPMENT RELATED TO AfL	Participates in professional development related to assessment	Responds to recommendations to undertake professional development	Independently recognises the need to undertake professional development	Identifies own literacy needs	Actively seeks opportunities for further learning	Provides leadership in professional development at school
	Engages in self-assessments/ reflection	Examines teaching and assessment practices	Regularly engages in self-reflection to identify what went wrong and what went well in teaching	Develops a teaching portfolio to record and analyse teaching and assessment practices (annotated lesson plans, assessment results, etc.)	Evaluates the suitability of teaching portfolio for reflecting teaching and assessment practices	Assists colleagues in documenting teaching and assessment practices and in analysing the teaching portfolio
	Engages in peer-review of teaching performance	Responds to the results of peer review conducted by colleagues, department heads and principals	Seeks feedback from colleagues about teaching and assessment practices	Establishes constructive dialogue with colleagues, department heads and principals to identify strengths and weaknesses and areas for improvement	Develops a comprehensive self-report of the results of own peer-review of teaching performance	Leads the school community in engaging in peer- review of teaching performance
ENGAGE IN PD RELATED TO CURRICULUM-CONTENT	Searches new and relevant subject-content information	Recognises the need to seek out new and relevant information	Finds new and relevant information from available resources in school	Distinguishes the best sources of needed information	Seeks out new and relevant information through networking with other institutions	Provides assistance to colleagues in seeking out new and relevant information
	Identifies subject-content knowledge needs	Identifies curriculum content knowledge needs based on own classroom experience	Discusses and confirms own curriculum content knowledge with department heads/principals	Develops strategies to address curriculum content knowledge needs	Assesses the effectiveness of the strategies used to address curriculum content needs	Assists other colleagues in identifying and responding to their content knowledge needs

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
	Undertakes further education/ training	Acknowledges the need to enhance own knowledge of the curriculum content area	Expresses the interest/desire to undertake further education or training	Participates in appropriate education or training	Maintains continuity of participation in education or training leading to certification for higher teacher qualifications	Assists other teachers to identify and respond to their professional learning and development needs

DOMAIN 6: Teacher as a STAKEHOLDER PARTNER –this domain requires the teacher to write reports tailored to the information needs of various stakeholder groups. Also, the teacher needs to establish a continuing dialogue with these groups to inform them of the assessment results in ways that they can be easily understood and to discuss and address issues arising. The reports and the feedback gathered are used to identify ways to improve learning and teaching.

JOB ROLE	CRITERIA/ INDICATORS	PERFORMANCE STANDARDS				
		A	B	C	D	E
ENHANCEMENT OF COMMUNITY'S TRUST	Informs community of the assessment practices	Articulates the principles of AL Policy	Engages in discussion with any member of the community about school's AfL practices	Discusses comprehensively the principles and rationale of school's AfL practices	Clarifies issues related to school's AfL practices	Provides expert advice to colleagues about effective ways of informing community of school's AfL practices
	Reports to community about students' performance	Provides a report to parents about student overall achievement	Establishes dialogue with the parents to discuss the report	Uses various strategies of reporting to and discussion with the parents to highlight achievement, learning progress and accountability	Evaluates the suitability and effectiveness of reporting strategies to parents	Leads other teachers to develop reports for and establish dialogue with the parents
UTILISATION OF ASSESSMENT INFORMATION	Identifies key assessment and teaching issues for review	Identifies assessment and teaching issues	Initiates discussion with colleagues about key assessment and teaching issues	Provides information to support curriculum issues for review	Evaluates evidence to support curriculum issues for policy articulation	Drafts a curriculum program/policy for deliberation
	Collaborates with family to establish home activities to support student learning	Initiates dialogue with parents/guardians regarding home activities	Suggests specific home activities tailored to supporting specific students' learning needs	Works with parents to identify and discuss home activities that enhance students' learning	Assesses the effectiveness of suggested home activities in assisting students in their learning	Helps colleagues to effectively establish home-school collaboration

6.9. Latent Profile Analysis

6.9.1. Overview

This subsection of the study answers the third research question, that is:

What are the latent classes of teachers based on their AfL literacy?

To answer this question, a latent profile analysis (LPA) was conducted to investigate the teachers' patterns of AfL literacy using the factor scores derived from CFA. This analysis also reveals the number of groups of teachers that can be identified by their use of the AfL concept.

6.9.2. Fit Indices of Latent Profiles

The decision made as to how many numbers of latent profiles exist amongst teachers was based on the results of LPA modelling presented in Table 6.17.

Table 6.17

Model Fit Indexes of Tested Latent Profile Models of Teacher AfL Literacy (N=354)

<i>Fit Index</i>	<i>1 class</i>	<i>2 classes</i>	<i>3 classes</i>	<i>4 classes</i>	<i>5 classes</i>	<i>6 classes</i>
AIC	7501.984	5123.074	4131.131	3303.836	2885.873	7853.715
BIC	7556.493	5209.381	4249.236	3453.737	3067.572	8421.524
Adjusted BIC	7518.391	5149.053	4166.681	3348.957	2940.565	8024.627
Entropy		0.912	0.917	0.921	0.925	.920
Vong-LMR		-3738.992 (<i>p</i> <0.001)	-2542.537 (<i>p</i> =0.0886)	-2039.566 (<i>p</i> = 0.3830)	-1618.918 (<i>p</i> =0.0540)	-4019.906 (<i>p</i> =0.5043)
Adjusted LMR		2341.776 (<i>p</i> =0.0001)	984.447 (<i>p</i> = 0.0925)	823.318 (<i>p</i> = 0.3882)	422.732 (<i>p</i> =0.0559)	432.790 (<i>p</i> =.5045)
BLRT		-3738.992 (<i>p</i> <0.001)	-2542.537 (<i>p</i> <0.001)	-2039.566 (<i>p</i> <0.001)	-1618.918 (<i>p</i> <0.001)	-4019.906 (<i>p</i> =0.062)

Note: AIC = Akaike information criteria; BIC = Bayesian information criteria; LMR = Lo-Mendell-Rubin likelihood ratio test; BLRT = Bootstrap parametric likelihood ratio test.

As can be seen from Table 6.17, the information criteria indexes (AIC, BIC, and Adjusted BIC) showed a continual improvement (decreasing) as the number of classes were added to the models except in the case of the 6-class model where all values had significantly increased and these values were even higher than the 1-class model. The results initially suggested that amongst the tested models, the 5-class model was the best fit model because it had the lowest information criteria. Further, the results initially suggested that the 6-class model was not needed to account for the existing classes of teachers.

Analysis of the log-likelihood tests showed that only the 2-class model had significant *p*-values for both Lo-Mendell likelihood ratio tests. However, whilst the LMRs (V-LMR and adjusted LMR) showed that the model with two classes performed better than the rest of the models, the BLRT test results indicated otherwise. That is, up to 4-class models were not sufficient to account for the existing classes found in the data set as indicated by the significant *p*-values of BLRT estimates. These results, along with the

information criteria and the poor fit indexes of the 6-class model, led to the decision that the 5-class model was the best-fitting model to the data. The decision to accept the 5-class model was also supported by the entropy value, which was best in the 5-classes model (i.e., closest to 1). Research has shown that entropy values (Celeux & Soromenho, 1996) and the BLRT (Nylund et al., 2007) are better indicators of latent classes.

Figure 6.3 shows the graphical representation of the LPA model with 5-classes. As can be seen, this LPA modelling shows the differentiated levels of teachers in their use of the AfL concept. However, the pattern of the use of different aspects of the AfL concept was consistent across the six identified dimensions (i.e., the lines are fairly parallel to each other).

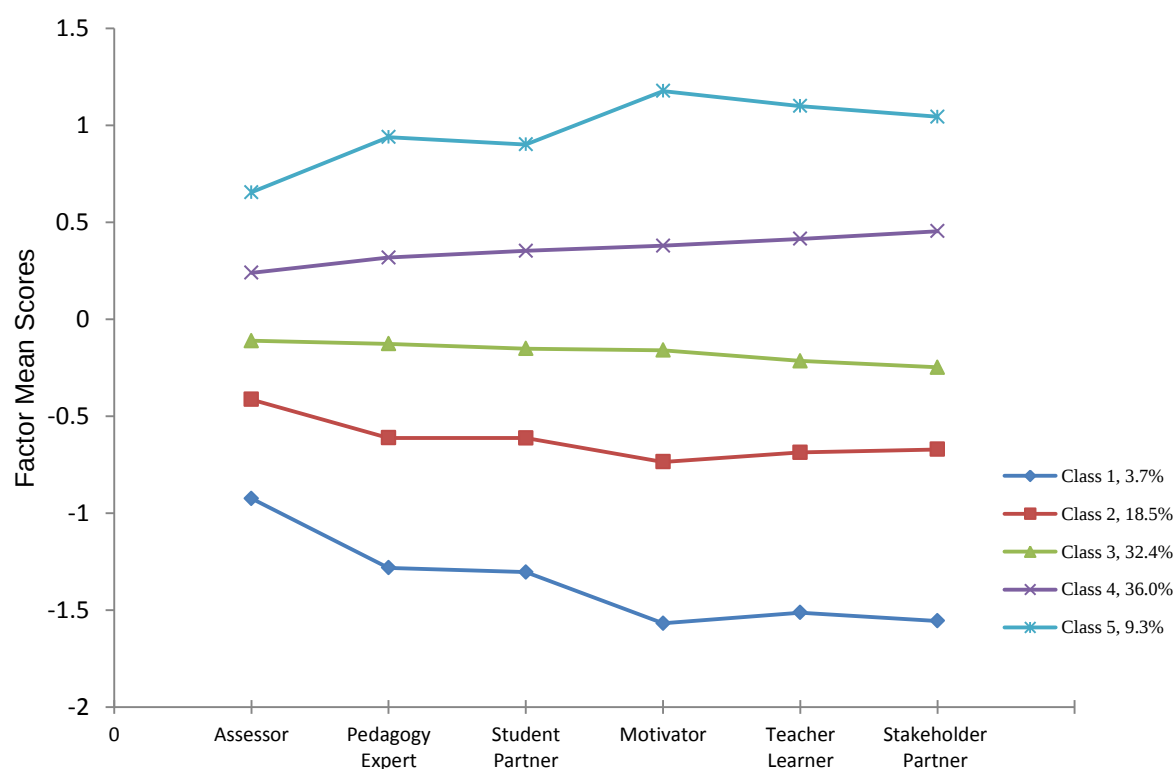


Figure 6.3. The latent profiles of teachers based on their AfL literacy.

6.9.3. Composition of Each Profile

One of the criteria in accepting the best latent profile model is the existence of a reasonable population in each class. From the results, majority of the teachers fell into Classes 3 and 4, comprising 32.4% ($n = 115$) and 36.0% ($n = 127$) respectively. Class 1 had the lowest number of samples, 3.7% (13) followed by Class 5 with 9.3% ($n = 33$) only and then Class 2 composed of 18.5% ($n = 66$). Although there is no strict rule as to the percentage of the total sample in each class, the percentages of all classes above were reasonable, especially when these results are applied to a larger population size.

6.9.4. Labels of Latent Profiles

Since the descriptions of performance standards in the tool were based on the levels of use of AfL practices, the labels of the classes observed were based on the levels of performance used to describe the stages of individual teacher's AfL development. The classes were labelled as follows:

1. Class 1, novice teachers who are unfamiliar with AfL principles and practices. They perform basic assessment tasks without understanding that supporting student learning is the central philosophy for the use of assessment and assessment information.
2. Class 2, developing teachers who show only limited skills and knowledge in using key AfL concepts and strategies. Their performance, although acceptable, is somewhat limited.

3. Class 3, proficient teachers who can use a range of AfL-related activities in the classroom. Their assessment practices are more effective and more coherent as a result of clear mechanisms and strategies.
4. Class 4, skilled teachers who consistently use AfL practices and reflect on and evaluate their own AfL knowledge and skills. Teachers in this class are able to modify and adjust their AfL practices to fit their own context. They have a high level of confidence in making important decisions about different issues in learning and teaching to help better develop assessment strategies and approaches.
5. Class 5, expert teachers who demonstrate leadership by providing professional advice and support to their colleagues in incorporating key AfL concepts and strategies into classroom practice.

Further analysis of these five classes of teachers revealed the general pattern of teacher AfL practices. Although the pattern of use of AfL practices was fairly consistent across the groups, a couple of points are noteworthy. Firstly, there was more consistency in the use of different AfL skills amongst the middle groups of teachers, Classes 2, 3, and 4, compared to Class 1 or 5. Secondly, the smallest gap across the five groups was observed in their use of AfL in the development of assessment tasks (Factor 1). In other words, there was relatively little difference in teachers' ability as assessors of student learning, regardless of their AfL literacy levels. These competencies include the ability of the teachers to develop and use assessment tasks and to ensure the trustworthiness of assessment activities and results. What appeared to separate the groups of novice and expert

teachers was the dimension of teachers as motivators (Factor 4).

The expert group (Class 5) and the novice group (Class 1) showed a further increase and decrease respectively in Factor 4 (compared to that of Class 4 and Class 2), which broadens the gap on Factor 4 between Class 1 and Class 5. This factor, which requires teachers to use assessment and assessment data to address the emotional investment of students in teaching and learning, appeared to be the factor that discriminated expert from novice teachers. Other factors discriminated teachers well according to their level of AfL literacy are Factor 5 (the ability of teachers to reflect on their assessment experiences to identify and address their professional development needs) and Factor 6 (the ability of the teachers to enhance the assessment literacy of stakeholders).

6.9.5. Mean Scores Differences between Teacher Classes

The five classes of teachers were further investigated using analysis of variance (ANOVA) to determine if the mean scores of each class of teachers were significantly different from the other classes across the six factors. Table 6.18 presents the results of ANOVA.

Table 6.18

One-Way Analysis of Variance of Mean Scores of Five Classes of Teachers in Six Factors
(*N*=354)

	<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Factor 1	Between groups	156.603	4	39.151	109.551	<.001
	Within groups	246.231	310	.357		
	Total	402.834	314			
Factor 2	Between groups	172.129	4	43.032	232.214	<.001
	Within groups	127.681	310	.185		
	Total	299.810	314			
Factor 3	Between groups	240.137	4	60.034	333.528	<.001
	Within groups	124.018	310	.180		
	Total	364.156	314			
Factor 4	Between groups	169.462	4	42.365	215.787	<.001
	Within groups	135.271	310	.196		
	Total	304.733	314			
Factor 5	Between groups	207.830	4	51.958	136.968	<.001
	Within groups	261.366	310	.379		
	Total	469.196	314			
Factor 6	Between groups	175.052	4	43.763	82.210	<.001
	Within groups	366.776	310	.532		
	Total	541.828	314			

*Note: SS = sum of squares; df = degrees of freedom; MS = mean square;
F = f-statistic.*

As can be seen from Table 6.18, there were significant differences amongst the mean scores of five classes of teachers across the six factors. Post-hoc comparisons (Table 6.19)

amongst classes of teachers in the six factors using Tukey HSD test indicated that there were significant differences between the means of:

- a) Class 1 and Class 2 teachers;
- b) Class 1 and Class 3 teachers;
- c) Class 1 and Class 4 teachers;
- d) Class 1 and Class 5 teachers;
- e) Class 2 and Class 3 teachers;
- f) Class 2 and Class 4 teachers;
- g) Class 2 and Class 5 teachers;
- h) Class 3 and Class 4 teachers;
- i) Class 3 and Class 5 teachers; and
- j) Class 4 and Class 5 teachers.

Further, it can be seen that Class 5 teachers have higher means in all six dimensions of teacher AfL literacy, followed by Class 4, then Class 3, Class 2 and the lowest means were observed in Class 1 teachers.

The results of ANOVA and the post-hoc test supported the existence of five classes of teachers because each class was distinct from the other classes of teachers in terms of levels of performance in all areas of AfL literacy. Hence, each class of teachers differed in their use of AfL principles and practices in the classroom.

Table 6.19

Post-hoc Comparison Test of the Means of Five Classes of Teachers for Six Factors using Tukey HSD (N=354)

Variables	Difference of Means									
	C1-C2	C1-C3	C1-C4	C1-C5	C2-C3	C2-C4	C2-C5	C3-C4	C3-C5	C4-C5
Factor 1	-.399**	-.757**	-1.303**	-1.883**	-.357**	-.904**	-1.484**	-.547**	-1.127**	-.580**
Factor 2	-.502**	-.971**	-1.441**	-2.095**	-.470**	-.940**	-1.594**	-.470**	-1.124**	-.654**
Factor 3	-.458**	-1.002**	1.1659**	-2.263**	-.544**	-1.201**	-1.804**	-.657**	-1.261**	-.604**
Factor 4	-.514**	-1.025**	-1.460**	-2.093**	-.511**	-.945**	-1.579**	-.435**	-1.068**	-.634**
Factor 5	-.450**	-.795**	-1.476**	-2.129**	-.345**	-1.026**	-1.679**	-.681**	-1.334**	-.653**
Factor 6	-.469**	-.689**	-1.392**	-1.916**	-.220**	-.923**	-1.446**	-.703**	-1.227**	-.524**

*Note: ** = mean difference is significant at the 0.01 level.*

6.10. Item Characteristics of the Teacher AfL Literacy Tool

6.10.1. Overview

This section answers the fourth research question, that is:

What are item characteristics of the final teacher AfL literacy tool?

This analysis was undertaken to analyse the tool at the item level. To answer this question, the psychometric properties of the tool derived from the Rasch analysis are discussed. The item characteristics of the teacher AfL literacy tool based on a probabilistic model of competence are presented to further support the reliability and validity of the tool. The analysis specifically used generalised partial credit modelling due to the fact that there is no guarantee that all items will have the same discrimination index (Muraki, 1992). I did not aim to produce a tool with items which have the same discrimination indexes otherwise it would contradict the overall aim for the use of the tool to discriminate between individual teachers so that appropriate professional development programs/interventions could be delivered to appropriately support each teacher in their AfL development.

In this analysis, the variables were numbered 1 to 54, which corresponds to the order of the teacher AfL skills in the final assessment tool presented in section 6.8.

6.10.2. The Reliability Index

The tool, with 42 teacher AfL practices, had a Cronbach's alpha equal to 0.98. This reliability index was relatively higher than those individual alpha values of each factor shown in Chapter 6.4.1. This was expected because the reliability index is partly dependent

to the total number of items. The high reliability index suggested that each item in the tool has a strong AfL component. Also, it gives high confidence to the user of the tool that it will give consistent results over time. This is an important characteristic of the tool as it is intended to be used regularly across career stages of teacher until they reach the highest standards in every area of AfL literacy.

6.10.3. Item Fit

Table 6.20 shows the items' weighted mean fit square estimates and their corresponding confidence intervals and t-values. As can be seen in the table, all weighted MNSQ values are within the given confidence interval and the residuals given by t-statistics are within the acceptable range of -2 to 2. These values indicate that all items in the scale are tapping into the construct of teacher assessment for learning literacy, and hence, the tool meets the requirements for construct validity. These results further corroborate the results of factor analyses discussed earlier.

The item fit measures of the tool showed that the items were unidimensional to some extent (as absolute unidimensionality does not exist). These results of unidimensionality should not be confused with the results obtained in factor analyses supporting the multidimensionality (six dimensions) of teacher assessment literacy. The item fit statistics presented here are equivalent to construct validity, which measures if each item in the tool contributes to the measurement of the construct.

Table 6.20

Item Parameter Estimates and Item Fit of the 42 Teacher AfL Skills

VARIABLES			UNWEIGHTED FIT			WEIGHTED FIT		
item	ESTIMATE	ERROR [^]	MNSQ	CI	T	MNSQ	CI	T
1 1	0.127	0.052	1.12 (0.86, 1.14)		1.2	1.12 (0.86, 1.14)		1.3
2 2	0.077	0.050	1.20 (0.86, 1.14)		1.3	1.06 (0.87, 1.13)		1.5
3 3	0.457	0.051	1.11 (0.86, 1.14)		1.3	1.10 (0.85, 1.15)		1.2
4 4	0.429	0.050	1.11 (0.86, 1.14)		1.9	1.09 (0.87, 1.13)		1.6
5 5	-0.373	0.048	1.12 (0.86, 1.14)		1.1	1.12 (0.86, 1.14)		1.4
6 6	-0.234	0.054	1.03 (0.86, 1.14)		0.5	1.03 (0.85, 1.15)		0.4
7 7	-0.112	0.050	1.08 (0.86, 1.14)		1.1	1.08 (0.86, 1.14)		1.1
8 8	-0.366	0.051	0.86 (0.86, 1.14)		-1.9	0.87 (0.86, 1.14)		-1.9
9 9	-0.146	0.053	0.93 (0.86, 1.14)		-0.9	0.93 (0.86, 1.14)		-0.9
10 10	-0.244	0.052	0.91 (0.86, 1.14)		-1.2	0.93 (0.86, 1.14)		-1.0
11 11	-0.080	0.053	0.92 (0.86, 1.14)		-1.1	0.92 (0.86, 1.14)		-1.0
12 12	-0.227	0.056	1.03 (0.86, 1.14)		0.4	1.05 (0.84, 1.16)		0.6
13 13	-0.596	0.052	1.08 (0.86, 1.14)		1.1	1.08 (0.85, 1.15)		1.0
14 14	-0.091	0.052	1.08 (0.86, 1.14)		1.0	1.07 (0.86, 1.14)		1.0
15 15	0.154	0.050	1.03 (0.86, 1.14)		0.4	1.02 (0.86, 1.14)		0.3
16 16	-0.149	0.054	1.00 (0.86, 1.14)		0.1	0.99 (0.86, 1.14)		-0.1
17 17	-0.154	0.052	0.89 (0.86, 1.14)		-1.3	0.95 (0.86, 1.14)		-1.2
18 18	0.471	0.049	1.04 (0.86, 1.14)		0.6	1.03 (0.87, 1.13)		0.4
19 19	0.393	0.050	0.96 (0.86, 1.14)		-0.5	0.94 (0.87, 1.13)		-1.0
20 20	0.230	0.050	0.90 (0.86, 1.14)		-1.4	0.89 (0.87, 1.13)		-1.1
21 21	0.204	0.049	0.92 (0.86, 1.14)		-1.0	0.93 (0.87, 1.13)		-1.1
22 22	0.180	0.050	0.89 (0.86, 1.14)		-1.5	0.89 (0.86, 1.14)		-1.6
23 23	0.576	0.049	0.92 (0.86, 1.14)		-1.2	0.88 (0.87, 1.13)		-1.8
24 24	-0.341	0.055	0.96 (0.86, 1.14)		-0.5	0.97 (0.85, 1.15)		-0.4
25 25	-1.153	0.049	1.11 (0.86, 1.14)		1.7	1.10 (0.86, 1.14)		1.7
26 26	-0.414	0.048	1.12 (0.86, 1.14)		1.6	1.11 (0.86, 1.14)		1.5
27 27	-0.784	0.052	1.00 (0.86, 1.14)		0.1	1.00 (0.85, 1.15)		-0.0
28 28	-0.412	0.050	0.90 (0.86, 1.14)		-1.3	0.90 (0.86, 1.14)		-1.4
29 29	-0.310	0.050	0.90 (0.86, 1.14)		-1.4	0.91 (0.86, 1.14)		-1.3
30 30	-0.311	0.050	0.91 (0.86, 1.14)		-1.3	0.92 (0.86, 1.14)		-1.2
31 31	0.257	0.052	0.88 (0.86, 1.14)		-1.6	0.89 (0.86, 1.14)		-1.6
32 32	0.071	0.052	0.95 (0.86, 1.14)		-0.6	0.96 (0.86, 1.14)		-0.5
33 33	0.196	0.047	1.09 (0.86, 1.14)		1.8	1.06 (0.87, 1.13)		1.7
34 34	0.352	0.052	0.96 (0.86, 1.14)		-0.5	0.97 (0.85, 1.15)		-0.4
35 35	0.009	0.051	1.08 (0.86, 1.14)		1.1	1.05 (0.86, 1.14)		0.6
36 36	0.569	0.050	1.01 (0.86, 1.14)		0.2	1.00 (0.87, 1.13)		0.0
37 37	0.209	0.051	1.06 (0.86, 1.14)		0.8	1.04 (0.85, 1.15)		0.6
38 38	-0.433	0.050	1.13 (0.86, 1.14)		1.2	1.05 (0.86, 1.14)		1.2
39 39	0.302	0.050	1.05 (0.86, 1.14)		0.7	1.05 (0.87, 1.13)		0.8
40 40	0.680	0.051	1.03 (0.86, 1.14)		0.4	1.04 (0.85, 1.15)		0.5
41 41	0.336	0.049	1.00 (0.86, 1.14)		0.1	1.02 (0.86, 1.14)		0.2
42 42	0.648*	0.326	0.97 (0.86, 1.14)		-0.4	0.98 (0.86, 1.14)		-0.2

An asterisk next to a parameter estimate indicates that it is constrained

Chi-square test of parameter equality = 2460.62, df = 41, Sig Level = 0.000

At the level of standards, where each standard is considered as a distinct item, the fit statistics were also within the acceptable MNSQ range and t-values (see Table 6.21). This means that each standard fits the construct well.

Table 6.21

Item Parameter Estimates and Fit of the Standards

VARIABLES			WEIGHTED FIT				
item	step	ESTIMATE	ERROR [^]	MNSQ	CI	T	
1	1	0		1.02 (0.79, 1.21)		0.2	
1	1	1	-2.782 0.166	1.03 (0.92, 1.08)		0.9	
1	1	2	-1.435 0.119	1.03 (0.94, 1.06)		0.9	
1	1	3	1.324 0.193	1.08 (0.75, 1.25)		0.7	
1	1	4	2.893*	1.53 (0.00, 2.01)		1.1	
2	2	0		1.04 (0.83, 1.17)		0.4	
2	2	1	-2.381 0.147	1.05 (0.93, 1.07)		1.4	
2	2	2	-1.044 0.120	1.06 (0.93, 1.07)		1.4	
2	2	3	0.645 0.165	1.09 (0.81, 1.19)		1.0	
2	2	4	2.780*	1.33 (0.24, 1.76)		0.9	
3	3	0		1.08 (0.85, 1.15)		1.1	
3	3	1	-2.822 0.139	0.99 (0.95, 1.05)		-0.3	
3	3	2	-0.238 0.131	1.03 (0.86, 1.14)		0.5	
3	3	3	0.394 0.200	1.00 (0.76, 1.24)		0.0	
3	3	4	2.665*	1.42 (0.01, 1.99)		0.9	
4	4	0		1.05 (0.88, 1.12)		1.4	
4	4	1	-2.112 0.130	1.02 (0.93, 1.07)		0.6	
4	4	2	-0.999 0.125	1.05 (0.91, 1.09)		1.1	
4	4	3	0.759 0.199	1.15 (0.75, 1.25)		1.1	
4	4	4	2.352*	1.34 (0.15, 1.85)		0.9	
5	5	0		1.02 (0.82, 1.18)		1.1	
5	5	1	-2.190 0.138	1.05 (0.93, 1.07)		1.2	
5	5	2	-0.385 0.116	1.05 (0.91, 1.09)		1.2	
5	5	3	0.952 0.166	1.00 (0.81, 1.19)		0.0	
5	5	4	1.623*	1.15 (0.67, 1.33)		0.9	
6	6	0		1.04 (0.64, 1.36)		0.3	
6	6	1	-3.557 0.208	0.96 (0.93, 1.07)		-1.1	
6	6	2	-0.638 0.115	0.95 (0.93, 1.07)		-1.5	
6	6	3	1.788 0.209	1.02 (0.72, 1.28)		0.2	
6	6	4	2.406*	1.14 (0.33, 1.67)		0.5	
7	7	0		0.97 (0.80, 1.20)		-0.3	
7	7	1	-2.498 0.160	1.05 (0.92, 1.08)		1.3	
7	7	2	-1.173 0.119	1.02 (0.94, 1.06)		0.5	
7	7	3	0.880 0.161	0.02 (0.82, 1.18)		0.2	
7	7	4	2.790*	1.10 (0.31, 1.69)		0.4	
8	8	0		0.93 (0.74, 1.26)		-0.5	
8	8	1	-2.557 0.178	0.97 (0.90, 1.10)		-0.5	
8	8	2	-1.392 0.120	0.97 (0.95, 1.05)		-1.2	
8	8	3	1.240 0.159	0.95 (0.82, 1.18)		-0.6	
8	8	4	2.709*	1.07 (0.42, 1.58)		0.3	
9	9	0		0.95 (0.71, 1.29)		-0.3	
9	9	1	-3.137 0.189	0.97 (0.93, 1.07)		-0.8	
9	9	2	-1.086 0.117	0.96 (0.94, 1.06)		-1.3	
9	9	3	1.564 0.192	0.88 (0.75, 1.25)		-0.9	
9	9	4	2.659*	1.35 (0.24, 1.76)		0.9	
10	10	0		0.98 (0.66, 1.34)		-0.1	
10	10	1	-3.495 0.206	0.94 (0.92, 1.08)		-1.7	
10	10	2	-0.534 0.117	0.95 (0.93, 1.07)		-1.4	
10	10	3	0.892 0.163	0.95 (0.82, 1.18)		-0.5	
10	10	4	3.137*	1.32 (0.24, 1.76)		0.9	
11	11	0		0.93 (0.72, 1.28)		-0.5	

11	11	1	-3.284	0.186	0.96 (0.93, 1.07)	-1.1
11	11	2	-0.775	0.116	0.94 (0.93, 1.07)	-1.7
11	11	3	1.509	0.201	1.00 (0.74, 1.26)	0.0
11	11	4	2.550*		1.27 (0.24, 1.76)	0.8
12	12	0			1.07 (0.57, 1.43)	0.4
12	12	1	-3.412	0.255	1.00 (0.88, 1.12)	0.0
12	12	2	-1.970	0.128	1.02 (0.92, 1.08)	0.5
12	12	3	1.643	0.177	0.97 (0.77, 1.23)	-0.2
12	12	4	3.739*		1.74 (0.00, 2.26)	1.2
13	13	0			1.09 (0.66, 1.34)	0.6
13	13	1	-2.764	0.189	1.00 (0.90, 1.10)	0.1
13	13	2	-1.227	0.118	1.03 (0.95, 1.05)	0.9
13	13	3	1.583	0.163	0.98 (0.81, 1.19)	-0.1
13	13	4	2.408*		1.17 (0.55, 1.45)	0.8
14	14	0			0.99 (0.73, 1.27)	-0.0
14	14	1	-3.259	0.181	1.04 (0.93, 1.07)	1.2
14	14	2	-0.436	0.119	1.04 (0.91, 1.09)	0.8
14	14	3	0.719	0.167	0.98 (0.81, 1.19)	-0.2
14	14	4	2.977*		1.20 (0.23, 1.77)	0.6
15	15	0			0.96 (0.84, 1.16)	-0.5
15	15	1	-2.499	0.143	0.99 (0.94, 1.06)	-0.3
15	15	2	-0.681	0.122	0.96 (0.90, 1.10)	-0.7
15	15	3	0.433	0.168	1.07 (0.81, 1.19)	0.7
15	15	4	2.747*		1.24 (0.23, 1.77)	0.7
16	16	0			0.99 (0.55, 1.45)	0.0
16	16	1	-4.041	0.257	0.96 (0.92, 1.08)	-1.0
16	16	2	-0.834	0.118	0.95 (0.94, 1.06)	-1.5
16	16	3	0.975	0.166	1.04 (0.81, 1.19)	0.4
16	16	4	3.901*		1.72 (0.00, 2.26)	1.1
17	17	0			0.95 (0.72, 1.28)	-0.3
17	17	1	-3.222	0.184	0.96 (0.93, 1.07)	-1.7
17	17	2	-0.609	0.117	0.98 (0.93, 1.07)	-1.8
17	17	3	1.136	0.178	0.95 (0.79, 1.21)	-0.4
17	17	4	2.695*		1.20 (0.31, 1.69)	0.6
18	18	0			1.01 (0.83, 1.17)	0.1
18	18	1	-2.442	0.155	1.02 (0.88, 1.12)	0.3
18	18	2	-2.093	0.126	1.03 (0.94, 1.06)	1.0
18	18	3	-0.084	0.146	0.96 (0.85, 1.15)	-0.5
18	18	4	4.619*		2.57 (0.00, 3.23)	1.4
19	19	0			0.99 (0.87, 1.13)	-0.2
19	19	1	-1.990	0.137	0.98 (0.89, 1.11)	-0.4
19	19	2	-1.763	0.122	0.98 (0.93, 1.07)	-0.5
19	19	3	1.067	0.197	0.93 (0.75, 1.25)	-0.5
19	19	4	2.687*		1.57 (0.00, 2.05)	1.1
20	20	0			0.86 (0.83, 1.17)	-1.7
20	20	1	-2.653	0.148	0.96 (0.94, 1.06)	-1.9
20	20	2	-0.758	0.123	0.93 (0.90, 1.10)	-1.4
20	20	3	0.091	0.159	1.01 (0.83, 1.17)	0.1
20	20	4	3.321*		1.49 (0.00, 2.03)	1.0
21	21	0			0.91 (0.86, 1.14)	-1.3
21	21	1	-1.851	0.136	0.97 (0.89, 1.11)	-0.5
21	21	2	-1.508	0.122	0.96 (0.93, 1.07)	-1.1
21	21	3	0.760	0.173	1.05 (0.80, 1.20)	0.5
21	21	4	2.599*		1.19 (0.22, 1.78)	0.6
22	22	0			0.95 (0.84, 1.16)	-0.6
22	22	1	-2.445	0.142	0.96 (0.94, 1.06)	-1.1
22	22	2	-0.863	0.121	0.95 (0.92, 1.08)	-1.1
22	22	3	0.713	0.178	0.99 (0.79, 1.21)	-0.0
22	22	4	2.595*		1.20 (0.23, 1.77)	0.6
23	23	0			0.86 (0.86, 1.14)	-1.8
23	23	1	-2.659	0.139	0.99 (0.94, 1.06)	-1.1

