Introducing Dynamic Testing to Teachers in Malaysia: An Experimental Investigation of Its Effects on Teachers' Beliefs and Practices about Assessment

RUSILAH YUSUP

Introducing Dynamic Testing to Teachers in Malaysia: An Experimental Investigation of Its Effects on Teachers' Beliefs and Practices about Assessment

Rusilah Yusup

Thesis submitted for the degree of Doctor of Philosophy
School of Education
Durham University

August 31, 2020

Abstract

Assessment is central to the effectiveness of teaching and improvement of learning. In the context of Malaysia, however, it appears that assessment has not fulfilled its promises, as the increase of low-performing schools and the urban-rural performance gap remains prevalent. The concern is, why does assessment "fail" to bring the intended positive impacts on instructional improvements to foster progress in students' learning experience? Research presented in this thesis argues that it is potentially caused by two factors: (i) teachers' lack of understanding of assessment; and (ii) the unsuitability of the currently used assessment tools. Specifically, this study aims to investigate the attitude of teachers towards the implementation of the existing assessment tool, i.e., the Form 1 Diagnostic Test (F1DT), particularly looking at their assessment beliefs and practices. In addition, critically reflecting upon the limitations of FIDT, this study intends to introduce an alternative assessment approach, i.e., dynamic testing (DT). Deploying an intervention-control group and pre-test-post-test experimental design, the answers to the formulated eight research questions were obtained through a self-developed questionnaire, the Survey of Educational Assessment (SEA), and teachers' written feedback. Due to the nested structure of the data, sampled from 862 teachers from six educational zones, the analysis of the questionnaire responses was largely conducted using Hierarchical Linear Modelling (HLM). A thematic analysis was used to analyse the data collected from teachers' written comments. The findings revealed that teachers still viewed F1DT as a useful diagnostic tool particularly in measuring prior students' attainment and identifying learning problems. The relationship between teachers' beliefs and practices regarding the implementation and utilisation of F1DT was found to be strong. However, after attending the educational workshop on DT, teachers indicated a lower level of agreement regarding their beliefs and practices about the purposes and uses of F1DT. Accordingly, this positive appraisal for DT implies that teachers became more aware of the limitations of the information provided by F1DT, especially for the purposes intended (e.g., identifying causes for unsatisfactory academic performance resulting in potentially ineffective instruction). As this is a pioneering study, in terms of its large scale of sample and the employment of an experimental design, it offers novel insights to the field of assessment beliefs and practices and DT application. It therefore has the potential to make a significant contribution in improving professional practices in assessment-related activities and ultimately, in addressing the developmental challenges of the education system in Malaysia.

Table of Content

List of Tables

List of Figures

List of Abbreviations

ACIL : Adaptive Computer-Assisted Intelligence Learning Test Battery

ADAFI : Adaptive Sequential Figure Learning Test

AfL : Assessment for Learning

AoL : Assessment of Learning

CFA : Confirmatory Factor Analysis

DT : Dynamic Testing

EFA : Explanatory Factor Analysis

FA : Factor Analysis

FA : Formative Assessment

FIDT : Form 1 Diagnostic test

GPA : Grade Point Average

GTP : Government Transformation Programme

KMO : Kaiser-Meyer-Olkin

LT : Learning Test

MEB : Malaysia Education Blueprint

MES : Malaysia Education Syndicate

MOE : Ministry of Education

NEP : National Education Philosophy

OECD : Organisation for Economic Cooperation and Development

PoA : Perception of Assessment

PCA : Principal Component Analysis

PISA : Programme for International Student Assessment

ProA : Practice of Assessment

PS1 : Pilot Study 1

PS2 : Pilot Study 2

PT3 : *Pentaksiran Tingkatan 3* or the Form 3 Assessment

SA : Summative Assessment

SBA : School-based Assessment

SD : Standard Deviation

SEA : Survey of Educational Assessment

SEP : Special Education Programme

SPM : *Sijil Pelajaran Malaysia* or the Malaysian Certificate of Education

SPMV : *Sijil Pelajaran Malaysia* or the Malaysian Certificate of Education (Vocational)

SQEM : Standard for Quality Education in Malaysia

T-APrI : Teacher Assessment Practices Inventory

TCoA : Teacher Conception of Assessment

TIMSS : Trends in International Mathematics and Science Study

UPSR : *Ujian Penilaian Sekolah Rendah* or the Primary School Achievement Test

ZPD : Zone of Proximal Development

Declaration

I declare that this thesis is my own work. No material contained in this thesis has previously been submitted for a degree in this or any other institution.