23	23	2	-0.868	0.127	0.94 (0.88, 1.12)	-0.9
23	23	3	-0.218	0.166	1.13 (0.82, 1.18)	1.4
23	23	4	3.745*		1.97 (0.00, 2.59)	1.2
24	24	0			1.00 (0.53, 1.47)	0.1
24	24	1	-3.746	0.255	0.98 (0.91, 1.09)	-0.5
24	24	2	-1.233	0.118	0.97 (0.94, 1.06)	-0.9
24	24	3	1.903	0.194	0.99 (0.74, 1.26)	-0.0
24	24	4	3.075*		1.36 (0.11, 1.89)	0.9
25	25	0			1.15 (0.66, 1.34)	0.9
25	25	1	-1.973	0.157	1.01 (0.85, 1.15)	0.2
25	25	2	-0.946	0.120	1.04 (0.93, 1.07)	1.0
25	25	3	0.365	0.117	1.07 (0.92, 1.08)	1.7
25	25	4	2.554*		0.97 (0.74, 1.26)	-0.1
26	26	0			1.16 (0.82, 1.18)	1.8
26	26	1	-1.707	0.137	0.98 (0.89, 1.11)	-0.3
26	26	2	-1.237	0.115	1.06 (0.94, 1.06)	2.0
26	26	3	1.596	0.182	0.97 (0.76, 1.24)	-0.2
26	26	4	1.349*		0.96 (0.68, 1.32)	-0.2
27	27	0			1.11 (0.63, 1.37)	0.6
27	27	1	-2.568	0.186	0.97 (0.87, 1.13)	-0.5
27	27	2	-1.402	0.120	0.97 (0.94, 1.06)	-0.9
27	27	3	1.812	0.163	0.97 (0.80, 1.20)	-0.3
27	27	4	2.157*		0.92 (0.63, 1.37)	-0.4
28	28	0			0.94 (0.72, 1.28)	-0.4
28	28	1	-2.951	0.172	0.98 (0.93, 1.07)	-0.6
28	28	2	-0.335	0.116	0.98 (0.92, 1.08)	-0.5
28	28	3	0.895	0.159	0.94 (0.83, 1.17)	-0.7
28	28	4	2.391*		1.05 (0.55, 1.45)	0.3
29	29	0			0.98 (0.80, 1.20)	-0.2
29	29	1	-2.070	0.153	0.99 (0.89, 1.11)	-0.1
29	29	2	-1.452	0.118	1.00 (0.94, 1.06)	-0.1
29	29	3	1.596	0.181	0.88 (0.77, 1.23)	-1.0
29	29	4	1.927*		1.05 (0.55, 1.45)	0.3
30	30	0			1.02 (0.77, 1.23)	0.2
30	30	1	-2.475	0.164	0.94 (0.91, 1.09)	-1.4
30	30	2	-1.044	0.118	0.95 (0.94, 1.06)	-1.7
30	30	3	0.841	0.152	0.99 (0.84, 1.16)	-0.1
30	30	4	2.678*		1.17 (0.46, 1.54)	0.7
31	31	0			1.02 (0.77, 1.23)	0.2
31	31	1	-3.353	0.173	0.99 (0.94, 1.06)	-1.2
31	31	2	-0.730	0.120	0.95 (0.91, 1.09)	-1.2
31	31	3	0.775	0.187	0.88 (0.77, 1.23)	-1.0
31	31	4	3.307*		1.70 (0.00, 2.24)	1.1
32	32	0			0.96 (0.77, 1.23)	-0.3
32	32	1	-3.173	0.172	0.96 (0.94, 1.06)	-1.3
32	32	2	-0.649	0.119	0.98 (0.92, 1.08)	-0.5
32	32	3	1.004	0.187	1.00 (0.77, 1.23)	0.0
32	32	4	2.818*		1.31 (0.13, 1.87)	0.8
33	33	0			1.11 (0.88, 1.12)	1.2
33	33	1	-1.720	0.128	1.06 (0.91, 1.09)	1.4
33	33	2	-0.801	0.126	1.01 (0.89, 1.11)	0.1
33	33	3	0.044	0.159	1.12 (0.84, 1.16)	1.5
33	33	4	2.478*		1.10 (0.37, 1.63)	0.4
34	34	0			1.03 (0.85, 1.15)	0.4
34	34	1	-2.744	0.137	0.99 (0.95, 1.05)	-0.5
34	34	2	-0.124	0.131	0.96 (0.87, 1.13)	-0.5
34	34	3	0.771	0.217	0.99 (0.72, 1.28)	-0.0
34	34	4	2.097*		1.30 (0.26, 1.74)	0.8
35	35	0			1.04 (0.83, 1.17)	0.5
35	35	1	-2.571	0.144	0.99 (0.95, 1.05)	-0.2
35	35	2	-0.528	0.119	0.99 (0.91, 1.09)	-0.2

35	35	3	1.348	0.209	0.95 (0.72, 1.28)	-0.3
35	35	4	1.751*		1.15 (0.48, 1.52)	0.6
36	36	0			1.07 (0.88, 1.12)	1.1
36	36	1	-2.126	0.128	0.98 (0.93, 1.07)	-0.6
36	36	2	-1.120	0.126	1.01 (0.90, 1.10)	0.2
36	36	3	0.885	0.216	0.91 (0.72, 1.28)	-0.6
36	36	4	2.361*		1.40 (0.00, 2.01)	0.9
37	37	0			1.06 (0.85, 1.15)	0.8
37	37	1	-2.475	0.135	0.98 (0.95, 1.05)	-0.8
37	37	2	-0.450	0.124	0.93 (0.90, 1.10)	-1.5
37	37	3	1.648	0.260	0.99 (0.62, 1.38)	0.0
37	37	4	1.277*		1.22 (0.46, 1.54)	0.8
38	38	0			1.15 (0.77, 1.23)	1.3
38	38	1	-1.963	0.163	1.10 (0.86, 1.14)	1.3
38	38	2	-1.751	0.123	1.05 (0.94, 1.06)	1.2
38	38	3	1.263	0.155	1.01 (0.83, 1.17)	0.1
38	38	4	2.451*		1.04 (0.52, 1.48)	0.2
39	39	0			1.04 (0.85, 1.15)	0.5
39	39	1	-2.199	0.143	1.02 (0.90, 1.10)	0.5
39	39	2	-1.663	0.121	1.04 (0.93, 1.07)	1.1
39	39	3	1.132	0.194	0.95 (0.75, 1.25)	-0.4
39	39	4	2.731*		1.45 (0.00, 2.01)	0.9
40	40	0			0.99 (0.89, 1.11)	-0.2
40	40	1	-2.374	0.124	0.97 (0.95, 1.05)	-1.2
40	40	2	-0.252	0.140	0.97 (0.84, 1.16)	-0.3
40	40	3	0.322	0.221	1.05 (0.72, 1.28)	0.4
40	40	4	2.304*		1.37 (0.02, 1.98)	0.8
41	41	0			1.01 (0.89, 1.11)	0.2
41	41	1	-1.868	0.124	0.99 (0.93, 1.07)	-0.3
41	41	2	-0.701	0.127	0.99 (0.89, 1.11)	-0.2
41	41	3	0.701	0.199	0.95 (0.76, 1.24)	-0.4
41	41	4	1.868*		1.23 (0.40, 1.60)	0.8
42	42	0			0.98 (0.89, 1.11)	-0.3
42	42	1	-2.268	0.126	0.99 (0.95, 1.05)	-0.2
42	42	2	-0.840	0.130	0.98 (0.89, 1.11)	-0.4
42	42	3	1.046	0.246	1.01 (0.66, 1.34)	0.1
42	42	4	2.062*		0.46 (0.00, 2.00)	1.0

6.10.4. Item Discrimination

The discrimination indexes of the items ranged from 0.48 to 0.91 as shown in Table 6.22.

This range of item discrimination values met the minimum requirement of .40 to usefully discriminate performance of individuals at a given trait level (Lee & Paek, 2014).

Methodologically, the varying discrimination indexes of the 42 items in the tool supported the use of generalised partial credit model, which assume that discrimination indexes are not necessarily the same for all items.

Table 6.22

Discrimination Indexes of the Final Items in Tool

Item	Discrimination Index
1 Designs assessment tasks	0.68
2 Uses rubrics to assess students' learning	0.69
3 Considers factors that affect students' performance	0.65
4 Avoids interference in task completion	0.59
5 Engages in social moderation	0.80
6 Translates learning standards to learning outcomes	0.74
7 Identifies appropriate teaching methods	0.63
8 Considers students' prior knowledge in lesson planning	0.71
9 Considers students' current level of abilities	0.62
10 Considers students' interest	0.63
11 Plans lessons according to students' learning needs	0.67
12 Tailors lessons to available resources	0.62
13 Develops teaching and learning resources	0.82
14 Involves students in the development of learning outcomes	0.86
15 Involves students in the development of criteria and standards	0.74
16 Makes students understand the learning outcomes	0.68

17	Explains the success criteria and standards (rubrics)	0.89
18	Develops students' capabilities in self and peer assessments	0.91
19	Engages students in self-assessment	0.78
20	Engages students in peer-assessment	0.68
21	Moderates feedback and results of self and peer assessment	0.78
22	Gives feedback on students' strengths and weaknesses	0.78
23	Assists students in using feedback to feed forward	0.68
24	Conducts assessment with consideration of student background and culture.	0.69
25	Uses a variety of teaching and learning methodologies	0.82
26	Develops an environment of trust	0.84
27	Ensures openness in the class	0.84
28	Uses assessment to build students' interest to learn	0.78
29	Demonstrates belief in the ability of every student to improve	0.82
30	Affirms students' good performance	0.76
31	Clarifies students misconceptions	0.81
32	Reinforces positive learning attitude of students	0.78
33	Participates in professional development related to assessment	0.71
34	Engages in self- assessments/ reflection	0.82
35	Engages in peer-review of teaching performance	0.86
36	Searches new and relevant subject-content information	0.84
37	Identifies subject-content knowledge needs	0.79
38	Undertakes further education/ training	0.81
39	Informs community of the assessment practices	0.85
40	Reports to community about students' performance	0.68
41	Identifies key assessment and teaching issues for review	0.81
42	Collaborates with family to establish home activities to support student learning	0.75

The values above indicated that all items in the scale could potentially determine the level of individual teachers' AfL literacy. This characteristic of the scale is critical to its purpose of generating data for the development of a needs-based professional development

program because it provides empirical evidence of the proper placement of teachers at each assessment skill. In other words, the results of teachers' self-assessment using the tool reflect their actual AfL literacy level. Therefore, the information gathered using the tool could be used to identify what the individual teachers can actually do and what they need to do further to enhance their AfL literacy. This gap between teachers' actual performance and the next level of performance (i.e., adjacent higher level of performance) gives the zone of proximal development, which is a critical input for developing a needs-based professional development programs and for providing support and enabling mechanism for individual teachers to help them advance in their AfL literacy development.

6.10.5. Difficulty Index

The IRT-based difficulty index is typically identified as “item delta” in a polytomous scale (see Table 6.20). Results showed that the 42 AfL skills were evenly dispersed along a continuum of difficulty from -1.0 to +0.85 logits. Figure 6.4 is a graphical presentation of the distribution of items from easiest (bottom most) to the most difficult one (topmost) whilst Figure 6.5 shows the distribution of standards also based on their individual difficulty level.

Looking at the item thresholds for the response categories of each item (i.e., given as estimates in Table 6.21), it can be seen that there was no category disordering observed. This means that the response categories functioned as intended, that is, the progression of standards of each assessment skill in the tool followed the concept of underlying growth continuum. In other words, the standards were progressing according to levels of sophistication of performance.

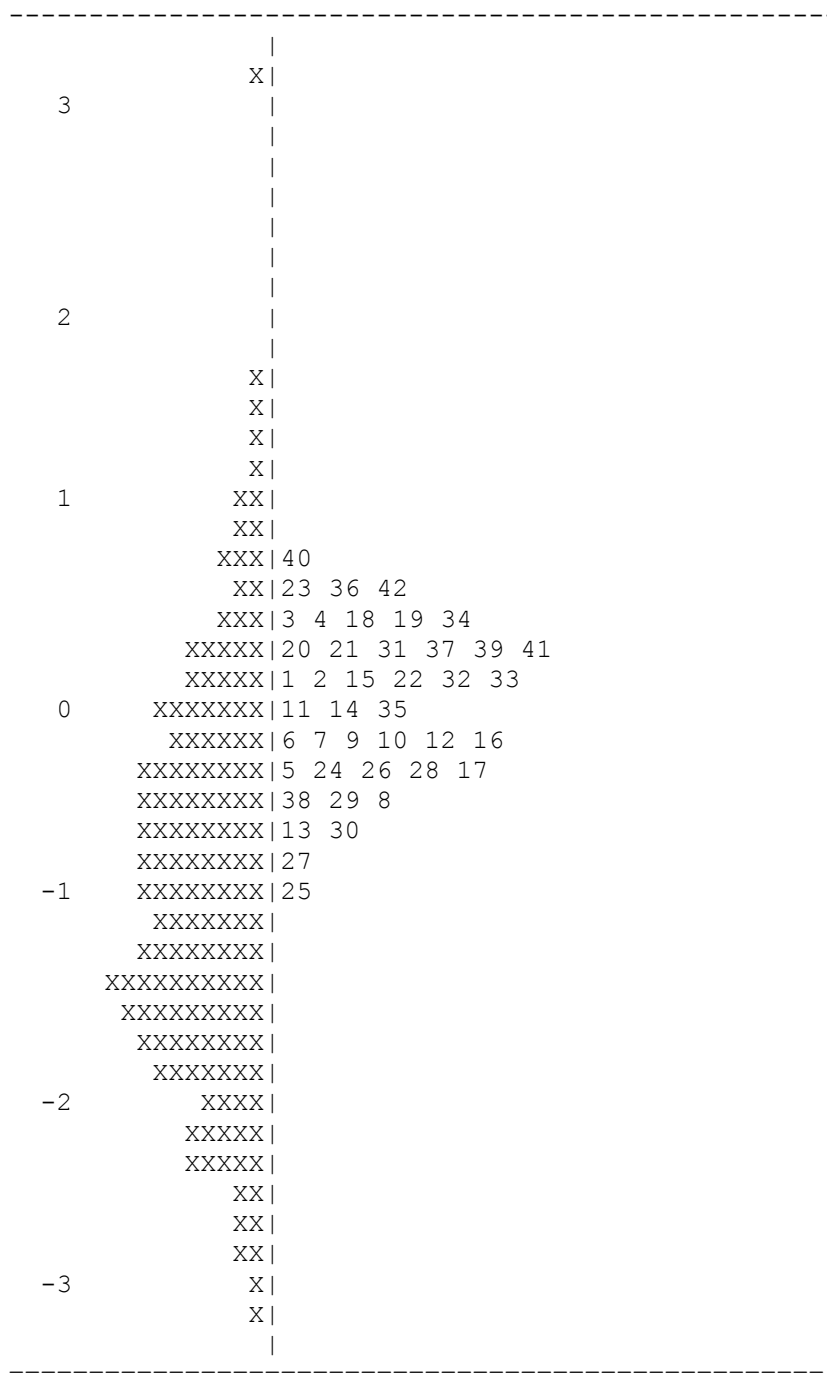


Figure 6.4. Map of latent distributions of teacher AfL practices based on difficulty level, where X represents 2.4 numbers of teachers.

As to the order and locations of the standards, which could further support the validity of the information gathered using the tool, it can be seen in Figure 6.5 that the standards were evenly distributed along the logit continuum ranging from -3.15 to 3.0 logit scores, which is within the recommended spread of the items for an excellent scale (Andrich & Styles, 2004). This means that the tool had reasonably good proportions of items showing easy, medium, and difficult levels of the AfL-related tasks.

Another finding that can be seen in the latent map distribution is the ability of the tool to account for the range of teachers' ability. Since the Rasch model calibrates item difficulty and participants' ability in the same scale, the match between the range of item difficulty and the range of teachers' level of AfL literacy could be determined. From Figure 6.5, the spread of the standards showed some teachers (labelled as Xs) located at the top (representing those who have very high level AfL literacy) and bottom (representing those who have very low level AfL literacy) of the scale but mostly clustering at the mid-ability level. It can be seen that there were teachers whose ability matched the logit of the easiest performance standard and similarly, there were teachers whose ability matched the logit of the most difficult performance standard.

Both the item map (criteria) and threshold map (standards) show that teachers had varying levels of AfL literacy. The distribution of teachers, given by the Xs on the left, showed that there were low performing and high performing teachers. From the point of view of the Rasch model, the range of the item difficulty matches the ability range of the teachers. Hence, the tool effectively measures the AfL ability of the teachers.

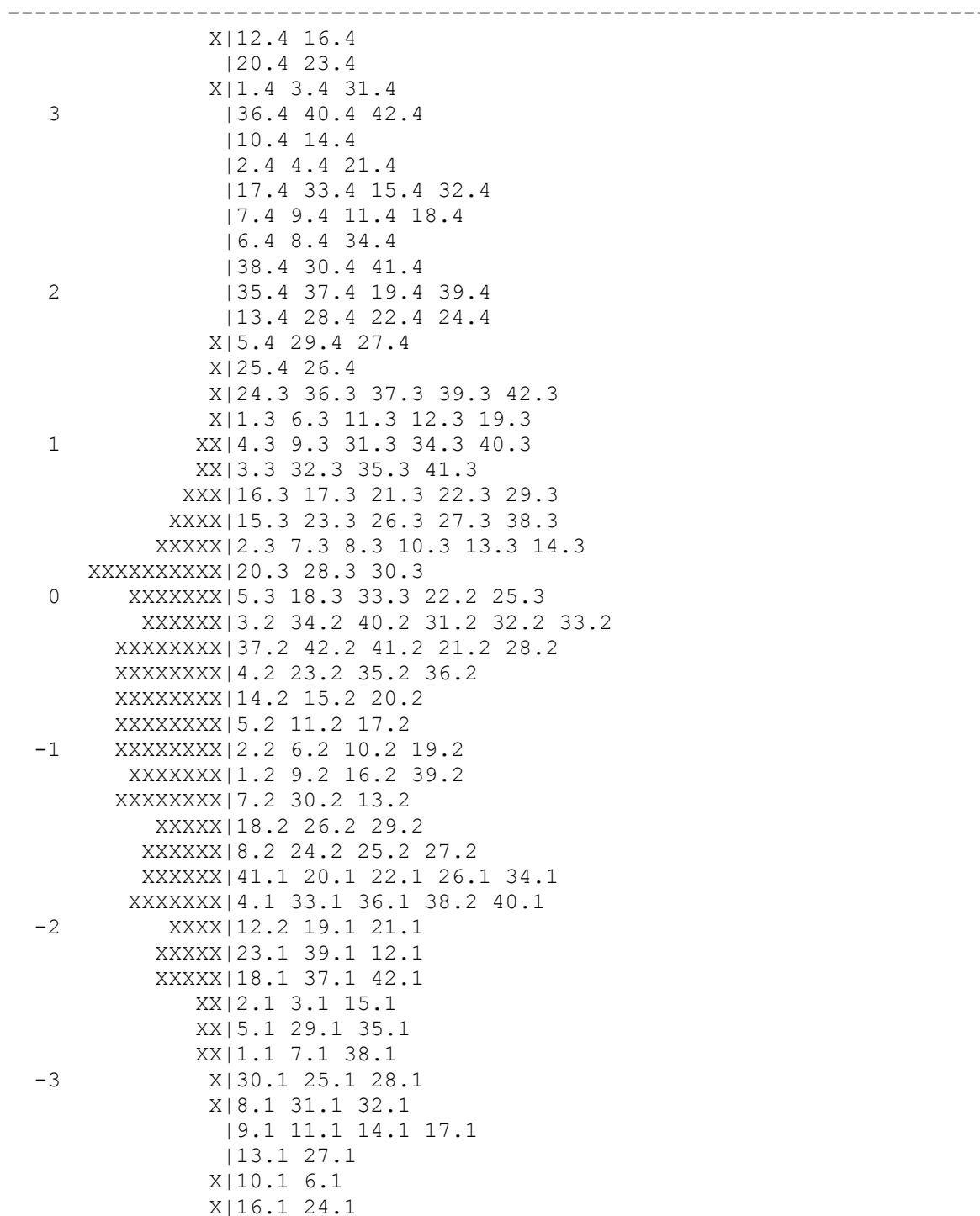


Figure 6.5. Map of latent distributions of teacher AfL standards based on difficulty level,

where X represents 4.8 numbers of teachers

6.10.6. Hierarchy of Teacher AfL Skills

This section shows the arrangement of the criteria based on their difficulty indexes (see Figure 6.6). The hierarchy indicates the relative placement of criteria along the continuum of AfL practices starting from the easiest (or the most common) AfL practice (at the bottom) to the most difficult (or rarely done) AfL practice (at the top). This continuum can indicate teachers' assessment literacy development across time.

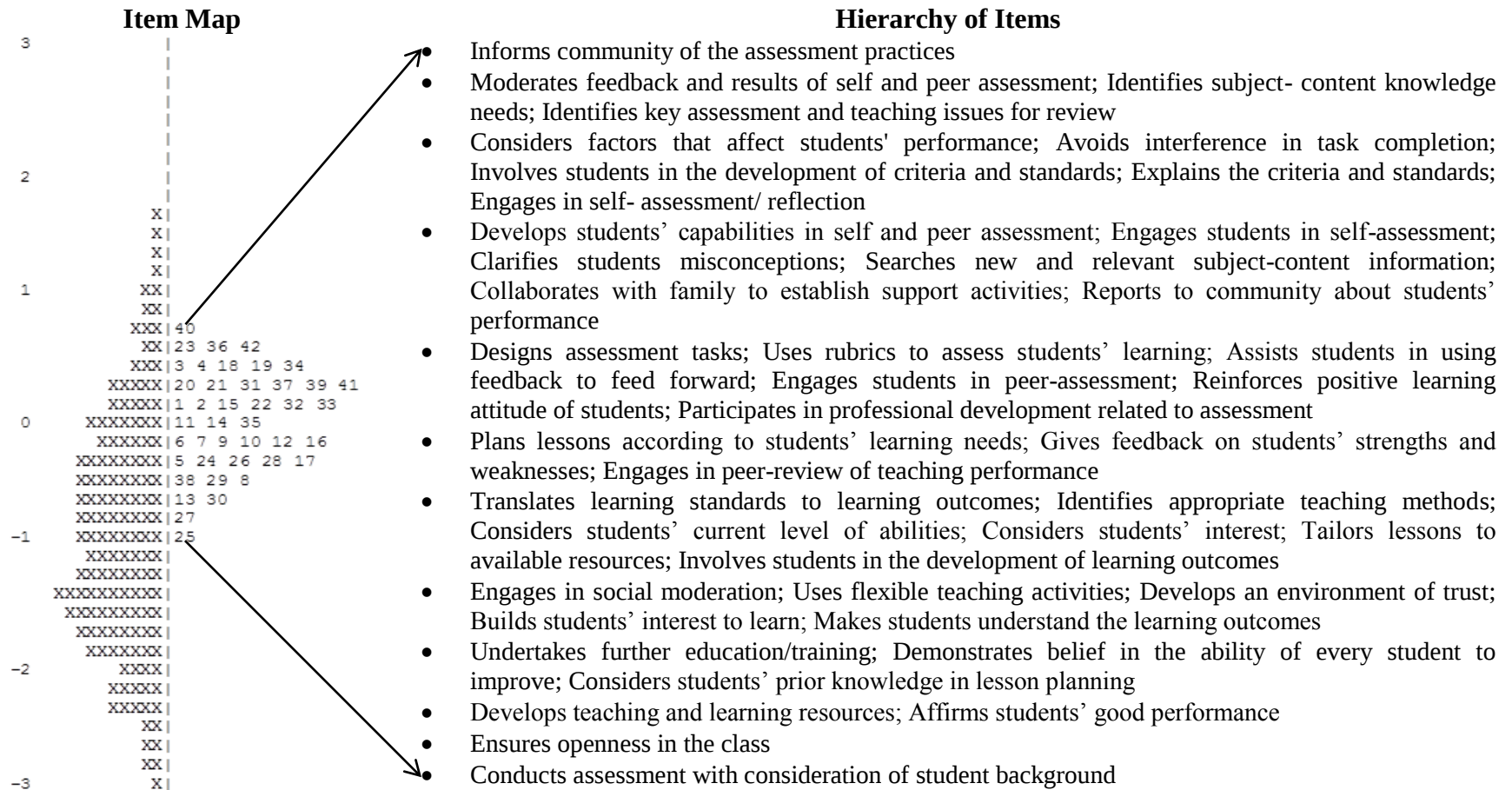


Figure 6.6. Hierarchical order of items based on difficulty level. Items located below are the most common teacher AfL skills used by teachers whilst items on top are rarely performed by teachers.

Apart from the hierarchy of indicators shown above, a hierarchy of individual standards was also generated. The item map from the generalised particle credit analysis showed the relative placement of each standard using their threshold value. As a guide, the clusters of standards are identified using an arbitrary lower and upper cut-off threshold values.

The identification of clusters to establish the cut off points for different levels of AfL performance was based on the two criteria used by Bateman and Griffin (2003). First, the items that cluster together in the same difficulty range are identified; second, the common theme of the clustered items are determined by looking at their similarities or common thematic lines. Each cluster of items should exhibit demand of ability level with their adjacent level. As shown in the map of latent distributions and threshold in Figure 6.7, thresholds between the adjacent standards were clustered altogether. For example, the threshold values between standards A and B, as indicated by Y.1s, *where Y is a number of the criteria*, were all clustered at the lower level between -3.38 to -1.55 logits; thresholds between standards B and C, as indicated by Y.2s, were between -1.56 to -0.19 logits; thresholds between standards C and D, as indicated by Y.3s were between -0.18 to 1.41 logits; and thresholds between standards D and E, as indicated by Y.4s, were between 1.42 to 3.12 logits.

There are four distinct cut-off points, hence, five major clusters in the data set. These clusters could indicate stages of teacher AfL literacy development. This result validated the results of latent profile analysis, showing five classes of teachers based on their AfL practice.

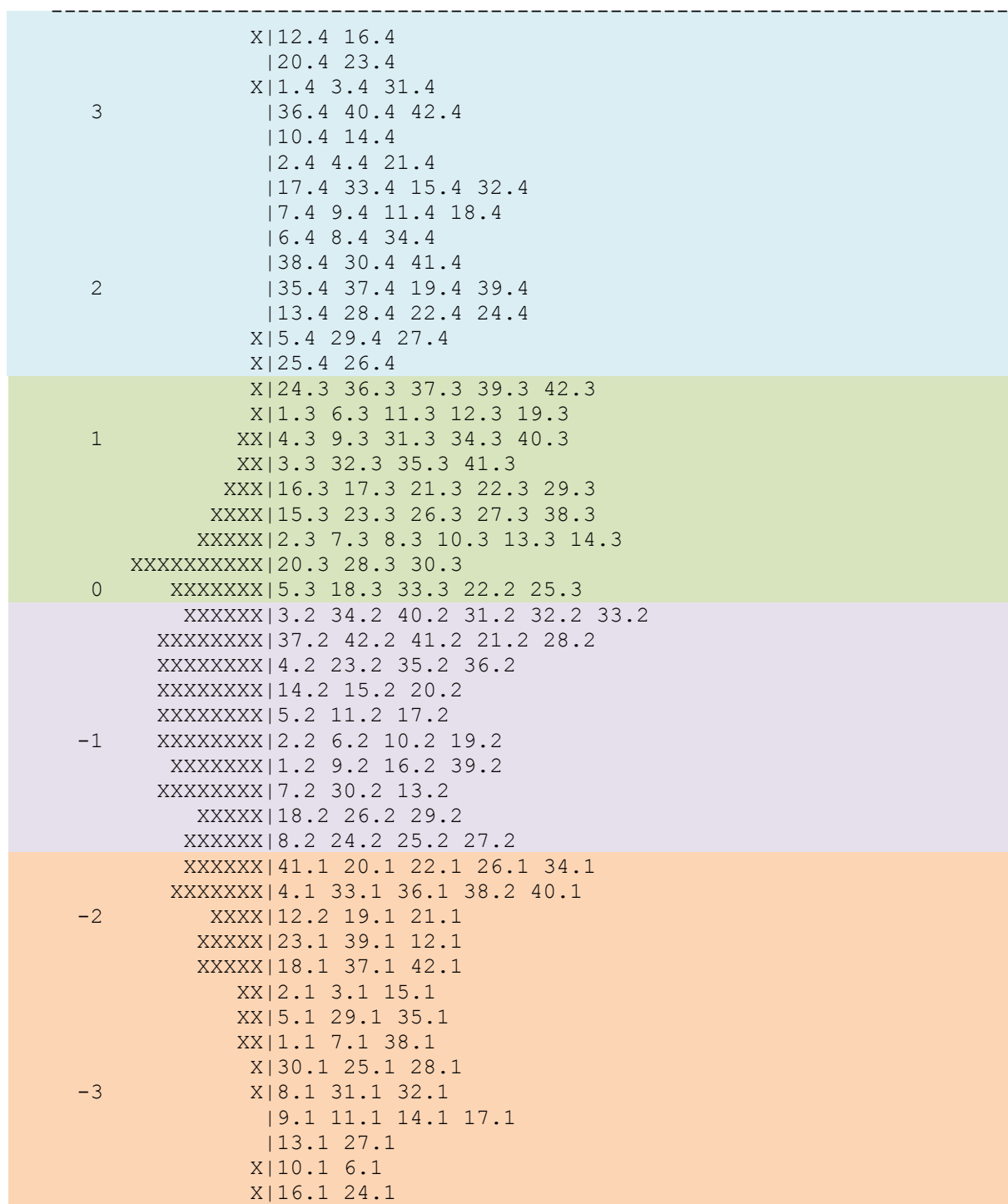


Figure 6.7. The distribution of standards showing the arbitrary cut-off logit scores.

6.11. Summary

In this section I have presented the results of the main study. Factor analyses and IRT have provided empirical evidence for the dimensionality and psychometric properties of the teacher A/fL literacy tool. Results showed that the tool could be used with a high level of confidence and that any results obtained from using it as a tool for teacher self-assessment reflect the actual self-perceived A/fL literacy level of teachers. Furthermore, latent profile analysis and IRT parametisation led to the establishment of a typology for teacher A/fL literacy, which can be used to determine individual teachers A/fL literacy needs, which could be used to design and implement a more needs-based professional development program.

CHAPTER 7 - CROSS-COUNTRY VALIDATION

7.1. Introduction

The aim of this chapter is to show how further evidence was gathered to support the results of the main study. A new data set drawn from Philippine teachers' self-assessment was used to verify the factor structure and typology of the A/fL literacy levels extracted from Brunei data. This part of the study replicated the method used in the main study, except that EFA was not included as an initial analysis. The results of the cross-country validation help show the applicability of the teacher A/fL literacy tool to other contexts. This section addresses the fifth and final research question given below:

Can the tool be applied and used in another context?

To achieve the aim of this chapter, I present the results of confirmatory factor analysis (CFA), hierarchical-CFA, exploratory structural equation modelling (ESEM), and latent profile analysis (LPA) of the Philippines data. There was no determination of measurement invariance property of the tool to adhere to the prior agreement with the Brunei Ministry of Education.

7.2. Population Sample

There were 383 school teachers from the Philippines who participated in this study, 54.83% (210) from secondary schools and the rest from the primary schools. Also, 110 of the participants were from private schools, 40 primary school teachers and 70 secondary school teachers.

7.3. Teachers' Self-Assessment

The final version of the teacher AfL literacy tool derived from the main study with six factors and 42 items was used for teachers' self-assessment. Teachers were gathered in three different venues in the Philippines. As in the main study, before teachers engaged in self-assessment, the individual AfL indicators and standards were explained to them to check their understanding and to clarify the task requirement for each standard.

7.4. Factor Analyses

7.4.1. Overview

The data gathered in the Philippines were subjected to CFA, second-order CFA and ESEM using the weighted least squares means and variance adjusted (WLSMV) with oblique geomin rotation. The same fit indexes were used to evaluate the model fit of the 6-factor model in the Philippine data set.

7.4.2. CFA Model

Results of CFA show that the 6-factor model fit the data quite well as shown in Table 7.1. The model fit statistics of the 6-factor model of teacher's AfL literacy, given by chi-square value ($X^2=924.372$; $df=676$), although significant, its ratio to degrees of freedom was 1.367, which indicated a perfect fit model (Kline, 2005, 2010). Analysis of fit statistics showed that all indexes were excellent with RMSEA of 0.036, CFI of 0.963, and TLI of 0.961, except for WRMR of 0.987, which indicated a good fitting model.

Table 7.1

Evaluation of Fit Indexes of the 6-Factor Confirmatory Model (N=383)

Indexes	Perfect Fit	Good Fit	Sample Statistics	Decision	Reference
χ^2/df	≤ 2	≤ 3	1.367	Perfect fit	Kline (2005, 2010)
RMSEA	$\leq .05$	$\leq .08$	0.036	Perfect fit	Stegier & Lind (1980); Hooper, Coghlan & Mullen (2008)
CFI	$\geq .95$	$\geq .90$	0.963	Perfect fit	Hu & Bentler (1999) Tabachnick & Fidell (2007)
TLI	$\geq .95$	$\geq .90$	0.961	Perfect fit	Hu & Bentler (1999)
WRMR		≤ 1.0	.987	Slightly over fit	Hancock & Mueller (2006); Yu & Muthen (2002)

Note: $\chi^2 = 924.372$; $df = 676$.

Table 7.2 presents the unstandardised and standardised estimates along with their standard errors. This result further supports the 6-factor model as all unconstrained loading estimates were significant at 0.05 level, as supported by the absolute values of all ratios between standards estimates and their corresponding standards errors, which were all greater than or equal to 1.96.

As can be seen in Table 7.2, the factor loadings of the indicators of teacher A/L literacy ranged from .707 to .820 with variance explained between 49.98% to 67.24%. Also, as can be seen from the table, all standardised loadings have strong magnitude, which shows that each variable contributed to the measurement of its corresponding factor of teacher A/L literacy. The high factor loadings observed supported the convergent validity of the tool. Figure 7.1 is the visual representation of the factor structure extracted from Philippine data.

Table 7.2

Unstandardised and Standardised Loadings for the 6-Factor Confirmatory Model of Teacher AfL Literacy (N=383)

Factors	Variables	Unstandardized		Standardized		R ²
		Estimate	S.E.	Estimate	S.E.	
Assessor	v3 Designs assessment tasks	1.000	0.000	0.763	0.024	0.5822
	V7 Uses rubrics to assess students' learning	1.005	0.056	0.707	0.030	0.4998
	V9 Considers factors that affect students' performance	1.005	0.056	0.707	0.030	0.4998
	V10 Avoids interference in task completion	1.225	0.073	0.798	0.029	0.6368
	V12 Engages in social moderation	1.211	0.063	0.773	0.023	0.5975
Pedagogy Expert	V25 Translates learning standards to learning outcomes	1.000	0.000	0.762	0.024	0.5806
	V28 Identifies appropriate teaching methods	1.047	0.065	0.709	0.030	0.5027
	V29 Considers students' prior knowledge in lesson planning	1.181	0.063	0.800	0.024	0.6400
	V30 Considers students' current level of abilities	1.127	0.059	0.764	0.024	0.5837
	V31 Considers students' interest	1.150	0.060	0.779	0.024	0.6068
	V32 Plans lessons according to students' learning needs	1.144	0.059	0.775	0.024	0.6006
	V33 Tailors lessons to available resources	1.089	0.051	0.766	0.023	0.5868
	V34 Develops teaching and learning resources	1.144	0.067	0.745	0.028	0.5550
Student Partner	V17 Gives feedback on students' strengths and weaknesses	1.000	0.000	0.745	0.028	0.5550
	V19 Assists students in using feedback to feed forward	1.098	0.059	0.711	0.029	0.5055
	V27 Involves students in the development of learning outcomes	1.055	0.040	0.820	0.024	0.6724
	V35 Makes students understand the learning outcomes	1.211	0.063	0.773	0.023	0.5975
	V37 Involves students in the development of criteria and standards	1.133	0.069	0.724	0.025	0.5242
	V39 Explains the criteria and standards	1.193	0.071	0.762	0.024	0.5806
	V40 Develops students' capabilities in self and peer assessment	1.233	0.070	0.787	0.024	0.6194
	V41 Engages students in self-assessment	1.195	0.068	0.763	0.024	0.5822
	V42 Engages students in peer-assessment	1.186	0.071	0.758	0.024	0.5746
	V43 Moderates feedback and results of self and peer assessment	1.273	0.073	0.813	0.021	0.6610
Motivator	V46 Uses flexible teaching activities	1.000	0.000	0.709	0.030	0.5027
	V47 Conducts assessment with consideration of student background	1.014	0.046	0.788	0.030	0.6209
	V48 Develops an environment of trust	1.006	0.057	0.708	0.028	0.5013
	V49 Ensures openness in the class	1.005	0.056	0.707	0.030	0.4998
	V50 Builds students' interest to learn	1.128	0.053	0.794	0.022	0.6304
	V51 Demonstrates belief in the ability of every student to improve	1.123	0.052	0.790	0.023	0.6241
	V53 Affirms students' good performance	1.089	0.051	0.766	0.023	0.5868
	V54 Clarifies students misconceptions	1.106	0.058	0.778	0.026	0.6209
	V55 Reinforces positive learning attitude of students	1.042	0.055	0.733	0.027	0.5373
Teacher	V58 Participates in professional development related to assessment	1.000	0.000	0.762	0.024	0.5806
Learner	V59 Engages in self- assessment/ reflection	1.201	0.067	0.783	0.025	0.6131
	V60 Engages in peer-review of teaching performance	1.129	0.066	0.736	0.028	0.5417
	V62 Identifies subject- content knowledge needs	1.225	0.073	0.798	0.029	0.6368
	V63 Searches new and relevant subject-content information	1.144	0.067	0.745	0.028	0.5550
	V64 Undertakes further education/ training	1.089	0.051	0.766	0.023	0.5868
Stakeholder Partner	V66 Collaborates with family to establish support activities	1.000	0.000	0.777	0.026	0.6037
	V67 Informs community of the assessment practices	1.014	0.046	0.788	0.030	0.6209
	V68 Reports to community about students' performance	1.055	0.040	0.820	0.024	0.6274
	V70 Identifies key assessment and teaching issues for review	1.050	0.046	0.816	0.026	0.6659

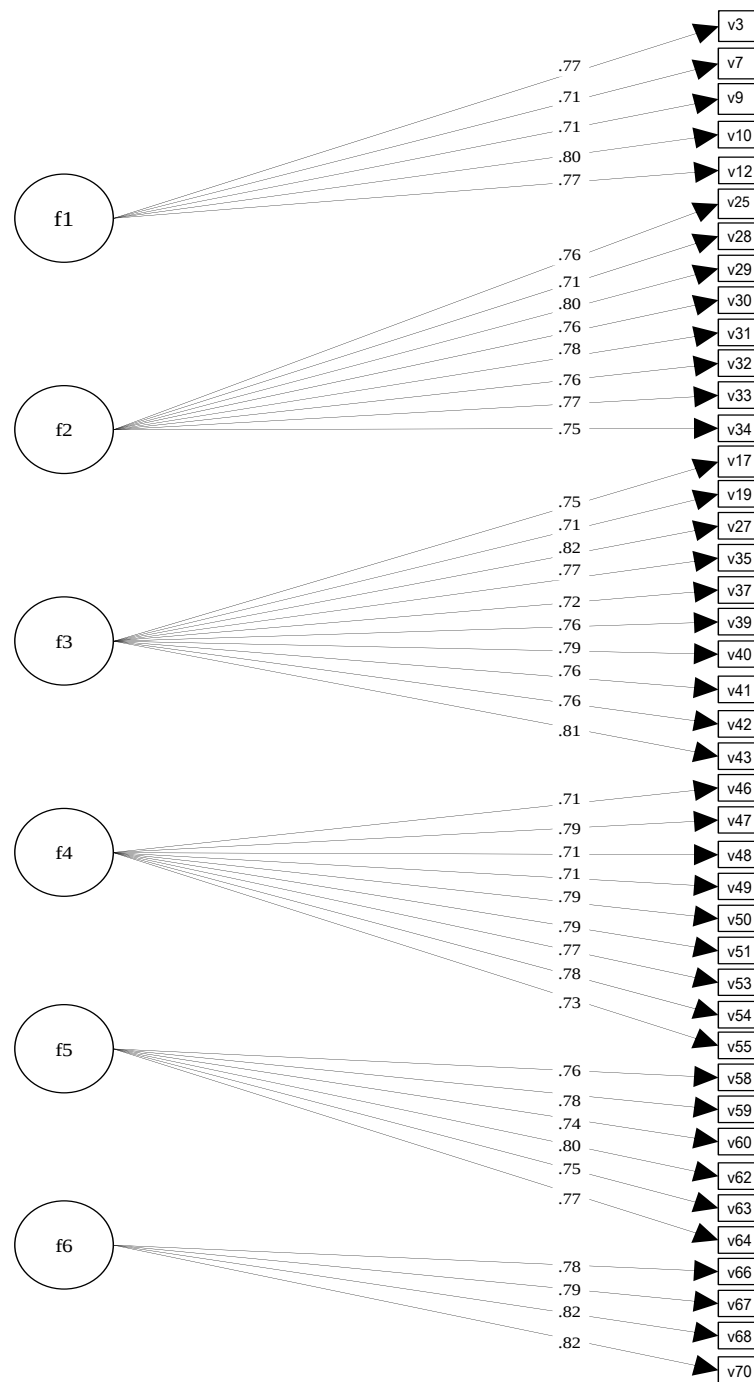


Figure 7.1. The 6-factor model of teacher AfL literacy using data from Philippine teachers.

The numbers shown are the standardised factor loadings of variables to their corresponding factor.

Table 7.3 presents the correlations amongst the six factors. All factor correlations are high (0.758 – 0.941), which makes the divergent validity problematic. This result is similar to the result obtained in the main study, which suggested that the six factors could be measuring the same latent construct. There was no modification index given to adjust the model, and hence, further investigation was needed using second-order CFA.

Table 7.3

Correlations amongst the Six Factors of the Confirmatory Model of Teacher AfL Literacy (N=383 Philippine Teachers)

	f1	f2	f3	f4	f5	f6
f1 Teachers as Assessors	1					
f2 Teachers as Pedagogy Experts	0.941	1				
f3 Teachers as Student Partners	0.883	0.932	1			
f4 Teachers as Motivators	0.855	0.867	0.937	1		
f5 Teachers as Teacher Learners	0.758	0.800	0.900	0.940	1	
f6 Teachers as Stakeholder Partners	0.809	0.804	0.865	0.938	0.924	1

7.4.3. Second-Order Factor

The existence of a second-order factor was also tested to determine if the first-order factors loaded onto a generic factor derived in the main study. Result showed that the chi-square value ($X^2 = 1494.082$; $df = 854$) was significant but its ratio to degrees of freedom was less than 2, and hence, the result still supported the model. Analysis of the fit indexes (Table 7.4) indicated a good fit of the specified model to the observed data. The RMSEA (0.040) was below .05. The CFI (0.953), and TLI (0.951) were all above 0.95, suggesting a perfect fit model of the second-order factor structure of teacher AfL literacy. Similarly, the result of WRMR (0.989) showed a good fitting model.

Table 7.4

Evaluation of Fit Indexes of the Second-order Confirmatory Factor Model (N=383)

Indexes	Perfect Fit	Good Fit	Sample Statistics	Decision	Reference
χ^2/df	≤ 2	≤ 3	1.750	Perfect fit	Kline (2005, 2010)
RMSEA	$\leq .05$	$\leq .08$	0.040	Perfect fit	Stegier & Lind (1980); Hooper, Coghlan & Mullen (2008)
CFI	$\geq .95$	$\geq .90$	0.953	Perfect fit	Hu & Bentler (1999) Tabachnick & Fidell (2007)
TLI	$\geq .95$	$\geq .90$	0.951	Perfect fit	Hu & Bentler (1999)
WRMR		≤ 1.0	.989	Slightly over fit	Hancock & Mueller (2006); Yu & Muthen (2002)

Note: $\chi^2 = 1494.082$; $df = 854$.

The validity measures of the second-order factor model were also investigated. Table 7.5 presents the unstandardised and standardised loadings and the corresponding R^2 values of each variable. From the table, all factor loadings were considerably high (.706 to .963), which meant that indicators were correlating highly with their respective factor. In other words, the individual items were contributing to the measurement of teacher A/L literacy construct. Hence, the model tested had high convergent validity.

Furthermore, it can be seen from the table that all absolute values of ratios between standards estimates and their corresponding standards errors were all greater than or equal to 1.96, which meant that all unconstrained loading estimates were significant at 0.05. Figure 7.2 shows the second-order CFA model where all six factors are loading onto one second-order general factor.

Table 7.5

Unstandardised and Standardised Loadings for the 6-Factor Second-order Confirmatory Model of Teacher AfL Literacy (N=383

Philippine Teachers)

Factors	Variables	Unstandardized		Standardized		R ²
		<i>Estimate</i>	<i>S.E.</i>	<i>Estimate</i>	<i>S.E.</i>	
Assessor	V3 Designs assessment tasks	1.000	0.000	0.737	0.028	0.5432
	V7 Uses rubrics to assess students' learning	1.135	0.069	0.724	0.025	0.5242
	V9 Considers factors that affect students' performance	1.194	0.071	0.762	0.024	0.5806
	V10 Avoids interference in task completion	1.148	0.068	0.746	0.028	0.5565
	V12 Engages in social moderation	1.017	0.047	0.787	0.024	0.6194
Pedagogy Expert	V25 Translates learning standards to learning outcomes	1.000	0.000	0.766	0.023	0.5868
	V28 Identifies appropriate teaching methods	1.047	0.065	0.709	0.030	0.5027
	V29 Considers students' prior knowledge in lesson planning	1.182	0.064	0.801	0.024	0.6416
	V30 Considers students' current level of abilities	1.128	0.059	0.764	0.024	0.5837
	V31 Considers students' interest	1.152	0.061	0.780	0.024	0.6084
	V32 Plans lessons according to students' learning needs	1.145	0.059	0.775	0.024	0.6006
	V33 Tailors lessons to available resources	1.234	0.071	0.787	0.024	0.6194
	V34 Develops teaching and learning resources	1.148	0.068	0.746	0.028	0.5565
Student Partner	V17 Gives feedback on students' strengths and weaknesses	1.000	0.000	0.737	0.028	0.5432
	V19 Assists students in using feedback to feed forward	1.100	0.059	0.702	0.029	0.4928
	V27 Involves students in the development of learning outcomes	1.134	0.067	0.737	0.028	0.5432
	V35 Makes students understand the learning outcomes	1.212	0.063	0.773	0.023	0.5975
	V37 Involves students in the development of criteria and standards	1.135	0.069	0.724	0.025	0.5242
	V39 Explains the criteria and standards	1.194	0.071	0.762	0.024	0.5806
	V40 Develops students' capabilities in self and peer assessment	1.234	0.071	0.787	0.024	0.6194
	V41 Engages students in self-assessment	1.196	0.068	0.763	0.024	0.5822
	V42 Engages students in peer-assessment	1.148	0.068	0.746	0.028	0.5565
	V43 Moderates feedback and results of self and peer assessment	1.274	0.074	0.813	0.021	0.6610
Motivator	V46 Uses flexible teaching activities	1.000	0.000	0.737	0.028	0.5432
	V47 Conducts assessment with consideration of student background	1.017	0.047	0.787	0.030	0.6194
	V48 Develops an environment of trust	1.007	0.057	0.708	0.028	0.5013
	V49 Ensures openness in the class	1.004	0.055	0.706	0.029	0.4984
	V50 Builds students' interest to learn	1.128	0.053	0.794	0.022	0.6304
	V51 Demonstrates belief in the ability of every student to improve	1.123	0.052	0.790	0.023	0.6241
	V53 Affirms students' good performance	1.089	0.051	0.766	0.023	0.5868
	V54 Clarifies students misconceptions	1.107	0.058	0.779	0.026	0.6068
	V55 Reinforces positive learning attitude of students	1.041	0.055	0.733	0.027	0.5373
Teacher	V58 Participates in professional development related to assessment	1.000	0.000	0.709	0.030	0.5027

Learner	V59 Engages in self- assessment/ reflection	1.203	0.068	0.782	0.025	0.6115
	V60 Engages in peer-review of teaching performance	1.134	0.067	0.737	0.028	0.5432
	V62 Identifies subject- content knowledge needs	1.234	0.074	0.802	0.029	0.6432
	V63 Searches new and relevant subject-content information	1.148	0.068	0.746	0.028	0.5565
	V64 Undertakes further education/ training	1.194	0.071	0.762	0.024	0.5806
Stakeholder	V66 Collaborates with family to establish support activities	1.000	0.000	0.774	0.026	0.5535
Partner	V67 Informs community of the assessment practices	1.017	0.047	0.787	0.030	0.6194
	V68 Reports to community about students' performance	1.062	0.041	0.823	0.024	0.6773
	V70 Identifies key assessment and teaching issues for review	1.054	0.047	0.817	0.026	0.6675
G Teacher as an AfL Expert	F1 Teacher as an assessor	1.000	0.000	0.860	0.027	0.7396
	F2 Teacher as a pedagogy expert	1.113	0.083	0.895	0.014	0.8010
	F3 Teacher as a student partner	1.129	0.084	0.963	0.009	0.9216
	F4 Teacher as a motivator	1.163	0.087	0.901	0.013	0.8118
	F5 Teacher as a teacher learner	1.026	0.081	0.860	0.020	0.7396
	F6 Teacher as a stakeholder partner	1.195	0.083	0.840	0.019	0.7056

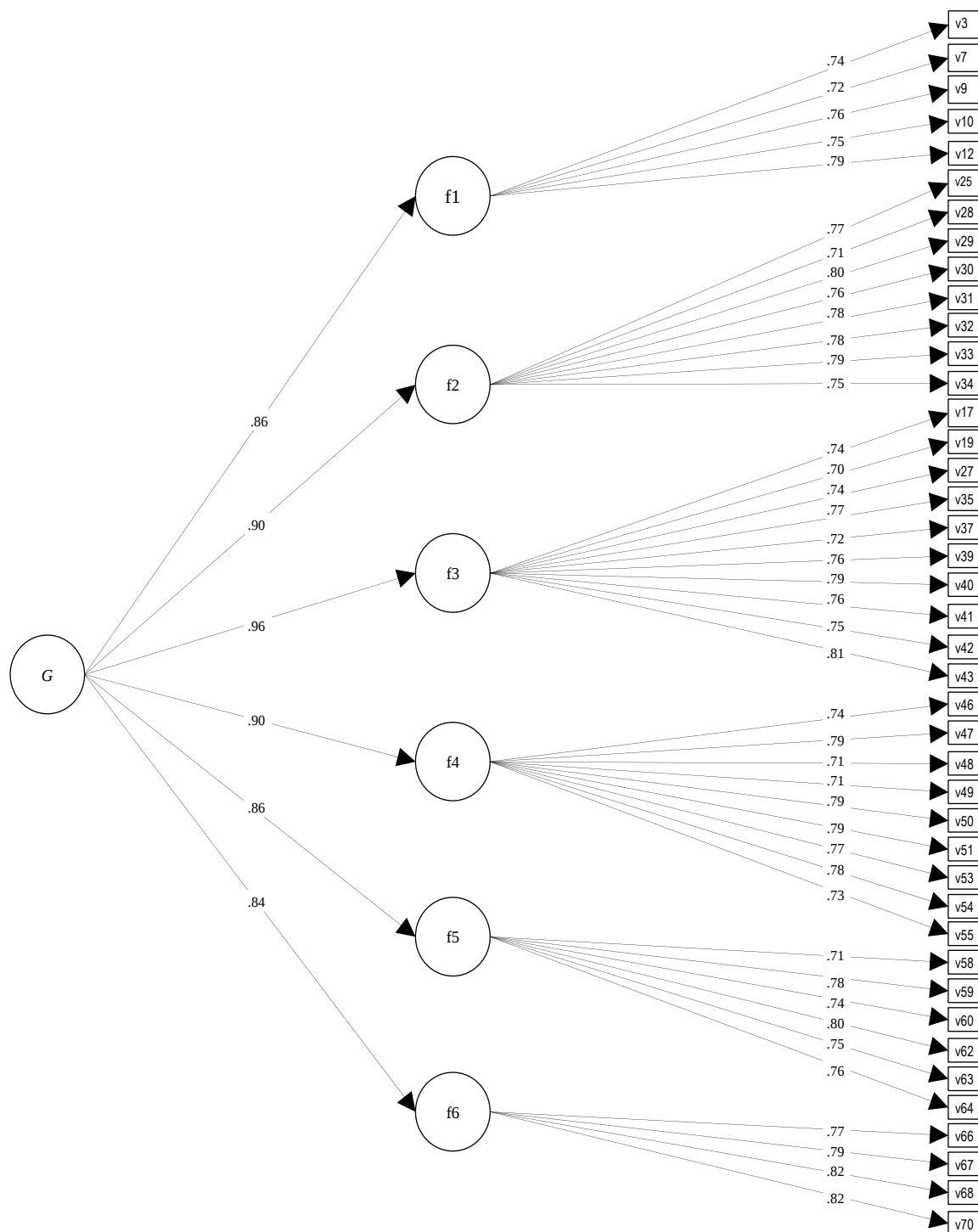


Figure 7.2. The 6-factor model with second-order general factor of teacher A_fL literacy using data from Philippine teachers. The numbers shown are the standardised factor loadings of variables to their corresponding factor.

In terms of divergent validity, Table 7.6 shows that most factor correlations were generally high. These values did not support well the divergent validity of the construct, and may have indicated that these factors were measuring the same dimensions of teacher AfL literacy. Hence, further analysis is needed to ensure that the divergent validity for the 6-factor model would be supported.

Table 7.6

Correlations amongst the Six Factors of the Second-order Confirmatory Model of Teacher AfL Literacy (N=383 Philippine Teachers)

	f1	f2	f3	f4	f5	f6
f1 Teachers as Assessors	1					
f2 Teachers as Pedagogy Experts	0.941	1				
f3 Teachers as Student Partners	0.883	0.932	1			
f4 Teachers as Motivators	0.855	0.867	0.937	1		
f5 Teachers as Teacher Learners	0.758	0.800	0.900	0.940	1	
f6 Teachers as Stakeholder Partners	0.809	0.804	0.865	0.938	0.924	1

7.4.4. Exploratory Structural Equation Modelling

Further validation of the model tested was conducted using exploratory factor analysis (ESEM). Results in Table 7.7 showed that the chi-square test ($\chi^2=811.072$, $df=624, p=0.000$) was significant, yet its ratio to degrees of freedom was just slightly above 1, which indicates a perfectly fitting model. The RMSEA value (.028) gave a 90% confidence that the model indicated an excellent fit to the data. Also, the CFI (0.989) and TLI (0.985) were very close to 1.000, which indicated an excellent model fit. Also, the WRMR was 0.896, which is smaller than the recommended value (1.0). These fit indexes showed that the model perfectly fitted the data well. Furthermore, the unstandardised loadings were all significant, given the absolute values of the ratios between estimates and their corresponding standard errors, which were all greater than 1.96.

Table 7.7

Evaluation of Fit Indexes of the 6-Factor Exploratory Structural Equation Model (N=345 Philippine Teachers)

Indexes	Perfect Fit	Good Fit	Sample Statistics	Decision	Reference
χ^2/df	≤ 2	≤ 3	1.300	Perfect fit	Kline (2005, 2010)
RMSEA	$\leq .05$	$\leq .08$	0.028	Perfect fit	Stegier & Lind (1980); Hooper, Coghlan & Mullen (2008)
CFI	$\geq .95$	$\geq .90$	0.989	Perfect fit	Hu & Bentler (1999) Tabachnick & Fidell (2007)
TLI	$\geq .95$	$\geq .90$	0.985	Perfect fit	Hu & Bentler (1999)
WRMR		≤ 1.0	0.896	Good fit	Hancock & Mueller (2006); Yu & Muthen (2002)

Note: $\chi^2 = 811.072$; $df = 624$.

Table 7.8 shows the structure matrix of the construct. As can be seen, all items were cross-loadings to all factors, ranging from -.094 to 0.308. These cross-loadings explain the true nature of the construct where a teacher's particular AfL skill is dependent to other AfL skills. This is further explained in the next chapter of this thesis. Looking at the results of for convergent validity of the model, the factor loadings (.696 to .780) were slightly reduced compared to the results of CFA and second-order CFA. However, the values were still sufficient to show the indicators factors were substantially correlated. The variance explained ranged from 48.44% to 62.4%, which showed that almost half of the variances were accounted for in each indicator.

Table 7.8

Unstandardised and Standardised Loadings for the 6-Factor Exploratory Structural Equation Model of Teacher AfL Literacy (N=383)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
v3 Designs assessment tasks	0.790	-0.038	0.260	0.036	0.074	-0.056
V7 Uses rubrics to assess students' learning	0.721	0.128	0.144	0.014	-0.063	-0.057
V9 Considers factors that affect students' performance	0.742	0.044	-0.013	0.131	0.028	0.020
V10 Avoids interference in task completion	0.751	-0.170	0.094	-0.094	0.026	0.115
V12 Engages in social moderation	0.700	0.120	0.063	0.049	0.132	-0.115
V25 Translates learning standards to learning outcomes	0.300	0.704	-0.016	-0.085	0.147	0.058
V28 Identifies appropriate teaching methods	0.111	0.780	-0.014	0.155	-0.041	0.021
V29 Considers students' prior knowledge in lesson planning	-0.003	0.773	0.134	-0.020	0.156	0.050
V30 Considers students' current level of abilities	0.007	0.742	0.001	0.005	0.078	-0.022
V31 Considers students' interest	0.069	0.766	0.005	0.280	0.025	0.072
V32 Plans lessons according to students' learning needs	-0.031	0.708	0.135	0.145	0.012	-0.032
V33 Tailors lessons to available resources	0.026	0.696	0.061	0.248	-0.068	0.023
V34 Develops teaching and learning resources	0.193	0.744	0.084	0.177	0.114	0.052
V17 Gives feedback on students' strengths and weaknesses	0.302	0.096	0.721	0.304	0.000	-0.047
V19 Assists students in using feedback to feed forward	0.308	0.058	0.742	0.072	0.016	0.060
V27 Involves students in the development of learning outcomes	0.310	0.273	0.719	0.036	-0.027	0.021
V35 Makes students understand the learning outcomes	0.183	0.281	0.786	0.204	0.021	-0.065
V37 Involves students in the development of criteria & standards	0.001	0.236	0.775	-0.042	-0.012	0.074
V39 Explains the criteria and standards	0.046	0.085	0.703	0.047	-0.044	0.164
V40 Develops students' capabilities in self/ peer assessment	0.139	-0.002	0.699	0.049	-0.002	0.033
V41 Engages students in self-assessment	0.045	-0.030	0.705	0.032	0.108	0.004
V42 Engages students in peer-assessment	0.153	0.036	0.747	0.018	0.115	0.007
V43 Moderates feedback and results of self and peer	0.030	0.055	0.775	-0.032	0.130	0.075
V46 Uses flexible teaching activities	-0.004	0.079	0.219	0.762	0.134	-0.115
V47 Conducts assessment with consideration of students' background	-0.047	-0.163	0.240	0.745	-0.012	-0.017
V48 Develops an environment of trust	0.080	0.026	-0.008	0.698	-0.016	0.100
V49 Ensures openness in the class	0.051	0.021	-0.077	0.722	0.058	0.095
V50 Builds students' interest to learn	0.022	0.108	0.088	0.757	0.094	-0.031
V51 Demonstrates belief in the ability of every student to improve	-0.026	0.087	0.308	0.710	0.036	0.126
V53 Affirms students' good performance	-0.018	0.083	0.258	0.709	0.066	0.124
V54 Clarifies students misconceptions	0.126	0.015	0.300	0.745	0.076	0.178
V55 Reinforces positive learning attitude of students	0.149	-0.009	0.249	0.759	-0.039	0.183
V58 Participates in professional development (assessment)	-0.030	-0.060	0.074	0.120	0.763	0.203
V59 Engages in self- assessment/ reflection	0.050	0.052	0.157	-0.044	0.695	0.074
V60 Engages in peer-review of teaching performance	-0.017	-0.028	0.004	0.207	0.716	-0.053
V62 Identifies subject- content knowledge needs	0.088	0.099	0.034	-0.020	0.729	0.214
V63 Searches new and relevant subject-content information	0.094	0.038	-0.097	0.121	0.698	0.016
V64 Undertakes further education/ training	0.008	-0.094	0.052	0.298	0.699	0.028
V66 Collaborates with family to establish support activities	0.004	0.024	0.059	0.120	0.142	0.716
V67 Informs community of the assessment practices	0.052	0.001	-0.035	0.042	0.259	0.756
V68 Reports to community about students' performance	0.052	0.011	0.071	0.021	0.296	0.745
V70 Identifies key assessment and teaching issues for review	-0.066	0.196	0.079	-0.012	0.238	0.714

Presented in Table 7.9 are the correlations amongst factors. As shown in the table, the factor correlations ranged from only 0.154 to 0.399, which were substantially reduced compared to the results in CFA and second-order CFA. Thus, this result provided evidence of the claim that each factor is measuring a separate dimension of teacher AfL literacy. These low correlations amongst factors supported the discriminant validity of the tool.

Table 7.9

Correlations amongst the Six Factors of the Exploratory Structural Equation Model of Teacher AfL Literacy (N=383 Philippine Teachers)

	f1	f2	f3	f4	f5	f6
f1 Teachers as Assessors	1.000					
f2 Teachers as Pedagogy Experts	0.399	1.000				
f3 Teachers as Student Partners	0.330	0.241	1.000			
f4 Teachers as Motivators	0.369	0.223	0.241	1.000		
f5 Teachers as Teacher Learners	0.294	0.245	0.232	0.213	1.000	
f6 Teachers as Stakeholder Partners	0.201	0.154	0.204	0.193	0.179	1.000

The results of CFA, second-order CFA, and ESEM were all consistent to the results obtained in the main study. Hence, the dimensionality of the teacher AfL literacy is further supported, and therefore the same factor model of teacher AfL literacy exists in similar context.

7.5. Latent Profile Analysis

Latent profile analysis was also conducted to determine if Philippine teachers exhibited the same latent profile based on their AfL practices. Table 7.10 below shows the results of LPA modelling with up to 6-class model was tested.

Table 7.10

Model Fit Indexes of Tested Latent Profile Models of Teacher AfL Literacy (N=383 Philippine Teachers)

Fit Index	1 class	2 classes	3 classes	4 classes	5 classes	6 classes
AIC	4654.989	3316.385	2627.463	2127.626	1779.541	2426.052
BIC	4702.365	3391.397	2730.112	2257.911	1937.463	2611.609
Adjusted BIC	4664.291	3331.113	2647.618	2153.207	1937.463	2462.486
Entropy		0.912	0.923	0.931	0.937	0.904
Vong-LMR		-2315.495 (<i>p</i> = 0.1531)	-1639.192 (<i>p</i> =0.7601)	-1287.732 (<i>p</i> = 0.0151)	-1056.215 (<i>p</i> =0.2298)	-812.526 (<i>p</i> =0.093)
Adjusted LMR		1320.880 (<i>p</i> =0.1581)	686.435 (<i>p</i> = 0.7619)	501.786 (<i>p</i> = 0.0153)	407.617 (<i>p</i> = 0.2332)	289.260 (<i>p</i> =.095)
BLRT		-2315.495 (<i>p</i> <0.001)	-1639.192 (<i>p</i> <0.001)	-1287.732 (<i>p</i> <0.001)	-1056.215 (<i>p</i> <0.001)	-812.526 (<i>p</i> =0.058)

Note: AIC = Akaike information criteria; BIC = Bayesian information criteria; LMR = Lo-Mendell-Rubin likelihood ratio test; BLRT = Bootstrap parametric likelihood ratio test.