Statement of Copyright

Acknowledgments

ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ

First and foremost, praises to Allah, the Most Gracious and the Most Merciful, for His Grace and Blessings to have made this journey possible.

I would like to express my heartiest gratitude to many people for their encouragement and contribution to the completion of this thesis. I am particularly thankful to my supervisor, Professor Jens F. Beckman for his patient guidance, continuous support and useful suggestions during the planning and development of this research. Though at times I felt over-challenged by his constructive critiques, I am so grateful as it enabled me to develop a better understanding of this research and to grow as an independent novice scholar. My appreciation is also extended to my secondary supervisor, Associate Professor Nadin Beckman whose encouraging support and expert advice on the application of HLM helped me to finish the writing up of this thesis.

I also owe my deepest gratitude to the Ministry of Education Malaysia for granting me study leave and scholarship in my pursuit of academic development. Also, I wish to thank Sabah Education Department and District Education Offices in facilitating logistic supports for my fieldwork and to Ustinov College for funding this research project through the Norman Richardson Postgraduate Research Fund. Most of all, I am forever indebted to teachers who participated in this project. I appreciate their willingness to spare time in answering the questionnaires and attending the educational workshop on dynamic testing.

My sincere gratitude also goes to my PhD comrades (especially Lan, Miyuki, Fieka, Esme, Marcel, Misa, Judith, Prathibha, Nadia, Linda, Dennis, Em, Opeti and Mattia) and my close friends (especially Fido, Pijan, Memet, Ziah, Shima, Farah, Lyney, Kak Nancy and Kak Ipa) for their tireless encouragement and moral support throughout this challenging journey. Finally, I thank my beloved family for their endless love and unwavering emotional support. To my late mother and father, this thesis is my special gift for you.

Thank you for having faith in me.

Dedication

To my late parents

Yusup bin Daud & Janidah binti Mujin

The fountain of my courage to finish this thesis.

To my brother

Mursid W Yusup Daud

The reason of what I become today.

1 Introduction

> Education in Malaysia is an ongoing effort towards further *developing the potential of individuals* [emphasis added] in a holistic and integrated manner, so as to produce individuals who are intellectually, spiritually, emotionally and physically balanced and harmonious, based on a firm belief in and devotion to God. Such an effort is designed to produce Malaysian citizens who are knowledgeable and competent, who possess high moral standards, and who are responsible and capable of achieving a high level of personal well-being as well as being able to contribute to the harmony and betterment of the family, the society and the nation at large. (Ministry of Education Malaysia, 2001, p. ix)

In Malaysia, the main objective of education is enshrined in the *Falsafah Pendidikan Negara* (the National Education Philosophy or NEP), which serves as guidance and direction for all educational matters in the country. The core aim of education, as stated above, outlines that the Malaysian education system is committed to developing individuals' potential holistically, in a balanced and harmonious manner to improve personal growth and to contribute to the betterment of the community and the nation.

Assessment is an important part of the education process, as it helps to ensure that the goals of education are met (Black & Wiliam, 1998a; Gordon, 2012; Travis, 1996). Research evidence indicates that effective assessments can have a powerful impact on students learning and ultimately leads to the improvement of education (Black & Wiliam, 1998b; Guskey, 2003; The Assessment Reform Group, 1999). This means that assessment has the potential to enable teachers, students, and parents to monitor students' learning and development and to guide policy makers in informed decision making when devising educational policies. Also, assessment provides teachers with a useful framework to reflect and to ask questions such as: Are we teaching what we are supposed to be teaching? Are students learning what they are supposed to be learning? Essentially, why and how teachers engage in assessment can have a powerful influence on their students' educational experience and to affect how and what they learn.

Questions such as "Is what we are doing aligned with the NEP's holistic vision to develop the potential of our students?", "Are we truly using assessments to improve the teaching and learning process?", "Are we utilising assessment information to scaffold students to

reach their potential?" are the starting point of this research project. These questions revolve around the issue of a sufficient utilisation of the potential that educational assessment promises. To address this issue and the subsequently derived research questions, the focus is on two key elements in this process: the user (i.e., the teachers) and the tool itself (i.e., the tests).