Almost the same trends as in the main study were observed in the results of the validation using data from the Philippines. The information criteria (AIC, BIC, and Adjusted BIC) showed a continual improvement as the classes were progressively added to the models, except in the case of 6-class model where all values significantly increased. Given these results, the 5-class model performed well. Although the log-likelihood ratio tests, the results of V-LMR and adjusted LMR supported the 4-class model given the

significant p-values, the BLRT test results indicated otherwise. That is, the 4-class model was not sufficient to account the existing classes found in the data set.

Considering the results of the information criteria and the log-likelihood ratio tests, together with the entropy value (highest in the 5-class model), a decision was made that the 5-class model was the best-fitting model for the data. This decision to accept the 5-class model was also supported by the poor fit indexes of the 6-class model.

Figure 7.3 shows the graphical representation of the LPA model with five classes. As can be seen, this LPA modelling showed the differentiated levels of the Philippine teachers in their use of the AfL concept. Like the results of the main study, the pattern of use of different aspects of the AfL concept was consistent across the six identified dimensions.

In terms of the number of teachers belonging in each class, unlike in the main study, where most participants were in Class 4, majority of the Philippine teachers fell into Class 2. As comparisons between the two countries were not allowed, no further implications were drawn. However, this result clearly supports the application of the typology of teachers' AfL literacy derived from the main study in similar contexts.

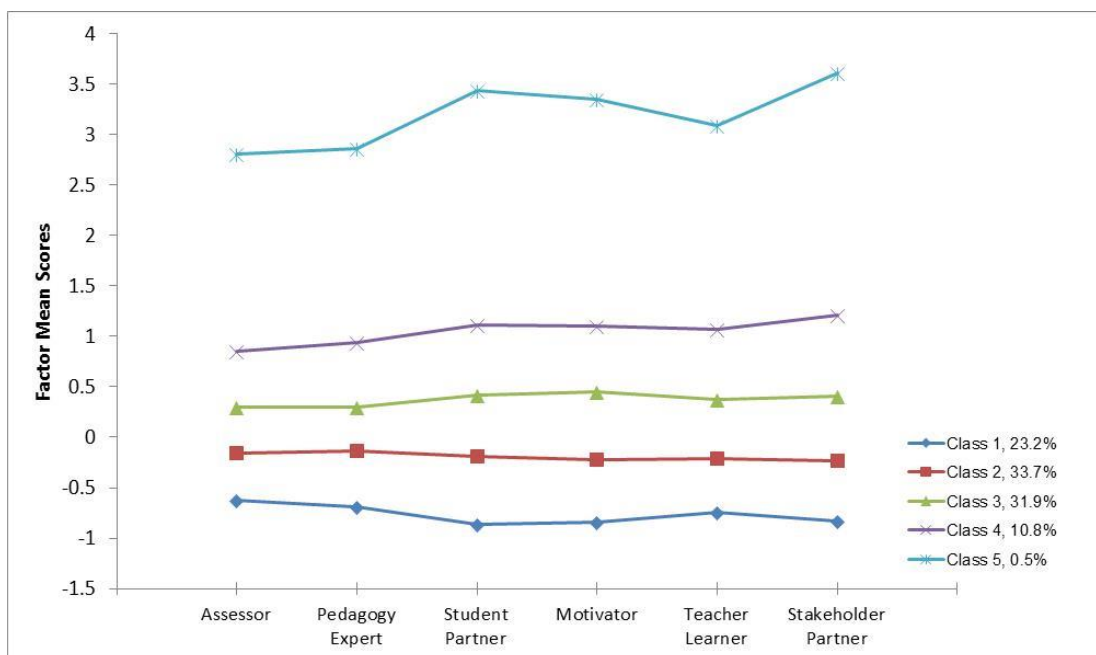


Figure 7.3. The latent profiles of teachers based on their AfL literacy.

7.6. Summary

This chapter has presented the results of a validation study using a data set from Philippine teachers' self-assessment following the same factor analyses and LPA modelling used in the main study. All analyses generated consistent results with the main study. The factor structures extracted both in CFA and second-order CFA, and the results of ESEM were consistent, except for slight differences in factor loadings and factor correlations. Similarly, the results of LPA modelling extracted the same number of classes, which supported the typology developed in the main study. There are differences in terms of the actual level of teachers' performance but a comparison between Brunei and Philippine teachers is beyond the scope of this study. Overall, this section of the study provides even further empirical support to answers the research questions.

CHAPTER 8 - DISCUSSION

8.1. Introduction

In this chapter, I discuss the major findings of my study related to the dimensionality of teacher assessment for learning (AfL) literacy and the item level characteristics of the teacher AfL literacy tool, with an emphasis on the contributions to theory and practice. The discussion builds on the results of the main study and cross-country validation. There are five major sections of this chapter. The first section describes the implications of the dimensions extracted from my study in describing and understanding the construct of teacher AfL literacy. To provide greater clarity, the dimensions were used to revise the conceptual framework of this study and are presented as a teacher AfL literacy framework. The second section highlights the ways in which the teacher AfL literacy tool can be used to effectively support teachers in AfL development and learning. The third section emphasises the implications of the measurement properties of the tool. The fourth section discusses the applications for establishing the teacher AfL profile in professional development. In the final section, I offer a more practical way on how to use the tool utilising the principles of classroom assessment.

8.2. The Teacher AfL Literacy Tool

8.2.1. Overview

The main purpose of this study is to develop a theoretically and empirically-driven tool for describing and understanding teacher AfL literacy; to explore the latent profiles of teachers based on their profiles of AfL literacy; and to determine the extent of generalizability of the

tool to other similar contexts. In this section I discuss the contribution of my study related to the conceptualisation of teacher AfL literacy, expansion of the existing teacher assessment literacy models, alignment of the tool to the AfL principles, and the existence of the general factor of teacher AfL skills.

8.2.2. A New Way of Conceptualising Teacher AfL Literacy

Although there have been numerous studies that have explored the dimensions of teacher assessment literacy, my study is the first to use an extensive methodology and the principles of AfL as the underlying philosophy to establish the dimensions of this construct. Most of the existing studies are at the item level only looking at individual teacher assessment skills (Pang & Leung, 2011; Popham, 2009; Thomson, 2012) whilst other studies use only a theoretical approach to establish the dimensions of teacher AfL literacy (Mertler & Campbell, 2005; Newfields, 2006; Stiggins, 1999b). Some of the dimensions used to describe teacher assessment literacy were derived from a theoretical approach using researcher professional judgment to cluster teacher assessment skills based on their similarities, whereas other dimensions were adopted from the existing professional assessment standards for teachers. Whilst these existing theoretical dimensions are useful to describe teacher assessment literacy, there are issues associated with their development, empirical support, and conceptualisations of teacher assessment literacy (see Chapter 3.4.), which my study has fully addressed.

In my study, the theoretical approach used by previous studies, combined with the factor analysis used by DeLuca and Kliger (2010), Fulcher (2012) and Newfields (2006), has been further expanded into a more robust methodology to gather empirical evidence to

support the theoretical conceptualisation of the construct drawn from the philosophical framework of AfL and assessment innovation and change. Hence, at the factor level, the use of exploratory, confirmatory factor and second-order factor analyses, as well as exploratory structural equation modelling, provided strong evidence that the 6-factor model extracted from my study is the most parsimonious, as it provides a more robust conceptualisation of teacher AfL literacy.

Specifically, empirical evidence derived from factor analyses supports the existence of the individual items as indicators of the construct explored by previous studies, and these items cluster together to form six broader overarching dimensions of teacher AfL literacy. The six dimensions of teacher AfL literacy present a new conceptualisation of teacher AfL literacy highlighting the major roles of teachers in using assessment to effectively support student learning and to ensure the assessment literacy development of stakeholders.

The previous conceptualisations of teacher assessment literacy as a construct with overarching dimensions is supported by the findings of this study, but the dimensions extracted are somewhat different to those previously used and described by other studies (e.g., Fulcher, 2012; Mertler & Campbell, 2005; Newfields, 2006; Stiggins, 1999b). This is because the definitions of teacher assessment literacy they used were not aligned so well to the AfL principles. As shown in Chapter 3, much of the emphasis of the existing frameworks and models is on the teacher knowledge and skills related to measurement principles rather than on the undervalued role of teachers in classroom assessment aimed at supporting students to take more responsibility for their learning. In my literature review, I argued that the strict adherence of teachers to the measurement principles does not account for the context-dependent nature of assessment (Brookhart, 2003; Moss, 2003; Smith,

2003), and there are other important teacher assessment skills that are more valuable in enhancing student learning (Black & Wiliam, 1998; Hattie, 2008).

The dimensions extracted from my study reinforce this broader conceptualisation of teacher AfL literacy. In particular, the ability of teachers related to measurement principles cluster in one factor only (*Factor 1: Teachers as Assessors*), and there are other five dimensions that contribute to the measurement of teacher AfL literacy. These include the role of teachers in using assessment information to plan learning and teaching activities (*Teachers as Pedagogy Experts*); using assessment to ensure high level of student motivation (*Teachers as Motivators*); engaging students in learning and teaching (*Teachers as Student Partners*); reflecting on their assessment experience to identify their professional development needs (*Teachers as Teacher Learners*); and ensuring the assessment literacy of parents/carers and the community in general (*Teachers as Stakeholder Partners*). These six dimensions provide the explicit link between teacher ability in assessing student learning and other abilities that require the use of assessment information to support student learning including teachers' gains in their assessment experience and stakeholder's assessment literacy.

These six underlying factors in teacher AfL literacy were used to revise the theoretical framework of my study as discussed in Chapter 2.6 and as presented as the teacher AfL literacy framework in Figure 8.1. The four dimensions of teacher AfL literacy embedded within the square are those competencies of teachers that are directly related to the actual classroom practices and student engagement. The two others on the periphery are those competencies of teachers that use their assessment experience to identify their professional development needs and to enhance the assessment literacy of other

stakeholders. The required assessment skills are listed below each dimension. These assessment skills are labelled as indicators in the teacher AfL literacy tool.

The rationale behind this presentation of the framework is to provide clarity to teachers in terms of the way they think of their roles, not only for ensuring student learning, but also for ensuring the assessment literacy of other stakeholders as well as their own AfL literacy. Thus, this structure provides a better framework for understanding and describing teachers' roles in AfL implementation, and AfL literacy is seen from this wider perspective.

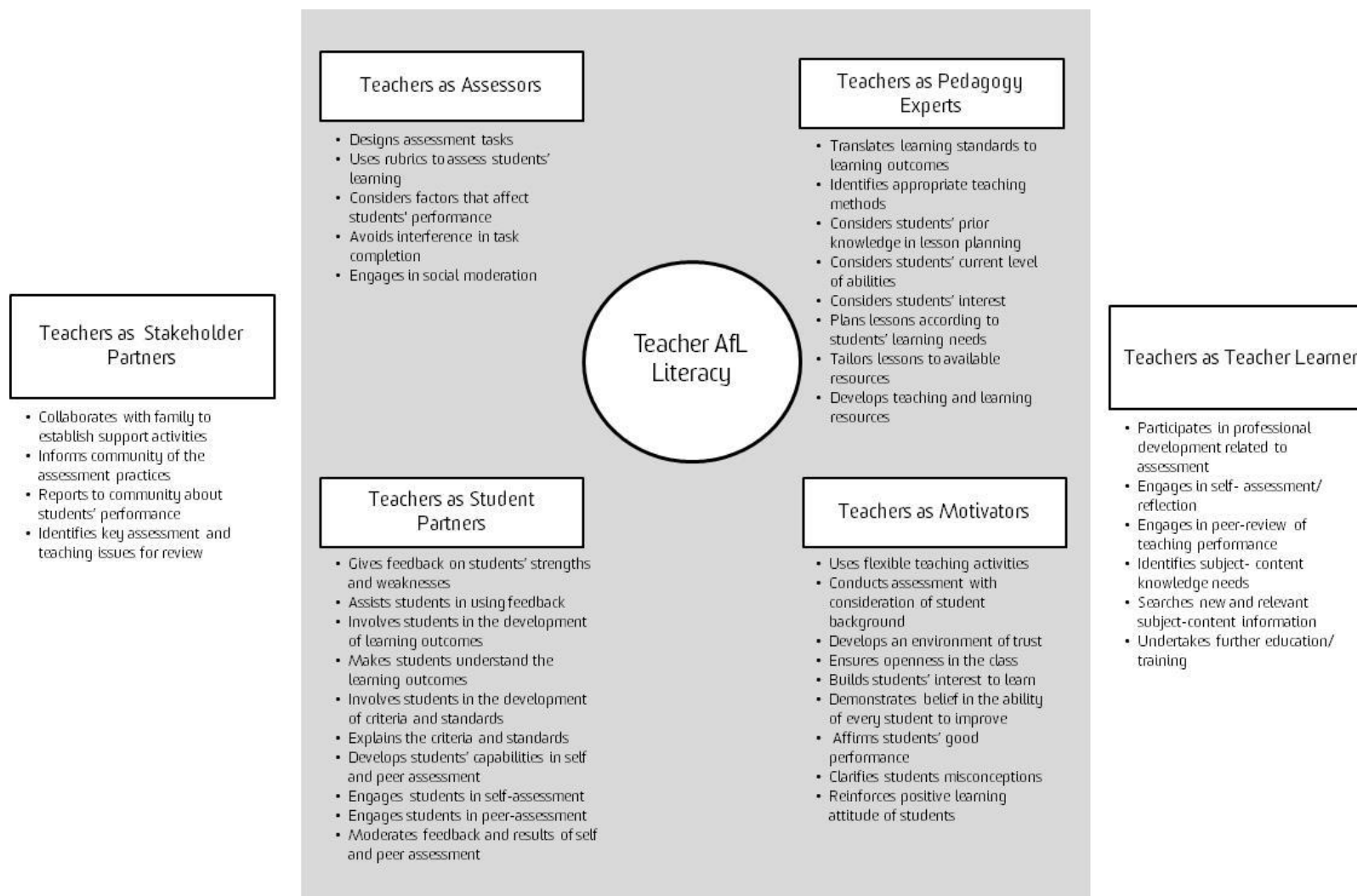


Figure 8.1. Revised theoretical framework of teacher AfL literacy.

8.2.3. Expanding the Existing Theoretical Models of Teacher Assessment Literacy

Apart from having strong empirical evidence to support the six dimensions of teacher AfL literacy that emerged from this study, its multidimensional and multifunctional features weave together the various theoretical models of teacher assessment literacy, thus creating a more comprehensive model of teacher AfL literacy. As discussed in Chapter 3, existing theoretical models of teacher assessment literacy are used, either for instructional purposes to guide teachers' classroom practices (e.g., Brookhart, 2006; Cowie & Bell, 1999; Davison, 2008; Harlen's time dimension, 2007; Wiliam & Thompson, 2008), or for the evaluation of teacher assessment practices (e.g., Harlen's component dimensions, 2007; Hill & McNamara, 2012). Unlike models that are function-specific, the teacher AfL literacy tool has combined these functions and can be used both as instructional guides and for evaluation purposes.

The teacher AfL literacy tool contributes to the expansion of the existing theoretical models, not only in terms of establishing what dimensions constitute teacher AfL literacy but also with the inclusion of performance descriptions of the five levels of performance for each indicator. The descriptions of the dimensions define the key roles of teachers involved in using assessment to effectively support student learning, whilst the criteria and standards provide a tool that can be used for self-reflection and evaluation of assessment practices and for professional development. The levels and descriptions of standards in the tool address one of the gaps in existing teacher assessment tools, particularly in relation to the absence of descriptions of what teachers can actually do, as identified by Bailey and Brown (1995) and Inbar-Lourie (2008). Also, the adherence of the tool to the principles of AfL

addresses the issue raised by Brookhart (2011) about the absence of a tool that embraces the philosophy of AfL. Hence, the criteria and standards in the tool help in clarifying the expected teacher assessment performance across stages of development. As such, the tool can be used for describing teacher AfL literacy development over time with greater specificity compared to the existing frameworks and models.

To make a more detailed comparison, the multidimensional features of the teacher AfL literacy tools specifies the assessment components or the overall AfL processes. Factors 1 to 4 (*roles of teachers as assessors, pedagogy experts, student partners, motivators respectively*) capture the dimensions used in the existing models or frameworks of teacher assessment literacy (e.g., Black & Wiliam, 2006; Brookhart, et al., 2006; Davison, 2008; Wiliam & Thompson, 2008). Also, these four factors are directly linked to learning and teaching activities that are closely related to the widely used categorisation of teachers' assessment skills, which capture the key features of the existing models as summarised in Table 8.1.

Table 8.1

A Summary of Existing Models of Teacher Assessment Practices

Models	Key Features
Cowie and Bell (1999)	Inclusion of planned and interactive formative assessment
Harlen (2007) Model A	Assessment is seen as a system with seven components operating: 1. Purpose 2. Use 3. Type of task 4. Agent of judgment 5. Basis of judgment 6. Form of report or feedback 7. Moderation
Harlen (2007) Model B	Focuses on the time dimension of assessment episodes from setting lesson goals to identifying learning activities and reporting levels of broad goals. The emphasis is on how the evidence of student learning collected in a formative way can be used for summative purposes.
Brookhart (1997)	Highlights the connection between teacher assessment practices and student effort and subsequently, to student learning.
Brookhart et al. (2006)	Retains the causal relationship amongst teacher assessment practices, student effort, and student learning, but includes the function of student motivation (the desire to do something) as directly affecting the level of students' effort exerted to complete the task. Another feature of this model is the expansion of teacher assessment practices to the classroom assessment environment, considering the unique context of every classroom thereby providing different context-dependent assessment experiences.
William & Thompson (2008)	Illustrates how formative assessment strategies can achieve the key three purposes of the AfL definition (<i>where the learning is going, where the learner is right now, how to get there</i>). This model shows how instruction is neatly integrated into assessment practices.
Davison (2008)	Demonstrates a step-by-step process from planning assessment, collecting information, making professional judgement to providing appropriate feedback or advice to help students achieve the learning outcomes. For each stage, there are suggested assessment strategies/activities
Black & Wiliam (2006)	Uses the framework of Activity Theory of Engestrom (2001) to conceptualise formative assessment. It highlights the role of socio-constructivist approach in achieving learning outcomes.
Hill & McNamara (2012)	Includes four dimensions 1. Teacher assessment practices (planning assessment, framing assessment, conducting assessment, using assessment data) 2. The underlying assessment construct 3. The epistemological basis for teachers' belief in assessment 4. Learners theories and beliefs in assessment

The dimensions of the teacher AfL literacy tool have expanded the conceptualisations of the construct of teacher assessment literacy of existing frameworks. For example, the limitations of the works of Cowie and Bell (1999) and Harlen's models (2007) in terms of including only some of the dimensions and processes of teacher assessment practices, leaving out some important features of using assessment to ensure student learning, are addressed in the teacher AfL literacy tool. Cowie and Bell's focus on planned and interactive formative assessments, too simplistic to capture teacher assessment literacy, is expanded with teacher assessment skills related to measuring student learning are conceptualised as a continuum of assessment skills from informal formative assessment to the most formal summative assessment (Davison, 2007), and all these assessments and assessment data obtained are used to support student learning (Brookhart, 2003; Kane, 2006; Nichols, 2009; Sadler, 2012).

Apart from the four dimensions that capture all the existing frameworks and models, two factors emerged which were only implied by some authors in their models. These two factors highlight the least explored assessment skills of teachers. Factor 5 (*Teachers as Teacher Learners*) accounts for the ability of the teacher to reflect on assessment data to identify their professional needs, which relates to what Timperley (2011) describes as adaptive expertise where teachers are actively engaged in a self-reflective practice to identify opportunities to innovate and to try new assessment approaches as discussed in Chapter 2. The framework for teacher-based assessment developed by Davison (2008) is comprehensive and is centred on the classroom assessment practices of teachers. However, the teachers' capacity to benefit from their assessment experience and their ability to reflect on assessment data to identify their professional needs

is not included. This teacher competency emerged as one factor in my study. The existence of this factor is supported by a range of literature demonstrating the ability of teachers to reflect on their assessment and teaching experiences and to identify their professional needs for both content and assessment skills. The inclusion of this factor in teacher AfL literacy is critical because this is the most effective method of increasing teacher assessment literacy (Timperley, 2008, 2011; Timperley et al., 2008), and that this increase in teachers' assessment literacy brought about by engagement in professional development, either formal or informal, contributes greatly to the improvement of student learning (Hattie, 2008). With this factor in the tool, teacher can develop their skills to respond appropriately to the gap between their current levels of ability both in assessment (Hattie, 2008) and curriculum-content and identify what capabilities are needed to respond to student learning needs appropriately (Timperley et al., 2008). As was shown in Chapter 2, the continuous engagement of teachers in professional learning is one of the factors which has the highest effect on improving student learning (Black & Wiliam, 1999; Hattie, 2008; Stiggins, 1999b).

The other factor (*Factor 6: Teachers Stakeholder Partners*) is the ability of the teacher to work closely with stakeholders to improve their assessment literacy to ensure that their expectations and beliefs are consistent with the teachers' AfL practices. Although the models presented by Brookhart (1997 & 2006), Harlen's component model (2007), Davison (2008) and Wiliam & Thompson (2008) clearly illustrate the role of teacher assessment literacy in helping students achieve specific learning outcomes, the role of teachers in addressing and enhancing the assessment literacy of other stakeholders is not explicitly described in these models. Although this competency of teachers is implied in

Davison's model, which mentions providing feedback to parents, and Harlen's model, which includes external agents as one of the agents for judging students' performance, it was not made explicit in either model that it is part of teachers' responsibility to address the assessment literacy needs of various stakeholders. This role of teachers lends strong support to the argument of Popham (2008) that each group of stakeholders has different AfL literacy needs because they have different assessment knowledge, assumptions and responsibilities, and excluding this teacher role in the assessment framework leaves the assessment system restricted to the classroom level only.

The inclusion of this factor in the tool helps develop teachers' ability to address the assessment literacy of all stakeholders (Davison, 2013; Taylor, 2009). Thus, the tool provides a mechanism for the successful implementation of AfL reforms within and across schools. Davison (2013) argues this as a critical component of AfL reform where it is the responsibility of teachers to develop a shared understanding and a common belief system amongst all stakeholders from the school level, including students, parents and schools heads, and across all levels of the education bureaucracy. The failure of teachers to develop a shared understanding and change the assessment belief system held by stakeholders will put pressure on teachers' assessment practices, which will inevitably result in the unsuccessful implementation of AfL (Davison, 2013).

As shown, the teacher AfL literacy framework offers a new way of describing and understanding teacher AfL literacy, which has a better potential to develop teacher AfL and teaching practices compared to existing models and frameworks. The six dimensions of teacher AfL literacy explicitly define the use of assessment and/or assessment information

in various learning and teaching activities, including for teacher professional development and for enhancing assessment literacy of stakeholders.

8.2.4. The Teacher AfL Literacy Tool as the Only Tool Aligned with AfL Principles

The weaknesses related to the conceptualisations of teacher assessment literacy and the need for a more useful and AfL-driven tool to support teachers resulted in the development of the teacher AfL literacy tool. The depth of empirical evidence supporting the factor level and psychometric requirements and the use of AfL as the overarching philosophy both in the operationalisation of the construct and in the development of the tool address the methodological limitations of the existing teacher assessment literacy tool respectively. The tool I developed provides a high level confidence for the users (i.e., teachers and school leaders) that it can effectively measure the assessment literacy of teachers over time.

The dimensions of the teacher AfL literacy tool established in my study reduce the large number of assessment skills described in the literature into six distinct but related factors with teacher assessment skills that are consistent with AfL principles. In other words, the tool I developed, along with its dimensions, includes only those teacher AfL practices that can improve student learning that have been described and empirically supported by previous studies. The excellent fit of the 6-factor model provides support for the claim that the dimensions in the teacher AfL literacy tool better accounts for the latent constructs of teacher AfL literacy compared to dimensions presented by other authors, which indicated a poor fitting model (Hailaya, Alagumalai, & Ben, 2014).

As described earlier, the six dimensions of teacher AfL literacy are different from other dimensions described and used in existing tools. Table 8.2 shows a comparison of the

dimensions used by existing tools by matching them with the recent dimensions extracted from this study. It can be seen from the table that whilst all the various dimensions used by previous tools fit into the six dimensions of teacher AfL literacy tool, there are several noteworthy inconsistencies.

Table 8.2

Comparison of the Dimensions of Teacher Assessment Literacy

Teacher A/L Literacy Tool	Mertler & Campbell (2005)*	Fulcher (2012)	Stiggins (1999b)	Newfields (2006)
Teachers as Assessors	• Choosing assessment methods • Developing assessment methods • Administering, scoring and interpreting • Developing valid pupil grading procedures • Recognizing unethical, illegal, and otherwise inappropriate assessment methods	• Test design and development • Large-scale standardised testing • Classroom testing and washback • Validity and reliability	• Connecting assessment to clear purposes • Applying proper assessment methods • Avoiding bias in assessment • <i>Developing quality assessment exercises and scoring criteria and sampling appropriately</i>	• Terminology • Procedures • Test interpretation • Assessment Ethics
Teachers as Pedagogy Experts	• <i>Using assessment results when making decisions</i>		• Using assessment as instructional intervention	
Teachers as Motivators				
Teachers as Student Partners	• <i>Communicating assessment results</i> • <i>Using assessment results when making decisions</i>		• Clarifying achievement expectations • <i>Developing quality assessment exercises and scoring criteria and sampling appropriately</i> • <i>Communicating effectively about student achievement</i>	
Teachers as Teacher Learners				
Teachers as Stakeholder Partners	• <i>Communicating assessment results</i>		• <i>Communicating effectively about student achievement</i>	

*Based on Standards for Teacher Competence in the Educational Assessment of Students (1990)

One of the most obvious inconsistencies is the absence of two dimensions related to the assessment skills of teachers in using assessment to enhance student motivation (*Teachers as Motivators*) and in teachers' ability to reflect on their assessment experience to identify and address their professional development needs (*Teachers as Teacher Learners*) in other tools (e.g., Fulcher, 2012; Mertler & Campbell, 2005; Newfields, 2006; Stiggins, 1999b). These two dimensions are critical in improving student learning, reinforcing the argument of Dweck (2007) about the skills of teachers in using assessment to enhance and sustain motivation of students, and support the work of Timperley, Wilson, Barrar and Fung (2008) and Timperley (2008, 2011) on the importance of using student assessment data and teacher reflection to develop a needs-based professional development program.

Another inconsistency is the overlap of some teacher assessment skills in more than one factor. In the existing tools, some teacher assessment skills are very broad, and need to be further broken down to provide better specificity of assessment skills. The teacher AfL literacy tool provides greater specification of teacher assessment skills. For example, in the works of Stiggins (1999b) and Mertler & Campbell (2005), the communication of student achievement is stated in a generic way. However, this skill can only be effectively undertaken if teachers prepare reports according to the assessment literacy needs of the stakeholders (Davison, 2007; Guskey & Bailey, 2001). In the tool, this skill is highly specific for students, who need more detailed descriptions of their strengths and weaknesses and advice on how to develop their learning goals to address their learning needs (O'Connor & Wormeli, 2011), and for parents, who need a more generic report

showing how to establish home-school activities to further support students' engagement in their learning is more effective.

In general, the tool I developed is based on a broader conceptualisation of teacher assessment literacy. The tools developed by Newfield (2006) and Fulcher (2012) that used factor analysis to establish the dimensions of teacher assessment literacy, both extracted four factors. These tools are too simplistic as the factors they extracted and used in their tools are all combined in Factor 1 in the teacher AfL literacy tool. The teacher AfL literacy tool includes a broader and detailed picture of teacher AfL literacy that it is not only about previously proposed skills like developing assessment tasks, assessing student learning and adhering to the principles of reliability and ethical practices in assessment, but more about adhering to the principles of AfL. In this tool, teachers' ability to use assessment and assessment information to engage students in learning and teaching activities and teacher professional development and stakeholders' assessment literacy needs are measured as part of teacher AfL literacy. In addition, apart from the dimensions described in the existing tools, there are other important teacher assessment skills that are measured by the teacher AfL literacy tool.

8.2.5. The General Factor of Teacher AfL Literacy

As seen in Chapter 6.5, the six factors in the teacher AfL literacy tool loaded significantly to a single, second-order factor, which was labelled as *Teachers as AfL literate professionals*. Conceptually, although the divergent validity of the second order factor is a little problematic, the second-order factor makes sense as the AfL practices represented in each of the six dimensions are conceptually interrelated with each other. The second-order

structure sheds light to the true nature of the dimensions of teacher AfL literacy. This is consistent with the conceptualisation of Glasson (2009) of teacher assessment skills as interconnected abilities. To illustrate, teachers' competency in developing lessons appropriate for learners (Factor 2) is best built from their knowledge of their students' ability levels, which is typically informed by student assessment information (Factor 1). Similarly, teachers' ability to engage students in learning and teaching activities (Factor 3) is closely linked to their ability to develop and use engaging lessons (Factor 2), as well as good knowledge of assessment strategies (Factor 1). Moreover, incorporating self and peer assessment in student learning, an aspect of Factor 3, is known to enhance students' motivation, a key theme in Factor 4 (Hanrahan & Isaacs, 2001; Patri, 2002). Furthermore, according to Hattie (2003), competent teachers gather relevant information (Factor 1) as they monitor students' learning and provide better feedback (Factor 3). Thus, the single, second-order factor demonstrates an underlying assumption of one of the AfL principles that at the heart of learning and teaching processes in the classroom lies teachers' expertise in using assessment to improve student learning (ARG, 1999). This interrelatedness of teacher assessment skills expanded the most common linear conceptualisation of the relationship of teacher assessment activities, where one process leads to the next process as demonstrated in the existing models.

The results of the second-order confirmatory factor analysis further revealed that teachers' competence in engaging students in all aspects of assessment, learning, and teaching (Factor 3) has the highest factor loading onto the second-order factor. This finding lends strong support from the results of previous studies emphasising the necessity to actively engage students in their learning (Black & Wiliam, 1998; Hounsell, McCune,

Hounsell, & Litjens, 2008). Moreover, this result aligns well with Hattie's (2008) argument that the most fundamental role of teachers in the classroom is activating student learning, which is strongly supported by a socio-cultural theorisation of AfL. This emphasises the view that teachers need to consider students as co-creators of knowledge whereby students are actively constructing their knowledge and skills through finding meanings in what they do (DeCorte, 1996; Nicol, 1997) as influenced by their inner self and the external environment (Hattie, 2008).

Other factors that have high factor loadings onto the second-order factor are those competencies of teachers that ensure assessment is used to effectively plan learning and teaching activities (Factor 2) and to build students' interest to learn (Factor 4). The former factor highlights the pedagogical focus of AfL practices as consistent to the principles of AfL, where assessment should be embedded in learning and teaching activities. The arguments of Biggs and Tang (2007) and Glasson (2009) about the links and interplays of assessment, learning and teaching are supported empirically by this high factor loading.

The latter factor sheds light on the long argued causal effect of assessment on motivation. This finding provides empirical support for the views of Brookhart et al. (2006), Dweck (2007), and Pekrun et al. (2002) that the careful selection of content that is relevant to students' interests, and the use of teaching methods that addresses their learning needs can motivate students to engage in the learning process. Hence, this finding emphasises the most critical roles of teachers in developing student self-regulation (Sadler, 1989a; Pintrich & Zusho, 2002). As such, in ensuring learning, teachers should put emphasis on looking at the motivation level of students.

8.3. The Criteria in Teacher AfL Literacy Tool and Classroom Practices

8.3.1. Overview

In this section, I discuss another contribution of my study related to the application of the criteria and standards in actual teachers' classroom practices. The criteria and standards provide descriptions of the overall process of assessment and what teachers need to do. Unlike the assessment literacy inventory tool developed by Mertler and Campbell (1995), which was shown by Hailaya, Alagumalai, and Ben (2014) to have items with low factor loadings to their corresponding factors, and some other tools with no data reported regarding factor loading, the items (labelled as criteria) of the teacher AfL literacy tool I developed have high factor loadings. The high factor loadings of individual teacher assessment skills give confidence to teachers (and other stakeholders) that these assessment skills being measured by the tool are those that contribute greatly to improving student learning. Hence, students would be most likely to benefit from the high positive effect of using assessment to improve their learning.

8.3.2. The Criteria and Standards as a Way to Operationalise AfL Principles

The first way to examine the applicability of the teacher AfL literacy tool is to see if it is aligned to the principles of AfL as set by ARG (2002). In Table 8.3, the 42 teacher assessment skills are mapped to the 10 principles of AfL. It is clearly shown that the criteria included in the teacher AfL literacy tool strongly align with these principles, thus the tool provides a new way of operationalising the principles of AfL in actual classroom practices.

The individual assessment skills comprising each dimension are useful guides for teachers to ensure that their practices adhere to the principles of AfL. Although assessment practices comprising each dimension in the teacher AfL literacy tool are clear and comprehensive, teachers need to be reassured that these assessment practices are not intended to be prescriptive. Teachers are encouraged to continuously innovate and improve their assessment practices and constantly evaluate the effectiveness of their practices to organise the emerging sets of teacher assessment skills. In a way, although the teacher AfL literacy tool is a useful guide for teachers, it must be recognised by teachers that significant criteria and standards in the teacher AfL literacy tool are context-dependent. For example, some of the assessment skills included might not be appropriate for a specific cohort or context in which learning occurs because they are not age or culturally appropriate. Furthermore, there might be other assessment skills that are context-driven (e.g., use of national examination results, use of technology in assessment, inclusiveness of students with special learning needs), which need to be included for a particular education system.

Table 8.3

The Alignment of Criteria of Teacher AfL Literacy Tool to AfL Principles

	Is part of effective planning	Focuses on how student learn	Is central to classroom practice	Is a key professional skill	Is sensitive and constructive	Fosters motivation	Promotes understanding of goals and criteria	Help learners how to improve	Develops the capacity for self-assessment	Recognises all educational achievement
Teachers as Assessors	Design assessment tasks	Consider factors that affect students' performance		Engage in social moderation	Avoid interference in task completion		Use rubrics to assess students' learning			
Teachers as Pedagogy Experts	Translate learning standards into learning outcomes	Consider students' prior knowledge in lesson planning	Identify appropriate teaching methods Develop teaching and learning resources	Tailor lessons to available resources	Consider students' current level of abilities	Consider students' interest		Plan lessons according to students' learning needs		
Teachers as Motivators		Reinforce positive learning attitude of students			Conduct assessment with consideration of student background	Use flexible teaching activities Develop an environment of trust Affirm students' good performance		Clarify students misconceptions Build students' interest to learn Ensure openness in the class		Demonstrate belief in the ability of every student to improve
Teachers as Student Partners				Moderate feedback and results of self and peer assessment	Give feedback on students' strengths and weaknesses		Involve students in the development of learning outcomes	Assist students in using feedback to feed forward	Develop students' capabilities in self/peer assessment	

			Make students understand the learning outcomes Involve students in the development of criteria and standards Explain the criteria and standards	Engage students in self-assessment Engage students in peer-assessment
Teachers as Teacher Learners	Engage in self-assessment/ reflection Engage in peer-review of teaching performance Search new and relevant subject-content information	Participate in professional development related to assessment Identify subject-content knowledge need Undertake further education/ training		
Teachers as Stakeholder Partners	Collaborate with family to establish support activities Report to community about students' performances	Inform community of the assessment practices Identify key assessment and teaching issues for review		

The teacher AfL literacy tool can be useful for teachers as an initial stimulus to get organised and involved in a community of practice (CoP) to put their assessment knowledge into practice (Schlager & Fusco, 2004). The dimensions and assessment skills in the framework can validate teachers' ideas and beliefs, and can serve as a common assessment language for teachers to initiate discussions with colleagues (Rhodes & Beneike, 2002). The interactions facilitated by the framework can facilitate co-construction of new knowledge and skills in assessment amongst teachers (Flagg & Ayling, 2011; Harrison, 2005), which would eventually enhance student learning (Goddard, Goddard, & Tschannen-Moran, 2007). Thus, it can be said that the framework is strongly underpinned by socio-cultural theory, which further provides confidence for teachers and other users of the teacher AfL literacy framework to help students learn more effectively.

8.3.3. The Potential Use of the Criteria and Standards of the Teacher AfL Literacy Tool in Teacher Standards

The teacher AfL literacy tool can also contribute to the future development of teacher standards. The relationships between and amongst the dimensions of the teacher AfL literacy tool offer a broader but clearer lens to see the link between assessment, learning and teaching, professional development, and the role played by other stakeholders. For example, in most teacher standards presented in Table 8.4 (where the assessment competencies described in some teacher professional standards are matched to the six dimensions and indicators of teacher AfL literacy tool) each assessment focus area appears to be isolated from other classroom activities. However, in reality, this is not the case because teachers need to use any specific assessment skill in conjunction with other

assessment skills to operate effectively. To illustrate, the dimension related to teachers as pedagogy experts requires teachers to use student assessment information to inform lesson planning. Although this dimension is captured in the Standards for Teacher Competence in the Educational Assessment of Students (1990) in the focus area of using assessment results when making decisions, the description is quite broad. It is unclear which specific aspects of learning and teaching will be informed by student assessment data.

Table 8.4

Comparison between the Dimensions of the Teacher AfL Literacy Tool and the Selected Professional Teacher Standards

Teachers' Assessment for Learning Literacy Tool	Australian Professional Standards for Teachers	UK Teachers' Standards	Philippines National Competency Based Teacher Standards	Teacher Competence in the Educational Assessment of Students (1990)
Teachers as Assessors • Designs assessment tasks • Uses rubrics to assess students' learning • Considers factors that affect students' performance • Avoids interference in task completion • Engages in social moderation	• Assess student learning • Make consistent and comparable judgment • Interpret student data	• Know and understand how to assess the relevant subject and curriculum areas, including statutory assessment requirements • Make use of formative and summative assessment to secure pupils' progress	• Develops and uses a variety of appropriate assessment strategies to monitor and evaluate learning	• Choosing assessment methods • Developing assessment methods • Administering, scoring and interpreting • Developing valid pupil grading procedures • Recognizing unethical, illegal, and otherwise inappropriate assessment methods
Teachers as Pedagogy Experts • Translates learning standards to learning outcomes • Identifies appropriate teaching methods • Considers students' prior knowledge in lesson planning • Considers students' current level of abilities • Considers students' interest • Plans lessons according to students' learning needs • Tailors lessons to available resources • Develops teaching and learning resources				• <i>Using assessment results when making decisions</i>
Teachers as Motivators • Uses flexible teaching activities • Conducts assessment with consideration of student background • Develops an environment of trust • Ensures openness in the class • Builds students' interest to learn • Demonstrates belief in the ability of every student to improve • Affirms students' good performance				

- Clarifies students misconceptions
- Reinforces positive learning attitude of students

Teachers as Student Partners

- Gives feedback on students' strengths and weaknesses
- Assists students in using feedback
- Involves students in the development of learning outcomes
- Makes students understand the learning outcomes
- Involves students in the development of criteria and standards
- Explains the criteria and standards
- Develops students' capabilities in self and peer assessment
- Engages students in self-assessment
- Engages students in peer-assessment
- Moderates feedback and results of self and peer assessment

- Provide feedback to students on their learning
- *Report on student achievement*

- Use relevant data to monitor progress, set targets, and plan subsequent lessons
- Give pupils regular feedback, both orally and through accurate marking, and encourage pupils to respond to the feedback.

- *Communicates promptly and clearly to learners, parents, and superiors about the progress of learners*
- Monitors regularly and provides feedback on learners' understanding of content

- *Communicating assessment results*
- *Using assessment results when making decisions*

Teachers as Teacher Learners

- Participates in professional development related to assessment
- Engages in self- assessment/ reflection
- Engages in peer-review of teaching performance
- Identifies subject- content knowledge needs
- Searches new and relevant subject-content information
- Undertakes further education/ training

Teachers as Stakeholder Partners

- Collaborates with family to establish support activities
- Informs community of the assessment practices
- Reports to community about students' performance
- Identifies key assessment and teaching issues for review

- *Report on student achievement*

- *Communicates promptly and clearly to learners, parents, and superiors about the progress of learners*

- *Communicating assessment results*

One of the most obvious inconsistencies between the dimensions of the tool and other teacher standards is the absence of two factors (*Teachers as Motivators and Teachers as Teacher Learners*). Although these factors are described in other focus areas of teacher professional standards, they are not explicitly linked to teacher assessment literacy. The tool establishes the link between teacher assessment literacy and student motivation, and supports the argument of Dweck (2007) and Hattie and Timperley (2007) related to the critical role of assessment in motivation. The explicit link between assessment and motivation would help teachers develop better understandings of the role of assessment in motivation. Similarly, the engagement of teachers in professional development is mostly described by teacher standards as a separate focus area rather than linking it to the reflective practice of teachers on their assessment experience. The tool provides a mechanism for teachers to identify their professional development needs based on their assessment experience, which Timperley et al. (2008) argue as the strongest input and basis for developing a needs-based professional development program. They recommend that individual teachers should always ask themselves what knowledge and skills they need to have in order to address student learning needs including knowledge, skills, support, motivation and resources. The gap between what the students' need and what the teachers have in terms of knowledge and skills would be the basis for seeking and engaging in a professional development program to effectively support student learning.

Another inconsistency is that in most teacher standards, the key focus areas are identified and used to describe teacher functions rather than using the dimensions of teacher assessment literacy. This kind of conceptualisation of teacher assessment literacy is problematic as the teacher assessment skills are either presented broadly or are too detailed.

For example, the Australian Professional Standards for Teachers (AITSL, 2011) include an assessment standard referred to as assess, provide feedback and report on student learning with focus areas on 1) assess student learning, 2) provide feedback to students on their learning, 3) make consistent and comparable judgment, 4) interpret student data, and 5) report on student achievement. These five focus areas were further unpacked in this study making all indicators of teacher AfL literacy more explicit and clearer. The dimensions in the tool complement the AITSL assessment standards but provide a better guide to teachers when reflecting on how to operationalise the five focus areas alongside other teacher assessment skills. To illustrate, Focus Areas 1, 3 and 4 are considered in the AITSL standards as distinct competencies but in the tool they belong to Factor 1 (*Teachers as assessors*). The association of these three focus areas in one factor helps teachers to have a holistic understanding (which will eventually translate into their practices) that whilst planning assessment strategies they need to consider how these assessment strategies can ensure the consistency, and defend the trustworthiness, of their assessment decisions as informed by their interpretations of student assessment information. Similarly, Focus Area 2 in the AITSL standards is just one indicator of Factor 3 (*Teachers as student partners*) in the tool. This highlights that providing feedback to students in their learning is not an isolated skill, but in order to maximise the effect of feedback in improving student learning, it must be done in conjunction with other teacher assessment skills that involve students in assessment and teaching activities (i.e., involve students in the development of learning outcomes; make students understand the learning outcomes; involve students in the development of criteria and standards; explain the criteria and standards; develop students' capabilities in self and peer assessment; engage students in self-assessment; engage

students in peer-assessment; moderate feedback and results of self and peer assessment; and assist students in using feedback).

The comparison between the dimensions of the tools and the assessment skills included in the existing teacher professional standards show that the tool presents a new way to shape teachers' assessment practices. It provides a tool for teachers to operationalise the principles of AfL, particularly in putting assessment as the central component of classroom activities. The six dimensions explicitly define the use of assessment and assessment information in various learning and teaching activities, including for teacher professional development and for enhancing AfL literacy of stakeholders.

8.4. The Measurement Properties of the Teacher AfL Literacy Tool

8.4.1. Overview

As argued by psychometricians and applied statisticians, and supported by the studies of Hailaya et al. (2014) and Lee and Paek (2014), analyses at the item level are critical to establish the psychometric properties of any scale or assessment tool. Although the results of factor analyses are substantial, there is a need to investigate the results of analyses at the item level. Hailaya's study shows that it is possible that at the item level the psychometric properties are convincing yet at the factor level, the results are problematic or the other way around. In principle, if the assessment tool does not meet the psychometric measures (i.e., validity, reliability, item fit, discrimination index, difficulty index) the data gathered would be problematic and the subsequent interpretations and utilisation of the assessment results are compromised (Field, 2009).

In this section, I discuss the implications of the results of my study related to the internal and external consistency of the tool, the continuum of AfL skills, and the consistency of measurement properties of the tools between similar contexts.

8.4.2. Internal and External Consistency of the Tool

The results of the Rasch analyses of the teacher AfL literacy tool provide evidence for the measurement properties required for a tool to have a lesser measurement error and higher discrimination power (Alagumalai & Curtis, 2005). The finding of this study revealed that the teacher AfL literacy tool has superior measurement properties compared to the existing teacher assessment tools. The high reliability provides confidence for the users that the tool can provide consistent measures of teacher AfL literacy at any point of time of assessment, and can identify the relative level of teacher assessment performance. The internal consistency of the tool is critical to achieving the purpose of its development as the tool is intended to be used at any stage of teacher AfL literacy development. Given this high internal consistency, any difference observed in the results between two periods of assessment can be attributed as teachers' progress towards higher AfL literacy.

Similarly, the evidence related to the fit of the items to the construct validates the results of factor analyses that the 42 teacher assessment skills are the true indicators of the teacher AfL literacy. This external consistency of the tool is also clearly demonstrated by the results of factor analysis.

8.4.3. The Teacher AfL Skills as a Continuum of Practice

The spread of the distribution of both criteria and standards in the logit continuum is an indicator of a well-developed tool or scale (Hendricks, Fyfe, Styles, Skinner, & Merriman, 2012). More importantly, the distribution of teacher assessment skills across the continuum of difficulty, which is approximately approaching the normal curve, provides evidence that the 42 items cover a good portion of a range of items from easy to difficult teacher assessment skills (Andrich & Styles, 2002). The same is true at the standards level, where the spread of individual standards taken as individual items cover a range of performance.

Unlike other tools, where there was no mention about the relative difficulty and the continuum of teacher assessment skills, the parametisation of the results in my tool revealed the difficulty of the items. The difficulty of the items established the continuum of teacher AfL practices. The empirical ordering of standards by difficulty was consistent with the theoretical ordering of standards and, therefore, supported the construct validity of the tool. The clustering of Level A standards for all items, and the Levels B, C, D, and E standards, in a particular logit range, make sense theoretically as these levels of standards require a particular set of skills. For example, in Level A teachers perform basic assessment skills, some of these inconsistent with AfL principles; in Level B, teachers demonstrate acceptable AfL performance; in Level C, teachers use a range of assessment strategies and skills to meet the diverse needs of students; in Level D, teachers are engaged in developing a range of assessment strategies and assessing the effectiveness of their innovation; and in Level E, teachers provide professional support and assistance to their colleagues on different aspects of assessment. These increasing threshold values for levels of standards indicate an orderly category structure for all items in the tool. As the purpose of developing the teacher AfL

literacy tool is to provide a tool that can discriminate teachers along stages of AfL development, the discrimination indexes of individual teacher assessment skills strongly support this property. The spread of indicators along the continuum is an important feature of the tool because it provides adequate information about the teachers' actual level of AfL literacy, which is important for the utilisation of the tool for professional development.

The distribution of standards based on the arbitrary limits set (Bateman & Griffin, 2003), which can be equated to the level of sophistication of teacher AfL performance, and the distribution of teachers based on their level of AfL literacy confirms the results of latent profile analysis that there are five groups of teachers based on their AfL literacy level. The match between the difficulty of the standards and teacher ability provides evidence that teacher AfL literacy comprises criteria and standards that can measure a range of teacher AfL literacy.

This characterisation of each assessment skill has strong implications for professional development. This result lends support to Griffin's (2007) argument that the standards of practice in any tool should follow the concept of the underlying growth continua (Glaser, 1983) to locate the zone of proximal development of individual teachers with reference to their current level of assessment ability.

8.4.4. Consistency of the Factor-Level and Measurement Properties of the Tool between Similar Contexts

The consistency of results of CFA, hierarchical-CFA and ESEM between teachers in Brunei and in the Philippines revealed that the same measurement model fits both data sets. However, this finding needs to be carefully interpreted given the purpose of this study. It is

not the intention of this study to develop a universal teacher AfL literacy tool because teacher AfL literacy is necessarily context-dependent with the social, cultural, political and economic factors associated with a particular educational system impacting on the overall practices and purposes of assessment.

For the purpose of this study, the results of cross-country validation are interpreted in the context that, *given the same teacher AfL skills required in another education setting and the same social context and cultural belief held by stakeholders*, the dimensions of teacher AfL literacy and the typology of teachers based on their AfL literacy can be considered as common overarching dimensions of teacher AfL literacy and groups of teachers. The individual teacher assessment skills comprising each dimension may change depending on the learning and teaching, but the overarching dimensions should be the same.

The context-dependency of teacher assessment practices, although not explicitly stated, is evident from the differences in teacher professional standards related to assessment. This is exemplified by the work of Brookhart (2011) in revising the 1990 National Standards for Teacher Competence in Educational Assessment in the USA where she pointed out “the Standards do not consider current conceptions of formative assessment knowledge and skills, and the Standards do not consider teacher knowledge and skills required to successfully work in the current accountability and ‘standards-based reform’ context” (p.3).

The context-driven nature of teacher AfL literacy is further evident in the existing assessment tools that were purposely developed for specific contexts (e.g., Mertler and Cambell’s (2005) ALI for USA teachers; Newfields (2006) for language teachers).

However, these tools are widely adopted and are being used by other education institutions and teacher cohorts.

8.5. Grouping of Teachers based on AfL Profiles

8.5.1. Overview

Although there have been a number of standards relating to teacher assessment practices proposed in the field, such as the AITSL framework, which uses four stages of teacher development (*graduate, proficient, highly accomplished, and lead*), my study is the first to use an empirical approach to establish the latent profiles of teachers based on their AfL literacy. The findings were used to develop a typology of teacher groupings based on their actual level of assessment literacy level, to illustrate the interdependence of teacher AfL skills, and to characterise high performing teachers, which are discussed below.

8.5.2. An Emerging Typology of Teacher AfL Literacy Development

There are five stages of teacher AfL development (i.e. novice, developing, proficient, skilled, expert) extracted from latent profile analysis as presented in Chapter 6.9. As shown from the results, the levels (as opposed to profile patterns) are prominent differences amongst teachers. Hence, the five groups indicate different stages of teacher AfL literacy. The naming of this typology is associated with Dreyfus' (1980) model of skills acquisition that moves from novice to expert. Looking at these stages of AfL practices, at the beginning level, teacher rely mostly on the abstract principles of AfL, and as they progress and become more skilled, they use more of their classroom experiences to quickly adjust their learning and teaching approaches. Thus, the standards in the tool show a hierarchy of task

difficulty across all dimensions in using AfL principles and practices.

These LPA results further support the results of parametisation in the Rasch analysis. The hierarchy of standards demonstrated in this study, where all Level A standards of teacher assessment practices across the six factors are aggregated in the lower level of the hierarchy map although they are not in the same difficulty level (and the same is true for all other standards) supports the claim that the tool follows the growth continuum of competence described by Glaser (1963/1981). This psychometric property of the tool is critical to its use for professional development because the tool can be used to discriminate between teachers based on their competency levels. Once the individual teacher's current level of assessment literacy is determined, this information can be linked to their zone of proximal development. Hence, appropriate interventions or professional development programs can be designed for individual teachers or group of teachers with the same professional development needs. Thus, the use of the tool draws heavily from Vygotsky's zone of proximal development as a critical starting point for any learning to occur.

As indicated, the teacher AfL framework is useful in conceptualising and understanding teacher AfL literacy from a Vygotskyian perspective. Knowing at what stage individual teachers are at in their AfL literacy helps to determine what is the next set of skills they need to learn to further improve their AfL literacy. The tool addresses a gap in existing assessment literacy tools, that is, the lack of descriptions of what teachers can do in every stage of their AfL development (Brookhart, 2011). The gap between the actual performance and the adjacent level of performance can define the teachers' zone of proximal development (Vygotsky, 1986), which can be translated to individual learning

needs. The level of performance equivalent to ZPD can be used to define the learning outcomes of the professional development program.

8.5.3. The Interdependency of Teacher AfL Skills

The five groups of teachers shown by the LPA suggest that teachers' use of AfL skills is demonstrated fairly consistently across all six dimensions. The results of this study support the notion that teachers undergo stages of skills acquisition, but it should be noted that the adoption of one skill is not isolated from other skills. This means, for example, that teachers who are competent in Factor 1 are also competent in the other five dimensions. In other words, it appears from this research that a particular teacher cannot be an expert in one skill but a novice in others. This finding highlights the interrelatedness of teacher assessment skills as argued by Glasson (2009), which was also demonstrated in the results of second-order factor analysis. Hence, the development of teacher AfL literacy needs to be viewed holistically. This finding has strong implications for the use of the tool for the development of professional development programs. The six dimensions should be seen as a set of interrelated skills and teachers need to be supported in developing their competence across these dimensions. In using the tool, it would not be an effective approach to isolate one or two dimensions and focus on these only. As suggested by this finding, professional development should be focused on across these six dimensions.

8.5.4. Characteristics of High Performing Teachers

The LPA results indicate that teachers' use of assessment to measure student learning is fairly close (denoted by small gaps) amongst groups. This shows that teachers' competencies related to the development of assessment strategies, and using assessment to measure student learning, are not what define effective teaching. Thus, contrary to the emphases on psychometric and measurement skills of some teacher professional standards and previous studies in teacher assessment literacy, the results of the LPA suggest that there are other dimensions, which are more important than teachers functioning as assessors. What discriminates teachers across competency levels is their ability to use assessment data to inform other assessment, learning and teaching activities (Factors 2, 3 & 4). Again, the result of second-order factor analysis, which highlights teachers' use of assessment information corroborates with the results of LPA modelling. The declining mean scores of Class 1 teachers in Factors 2 to 4 indicates that these teachers are not effectively using assessment data gathered from Factor 1 to help students in their learning. What is shown in my research is that the teachers' level of assessment data literacy defines their effectiveness in ensuring student learning (Chappuis, Stiggins, Chappuis & Arter, 2012). This finding highlights the long argued importance of teacher data literacy (Nichols et al., 2009) to support student learning. Hence, it is not only important that teachers are well-equipped in using various assessment types to measure student learning, but more importantly, they need to develop a high level of skills in using assessment information to plan and engage students in learning and teaching activities.

This finding confirms the results of a number of studies showing that high performing teachers are characterised by their abilities to use assessment data to plan

differentiated teaching and assessment activities (Long, Rivas, Light, & Mandinach, 2008); to address students' learning characteristics and needs (Boudett, City, & Murnane, 2007); to quickly adapt instructional methods based on what the assessment data suggest (Abbott, 2008); and to identify learning opportunities for each student (Love, Stiles, Mundry, & DiRanna, 2008).

8.5.5. Teachers' Actual Level of AfL Literacy

This study has also shown that whilst the upper four classes of teachers based on their AfL literacy levels resemble the stages described by AITSL (2011) as graduate, proficient, highly accomplished and lead teacher, the lowest stage is that of teachers who do not meet the minimum standards of AfL practice. This reality, cited by Stiggins as early as 1999, means that there are teachers who are using assessment contrary to the principles of AfL, which may have negative effects on student learning. Although these teachers were not identified in terms of their years of teaching experience, newly hired teachers are most likely to be in this lowest class. Volante and Fazio (2007) concluded in their study exploring the assessment literacy of pre-service teachers, it is unlikely that “teacher candidates are graduating with an acceptable level of assessment literacy to assess and evaluate students effectively” (p. 761). Hence, it is unrealistic to assume that teachers' minimum assessment literacy is within the acceptable level of performance.

Looking at the career stages of teacher development as described in Chapter 3.3.3, this group of teachers fall within the survival stage described by Katz (1972) and Burden (1990) where teachers have inadequate ability and are mostly unprepared to take on the role of effective teachers. This lowest stage of teacher development needs to be accounted for in

the teacher professional standards to reflect the range of teacher development. The lowest stage described by AITSL (2011) as graduate teachers who have completed initial teacher education qualification requirements does not capture this group of teachers. The existence of this class of teachers has strong implications for professional development and student learning, which further reinforces the call of DeLuca and Klinger (2010) and Popham (2009, 2011) for institutionalisation and promotion of teacher AfL literacy. This group of teachers need to be identified and a strong enabling developmental mechanism is needed to fully support them to ensure that student learning is not compromised.

8.6. The Trustworthiness of the Tool

8.6.1. Overview

The rigorous methods used in this research, as described in Chapters 3 and 4, established the depth of empirical evidence supporting the psychometric and factor analytic requirements of the teacher AfL literacy tool (Ewing, Salzberger, & Sinkovics, 2005). However, to fully realise the potential uses of the tool, there is a need to extend the concepts of reliability and validity to a more practical approach (Brookhart, 2003; Kane, 2001; Moss, 2003; Smith, 2003) and ensure that the use of the tool results in effective teacher learning (McNamara & Roever 2006).

In this section, I explain how to enhance the overall trustworthiness of the teacher AfL literacy tool by extending the traditional concept of reliability and validity to a more practical approach. I highlight that ensuring the reliability of an assessment tool does not only require the computation of a reliability index based on psychometric principles and the establishment of the dimensions of the tool, but equally important is collecting sufficient

information and evidence to form a coherent picture of individual teachers' AfL literacy profiles. I then go on to argue that the concept of validity does not only require exploring the dimensionality of the teacher AfL literacy tool by determining the item fit and factor structure, but that it should be extended to social moderation activity where inferences are drawn from the results of assessments (i.e., teacher self-assessment, collegial peer-assessment, school leaders' assessment of teaching performance). Lastly, I propose that the social consequences of teachers' AfL literacy assessment, one measure of validity identified by McNamara and Roever (2006), should be focused on enhancing teachers' assessment literacy, rather than on making evaluative judgments about their performance.

8.6.2. Reliability and Validity: Practical Perspectives

A practical approach to reliability includes evaluating the actual use of the teacher AfL literacy tool and the quality of information gathered (Brookhart, 2003; Smith, 2003). It is recommended that in the actual use of this tool, teachers' self-assessment should be supported by colleagues' peer-assessment and school leaders' evaluation of teaching performance using the same tool. The results of assessment by different people contribute to the information needed to make a decision about individual teachers' actual level of AfL literacy. This practice increases the reliability of the assessment instrument, as according to Smith (2003), 'at the rudimentary level, reliability theory is based on the notion of having enough information to make decisions or draw inferences' (p. 26).

On the issue of validity, Brookhart (2003) and Moss's (2003) concept of classroom validity, which is the reasonableness of the decisions made with regards to learning engagement and learning enhancement, needs to be adopted in evaluating the actual use of

the teacher AfL literacy tool. According to Moss, assessment activities should elicit information that can be used to support learning progress. In other words, validity criteria should include whether the teachers' AfL literacy assessment contributes successfully to a professional development program and enhancement of assessment practices, and how well the assessment serves multiple purposes and audiences. Hence, evidence related to the changes of practices and actual effects on student learning need to be accounted for to support the validity evidence of the tool.

In the actual use and application of the teacher AfL literacy tool, its trustworthiness would be influenced by several factors, which include the nature of the assessment process, the sources of discrete information, the moderation activity and the social consequence of the assessment activity. Each of these factors is discussed below.

8.6.3. The Nature of Teachers' AfL Literacy Assessment Tool

The nature of an assessment tool determines the quality of information that can be gathered. The teacher AfL literacy tool utilised rubrics where a description of performance for each category level is provided. With this feature of the tool, in which the description of each individual category per item reflects the quality of performance arranged in increasing order of competence, the data gathered about individual teachers' AfL literacy level reflect their actual self-perceived performance, which enhances the validity of the assessment tool. The clear and explicit teacher assessment skills (labelled as criteria) and the levels of performance used in the teacher AfL tool provide more detailed guidelines for teachers.

The inclusion of 42 indicators of teacher AfL skills, the explicit criteria for evaluating teacher AfL literacy, and the use of explicit standards arranged in increasing levels of

sophistication has intended explicit benefits for designing and evaluating teacher A/fL literacy program. In the same way as those benefits are in using rubrics to improve student performance, it is argued that the use of criteria and standards for teacher assessment literacy will ultimately enhance teacher assessment practices. The teacher A/fL literacy tool is discussed with reference to Wolf & Steven's (2007) propositions on the benefits of using rubrics.

1. Criteria and standards explicitly define the learning outcomes of the teacher A/fL literacy program. The use of criteria addresses one of the recommendations of Kahl, Hofman and Bryant (2013) to "flesh out the domain of assessment literacy into objectives and learning targets to provide the specificity needed" (p.3). The results of individual teachers' A/fL literacy assessment can provide evidence of their level of performance for each teacher A/fL criterion. The gap between the actual performance and the adjacent level of performance defines the teachers' zones of proximal development (Vygotsky, 1986), which can be translated to their individual learning needs. The level of performance where ZPD sits can be used to define the learning outcomes of a professional development program.
2. The criteria and standards allow the selection of appropriate instructional design and delivery strategies. The key learning targets identified by the teacher A/fL literacy tool will assist the designers of the professional program to identify the best training approach to help teachers learn best. For example, using the typology of teachers derived in LPA, for Class 1 teachers, a seminar will suffice to develop their understanding about the principles of A/fL together with mentoring to help them translate these principles into actual assessment practices. For Class 3

teachers, who are using a range of assessment strategies and trying to innovate and develop more creative approaches to AfL, the focus of their professional development programs should be to enhance their research skills, particularly in evaluating the effectiveness of their practices to help them move to Class 4 teachers. The most effective approach to professional development for Class 3 teachers might be authentic research mentoring. For Class 4 teachers, the focus of the professional development program should be on developing teachers' mentoring and leadership skills to help them move to the next class. Teachers in Class 4 may undergo a program of on the job training in mentoring and facilitating skills in conducting professional development programs.

3. The use of standards enhances the mechanism for monitoring individual teachers' AfL literacy development. The descriptions of performance can accurately and fairly identify teachers' current level of AfL literacy and their AfL literacy needs. The tool designed this way gives a better understanding of the construct being measured and developed. The criteria describe the levels of task demand and the capacity required to execute the competence. The progression of standards and item difficulties allows the placement of the teachers based on their ability level. Thus, teachers can be accurately placed at a particular level on the continuum of AfL performance.
4. The criteria and standards help teachers to engage in self and peer assessment. True to the philosophy of AfL, the use of the teacher AfL literacy tool will provide teachers with a tool for self-evaluation. Hafner and Hafner (2003) stress that when assessment criteria are available, individuals are able to better critique their own

performance. Similarly, the same tool can be used by other teachers to engage in collegial peer-assessment.

5. The criteria and standards make the teacher AfL literacy program more transparent giving equal success for all teachers coming from different backgrounds. The use of the teacher AfL literacy tool makes the AfL practices more explicit and all teachers can develop a common understanding of the required performance. Hence, regardless of teachers' backgrounds and existing level of performance, the chance of improving their own practices is facilitated by the use of the framework.
6. The use of criteria and standards, which were established using the Rasch model, to measure teacher AfL literacy makes a significant contribution to professional development. As argued by Griffin (2004), a tool that uses a probabilistic model of competence impacts teaching and learning and provides an appropriate framework to inform stakeholders about the progress of users. The results of teachers' assessment using the teacher AfL literacy tool can be linked to the content and design of an assessment literacy program (Glaser, 1963; Rasch, 1980; Vygostky, 1986). The Rasch notion of underlying growth continua or latent traits specifies that any construct can be observed through people's performance of tasks. The indicators of competence, as defined by the criteria in the tool, are arranged in increasing order of difficulty or task demands. This arrangement has an important implication in teacher learning because it facilitates the development of skills along the continuum of competence. This is consistent with the work of Glaser where performance and development were also described in terms of the nature and difficulty of the tasks. Drawing together the works of Glaser, Rasch and

Vygotsky the teacher AfL literacy tool can provide a more meaningful interpretation of what individual teachers can actually do in relation to their AfL literacy, and what they need to do to move towards higher levels of AfL performance.

8.6.4. The Components of Teacher AfL Literacy Assessment

The assessment tool developed in this study is neither a checklist against which teachers should tick off their assessment practices, nor a technical set of procedures for teacher training. Rather, it is a systems tool that draws on the sharing of knowledge and practices, forming a community of teacher learners, focusing stakeholder attention and effort around a common goal. The model for assessing teachers' AfL literacy is presented in Figure 8.2.

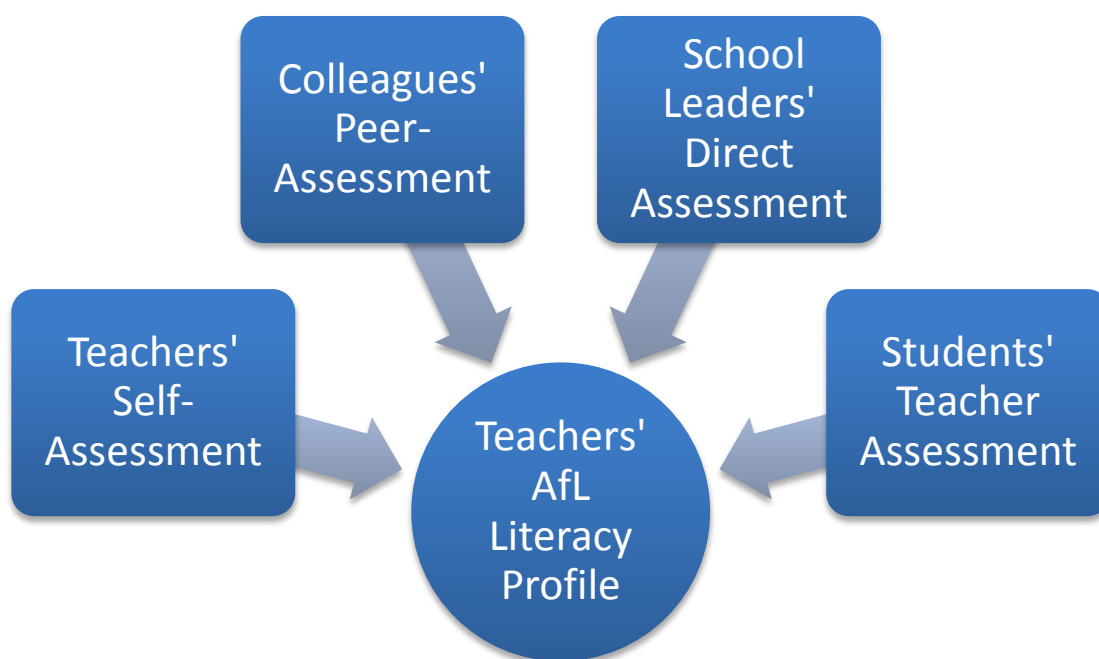


Figure 8.2. The components of teacher's AfL literacy assessment.

As shown in Figure 8.2, the establishment of an individual teacher's AfL literacy profile is based on four sources of information. Teachers, colleagues, and school leaders use the teacher AfL literacy tool, whilst students could use another tool for teacher assessment (Alonzo & Davison, in preparation). The results of these four assessments is then standardised through social moderation. Each component is described below, together with the social moderation process needed to reach consensus about the actual level of a teacher's AfL literacy.

8.4.3.1. Teacher Self Assessment

Teachers know their assessment practices better than any other person in the school system. However, the process of knowing oneself is best facilitated through self-reflection using the

teacher A/FL literacy tool. It is emphasised that the assessment tool should not be used as a checklist where teachers “tick” only the standards met. To realise the full potential of the tool, teachers’ engagement in self-assessment should be grounded in learning activities where teachers gain a deeper understanding of the criteria and performance standards (Torrie & Van Buren, 2008), developing reflective, critical, and evaluative skills, setting professional development goals, gathering evidence of performance, initiating and maintaining professional discussion with colleagues (Lave & Wenger, 1991; Inbar-Lourie, 2008), and establishing dialogue with their immediate superior or principal (Davison, 2013).

In essence, the nature of self-assessment is continuous and iterative. It is recommended that in the initial stage, the school leaders initiate dialogue amongst teachers to explain the criteria and standards to check their understanding. For those criteria that might be interpreted in other ways, the use of exemplar materials and videos of teachers’ actual assessment practices are needed to demonstrate what those criteria and performances look like in actual classroom practice. Following this, teachers need to be given enough time to find opportunities to demonstrate their understanding of the criteria and standards. Teachers are encouraged to regularly engage in self-reflection using the tool to continue to monitor their assessment literacy development and to gather a range of evidence to support the results of their self-assessment. This evidence can be teacher’s output, students’ output or comments about teaching and assessment activities, certification, observation notes, reflection notes, video tapes of classroom assessment activities, lesson plans, records of assessment tasks, feedback sheets or any other document, all organized into a teacher portfolio. This portfolio can be used to show teachers’ understanding of the criteria and

performance standards, their progress over time, and their actual performance (Darling-Hammond & Snyder, 2000). Once teachers are comfortable about their self-assessment rating and have gathered enough evidence, they can submit their portfolios to their school leaders for comments and feedback.

The portfolios are not only useful for teachers to document their assessment and teaching practices but also needed during for regular social moderation activity. Whenever there is disagreement between the ratings of the teacher, colleagues, school leaders, the teachers can use the evidence in their portfolios to support their rating. In this case, their portfolios serve as the document that will inform decisions in regards to their actual level of AfL literacy. Hence, teachers' need to present an organized and coherent portfolio where all artefacts are carefully selected and are linked to specific criteria and performance standards included in the teacher in the teacher AfL literacy tool.

8.4.3.2. Colleagues' Peer Assessment

Teachers also need to engage colleagues' in peer evaluation and feedback as each teacher is a source of assessment information (Timperley, 2011; Timperley et al., 2008) and doing this facilitates constructive interactions and deeper understanding of criteria and standards, which further enhances learning. The value of peer-assessment in student learning has been proven by research Luca & Mclaoughlin, 2002; McDonald & Boud, 2003; Taras, 2003) to help students understand and engage in the learning process. When teachers are the learners, their engagement in peer-assessment has two direct effects. The assessee will learn from the feedback provided by his/her colleagues, whilst the assessor will have another opportunity to apply his critical and evaluative skills and to increase his/her understanding

of the criteria and standards. Also, peer assessment interactions can help initiate the building of a community of assessment literate teachers in which the actual process of colleagues' peer assessment requires is not just limited to classroom observation, but includes everyday interaction and discussion amongst teachers.

Collegial peer assessment activity, however, is not a straightforward activity. As we are dealing with emotive activity, there is a need to establish an environment where the motivation of an individual teacher is not undermined. It requires, therefore, as in student peer assessment, the establishment of an environment of trust, respect, and confidentiality before peer assessment commences.

8.4.3.3. School Leaders' Direct Assessment

There are two main purposes of assessment of teachers' A/fL literacy by school leaders: to determine an individual teacher's A/fL literacy level and to determine the appropriate support needed by individual teachers to further enhance their A/fL literacy. In the context of a teacher's A/fL literacy, the philosophy underpinning a school leader's direct assessment should not be about appraisal but rather a point of entry for school leaders to assume their roles as instructional leaders. Whilst school leaders are engaged in direct performance assessment of a teacher, they can check the teacher's understanding of criteria and standards of A/fL literacy. The alignment between their understanding and teachers' understanding of the criteria and standards is a crucial factor for the successful implementation of A/fL (Davison, 2013; Timperley, Wilson, Barrar, & Fung, 2008).

School leaders' direct assessment should take the form of authentic assessment where evaluation is not limited to classroom observation only but allows teachers to

demonstrate their performance in a variety of ways and requires teachers to provide evidence of their understanding of the criteria and standards and their actual performance.

In conducting classroom observation, as the presence of school leaders is often perceived by teachers as distracting their demonstration of their true level of performance, there is a need to establish an environment of trust and confidence where no teacher threatened by the presence of the school leader. Unless teachers see the presence of their school leader as more of a facilitator than an inspector, direct assessment should not proceed.

8.4.5.2. Students' Teacher Assessment

Students' assessment of their teachers' AfL literacy needs to be focused on the domains directly affecting their learning and motivation. These include the teacher's roles in enhancing student motivation, developing a classroom environment of trust, value, and respect, working closely with them, ensuring trustworthiness of the assessment and teaching activities and providing learning support. Based on the tool developed in this study, these are the competencies of teachers in Domains 2, 3, and 4. The indicators of these domains need to be contextualised for various cohorts of students according to their age level (discussed in Chapter 9). A different tool is needed for students to use, as the teacher AfL literacy tool developed in this study is inappropriate for students because the language used is not within the expected level of student understanding. If such a tool is developed using a more student-friendly criteria and standards, the result of students' assessment can be used to validate the results of teacher's self-assessment on these particular domains.

The perceptions of students about their teachers' AfL practices are important considerations in classroom improvement (Cowie, 2005, 2009). Since the students are the primary focus of AfL, it is necessary to understand how students perceive assessment and what they think is the best way to assess their learning. Insights from literature tell teachers how to shape and modify their approaches to assessment as well as teaching and learning. In Remesal's (2011) study, it was demonstrated that even very young students could perceive the nature of their teachers' assessment practices and consequently, their perceptions affect their learning behaviour. In the study, the second grade students in mathematics who were exposed to a teacher who advocated the use of AfL perceived assessment as pedagogical by nature and students were more engaged in their learning. In contrast, those students who were only accustomed to the use of assessment for marking tended to withdraw from their learning. Although the study has limitations in terms of sample size, so cannot be generalised, the results are consistent with the findings of G. Brown and Hirschfeld (2007) on secondary students in Mathematics. Generally, the achievement of students is higher for those who see assessment as a process that makes them accountable for their learning and who see the positive effect of assessment, with lower mathematics achievement observed in students who perceive assessment as interfering with their learning and those who withdraw from assessment.

8.4.5.3. The Social Moderation Process

A process of social moderation is also important to provide teachers with the opportunity to present evidence on how much they have achieved using the AfL literacy criteria and standards. This activity helps address the reliability requirement of sufficiency of

information and the reasonableness of the interpretation of teacher's assessment literacy profile.

The purpose of social moderation is to reach a consensus amongst the different sources of teacher assessments to establish the actual level of individual teachers' assessment literacy. Some factors that are needed for successful moderation include the development of a culture of trust and learning from each other, as well as openness in communication.

The interpretations of individual teachers' A/L literacy levels by pooling all the discrete pieces of evidence to form an interpretable overall performance profile (Moss, 2003). The interpretations that can be made can then be used for the teacher's professional development. In cases where school leaders and colleagues have found inconsistencies between a teacher's practices and what the standards require them to perform, they can work closely with the teacher to assist him/her to enhance his/her understanding, which may ultimately change teacher practices.

In any decision-making, school leaders must consider the welfare and integrity of the teacher. To move forward, all decisions should impact positively on teachers' learning and lead to the enhancement of their assessment and teaching practices. If the social moderation activity creates self-doubt in teacher, then it has failed to achieve its goal of helping individual teachers to understand their current level of assessment literacy, to establish goals to meet higher level of performance and to build a strong commitment for continuous engagement in self assessment and social moderation activities.

8.6.5. The Social Consequences of Using the Teacher AfL Literacy Tool

As argued by McNamara and Roever (2006), the validity of any assessment tool must also incorporate its social consequences. In this case, the social consequence of these teachers' assessment activities is deeper understanding of AfL principles and the assessment responsibilities of all participants in the assessment process. In the case of teachers, the social consequence of these assessment activities is the improvement of their AfL literacy, enabling them to develop a high level of confidence and expertise to make important decisions to further support student learning. In the case of the school leaders, their involvement in the assessment process should focus on understanding the standards and identifying and providing appropriate support for successful implementation of AfL in the school system.

8.6. Summary

In this chapter, I have presented a discussion of the findings of my study. I have highlighted the major contributions of my study to theory and practice related to the new conceptualisations of teacher AfL literacy, new ways of categorising teachers' AfL practices, understanding the interdependence of assessment skills, and developing a typology of stages of teacher AfL development. Also, I have discussed the advantages of the teacher AfL literacy tool as compared to existing teacher assessment literacy tools. Furthermore, I have suggested a collaborative assessment process for using the tool that emphasises the importance of establishing the trustworthiness of the assessment process and the confidence of people involved in the process.

CHAPTER 9 - SUMMARY AND CONCLUSION

9.1. Introduction

In this chapter, I present the conclusions of my study. I start with a synthesis of findings to demonstrate the major contributions of the study, and then I explore the key methodological limitations of the research. Finally, I outline the implications of the research for theory and practice and I discuss potential areas for future investigation.

9.2. Summary of Findings

The primary aim of this research was to develop a tool for teachers to help them improve their assessment for learning literacy “because assessment-literate teachers will typically make better decisions, and because we want students to be better taught, it should be obvious that today’s teachers must acquire more assessment literacy - and the sooner, the better” (Popham, 2009, p. 6).

This aim of the study was achieved by developing an AfL literacy tool and the results of teachers’ self-assessment were used to explore the dimensions of the construct of teacher AfL literacy using factor analyses, and to establish a typology of teacher AfL latent profiles using latent profile analysis. Also, within the context of the main study, a small additional study was conducted using a different data set from the Philippines to gather evidence related to the applicability of the tool to other similar education contexts. Because of its strong theoretical rationale, and the involvement of various stakeholders and experts in the process, this research can be considered an exemplar of good practice in measurement tool development (Johnston et al., 2003).

There are a number of noteworthy findings arising from both the theoretical and empirical processes of developing the teacher AfL literacy tool and the framework associated with it. Of primary importance was the establishment of six factors as the dimensions of teacher AfL literacy. This result supported two previous conceptualisations of teacher assessment literacy: the item level interpretation of teacher assessment literacy using teacher assessment skills as its indicators; and the dimensionality of teacher AfL literacy as described by other researchers. My study weaves together these two ways of thinking about teacher assessment practices so that individual teacher assessment skills are used to establish the dimensions of teacher AfL literacy. These six factors provide a new classification scheme for teacher AfL practices and better description of roles as *assessors*, *pedagogy experts*, *student partners*, *motivators*, *teacher learners*, and *stakeholder partners*.

These dimensions of teacher AfL literacy extracted expand the present conceptualisation of teacher assessment literacy by highlighting some of the dimensions not described in previous studies. One of these is the role of teachers to reflect on their assessment experience to identify and address their professional needs, and the other one is their responsibility to address stakeholders' assessment literacy. These two dimensions were not emphasised in existing teacher assessment models or frameworks although there is a range of research and arguments supporting these critical roles of teachers.

Similarly, the tool developed reduces the many teacher assessment skills to 42 items that have high factor loadings on their respective factors. Thus, this provides empirical evidence of the key teacher assessment skills that contribute to measuring teacher AfL literacy. Specifically, the results of second-order confirmatory factor analysis highlight that the teachers' role in providing various opportunities for students to take a more active

part and more responsibility in their learning is the factor that greatly contributes to measuring teacher AfL literacy. This highlights the socio-cultural theorisation of teacher AfL practices and the AfL paradigm in general, where interactions and engagement to learning are central to effective teaching and learning. This is followed by the teachers' role in using assessment and assessment information to plan learning and teaching activities and then in developing and sustaining student motivation. The former role highlights the pedagogical focus of teacher AfL practices, whilst the latter role strengthens the long-argued causal effects of assessment on motivation, thereby ensuring that teacher AfL practices build student motivation, hence, avoiding the backwash effect of assessment.

In addition, the results of this study offer a new way to operationalize the principles of AfL, as described by the ARG (2002), with the development of a tool containing teacher assessment skills that adhere to these principles. The tool describes teacher assessment processes within and outside the school. Hence, teachers who use the tool as their guide in teaching are assured that their assessment practices will not compromise student learning. The tool can be used as a measurement instrument and a professional development tool, both theoretically supported and empirically-driven. That is, individual teachers can engage in self-reflection of practice using the tool to determine their perceived level of AfL literacy. The use of explicit standards with five levels of performance guides teachers to identify the next level of performance they need to acquire to further improve their AfL literacy. Hence, the tool could also be used to identify and address teachers' professional development needs.

This study also established a typology of the stages of teacher AfL literacy development from novice, developing, competent, proficient to lead teachers. As this

typology can be linked to the specific professional development needs of teachers, these five stages can be used to cluster teachers who have the same AfL literacy latent profile to design a specific professional development program according to the needs of each class or group.

Furthermore, the results of LPA revealed the roles of teachers that can discriminate high performing from low performing teachers in terms of using AfL. The dimensions related to teachers as pedagogy experts, student partners, and motivators are amongst those teachers' roles that are critical to group teachers based on their AfL skills profile. These three dimensions highlight teachers' abilities in using assessment to engage students in their own learning and to motivate them and in using assessment data to inform other assessment, learning and teaching activities.

The results of the validation study show that the same factor structure of teacher AfL literacy and latent profile of teachers were found in another educational context. This means that the tool can be used in a similar contexts or settings, provided that the same assessment characteristics and cultural assumptions about AfL hold.

At the item-level characteristics of the tool, its trustworthiness is supported by the high reliability index and the fit of all items in the construct. Also, each teacher AfL skill representing the latent variable has a high discrimination index, which supports the overall aim of the tool to determine individual teachers' AfL literacy level. Furthermore, the individual criteria and standards used in the tool showed a range of difficulty and match teachers' range of AfL abilities. This provides evidence that the teacher AfL literacy contains criteria and standards that could measure a range of teacher AfL literacy.

9.3. Limitations of the Research

This study has a number of limitations, which need to be considered when interpreting the results and when using the tool.

In terms of the source of data, it was based only on the self-report of teachers, with no validation of the teachers' self-perceived level of assessment literacy undertaken, such as no collegial peer-assessment, classroom observation, or interview. Although surveys are widely used in social science research, there are issues associated with their use. These include the tendency to gather non-homogenous data, as each item in the survey instrument is prone to various interpretations by individual participants (Fadnes, Taube, & Tylleskär, 2008); the inaccuracies of responses due to carelessness or confusion and intentional false responses labelled as "jokesters" by Fan et al. (2006); and the respondents' tendencies to only report in highly positive performance in case any negative comments would reflect badly on their abilities (Donaldson & Grant-Vallone, 2002; van de Mortel, 2008). Although, in this study, measures were undertaken to address these issues (i.e. teachers were gathered in one venue and before the self-assessment, each criterion and individual standard was explained to them; it was made clear to the teachers that their responses will not be used to report their individual level of assessment literacy, but rather the data will be used to explore the dimensionality of the construct; and there was no information gathered that could identify them) there is still the need to explore the relationship between teacher self-perceptions and their actual assessment performance to establish if some groups of teachers consistently understate or overstate their levels of performance.

Another methodological limitation was the lack of comparisons made between Brunei and the Philippine teachers. This process was not allowed by the Ministry of Education of Brunei to protect the privacy and integrity of teachers from the two countries, but it begs the question as to how really applicable the instrument is across diverse cultural and linguistic contexts.

In the analysis of the data, the exploration of the latent profiles of teachers did not account for covariates such as gender, years of teaching, subjective feelings about assessment literacy, and other variables. This was due to the limited access to the data source owned by Brunei government. Hence, the results are limited only to findings about the general profile of teachers, and no further information as to what covariates influence teacher's placement in a particular latent class. The use of covariates could have been usefully modelled in the latent profile analysis, which could have provided information about which variables could predict the class membership of the teachers. The inclusion of covariates in LPA modelling would have provided additional information to understand the causal relationship of these variables to the stages of teacher AfL development. Hence, if this information is available, the pathway for teacher professional development could be better understood. For example, if gender predicts teacher class membership, then a different training program can be developed for male and female teachers. The same is true if the subjects being taught by teachers could predict their class membership, then the professional development program should be contextualised to the subjects taught by the teachers.

In the generalised partial credit analysis of the Rasch model, these covariates could have been also explored to determine if the tool exhibits measurement bias or differential

item functioning (DIF). In this analysis, these variables could have been taken as facets of the tool, which might have direct effects as to the probability of teachers with the same level of assessment literacy who come from different groups (variables as grouping entity) to place themselves on different levels of standards. The results that could be obtained from this analysis would ensure that the tool is measurement invariant, or in other words, these variables do not influence the responses of individual teachers in assessing their AfL literacy assessment skills. Once these facets of the framework have been thoroughly explored, users of the teacher AfL tool can be assured that teachers with the same AfL literacy level will have the same results in their self-assessment.

9.4. Implications of the Research

The findings of the research have significant implications for both research and theory and practice.

9.4.1. Implications for Theory

The presence of six factors in the teacher AfL literacy framework associated with the tool takes AfL conceptualisation and theorisation in a new direction. This broader categorisation provides a new point of discourse when describing the nature and the processes of teacher AfL literacy. In particular, the AfL literacy framework emerging from this study provides a holistic picture of what skills constitute teacher assessment literacy. The previous conceptualisations and theorisations of teacher assessment literacy that focused on using assessment to support students in their learning have been further expanded to how teachers can reflect on their assessment experience to identify and address their professional

development needs and to ensure that other key stakeholders share the same beliefs and expectations aligned to the principles of AfL.

The result of the second-order factor analysis highlights the role of teacher assessment data literacy. This highly valued teacher assessment skill is shown in this study to have the highest contribution to measuring teacher AfL literacy. Hence, teacher AfL literacy should be predominantly conceptualised based on the teachers' ability to use assessment data to inform various decisions related to learning and teaching, not simply on their ability to interpret and use statistical data. This is contrary to the widely-held concept of teacher assessment literacy as focusing on teachers' ability to use measurement principles to develop assessment tools, particularly in test design and development.

The typology established for teachers' AfL latent profiles, including the interdependency of all factors across the five stages of teacher assessment literacy development, suggests a more holistic professional development design as a requirement for teachers' AfL literacy learning. This finding implies that adoption of a particular assessment skill is not isolated from other skills, and hence, professional development programs should be holistically theorised and developed. This is contrary to the widely used practice of developing professional development programs where the four key aspects of teacher AfL practices (sharing of learning outcomes, using criteria and rubrics, engaging students in self and peer assessment, eliciting and giving of feedback) are used as themes. Therefore, when describing teacher AfL literacy, the relationships between teacher assessment skills need to be emphasised to provide complete descriptions of its nature and processes.

9.4.2. *Implications for Practice*

The development of the teacher AfL literacy tool was guided by the belief that, once it is put into practice, it would help teachers acquire a high level of assessment for learning literacy. This belief is rooted in the strong research evidence of the importance of AfL principles and practices in the classroom, which could also be applied to teacher professional learning and presumably have the same effect as in student learning. It is argued that the teacher AfL literacy tool can provide the necessary tool that integrates practice, theory and philosophy (DeLuca & Klinger, 2010). The robust methodology used in the development of the tool, including a theoretical approach and the deep empirical evidence from factor analysis and the generalised partial credit model analysis of the Rasch analysis, gives confidence to the teachers to use this for their self-assessment.

In a more specific application, the teacher AfL literacy tool has various uses, including but not limited to:

1. *Clarification purposes.* The tool can promote a shared understanding of the levels of performance required in each assessment skill within schools and across the education system. The framework derived from this study can be used to describe a comprehensive picture of teacher AfL literacy, which can be used as the basis for a conversation between teachers and school leaders to initiate discussions around teachers' individual and collective roles and accountabilities. Thus, the tool itself can facilitate the development of a shared understanding of what sound assessment principles and effective practices are, and can address any issues of inconsistent understanding of assessment in terms of task, process and product, concerns about reliability and validity, criteria and standards,

grading models, teachers' needs including support for the on-going development of assessment literacy, and inconsistency between teachers' knowledge and practices.

2. *Monitoring purposes.* The tool can be used by teachers for self and peer assessment. Teachers can use it both to identify their current level of performance and to engage in a collegial evaluation of their teaching practices. Also, the tool can be used by school leaders to monitor and further support the practices and professional development of teachers.
3. *Professional development purposes.* The results of self and peer assessment can be used by teachers to identify their professional development needs. The teacher AfL literacy tool demonstrates the sequence of individual teacher assessment skills, which can be used by teachers to identify their current level of performance and the skills they need to acquire and develop to further improve their assessment literacy. Similarly, school leaders can use these results of self and peer assessments, alongside the results of their evaluation of teacher performance, to develop an assessment literacy program within their school to further scaffold teachers' AfL practices.
4. *Sustainability purposes.* The teacher AfL literacy tool can be used by teachers to engage in on-going self-reflective practice. They can use the criteria and standards described to gather pieces of evidence to develop their teaching portfolios. School leaders can also use the tool to build a community of practice related to assessment to further provide professional development opportunities for teachers.

5. *Evaluation purposes.* Through the collection and analyses of the aggregated data, the tool can also be used by educational systems to evaluate the extent to which stated policies regarding AfL implementation and/or aims of assessment reform programs have been adopted in practice.

9.5. Areas for Future Investigation

The applicability of the tool to another context was explored using teachers from the Philippines, replicating the methods used in the main study. The cultural and political contexts of the two countries are somewhat related because both are in Southeast Asia and English is an official language for instruction together with their respective national languages, but major differences exist in terms of assessment reform as AfL is a key educational reform in Brunei, whilst it is not currently emphasised in the Philippines. The trialling of the tool amongst teachers from other countries, especially those that have no history of colonisation and Western influences, and also amongst English speaking countries such as the USA, Australia and the United Kingdom, would be desirable, yet time and resources did not allow the researcher to include samples of teachers from these countries. However, the evidence suggests that although the results of Brunei and the Philippines were highly consistent, the tool will always need to be recontextualised and revalidated each time to ensure that it accounts for the unique context of the users.

Since this study was completed, two studies have been initiated to re-contextualise the tool and determine its usefulness in the area of teaching English as an additional

language or dialect (EAL/D)² in Australian schools and in the university sector³ in Australia. These two studies provide support for the viability of adapting the teacher AfL literacy tool.

As mentioned in Chapter 8.6.4, this tool should have a companion assessment tool to be used by students to assess their own level of assessment literacy and that of their teachers. It is recommended that this tool should be contextualised from the perspective of students, particularly looking at Factors 2, 3, and 4, which contain AfL practices that are directly engaging and influencing students. This students' teacher assessment tool could provide an additional source of data, which could further validate and support the results of teachers' self-assessment. This would further enhance the trustworthiness of teachers' self-assessment as more information is used to establish the actual level of individual teachers' AfL literacy.

In a broader context, the evidence validity of the tool is yet to be established. A study to determine the impact of the tool on teachers' AfL practices, including changing their belief system towards and understanding of AfL is recommended. Also, there is a need to determine the impact of using the tool to enhancing student learning and increasing achievement. The results of these two recommended studies will provide further empirical evidence related to the utilisation of the tool.

² Alonzo, D. & Davison, C. (in preparation). An EAL/D teacher assessment for learning literacy framework.

³ Alonzo, D. Mirriahi, N. & Davison, C. (in preparation). Putting standards to academic standards-based assessment practices.

9.6. Summary

In this chapter, I have summarised the main findings, discussed the implications for research, theory and practice, presented the limitations of the study, and outlined the potential areas for future research. More importantly, I have shown the contribution of my research towards re-thinking teacher AfL literacy and the ways in which the tool I developed can be used both to guide individual teachers' AfL practices and to identify individual teachers' AfL professional needs to develop a needs-based professional development program.

In conclusion, I have developed a tool to measure the construct of teacher AfL literacy. The development process accounted for the context- dependent nature of assessment. Also, it gave sense of ownership to the system with teachers, school leaders, and Ministry of Education officials highly involved in the entire process of research. The tool was tested both at the item level using the Rasch model and factor level using factor analyses and exploratory structural equation modelling, resulting in the inclusion of 42 items with five factors, including teachers' tasks as assessors, pedagogy experts, student partners, motivators, teacher learners, and stakeholder partners. The teacher AfL literacy tool can be used for teachers' self-assessment to guide them in their reflective practice, as well as for collegial evaluation of teaching to help create a community of practice. It may also be used as part of teacher performance development to provide appropriate support and professional learning to further facilitate teachers in their AfL literacy development.

REFERENCES

- Abbott, D. V. (2008). A functionality framework for educational organizations: Achieving accountability at scale. In E. B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning*. New York, NY: Teachers College Press.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* 52, 317–332.
- 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.
- Aldridge, A. A., & Roesch, S. C. (2008). Developing coping typologies of minority adolescents: A latent profile analysis. *Journal of Adolescence*, 31(4), 499-517. doi: 10.1016/j.adolescence.2007.08.005
- Allen, S., & Sudweeks, R. R. (2001). *Identifying and managing local item dependence in context-dependent item sets*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.
- Allison, C., Baron-Cohen, S. Stone, M. H. & Muncer, S.J. (2015). Rasch modelling and confirmatory factor analysis of systemizing quotient-revised (SQ-R) scale. *Spanish Journal of Psychology*, 18(16), 1-11.
- Ambrosini, V., & Billsberry, J. (2007). *Person-organisation fit as an amplifier of tacit knowledge*. Paper presented at the 1st Global e-Conference on Fit.
- American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). The standards for competence in the educational assessment of students.

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573. doi: 10.1007/bf02293814
- Andrich, D., & Styles, I. (2004). Final report on the psychometric analysis of the Early Development Instrument (EDI) using the Rasch model: A technical paper commissioned for the development of the Australian Early Development Instrument (AEDI). Perth: Murdoch University.
- Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers* (2nd ed.). San Francisco: Jossey-Bass.
- Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment and Evaluation in Higher Education*, 32 (2), 159-181.
- Asghar, A. (2010). Reciprocal peer coaching and its use as a formative assessment strategy for first-year students. *Assessment & Evaluation in Higher Education*, 35(4), 403-417.
- Ash, D., & Levitt, K. (2003). Working within the zone of proximal development: Formative assessment as professional development. *Journal of Science Teacher Education*, 14(1), 1-26.
- Asparouhov, T., & Muthén, B. O. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3), 397-438.
- Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box*.
- Assessment Reform Group. (2002). *Assessment for learning: 10 principles*. In A. R. Group (Ed.): Nuffield Foundation.

- Arter, J. A., Stiggins, R. J., Duke, D., & Sagor, R. (1993). Promoting assessment literacy among principals. *NASSP Bulletin*, 77(556), 1-7. doi: 10.1177/019263659307755602
- Au, W. (2007). High-stakes testing and curricular control: A qualitative meta synthesis. *Educational Researcher*, 258-267.
- Australian Institute for Teaching and School headship. (2011). National professional standards for teachers. Education Services Australia.
- Bailey, K. M., & Brown, J. D. (1995). Language testing courses: What are they? In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 236-256). Clevedon, UK: Multilingual Matters.
- Bamber, M. (2015) The impact on stakeholder confidence of increased transparency in the examination assessment process, *Assessment & Evaluation in Higher Education*, 40(4), 471-487.
- Baroudi, Z. M. (2007). Formative assessment: Definition, elements and role in instructional practice. *Post-Script: Postgraduate Journal of Education Research*, 8(1), 37-48.
- Barr, R. B., & Tagg, J. (1995). From teaching to learning: A new paradigm for undergraduate education. *Change*, 27(6), 12-25.
- Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42, 815-824.
- Bateman, A., & Griffin, P. (2003). The appropriateness of professional judgment to determine performance rubrics in a graded in a competency based assessment framework. A paper presented at Australian Association for Research in Education. Auckland, New Zealand.

- Baumard, P. (1999). *Tacit knowledge in organizations*. London: Sage.
- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186-203. doi: 10.1207/s15328007sem1302_2
- Beaumont, C., O'Doherty, M., & Shannon, L. (2011): Reconceptualising assessment feedback: A key to improving student learning?, *Studies in Higher Education*, 36(6), 671-687.
- Bennett, R. E. (2009). Formative assessment: can the claims for effectiveness be substantiated? Proceedings of the 35th Annual Meeting of the International Association for Educational Assessment, Assessment for a Creative World. Brisbane, Australia.
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles Policy and Practice*, 18(1), 5-25.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.
- Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, 17, 303-316.
- Bevitt, S. (2015) Assessment innovation and student experience: A new assessment challenge and call for a multi-perspective approach to assessment research, *Assessment & Evaluation in Higher Education*, 40(1), 103-11.
- Biggs, J. (1996). Assessing learning quality: Reconciling institutional, staff and educational demands. *Assessment and Evaluation in Higher Education*, 21(1), 5-15.

- Biggs, J. (1998). Assessment and classroom learning: A role for summative assessment? *Assessment and Evaluation*, 5(1), 103-110.
- Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development*, 18(1), 57-75. doi: 10.1080/0729436990180105.
- Biggs, J., & Tang, C. (2007). *Teaching for quality learning at university* (3rd ed.). Berkshire, England: Open University Press.
- Bird, F. L., & Yucel, R. (2015) Feedback codes and action plans: Building the capacity of first-year students to apply feedback to a scientific report, *Assessment & Evaluation in Higher Education*, 40(4), 508-527.
- Birenbaum, M., Kimron, H., & Shilton, H. (2011). Nested contexts that shape assessment for learning: School-based professional learning community and classroom culture. *Studies in Educational Evaluation*, 37(1), 35-48. doi: <http://dx.doi.org/10.1016/j.stueduc.2011.04.001>
- Black, H. (1986). Assessment for learning. In D. L. Nuttall (Ed.), *Assessing educational achievement* (pp. 7–18). London: Falmer Press.
- Black, H., & Dockrell, W. B. (1984). *Criterion-referenced assessment in the classroom*. Edinburgh: Scottish Council for Research in Education.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assesment for learning: Putting it into practice*. Oxford, United Kingdom: Oxford University Press.
- Black, P., McCormick, R., James, M., & Pedder, D. (2006). Learning how to learn and assessment for learning: a theoretical inquiry. *Research Papers in Education*, 21(2), 119 - 132. [10.1080/02671520600615612](https://doi.org/10.1080/02671520600615612)

- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policies and Practice*, 5(1), 7-74.
- Black, P., & Wiliam, D. (1999). *Assessment for learning: Beyond the black box*. Cambridge: University of Cambridge School of Education.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5-31.
- Black, P., & Wiliam, D. (2006) Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81-100). London: Sage.
- Bloom, B. S., Hastings, J. T., & Madaus, G. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.
- Bloom, B. S., Madaus, G., & Hastings, J. T. (1981). *Evaluation to improve learning*. New York: McGraw-Hill.
- Board of Studies (2015). Assessment resource centre. Teaching and Educational Standards NSW for and on behalf of the Crown in right of the State of New South Wales.
Link: <http://arc.boardofstudies.nsw.edu.au/>
- Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology*, 54(2), 199-231.
- Boekaerts, M., Maes, S., & Karoly, P. (2005). Self-regulation across domains of applied psychology: Is there an emerging consensus? *Applied Psychology*, 53(2), 149-154.
- Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment and Evaluation in Higher Education*, 24(4), 413-426.

- Boudett, K. P., City, E. A., & Murnane, R. J. (Eds.). (2007). *Data wise: A step-by step guide to using assessment results to improve teaching and learning*. Cambridge, MA: Harvard Education Press.
- Brewer, D. J. (1993). Principals and student outcomes: Evidence from U.S. high schools. *Economics of Education Review*, 12, 281-292.
- Brookhart, S. M. (1997). A Theoretical Framework for the Role of Classroom Assessment in Motivating Student Effort and Achievement. *Applied Measurement in Education*, 10(2), 161-180. doi: 10.1207/s15324818ame1002_4
- Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, 22(4), 5-12. doi: 10.1111/j.1745-3992.2003.tb00139.
- Brookhart, S. M., Walsh, J. M., & Zientarski, W. A. (2006). The Dynamics of Motivation and Effort for Classroom Assessments in Middle School Science and Social Studies. *Applied Measurement in Education*, 19(2), 151-184. doi: 10.1207/s15324818ame1902_5
- Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practices*, 30(1), 3-12.
- Brookhart, S. M., & Durkin, D. T. (2003). Classroom assessment, student motivation, and achievement in high school social studies classes. *Applied Measurement in Education*, 16(1), 27-54. doi: 10.1207/S15324818AME1601_2
- Brown, F. G. (1983). *Principles of educational and psychological testing* (3 ed.). New York: Holt, Rinehart, & Winston.

- Brown, G., & Hirschfeld, G. (2007). Students' conception of assessment and mathematics: Self-regulation raises achievement. *Australian Journal of Educational & Development Psychology*, 7, 63-74.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.
- Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136-162). Thousands Oak, CA: Sage.
- Bruning, R. H., Schraw, G. J., & Ronning, R. R. (1999). *Cognitive psychology instruction* (3rd ed.). Upper Saddle River, N.J.: Merrill.
- Buhagiar, M. (2007). Classroom assessment within the alternative assessment paradigm: Revisiting the territory. *The Curriculum Journal*, 18(1), 39-56.
- Burden, P. (1990). Teacher development. In W. R. Houston (Ed.), *Handbook of research on teacher education* (pp. 311-328). New York, New York: Mcmillan.
- Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education*, 36, 395–407
- Carlson, J. E. (1993). *Dimensionality of NAEP instruments that incorporate polytomously-scored items*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.

- Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2), 195-212. doi: 10.1007/BF01246098
- Chappuis, J. (2009). *Seven strategies of assessment for learning*. Portland OR: ETS Assessment Training Institute.
- Chappuis, S., Stiggins, R. J., Arter, J., & Chappuis, J. (2005). *Assessment for learning: An action guide for school heads* (2nd ed.).
- Chappuis, J., Stiggins, R. Chappuis, S., Arter, J. (2012). Classroom assessment for student learning: Doing it right-using it well. *Assessment Training Institute*.
- Chater, P. (1984). *Marking and assessment in English*. London: Methuen.
- Clark, I. (2011). Formative assessment and motivation: Theories and themes. *Prime Research on Education*, 1(2), 26-36.
- Clemens, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, 42(2), 127-166.
- Cochran-Smith, M. (2011). The outcomes question in teacher education. *Teaching and Teacher Education*, 17, 527-546.
- Cockcroft, W. H. (1982). Mathematics counts. Report of the Committee of Inquiry into the Teaching of Mathematics in Primary and Secondary School sin England and Wales. HMSO: London.
- Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Larence Erlbau Associates, Publishers.

- Comrey, A. (1973). *A first course in factor analysis*. 1973: New York: Academic Press, Inc.
- Converse, J. M., & Preser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Newbury Park, CA: Sage.
- Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, 6, 147–168.
- Costello, A. B., & Osborne, J.W. (2005), Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7).
- Cowie, B. (2005). Pupil commentary on assessment for learning. *The Curriculum Journal*, 16(2), 137-151.
- Cowie, B. (2009). My teacher and my friends helped me learn. In D. M. McInerney, G. T. L. Brown & G. A. D. Liem (Eds.), *Student perspective on assessment: What students can tell us about assessment for learning* (pp. 85-105): Information Age Publishing.
- Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education*, 6(1), 102 – 116.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438-481.

- Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education *International Journal of Educational Research*, 9-19.
- Darling-Hammond, L. (1998). Teacher learning that supports student learning *Educational Leadership*, 55(5).
- Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teacher and Teacher Education*, 16, 523-545.
- Davies, A. (2008). *Assessing academic english: Testing english proficiency 1950–1989–The IELTS solution*. Cambridge, England: UCLES/Cambridge University Press.
- Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5(4), 194-197. doi: [http://dx.doi.org/10.1016/S0897-1897\(05\)80008-4](http://dx.doi.org/10.1016/S0897-1897(05)80008-4)
- Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing*, 21(3), 305-334.
- Davison, C. (2007). Views from the chalkface: English language school based assessment in Hong Kong. *Language Assessment Quarterly*, 4(1), 37-68.
- Davison, C. (2008). *Assessment for learning: Building inquiry-oriented assessment communities*. Paper presented at the 42nd Annual TESOL Convention and Exhibit, New York, NY.
- Davison, C. (2013). Innovation in assessment: Common misconceptions and problems. In K. Hyland & L. Wong (Eds.), *Innovation and change in English language education* (pp. 263-275). Oxon: Routledge.

- Davison, C., & Leung, C. (2009). Current issues in english language teacher-based assessment. *TESOL Quarterly*, 43(3), 393-415.
- Davison, C., & Michell, M. (2014). In press. *EAL assessment: What do Australian teachers want?*
- DeCorte, E. (1996). New perspective in learning and teaching in higher education. In A. Burgen (Ed.), *Goals and purposes of higher education in the 21st Century*. London: Jessica Kingsley.
- DeCoster, J. (1998). Overview of factor analysis. Retrieved on 23 February 2013 from <http://stat-help.com/notes.html>
- Dekker, T., & Feijs, E. (2005). Scaling up strategies for change: Change in formative assessment practices. *Assessment in Education*, 12(3), 237-254.
- DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice*, 17(4), 419–438.
- DeVellis, R. F. (2003). *Scale development: Theory and applications* (2 ed.). Sage: Thousands Oaks, CA.
- de Winter, J., & Dodou, D. (2015). Common factor analysis versus principal component analysis: A comparison of loadings by means of simulations, *Communications in Statistics - Simulation and Computation*, DOI:10.1080/03610918.2013.862274
- Dhanaraj, C., Lyles, M. A., Steensma, H. K., & Tihanyi, L. (2001). Managing tacit knowledge and explicit knowledge transfer in IJVs: The role of relational embeddedness and the impact on performance. *Journal of International Business Studies*, 35(5), 428-442.

- Diamantopoulos, A., & Siguaw, J.A. (2000). *Introducing LISREL*. London: Sage Publications.
- Dinham, S. (2005). Principal leadership for outstanding educational outcomes. *Journal of Educational Administration*, 43(4), 338-356.
- Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, 24(3), 331-350. doi: 10.1080/03075079912331379935
- Donaldson, S. I., & Grant-Vallone, E.J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, 17(2), 245-260.
- Doppelt, Y. (2003). Implementation and assessment of project-based learning in a flexible environment. *International Journal of Technology & Design Education*, 13(3), 255-272.
- Dreyfus, S. E., & Dreyfus, H. L. (1980). A five-stage model of the mental activities involved in directed skill acquisition. CA: Defense Technical Information System-California University Berkeley Operations Research Center.
- Dunn, K. E., & Mulvenon, 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).
- Dunn, R., Griggs, S. A., Olson, J., Beasley, M., & Gorman, B. S. (1995). A meta-analytic validation of the Dunn and Dunn model of learning-style preferences. *The Journal of Educational Research*, 88(6), 353-362. doi: 10.1080/00220671.1995.9941181

- Dweck, C. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press.
- Dweck, C. (2007). *Mindset: The new psychology of success*. New York: Ballantine Books.
- Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist*, 36(2), 103-112. doi: 10.1207/S15326985EP3602_5
- Evers, C. W., & Mason, M. (2011). Context based inferences in research methodology: the role of culture in justifying knowledge claims. *Comparative Education*, 47(3), 301-314. doi: 10.1080/03050068.2011.586763
- Ewing, M., 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,
- Facione, P. A. (2011). *Critical thinking: What it is and why it counts*. Milbrae, CA: Measured Reasons and The California Academic Press.
- Fadnes, L., Taube, A., & Tylleskär, T. (2008). How to identify information bias due to self-reporting in epidemiological research. *The Internet Journal of Epidemiology*, 7(2).
- Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovations in Education & Training International*, 32(2), 175-187. doi: 10.1080/1355800950320212
- Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods,, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, 6, 56-83.

- Fan, X., Miller, B. C., Park, K., Winward, B. W., Christensen, M., Grotevant, H. D., Tai, R.H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, 18, 223–244.
- Ferguson, R. F., & Brown, J. (2000). Certification test scores, teacher quality, and student achievement. In D. Grissmer & J. M. Ross (Eds.), *Analytic issues in the assessment of student achievement*. Washington, DC: National Center for Education Statistics.
- Field, A. (2009). *Discovering statistics using SPSS*. London: Sage.
- Flagg, E., & Ayling, D. (2011). Teacher engagement in a Web 2.0 world: Developing your online teaching and learning Community of Practice. In G. Williams, P. Statham, N. Brown & B. Cleland (Eds.), *Changing Demands, Changing Directions*. Proceedings ascilite Hobart 2011. (pp.386-391).
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466–491.
- Foos, T., Schum, G., & Rothenberg, S. (2006). Tacit knowledge transfer and the knowledge disconnect. *Journal of Knowledge Management*, 10(1), 6-18.
- Fredericksen, J., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27-32.
- Friedenberg, L. (1995). *Psychological testing: design, analysis and use*. Boston, MA: Allyn and Bacon. .
- Friedman, H., & MacDonald, D. (2006). Humanistic testing and assessment. *Journal of Humanistic Psychology*, 46(4), 510-529.

- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9(2), 113-132. doi: 10.1080/15434303.2011.642041
- Fullan, M. (2001). *Leading in a culture of change*. San Francisco: Jossey-Bass.
- Fullan, M. (2002). Principals as leaders in a culture of change. *Educational Leadership*, special issue, May.
- Fuller, F., & Brown, O. (1975). Becoming a teacher. In K. Ryan (Ed.), *Teacher education: Seventy-fourth yearbook of the national society for the study of education*. Chicago: University of Chicago Press.
- Gamire, E., & Pearson, G. (2006). *Tech tally: Approaches to assess technological literacy*. National Academies of Sciences.
- Gatewood, R. D., & Field, H. S. (2001). *Human resource selection* (5th ed.). United States of America: South-Western.
- George, D., & Mallery, P. (2005). *SPSS for windows step by step: A simple guide and reference* (5th ed.). Boston: Pearson Education.
- Gibbs, G. (1999). Using assessment strategically to change the way students learn. In S. Brown & A. Glasner (Eds.), *Assessment matters in higher education: Choosing and using diverse approaches*. Buckingham: SRHE/ Open University Press.
- Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education*, 24(355-392).
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519-521.
- Glaser, R. (1981). The future of testing. A research agenda for cognitive psychology and psychometrics. *American Psychologist*, 36, 923-936.

- Glasson, T. (2009). *Improving student achievement: A practical guide to assessment for learning*. Victoria, Australia: Curriculum Corporation (Education Services Australia Ltd).
- Glickman, C. (2002). *Leadership for learning: How to help teachers succeed*. Arlington, Virginia: Association for Supervision and Curriculum Development.
- Goddard, Y. L., Goddard, R. D., & Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary school. *Teachers College Record*, 109(4) 877-896.
- Goddard, R. D., Hoy, W. K., & Hoy, A. W. (2000). Collective teacher efficacy: Its meaning, measure, and impact on student achievement. *American Educational Research Journal*, 37(2), 479-507. doi: 10.3102/00028312037002479
- Gordon, S., & Reese, M. (1997). High stakes testing: Worth the price? *Journal of School headship*, 7, 345-368.
- Gray, E., & Tall, D. (1994). Duality, ambiguity and flexibility: A proceptual view of simple arithmetic. *Journal for Research in Mathematics Education*, 26(2), 115-141.
- Griffin, P. (2000). *Competency based assessment of higher order competencies*. Paper presented at the NSW ACEA State Conferences Mudgee.
- Griffin, P. (2004). Students test development calibration. *Monitoring Primary Achievement in Vietnam*, 2, 64-98.
- Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation*, 33, 87-99.

- Grimm, L. G., & Yarnold, P. R. (1995). *Reading and understanding multivariate statistics*. Washington, DC: American Psychological Association.
- Guadagnoli, E., & Velicer, W. F. (1998). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265-275.
- Guskey, T.R. (2003). How classroom assessment improve learning. *Educational Leadership*, 60(5), 6-11.
- Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting system for student learning*. Corwin Press, Inc.Thousands Oak: CA.
- Habing, B. (2003). Exploratory factor analysis. Retrieved on 20 March 2012 from <http://www.stat.sc.edu/~habing/courses/530EFA.pdf>
- Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*, 25(12), 1509-1528. doi: 10.1080/0950069022000038268
- Hailaya, W., Alagumalai, S., & Ben, F. (2014). Examining the utility of assessment literacy inventory and its portability to education systems in the Asia Pacific region. *Australian Journal of Education*, 58(3), 297-317. doi: 10.1177/0004944114542984
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective*. Upper Saddle River, New Jersey: Pearson.
- Hall, M. T., Howard, M. O., & McCabe, S. E. (2010). Subtypes of adolescent sedative/anxiolytic misusers: A latent profile analysis. *Addictive Behaviors*, 35(10), 882-889. doi: <http://dx.doi.org/10.1016/j.addbeh.2010.05.006>
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement* (3 ed.). New York: Macmillan.

- Hancock, G. R., & Mueller, R.O. (2006). *Structural equation modeling: A second course*. Greenwich CT: Information Age Publishing, Inc.
- Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Reserach and Development*, 20(1), 53-70.
- Harlen, W. (2005). Teachers' summative practices and assessment for learning-tensions and synergies. *British Curriclum Foundation*, 16(2), 207-223.
- Harlen, W. (2007). *Assessment of learning*. London: Sage.
- Harlen, W., & Deakin Crick, R. (2003). Teaching and motivation for learning. *Assessment in Education*, 10(2), 169-208.
- Harrison, C. (2005). Teachers developing assessment for learning: mapping teacher change. *Teacher Development*, 9(2), 255-263. doi: 10.1080/13664530500200251
- Hasking, P. A., & Oei, T.P.S. (2002). Confirmatory factor analysis of the COPE questionnaire on community drinkers and an alcohol-dependent sample, *Journal os Studies in Alcohol*, 63, 631-640.
- Hattie, J. (2003). *Teachers make a difference: What is the research evidence?* Paper presented at the Australian Council for Educational Research Annual Conference on Building Teacher Quality, Melbourne.
- Hattie, J. (2008). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*: Routledge, Hoboken.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112.

- Hawe, E. (2002). Assessment in a pre-service teacher education programme: The rhetoric and the practice of standards-based assessment. *Asia-Pacific Journal of Teacher Education*, 30(1), 93-106. doi: 10.1080/13598660120115002
- Hayward, L., & Hedge, N. (2005). Travelling towards change in assessment: policy, practice and research in education. *Assessment in Education*, 12(1), 55-75.
- Hayward, L., Priestley, M., & Young, M. (2004). Ruffling the clam of the ocean floor: Merging practice, policy and research in assessment in scotland. *Oxford Review of Education*, 30(3), 397-415.
- Hébert, T. P. (1993). Reflections at graduation: The long-term impact of elementary school experiences in creative productivity. *Roeper Review*, 16(1), 22-28. doi: 10.1080/02783199309553529
- Hendry, G. D., Armstrong, S., & Bromberger, N. (2011). Constructive guidance and feedback for learning: The usefulness of exemplars, marking sheets and different types of feedback in a first year law subject. *Assessment and Evaluation in Higher Education*, 36, 1-11.
- Hendriks, J., Fyfe, S., Styles, I., Skinner, S.R., & Merriman, G. (2012). Scale construction utilising the Rasch unidimensional measurement model: A measurement of adolescent attitudes towards abortion. *Australian Medical Journal*. 5(5), 251-261. <http://dx.doi.org/10.4066/AMJ.2012.952>.
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66(3).

- Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). The nature and impact of teachers' formative assessment practices. Los Angeles, CA: Center for Assessment and Evaluation of Student Learning, University of California, Los Angeles.
- Hill, K., & McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing*, 29(3), 395-420. doi: 10.1177/0265532211428317
- Hogarty, K., Hines, C., Kromrey, J., Ferron, J., & Mumford, K. (2005). The quality of factor solutions in exploratory factor analysis: the influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, 65(2), 202-206.
- Hong Kong Curriculum Development Council (2001). *Learning to learn: The way forward to curriculum development*. Hong Kong Curriculum Development Council.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1), 53-60.
- Hounsell, D. (1997). Contrasting conceptions of essay-writing. In F. Marton, D. Hounsell & N. Entwistle (Eds.), *The Experience of Learning*. Edinburgh: Scottish Academic Press.
- Hounsell, D., McCune, V., Hounsell, J., & Litjens, J. (2008). The quality of guidance and feedback to students. *Higher Education Research & Development*, 27(1), 55-67.
- Howard, T. (2003). Culturally relevant pedagogy: Ingredients for critical teacher reflection. *Theory into Practice*. 42(3), 195-203.

- Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, 64(2), 157-170. doi: 10.1111/j.1467-9574.2009.00445.x
- Hoy, W. K., & Miskel, C. G. (2001). *Educational administration: Theory, research, and practice* (6th ed.). New York, NY: McGraw-Hill.
- Hu, L. T., & Bentler, P. M. (1999). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.
- Hutchinson, S, R., & Olmos, A (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling: A Multidisciplinary Journal*, 5,344-364.
- Hutchinson, C., & Young, M. (2011). Assessment for learning in the accountability era: Empirical evidence from Scotland. *Studies in Educational Evaluation*, 37(1), 62-70. doi: <http://dx.doi.org/10.1016/j.stueduc.2011.03.007>
- Huynh, H., & Ferrera, S. (1994). A comparison of equal percentile and partial credit equating for performance-based assessment comprised of free-response items. *Journal of Educational Measurement*, 15-141.
- Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: a focus on language assessment courses. *Language Testing*, 25(3), 385-402.
- Jackson, D. L., Gillapsy, J.A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations, *Psychological Methods*, 14(1), 6-23.

- James, M. (1992). Assessment for learning. Paper presented at the Annual Conference of the Association for Supervision and Curriculum Development. New Orleans, LA.
- James, M. (2006). Assessment, teaching and theories of learning. In J. Gardner (Ed.), *Assessment and learning*. London: Sage Publication Ltd.
- James, M., Black, P., Carmichael, P., Drummond, M. J., Fox, A., MacBeath, J., . . . William, D. (2007). *Improving learning how to learn: Classrooms, schools and networks*. London: Routledge.
- James, M., & Pedder, D. (2006) 'Professional learning as a condition for assessment for learning'. In: J. Gardner (ed.) *Assessment for learning: theory, policy and practice*, London: Sage, 27-43.
- James, M., Black, P., Carmichael, P., Drummond, M. J., Fox, A., MacBeath, J., . . . William, D. (2007). *Improving learning how to learn: classrooms, schools and networks*. London: Routledge.
- Jensen, B., & Reichl, J. (2011). Better teacher appraisal and feedback: Improving performance. Melbourne: Grattan Institute.
- Johnston, J. M., Leung, G. M., Fielding, R., Tin, K., & Ho, L. (2003). The development and validation of a knowledge, attitude and behaviour questionnaire to assess undergraduate evidence-based practice teaching and learning. *Medical Education*, 37, 992–1000
- Johnston, J., & McClune, W. (2000). *Selection project sel 5.1. Pupil motivation and attitudes- self-esteem, locu sof control, learning disposition and the impact of selction tecahing and learning*. Belfast: Queen's University.

- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183-202. doi: 10.1007/bf02289343
- Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models* Cambridge, MA: Abt Books.
- Kahl, S. R., Hofman, P., & Bryant, S. (2013). Assessment literacy standards and performance measures for teacher candidates and practicing teachers. *A report prepared for the Council for Accreditation of Educator Preparation*. Measured Progress: Dover, New Hampshire.
- Kalyuga, S., Chandler, P., & Sweller, J. (2001). Learner experience and efficiency of instructional guidance. *Educational Psychology*, 21(1), 5-23. doi: 10.1080/01443410124681
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342. doi: 10.2307/1435453
- Katz, L. G. (1972). Developmental stages of preschool teachers. *The Elementary School Journal*, 73(1), 50-54. doi: 10.2307/1000851
- Kauffman, J., & Schunn, C. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science*, 39, 387–406.
- Kennedy, K. J., Sang, J. C. K., Wai-ming, F. Y., & Fok, P. K. (2006). *Assessment for productive learning: Forms of assessment and their potential for enhancing learning*. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore.

- Keppel, M., & Carless, D. (2006). Learning-oriented assessment: a technology-based case study. *Assessment in Education*, 13(2), 179-191.
- Killen, R. (2003). Validity in outcomes-based assessment. *Perspective in education*, 21(1), 1-14.
- Kim, H.J. (2008). Common factor analysis versus principal component analysis: Choice for symptom cluster research, *Asian Nursing Research*, 2(1), 17-24.
- Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. *Assessment in Education: Principles, Policy & Practice*, 2(2), 145-163. doi: 10.1080/0969594950020203
- Klenowski, V. (2011). Assessment for learning in the accountability era: Queensland, Australia. *Studies In Educational Evaluation*, 37(1), 78-83.
- Klenowski, V., Tobias, S., Funnell, B., Vance, F., & Kaesehagen, C. (2010) Culture-fair assessment: Challenging Indigenous students through effortful mathematics teaching. In *AARE International Education Research Conference*, 28 November - 2 December 2010, Melbourne.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.
- Kline, R. B. (2010). *Principles and practices of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254-284.

- Knight, P. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education*, 27(3), 275-286.
- Koh, K.H. (2011) Improving teachers' assessment literacy through professional development, *Teaching Education*, 22(3), 255-276,
- Koshy, S. (2013). Differentiated assessment activities: Customising to support learning. In P. Bartholomew, N. Courtney & C. Nygaard (Eds.), *Quality Enhancement of University Teaching and Learning*, Libri Publishing, Faringdon.
- Kreiner, S., & Christensen, K.B., (2011). Item screening in graphical loglinear Rasch models. *Psychometrika*, 76, 228-256.
- Kroh, M. (2007). Measuring left-right political orientation: The choice of response format. *The Public Opinion Quarterly*, 71(2), 204-220. doi: 10.2307/4500371.
- Kuh, G. D., & Ewell, P.T. (2010), The state of learning outcomes assessment in the United States, *Higher Education Management and Policy*, 22(1). DOI: <http://dx.doi.org/10.1787/hemp-22-5ks5dlhqbfr1>
- Lamoureux, E. L., Pallant, J. F., Pesudous, K., Rees, G., Hassell, J. B., & Keeffe, J.E. (2007). The impact of vision impairment questionnaire: An assessment of its domain structure using confirmatory factor analysis and Rasch analysis. *Investigative Ophthalmology & Visual Science*, 48(3), 1001-1006.
- Lantolf, J.P. (2000). Second language learning as a mediated process. *Language teaching*, 33(2), 79-96.
- Lantolf, J.P. (2007) Socio-cultural theory: A unified approach to L2 learning and teaching. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp.615-623). New York: Springer.

- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.
- Learning and Teaching Scotland (2006). *Assessment is for Learning - Self-Assessment Toolkit*. The Scottish Government: Education Scotland.
- Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, 12(3), 237-255. doi: 10.1207/s15324818ame1203_2
- Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment*, 32(7), 663-673. doi: 10.1177/0734282914522200
- Leonard, M., & Davey, C. (2001). *Thoughts on the 11 plus*. Belfast: Save the Children Fund.
- Lew, M. D. N., Alwis, W. A. M., & Schmidt, H. G. (2010). Accuracy of students' self-assessment and their beliefs about utility. *Assessment & Evaluation in Higher Education*, 35, 135–156
- Linacre, J. M. (1999). Category disordering (disordered categories) vs. threshold disordering (disordered thresholds). *Rasch Measurement Transactions*, 13(1), 675.
- Lindquist, R. (1991). Don't forget the pilot work. *Heart Lung*, 20, 91-92.
- Linn, M.C. (1992). *Gender differences in educational achievement, sex equity in educational opportunity, achievement and testing*. Princeton, NJ: Educational Testing Service.
- Lipman, M. (1995). *Critical thinking - what can it be?* . Boston, MA: Allyn & Bacon.

- Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the components in a normal mixture. *Biometrika*, 88(767-778).
- Long, L., Rivas, L., Light, D., & Mandinach, E. B. (2008). The evolution of a homegrown data warehouse: TUSDStats. In E. B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 209-232). New York, NY: Teachers College Press.
- Losada, A., Knight, B. G., Marquez-Gonzales, M., Montorio, I., Etxeberria, I., & Peñacoba, C. (2008). Confirmatory factor analysis of the familism scale in a sample of dementia caregivers, *Aging and Mental Health*, 12(4), 504-508.
- Loughran, J., Mulhall, P., & Berry, A. (2004). In search of pedagogical content knowledge in science: Developing ways of articulating and documenting professional practice. *Journal of Research in Science Teaching*, 41(4), 370-391. doi: 10.1002/tea.20007
- Love, N. (2013). *Data literacy for teachers*. Port Chester, NY: National Professional Resources, Inc.
- Love, N., Stiles, K. E., Mundry, S., & DiRanna, K. (2008). *A data coach's guide to improving learning for all students: Unleashing the power of collaborative inquiry*. Thousand Oaks, CA: Corwin.
- Lozano, L. M., García-Cueto, E., & Muñoz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 4(2), 73-79. doi: 10.1027/1614-2241.4.2.73

- Luca, J., & McLaoughlin, C. (2002). *A question of balance: Using peer and peer assessment effectively in teamwork. In wind sof change in the sea of learning: Charting the course of digital education*. Paper presented at the 19th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education., Auckland, New Zealand.
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3(4), 331-345.
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382-385. doi: 10.1097/00006199-198611000-00017
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84-99.
- Magin, D., & Helmore, P. (2001). Peer and teacher assessments of oral presentations: how reliable are they? . *Studies in Higher Education*, 26, 287-298.
- Mak, B. L., & Sockel, H. (2001). A confirmatory factor analysis of IS employee motivation and retention, *Information & Management*, 38(5), 265-276.
- Marsh, H. W., Hau, K., Balla, J., & Grayson, D. (1998). Is more ever too much? the number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33, 181-220.
- Marsh, H. W., Lüdtke, O., Muthén, B. O., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, 22(3), 471-491. doi: 10.1037/a0019227

- Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological Methods, 18*(3), 257-284. doi: 10.1037/a0032773
- Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85-110.
- Marsh, H. W., Muthén, B. O., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 439-476. doi: 10.1080/10705510903008220
- Marshall, B., & Drummond, M. J. (2006). How teachers engage with assessment for learning: lessons from the classroom. *Research Papers in Education, 21*(2), 133-149.
- Martens, M. P. (2005). The Use of Structural Equation Modeling in Counseling Psychology Research. *The Counseling Psychologist, 33*(3), 269-298. doi: 10.1177/0011000004272260
- Martin, A., & Dowson, M. (2009). Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current issues, and educational practice. *Review of Educational Research, 79*(1), 327-365.
- Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. doi: 10.1007/bf02296272.

- Mayer, R. (2002). Rote versus meaningful learning. *Theory Into Practice*, 41(4), 226.
- McCutcheon, A. C. (1987). *Latent class analysis*. Beverly Hills, CA: Sage.
- McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education: Principles, Policy & Practice*, 10(2), 209-220. doi: 10.1080/0969594032000121289.
- McDonald, R.P. (2005). Semiconfirmatory factor analysis: The example of anxiety and depression. *Structural Equation Modeling*, 12, 163-172
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- McMahon, T. (2010). Peer feedback in an undergraduate programme: Using action research to overcome students' reluctance to criticise. *Educational Action Research*, 18, 273-287.
- McMillan, J. H. (2000). Fundamental assessment principles for teachers and schools administrators. *Practical Assessment, Research & Evaluation*, 7(8).
- McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement, Issues and Practice*, 20(1), 20-32.
- McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice *Educational Measurement: Issues and Practice*(Winter), 34-43.
- McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, 0(0), 1-22.

- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA, and Oxford, England: Blackwell.
- McRobbie, C., & Tobin, K. (1997). A social constructivist perspective on learning environments. *International Journal of Science Education*, 19(2), 193-208. doi: 10.1080/0950069970190205
- Mehan, H. (1998). The study of social interaction in educational settings: Accomplishments and unresolved issues. *Human Development*, 41, 245-269.
- Mertler, C. A., & Campbell, C. (2005). *Measuring teachers' knowledge and applicaiton of classroom assessment concepts: Development of the assessment literacy inventory*. Paper presented at the American Educational Research Association, Montreal, Quebec, Canada.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: American Council on Education and Macmillan.
- Ministry of Education of Brunei (2008). SPN 21 interem guidelines. Brunei: Curriculum Development.
- Ministry of Education of Brunei (2011). School-based assessment for learning: Brunei SBAfL guidebooks for years 7 and 8 core subjects. Brunei: Curriculum Development.
- Mitchell, J. E., Crosby, R. D., Wonderlich, S. A., Hill, L., le Grange, D., Powers, P., & Eddy, K. (2007). Latent profile analysis of a cohort of patients with eating disorders not otherwise specified. *International Journal of Eating Disorders*, 40(S3), S95-S98. doi: 10.1002/eat.20459

- Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, 22(4), 13-25. doi: 10.1111/j.1745-3992.2003.tb00140.x
- Munns, G., & Woodward, H. (2006). Student engagement and student self-assessment: the REAL framework. *Assessment in Education: Principles, Policy & Practice*, 13(2), 193-213. doi: 10.1080/09695940600703969
- Muñiz, J., García-Cueto, E., & Lozano, L. M. (2005). Item format and the psychometric properties of the Eysenck personality questionnaire. *Personality and Individual Differences*, 38(1), 61-69. doi: <http://dx.doi.org/10.1016/j.paid.2004.03.021>
- Muoio, R., Wolcott, L., & Seigel, H. (1995). A win-win situation: The pilot program. *Journal of Continuing Education in Nursing*, 26, 230-233.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 186-176.
- Muthén, B. O. (2000). Mplus Home: Calculating the % variance explain, from <http://www.statmodel.com/discussion/messages/8/46.html?1403633947>
- Muthén, B.O, du Toit, S., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modelling with categorical and continuous outcomes*. Unpublished manuscript.
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- National Curriculum Task Group on Assessment and Teaching (1998). Department of Education and Science and the Welsh Office.

- Newfields, T. (2006). *Teacher development and assessment literacy*. Paper presented at the 5th Annual JALT Pan-SIG Conference, Shizuoka, Japan.
- Nicol, D. J. (1997). *Research on learning and higher education teaching*. Paper presented at the UCoSDA Briefing Paper 45, Sheffield, Universities and Colleges Staff Development Agency.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218. doi: 10.1080/03075070600572090
- Nichols, P. D., Meyers, J. L., & Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice*, 28(3), 14-23.
- Nonaka, I. (1991). The knowledge cretaing-cretaing company. *The Harvard Business Review*(November- December), 96-104.
- Norton, L. S. (1990). Essay writing: What really counts? *Higher Education* 20(4), 411-442.
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*, (3rd ed.). New York:McGraw-Hill.
- Nuthal, G. (1999). The way students learn: Acquiring knowledge from an integrated science and social studies unit. *Elementary School Journal*, 99(4), 303-342.
- Nylund, K., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and groth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, 14(4), 535-569.
- O'Connor, K., & Wormeli, R. (2011). Reporting student learning. Effective grading practices. *Educational Leadership*, 69(3), 40-44.

- O'Donovan, B., Price, M., & Rust, C. (2008). Developing student understanding of assessment standards: a nested hierarchy of approaches. *Teaching in Higher Education*, 13(2), 205-217. doi: 10.1080/13562510801923344
- Osterloh, M., & Frey, B. S. (2000). Motivation, knowledge transfer, and organizational forms. *Organisation Science*, 11(538-550).
- Pang, N. S., & Leung, Z. L. (2011). Teachers' competency in assessment for learning in early childhood education in Hong Kong. *Educational Research Journal*, 26(2), 199-222.
- Park, H. S., Dailey, R., & Lemus, D. (2002). The use of exploratory factor analysis and principal components analysis in communication research. *Human Communication Research*, 28, 562–577.
- Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, 32(1), 8-47.
- Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, 19(2), 109-131. doi: 10.1191/0265532202lt224oa
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis : An integrated approach*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. (2002) Academic emotions in students' self-Regulated learning and achievement: A program of qualitative and quantitative research, *Educational Psychologist*, 37(2), 91-105.

- Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes. Towards a wider conceptual field. *Assessment in Education: Principles, Policy & Practice*, 5(1), 85-102. doi: 10.1080/0969595980050105
- Pintrich, P. R., & Zusho, A. (2002). *The development of academic self-regulation: The role of cognitive and motivational factors*. San Diego, CA: Academic Press.
- Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95(4), 667-686.
- Pintrich, P. R., & Zusho, A. (2002). *The development of academic self-regulation: The role of cognitive and motivational factors*. San Diego, CA: Academic Press.
- 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(5), 489-497. doi: 10.1002/nur.20147
- Pollard, A., Triggs, P., Broadfoot, P., Mcness, E., & Osborn, M. (2002). *What pupils say: Changing policy and practice in primary education*. London: Continuum.
- Popham, W. J. (2008). *Transformative assessment*. United States of America: Association for Supervision and Curriculum Development
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, 48(1), 4-11. doi: 10.1080/00405840802577536
- Popham, W.J. (2011). Assessment literacy overlooked: A teacher educator's confession, *The Teacher Educator*, 46(4), 265-273. doi: 10.1080/08878730.2011.605048

- Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1), 1-15. doi: [http://dx.doi.org/10.1016/S0001-6918\(99\)00050-5](http://dx.doi.org/10.1016/S0001-6918(99)00050-5)
- Price, M., O'Donovan, B., & Rust, C. (2004). Orienting students in the assessment processes through peer review and feedback.
- Pryor, J., & Crossouard, B. (2005). *A sociocultural theorization of formative assessment*. Paper presented at the Sociocultural Theory in Educational and Practice Conference, University of Manchester.
- Pryor, J., & Lubisi, C. (2002). Reconceptualising educational assessment in South Africa-testing times for teachers. *International Journal of Educational Development*, 22, 673-686.
- Queensland Department of Education and the Arts (2005). Assessment and reporting. Queensland Government: Education.
- Quintana, S. M., & Minami, T. (2006). Guidelines for meta-analyses of counselling psychology research. *The Counseling Psychologist*, 34, 836-876.
- Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, 28, 4-13.
- Randall, J., & Engelhard Jr. G., (2010) Using confirmatory factor analysis and the Rasch model to assess measurement invariance in a high stakes reading Assessment, *Applied Measurement in Education*, 23(3), 286-306.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen Neilson & Lydiche.
- Rasch, G. (1980). *Some probabilistic models for the measurement of attainment and intelligence*. Chicago: MESA Press.

- Reay, D., & William, D. (1999). 'I'll be nothing': Structure, agency and the construction of identity through assessment. *British Educational Research Journal*, 25, 343-354.
- Reeves, D. B. (2004). *Assessing educational leaders*. Thousand Oaks, CA: Corwin Press.
- Reeves, D. (2013). From differentiated instruction to differentiated assessment. *ASCD Express*, Vol. 6, No. 20. Copyright by ASCD.
- Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education*, 27(2), 472-482. doi: 10.1016/j.tate.2010.09.017.
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17, 354-373.
- Rhodes, C., & Beneicke, S. (2002). Coaching, mentoring and peer-networking: Challenges for the management of teacher professional development in schools. *Journal of In-Service Education*, 28(2), 297-310
- Rietveld, T., & Van Hout, R. (1993). *Statistical techniques for the study of language and language behaviour*. Berlin – New York: Mouton de Gruyter.
- Rowntree, D. (1977). *Assessing students: how shall we know them?* London: Harper and Row.
- Ruiz-Primo, M. A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies in Educational Evaluation*, 37(1), 15-24.

- Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, 28, 1629–1646.
- Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment and Evaluation*, 28(2), 147-164.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119-144. doi: 10.1007/BF00117714
- Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation on Higher Education*, 30(2), 175-194.
- Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, 34(7), 807-826. doi: 10.1080/03075070802706553
- Sadler, D. R. (2012). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy and Practice*, 1-15.
- Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, 60(4), 549-572. doi: 10.1007/bf02294328
- Samson, J. F., & Collins, B.A. (2012). Preparing all teachers to meet the needs of English language learners: Applying research to policy and practice for teacher effectiveness. *Center for American Progress*.
- Sanders, W. L., & Rivers, J. C. (1996). Cumulative and residual effects of teachers on future student academic achievement. Tennessee Value-Added Assessment System: Maryland.

- School-Based Assessment for Learning Guidebook. (2011). *School based assessment for learning* Brunei Darussalam: Ministry of Education
- Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *Educational Research Journal*, 33(2), 359-382.
- Schunk, D. H., & Zimmerman, B. J. (1994). *Self-regulation of learning and performance: Issues and educational applications*. New Jersey: Lawrence Erlbaum Associates.
- Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(461-464).
- Schlager, M. S., & Fusco, J. (2004). Teacher professional development, technology, and communities of practice: Are we putting the cart before the horse? In S. Barab, R. Klin & J. Gray (Eds.), *Designing for virtual communities in the service of learning*. Cambridge MA: Cambridge University Press.
- Sclove, S.L. (1987), Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343.
- Scott-Parker, B., Watson, B., King, M. J., & Hyde, M.K. (2012). Confirmatory factor analysis of the behaviour of young novice driver scale (BYNDS). *Accident Analysis & Prevention*, 49, 385-391.
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*, 39-83. Chicago, IL: Rand McNally.
- Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.

- Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research in teaching* (4th ed., pp. 1066-1101). Washington, DC: American Educational Research Association.
- Shepard, L. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623-646). Westport, CT: American Council on Education/Praeger.
- Sick, J. (2011). Rasch measurement and factor analysis. *SHIKEN: JALT Testing and Evaluation SIG Newsletter*, 15(1), 15-17.
- Siegert, R. J., Jackson, D. M., Tennant, A., & Turner-Stokes, L. (2010). Factor analysis and Rasch analysis of the Zarit burden interview for acquired brain injury carer research. *Journal of Rehabilitation Medicine*, 42, 302-309.
- Singapore Ministry of Education (2008). Singapore Assessment and Examination Board.
- Simpson, M., & Goulder, J. (1997). Implementing the 5-14 development programme in secondary schools: Continuity and progression: Aberdeen, Northern College.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237-247. doi: 10.1111/j.1745-3984.1991.tb00356.
- Slavec, A., & Drnovsek, M. (2012). A perspective on scale development in entrepreneurship research. *Economic and business review*, 14(1), 39-62.
- Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, 22(4), 26-33. doi: 10.1111/j.1745-3992.2003.tb00141.

- Smith, E., & Gorard, S. (2005) The dissemination of formative assessment: A short response to Black, Harrison, Hodgen, Marshall and Wiliam. *Research Intelligence*, 4, 466-478.
- Snook, I., O'Neill, J., O'Neill, A., & Openshaw, R. (2009). Invisible learning? A commentary on John Hattie's book: Visible learning: A synthesis of over 800 meta-analyses relating to achievement.
- Spanierman, L. B., Oh, E., Heppner, P. P., Neville, H. A., Mobley, M., Wright, C. V., . . . Navarro, R. (2010). The Multicultural Teaching Competency Scale: Development and Initial Validation. *Urban Education*. doi: 10.1177/0042085910377442
- Stefanou, C., & Parkes, J. (2003). Effects of classroom assessment on student motivation in fifth-grade science. *The Journal of Educational Research*, 96(3), 152-162.
- Steiger, J. H., & Lind, J. (1980). *Statistically based tests for the number of common factors*. Paper presented at the Annual Meeting of Psychonomic Society, Iowa City, IA.
- Stiggins, R. J. (1997). *Student centered classroom assessment*. Upper Saddle River, NJ: Prentice Hall.
- Stiggins, R. J. (1999a). Are you assessment literate? *The High School Journal*, 6(5), 20-23.
- Stiggins, R. J. (1999b). Evaluating lassroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18(1), 23-27. doi: 10.1111/j.1745-3992.1999.tb00004.x
- Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Kappan Professional Journal*, 1-10.
- Stiggins, R. J. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *Phi Delta Kappan*, 87(4), 324-328.

- Stiggins, R. J., Arter, J., Chappuis, J., & Chappuis, S. (2007). *Classroom assessment for student learning: Doing it right-using it well*. Upper Saddle River, New Jersey: Pearson Education, Inc.
- Stiggins, R. J., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory into Practice*, 44(1), 11-18.
- Stiggins, R. J., & Duke, D. (2008). Effective instructional leadership requires assessment leadership. *Phi Delta Kappa*, 90(4), 285-291.
- Stoynoff, S., & Chapelle, C. (2005). *ESOL test and testing: A resource for teachers and program administrators*. Alexandria, VA: TESOL.
- Struyven, K., Dochy, F., Janssens, S., Schelfhou, W., & Giele, S. (2006). The overall effects of end-of-course assessment on student performance: A comparison between multiple choice testing, peer assessment, case-based assessment and portfolio assessment. *Studies in Educational Evaluation*, 32(3), 202-222.
- Sutton, R. (1995) *Assessment for learning*. Salford, RS Publications.
- Swanson, D., Case, S., & Van der Vleuten, C. (1991). Strategies for student assessment. In D. Boud & G. Feletti (Eds.), *The challenge of problem based learning* London: Kogan Page.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson Education Inc.

- Tan, K. (2011). Assessment for learning in singapore: Unpacking its meanings and identifying areas for improvement. *Educational Research Policy Practice*, 10, 91-103.
- Tang, K. L. (1996). *Polytomous item response models and their application in large-scale testing-programs: Review of literature*. TOEFL Monograph. Educational Testing Service, (RM-96-8). Princeton, New Jersey.
- Taras, M. (2003). To feedback or not to feedback in student self-assessment. *Assessment and Evaluation in Higher Education*, 28(5), 549-565.
- Taras, M. (2005). Assessment-summative and formative- some thoeretical refelctions. *British Journal of Educational Studies*, 53(4), 466-478. doi: 10.1111/j.1467-8527.2005.00307.x
- Taras, M. (2007). Terminal terminology: The language of assessment. In M. Reiss, R. Hayes, and A. Atkinson (Eds.), *Marginality and difference in education and beyond* (pp. 52-67). Stoke-on-Trent: Trentham Books.
- Taras, M. (2009). Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education*, 33(1), 57-69.
- Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21-36.
- Taylor, P., Mills, A., & O'Driscoll, M. (1993). Personnel selection methods used by New Zealand organisations and personnel consulting firms. *New Zealand Journal of Psychology*, 22, 19-31.

- Teaching Standards in Asia-Pacific Region (2008). Scoping study on the development of teaching standards in the broader Asia-Pacific region. Report to the Department of Education and Workplace Relations. Erebus International.
- Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning, *Journal of Curriculum Studies*, 43 (3) 425-438.
- Thompson, M., & Wiliam, D. (2007). *Tight but loose: A conceptual framework for scaling up school reforms*. Paper presented at the American Educational Research Association, Chicago, IL.
- Thomson, R. (2012). Implementation of criteria and standards-based assessment: an analysis of first-year learning guides. *Higher Education Research & Development*, 32(2), 272-286. doi: 10.1080/07294360.2012.676026
- Thorndike, E. L. (1918). The nature, purposes, and general methods of measurement of educational products. In S. A. Courtis (Ed.), *The measurement of educational products (17th Yearbook of the National Society for the Study of Education, Pt. 2)* (pp. 16-24). Bloomington, IL: Public School.
- Timperley, H. 2008, Teacher professional learning and development, *Educational Practices*, Series-18, International Bureau of Education, UNESCO
- Timperley, H. (2011). Using student assessment for professional learning: Focusing on students' outcomes to identify teachers' needs Victoria, Australia: Department of Education and Early Childhood Development

- Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2008). Teacher professional learning and development: Best evidence synthesis on professional learning and development, Ministry of Education, Wellington.
- Tomlinson, C. A., & McTighe, J. (2006). *Integrating differentiated instruction + understanding by design*. Association for Supervision and Curriculum Development: USA.
- Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education: Principles, Policy & Practice*, 14(3), 281-294.
- Torrie, M. C., & Van Buren, J. B. (2008). Student and program assessment: Assessment literacy, the basis for student assessment. *Journal of Family Consumer Sciences Education*, 26, 320-333.
- Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17(7), 783-805. doi: [http://dx.doi.org/10.1016/S0742-051X\(01\)00036-1](http://dx.doi.org/10.1016/S0742-051X(01)00036-1)
- Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10.
- van de Mortel, T.F. (2008). Faking it: Social desirability response bias in self-report research. *Australian Journal of Advanced Nursing*, 25(4), 40-48.
- van Teijlingen, E. R., & Hundley, V. (2001). The importance of pilot testing. *Social Research Update*, Winter Issue.

- Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*(3), 231-251.
- Vogt, F., & Rogalla, M. (2009). Developing Adaptive Teaching Competency through coaching. *Teaching and Teacher Education*, 25(8), 1051-1060. doi: <http://dx.doi.org/10.1016/j.tate.2009.04.002>
- 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.
- Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57, 307333.
- Vygotsky, L. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.
- Vygotsky, L. (1986). *Thought and language*. Boston: MIT Press.
- Walker, M. (2015). The quality of written peer feedback on undergraduates' draft answers to an assignment, and the use made of the feedback, *Assessment & Evaluation in Higher Education*, 40(2), 232-247.
- Wang, X., Kammerer, C. M., Anderson, S., Lu, J., & Feingold, E. (2009). A comparison of principle component analysis and factor analysis strategies for uncovering pleiotropic factors. *Genetic Epidemiology*, 33(4), 325–331.
- Webb, N. (2002). *Assessment literacy in a standard-based urban education setting*. Paper presented at the American Educational Research Association Annual Meeting, New Orleans.

- Weinert, F. E., & Helmke, A. (1998). The neglected role of individual differences in theoretical models of cognitive development. *Learning and Instruction*, 8(4), 309-323. doi: [http://dx.doi.org/10.1016/S0959-4752\(97\)00024-8](http://dx.doi.org/10.1016/S0959-4752(97)00024-8)
- West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues and Applications* (pp. 56-75). Thousand Oak: Sage.
- Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177–203). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wiliam, D. (2011). *Formative assessment: definitions and relationships*. London: Institute of Education, University of London.
- Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3–14.
- Wiliam, D., & Thompson, M. (2008). Integrating assessment with learning: What will it take to make it work? In C.A. Dwyer (Ed.), *The Future of Assessment: Shaping Teaching and Learning*, (pp. 53–82). New York: Erlbaum.
- Williams, B., Brown, T. & Osman, A. (2010). Exploratory factor anlaysis: A five-step guide for nivices, *Australasian Journal of Paramdedivine*, 8(3).
- Williams, K., & Williams, C. (2011). Five key ingredients for improving student motivation. *Research in Higher Education Journal*, 23, 1-23.

- Willey, K., & Gardner, A. (2010) Investigating the capacity of self and peer assessment activities to engage students and promote learning, *European Journal of Engineering Education*, 35(4), 429-443,
- Willis, J. (2009) Assessment for learning: A sociocultural approach. In: Proceedings of Changing Climates: Education for Sustainable Futures, 30 November - 4 December 2009, Australia, Queensland, Kelvin Grove.
- Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching*, 7(1), 3-14.
- Worthington, R. L., & Wittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, 34, 806-838.
- Wright, B. D. (1996. Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(1), 3-24.
- Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Educational Measurement Solutions. Melbourne.
- Yang, C. (2006). Evaluating latent class analysis in qualitative phenotype identification. *Computational Statistics & Data Analysis*, 50, 1090-1104.
- Yates, A. (1987). Multivariate exploratory data analysis: A perspective on exploratory factor analysis. Albany: State University of New York Press.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187-213. doi: 10.1111/j.1745-3984.1993.tb00423.x

- Yu, C.-Y. (2002). Evaluating cut-off criteria of model fit indices for latent variable models with binary and continuous outcomes (*Unpublished doctoral dissertation*).
- Yu, W. M., Kennedy, K. J., Fok, P. K., & Chan, K. S. (2006). *Assessment reform in basic education in Hong Kong: The emergence of assessment for learning*. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore.
- Yu, C.-Y., & Muthén, B. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Paper presented at the annual meeting of the *American Educational Research Association*, New Orleans, LA.
- Yung, H.-W. (2002). Same assessment, different practice: Professional consciousness as a determinant of teachers' practice in a school-based assessment scheme. *Assessment in Education*, 9(1), 97-116.

APPENDIXES

Appendix A: Support Letter from the Ministry of Education of Brunei

Fax : 2380101/2382398/2380050
 Telex : MIEDU BU 2602
 Cable : MIEDU BRUNEI
 Telephone : 2381133



كمنترين قنديديقتن

KEMENTERIAN PENDIDIKAN
 MINISTRY OF EDUCATION
 BANDAR SERI BEGAWAN BB3510
 NEGARA BRUNEI DARUSSALAM

KPE/TSUT(CE)/1/9.4

01 August 2012

Research Ethics Committee
 University of New South Wales
 Kensington, NSW
 Australia

This is to certify that the Brunei Darussalam Ministry of Education (MoE) fully supports the research study being undertaken by Mr Dennis Alonzo of the School of Education, University of New South Wales (UNSW), as part of the School Based Assessment for Learning Phase Two contract being undertaken for the MoE by Optimise International (B) Sdn Bhd (OIB).

Mr Alonzo's research study, entitled: **A Teachers' Assessment for Learning Competency Framework: Development and Application**, is a priority activity of the Ministry.

Research output will support and enhance implementation of the Ministry's School-Based Assessment for Learning (SBAfL) policy, a key element of Sistem Pendidikan Negara Abad ke-21 (SPN21), Brunei's new education system for the 21st century.

The research protocol has been reviewed for the Ministry of Education by the Ministry's School-Based Assessment for Learning consultants and advisers, Optimise International (B) Sdn Bhd (OIB), who have confirmed the protocol conforms to accepted international standards as reflected in UNSW's Human Research Ethics Committees (HRECs) Operations Manual.

To ensure continuous compliance with these standards, the MoE and OIB will jointly and directly supervise each stage of the entire research process and will ensure any ethical issues that may arise will be dealt with in accordance with these standards.


{HAJI SUHAILA BIN HAJI ABD KARIM}
 Deputy Permanent Secretary (CE)
 Ministry of Education
 Brunei Darussalam

cc : Optimise International (B) Sdn Bhd (OIB), Brunei Darussalam

Confirmed as above


S.B. bin Ibrahim
CEO and Chief Consultant
Optimise International
(B) Sdn Bhd (OIB)
Brunei Darussalam
28 August 2012

Appendix B: Ethics Approval from HREA Panel B of UNSW Australia

Human Research Ethics Advisory Panel B Arts, Humanities & Law

Date: 13.11.2012

Investigators: Mr Dennis Alonzo

Supervisors: Professor Chris Davison, Dr John Bennett

School: School of Education

Re: A Teachers' Assessment for Learning Competency Framework:
Development and Application

Reference Number: 12 096

The Human Research Ethics Advisory Panel B for the Arts, Humanities & Law is satisfied that this project is of minimal ethical impact and meets the requirements as set out in the National Statement on Ethical Conduct in Human Research*. Having taken into account the advice of the Panel, the Deputy Vice-Chancellor (Research) has approved the project to proceed.

Your Head of School/Unit/Centre will be informed of this decision.

This approval is valid for 12 months from the date stated above.

Yours sincerely



Associate Professor Anne Cossins
Panel Convenor
Human Research Ethics Advisory Panel B

Cc: Professor Chris Davison
Head of School
School of Education

* <http://www.nhmrc.gov.au/>

Appendix C. Participant Information Sheet

11 July 2013



School of Education
Approval No. 13 065

Dear Participant:

You are invited to participate in a study on the establishment of teachers' practices in making highly contextualised, consistent and trustworthy decisions to better support student learning. I, Dennis Alonzo, a PhD student of School of Education, University of New South Wales, is conducting a research entitled: "A teachers' assessment for learning competency framework: Development and application" with Ethics approval number 13 065 from the UNSW Human Research Ethics (HREA) Panel B.

The assessment instrument measures your assessment competency. The underlying principles in the abilities described in the instrument are based on your effectiveness in using assessment to make important learning and teaching decisions to help individual learners achieve sustained high performance.

Privacy. No personal details such as names and emails will be gathered in the survey.

Risks. Completing the survey is strictly voluntary and there is no risk of physical or emotional harm. Your decision whether or not to participate will not prejudice your future relations with UNSW. If you do participate, you can withdraw your participation at any stage without any consequences and you may refuse to answer any questions on the survey.

Confidentiality and disclosure of information. Any information that is obtained in connection with this study and can be identified with you will remain confidential and will be disclosed only with your permission, except as required by law. If you give me your permission by signing this document, I plan to publish the results in the forms of journal articles and/ or thesis. In any publication, information will be provided in such a way that you cannot be identified.

Results. A report on the findings of this study will be publicly available to the UNSW community. Aggregated results may also be reported in publications or conference proceedings.

Consent. You are making a decision whether or not to participate. Your consent is confirmed by the completion and return of this survey under the National Statement on Ethical Conduct in Research Guidelines, section 1.9.

If you have queries regarding the research or questionnaire. please contact me, Dennis Alonzo, *Tel. (61) 406420568 (in Australia) (+639) 298147609 (in Philippines)*, email: d.alonzo@unsw.edu.au, and I will be happy to answer them.

Complaints may be directed to the Ethics Secretariat, The University of New South Wales, SYDNEY 2052 AUSTRALIA (phone 9385 4234, fax 9385 6648, email ethics.sec@unsw.edu.au). Any complaint you make will be investigated promptly and you will be informed of the outcome.

Appendix D: Certification from Optimise International (B) Sdn Bhd (OIB)



Tuesday 8 January 2013

To Whom It May Concern

This document confirms Mr Dennis Alonzo, a researcher and PhD candidate at the School of Education, University of New South Wales, Australia, has been engaged by Optimise International (B) Sdn Bhd (OIB) to work with the company through 2012 and 2013 as a senior researcher/consultant in OIB's work with the Brunei Darussalam Ministry of Education.

A handwritten signature in black ink, appearing to read 'Brian Galbraith', enclosed within a circular stamp or seal.

Brian Galbraith
Chief Executive Officer and Coordinating Consultant

Whole system reform and transformation specialists – education, government and community

PO Box 522 MPC Old Airport Berakas BB 3577 BSB Negara Brunei Darussalam +6738935878 optimisebrunei@gmail.com www.optimise.net.au

Appendix E: Workshop Schedule

Brunei Darussalam
MINISTRY OF EDUCATION
Old Airport Road, Berakas BB3510



A Teachers' Assessment for Learning Competency Framework: Development and Application

Establishment of Domains, Capabilities, Indicators and Success Criteria
(First Workshop, 3-5 September 2012)

Schedule of Activities

Participants:

1. MoE Senior Officers
 - Permanent Secretary
 - Deputy Permanent Secretary Core Education
 - Director General of Education
 - Core education department heads and assistant heads (Directors/Acting Directors and Assistant Directors/Acting Assistant Directors)
 - Core education department unit heads
2. 30 teachers (Considered as Experts in AfL)
3. 15 Principals/ Deputy Principals/ Department Heads

Time	Activities	Expected Output
Day 1 8:00-8:30	Preliminary Activities Welcome Remarks <i>Brian Galbraith</i> <i>Kalthom Ahmad</i>	
8:30 - 9:30	Orientation: What is the research all about and how will it help the MoE of Brunei? <i>Dennis Alonzo</i> Together with the participants are: - Permanent Secretary - Deputy Permanent Secretary Core Education - Director General of Education - Core education department heads and assistant heads (Directors/Acting Directors and Assistant Directors/Acting Assistant Directors) - Core education department unit heads	MoE Senior Officers and Participants are clear about the aims of the study

9:30 – 10:30	Open Forum Discussion with Brunei MoE Senior Officers and other participants	Issues are clarified and expectations are leveled off
10:30 – 11: 00	Tea Break	
11:00 – 12:00	Workshop 1: Identifying the Indicators of Teachers' AfL Competence	List of indicators of AfL
12:00 – 2:00	LUNCH BREAK	
2:00 – 3:30	Workshop 2: Establishing AfL Capabilities of Teachers: Categorisation of indicators based on similarities	List of domains and their corresponding job functions.
3:30 – 4:00	Presentation of Output	
Day 2 8:00 – 12:00	Workshop 3: Establishing the Success Criteria	Descriptions of success criteria following Dreyfus model of competence
12:00 – 2:00	Lunch	
2:00 – 3:30	Continuation of Workshop 3	Descriptions of success criteria following Dreyfus model of competence
3:30 – 4:30	Presentation of Output and Giving of Feedback <i>Dennis Alonzo</i> <i>Brian Galbraith</i> <i>Kalthom Ahmad</i>	
Day 3 8:00 – 9:00	Presentation of the Revised Output	
9:00 – 12:00	Review and Finalisation of the Scale	Questionnaire ready for Expert Validation
12:00 – 2:00	Lunch	
2:00 – 3:30	Debriefing: Sharing of Experiences/ Insights/ Learning from the workshop.	Teachers' and principals' perception on the importance of their participation in the research process
3:30 – 4:00	Wrap up and closing	

Appendix F: Syntax for Exploratory Factor Analysis

```
TITLE: AfL EFA Set 1
DATA: FILE IS AfL EFA_Set1.dat;
VARIABLE: NAMES ARE v1-v72;
MISSING ARE all (9);
CATEGORICAL ARE V1-V72
ANALYSIS: TYPE = EFA 1 20;
ESTIMATOR = wlsmv;
OUTPUT:
sampstat;
```

Appendix G: Syntax for Confirmatory Factor Analysis

```

TITLE: AfL CFA Set 2 Data
DATA: FILE IS AfL_CFA_Set_2.dat;
VARIABLE: NAMES ARE v3 v4 v5 v7 v8 v9 v10 v12 v22 v23 v25 v26
v28 v29 v30 v31 v32 v33 v34 v16 v17 v18 v19 v27 v35 v37 v39 v40 v41 v42
v43 v38 v46 v47 v48 v49 v50 v51 v53 v54 v55 v58 v59 v60 v62 v63 v64
v66 v67 v68 v70 v71;
MISSING ARE all (9);
CATEGORICAL ARE v3 v4 v5 v7 v8 v9 v10 v12 v22 v23 v25 v26
v28 v29 v30 v31 v32 v33 v34 v16 v17 v18 v19 v27 v35 v37 v39 v40 v41 v42
v43 v38 v46 v47 v48 v49 v50 v51 v53 v54 v55 v58 v59 v60 v62 v63 v64
v66 v67 v68 v70 v71;
ANALYSIS:
MODEL: f1 BY v3 v4 v5 v7 v8 v9 v10 v12;
      f2 BY v22 v23 v25 v26 v28 v29 v30 v31 v32 v33 v34;
      f3 BY v16 v17 v18 v19 v27 v35 v37 v39 v40 v41 v42 v43;
      f4 BY v38 v46 v47 v48 v49 v50 v51 v53 v54 v55;
      f5 BY v58 v59 v60 v62 v63 v64;
      f6 BY v66 v67 v68 v70 v71;
OUTPUT: MODINDICES
      STDYX
      RESIDUAL
      TECH4

```

Appendix H: Syntax for Second-order Factor Analysis

```

TITLE: AfL Higher Order CFA
DATA: FILE IS AfL_CFA_Set_2.dat;
VARIABLE: NAMES are NAMES are v1-v72;
USEVARIABLES are v3 v7 v9 v10 v12
v17 v19 v25 v27 v28 v29 v30 v31 v32
v33 v34 v35 v37 v39 v40 v41 v42 v43 v46 v47 v48 v49 v50
v51 v53 v54 v55 v58 v59 v60 v62 v63 v64 v66 v67 v68 v70;

MISSING ARE all (9);
CATEGORICAL ARE v3 v7 v9 v10 v12
v17 v19 v25 v27 v28 v29 v30 v31 v32
v33 v34 v35 v37 v39 v40 v41 v42 v43 v46 v47 v48 v49 v50
v51 v53 v54 v55 v58 v59 v60 v62 v63 v64 v66 v67 v68 v70;

ANALYSIS:
MODEL: f1 BY v3 v7 v9 v10 v12;
      f2 BY v25 v28 v29 v30 v31 v32 v33 v34;
      f3 BY v17 v19 v27 v35 v37 v39 v40 v41 v42 v43;
      f4 BY v46 v47 v48 v49 v50 v51 v53 v54 v55;
      f5 BY v58 v59 v60 v62 v63 v64;
      f6 BY v66 v67 v68 v70;
      f7 BY f1-f6;

OUTPUT: MODINDICES
      STDYX
      RESIDUAL
      FSDETERMINACY
      TECH4

```

Appendix I: Syntax for Exploratory Structural Equation Modelling

TITLE: AfL ESEM

DATA: FILE IS AfL_CFA_Set_2.dat;

VARIABLE: NAMES are v1-v72;

USEVARIABLES are v3 v7 v9 v10 v12

v17 v19 v25 v27 v28 v29 v30 v31 v32

v33 v34 v35 v37 v39 v40 v41 v42 v43 v46 v47 v48 v49 v50

v51 v53 v54 v55 v58 v59 v60 v62 v63 v64 v66 v67 v68 v70;

MISSING ARE all (9);

CATEGORICAL ARE v3 v7 v9 v10 v12

v17 v19 v25 v27 v28 v29 v30 v31 v32

v33 v34 v35 v37 v39 v40 v41 v42 v43 v46 v47 v48 v49 v50

v51 v53 v54 v55 v58 v59 v60 v62 v63 v64 v66 v67 v68 v70;

ANALYSIS:

ESTIMATOR = WLSMV;

MODEL:

f1-f6 BY v3 v7 v9 v10 v12

v17 v19 v25 v27 v28 v29 v30 v31 v32

v33 v34 v35 v37 v39 v40 v41 v42 v43 v46 v47 v48 v49 v50

v51 v53 v54 v55 v58 v59 v60 v62 v63 v64 v66 v67 v68 v70 (*1);

OUTPUT: TECH1; STDYX; RESIDUAL; TECH4; sampstat;

Appendix J: Syntax for Latent Profile Analysis

```
TITLE: AfL LPA
DATA: FILE IS AfL_Data_LCA.dat;
VARIABLE: NAMES are v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11;
categorical are v1 v2 v3 v4 v5;
MISSING ARE all (9);
classes = c(5)
ANALYSIS:
    TYPE=MIXTURE;
    LRTSTARTS = 0 0 500 200;
PLOT: Type is plot3;
    series is v6 (1) v7 (2) v8 (3) v9 (4) v10 (5) v11 (6);
OUTPUT:
    tech1 tech11 tech14 tech8;
```