This introductory chapter aims to offer background information about the underlying rationales of why this study is particularly relevant in the context of Malaysia. It starts with a brief description of several challenges the Malaysian education system faces. This serves as the foundation for the formulation and the structuring of the research problem focussing on the specific issues around the utilisation of assessment information. Addressing a research problem related to the utilisation of assessment information requires the consideration of both the teachers (as users of such information) and the assessment tools (as the source of such information). The final section of this chapter summaries the objectives of the study, which are then translated into the research questions.

1.1 Challenges in the Malaysian Education System

Scholars have consistently advocated that assessment is an integral part of effective teaching and learning, and thus assessment data should be intentionally utilised to enhance teaching and to improve student achievement (Black & Wiliam, 1998a; Gordon, 2012; Hayward, 2015; Wiliam, 2011). This "idealistic" notion, however, remains debatable in practice. This is in agreement with the claim made by Black and Wiliam (2010), who acknowledged the problems and shortcomings of assessment practice despite a wealth of research evidence documenting its usefulness.

Educational practitioners often express concerns that the Malaysian education system is too exam-oriented (e.g., Abdul Wahid, Abdul Hamid, Low, & Mohd Ashhari, 2011; Hashim, Freddy, & Rosmatunisah, 2012; Mohamad Ali & Talib, 2013; Ong, 2010; Suhaili, 2014). The former Minister of Education, Tan Sri Muhyiddin Hj. Mohd Yassin, admitted that:

> For some reasons, Malaysia's education system has veered towards getting a string of A's and not much else. Parents were also swept by this wave and we do not blame them for reasons they themselves are not in control of. Getting scholarships was one of them, Government and private sector sponsor normally attach a significant number of A's as the main criteria. The effect of this then trickled down to how school principals and teachers managed their teaching, learning and timetabling… In short, the system became too exam-oriented. Ironically, this was against the spirit of the national education philosophy which stresses on a holistic mental, physical, emotional and spiritual well-being and development.
>
> (The Sports Digest, 2013)

The system is seen as exam-oriented in the sense that it puts too much emphasis on student performance, in national public examinations as well as international large-scale assessments, as the essence of quality education (Mohamad Ali & Talib, 2013; Ong, 2010). For instance, under the Government Transformation Programme[1] (GTP), the dominance of student academic performance is evidently documented in government educational policies, which are linked to examination-stringent evaluation. The Ministry of Education (MOE) has even introduced several monetary incentives for schools and teachers for impressive performances in public examinations (Jabatan Perdana Menteri, 2010). Moreover, the school band system[2] initiated by the MOE accounts for 70 percent of the evaluation criteria based on student achievement in high-stakes examinations (Jemaah Nazir dan Jaminan Kualiti, 2017). At the international level, student performance is measured by the outcomes of large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). It is the aspiration of the MOE to put Malaysia among the top third of high-performing countries by 2025 (Ministry of Education Malaysia, 2013).

The question arises whether enacting such policies has really worked. Should outstanding performance in standardised tests be interpreted as an indication of better teachers and schools or better quality of the education system? If that is so, then the financial rewards

[1] An effort by the government, launched in 2010, to address seven National Key Result Areas (NKRA) concerning the people of the country. The seven NKRAs are: addressing the cost of living, reducing crime, fighting corruption, assuring quality education, raising living standards of low-income households, improving rural development and improving urban public transport.

[2] In measuring performance, Malaysian schools are classified into seven bands. Bands 1 and 2 are high-performing schools, while bands 6 and 7 are low-performing (details of this category will be described in Chapter 2).

should have motivated teachers to work hard to better the achievement of students in public examinations as well as TIMSS and PISA. The most recent reports, however, show a trend to the contrary. Despite the – in some way - encouraging results in TIMMS and PISA, Malaysia is still lagging behind compared to the neighbouring countries, particularly Singapore, which tops the scores in both assessments. Furthermore, Malaysia is still struggling with the long-standing problem of the academic achievement gap where the high–low-performing school and urban–rural gaps continue to broaden. Reports in recent years demonstrated a relative increase in the number of low-performing schools, particularly in secondary schools (Ministry of Education Malaysia, 2015, 2016, 2017).

The above prevailing challenges call for a systematic investigation. What is happening and why it is happening? These are critical questions for discussion among the key stakeholders of education: the questions to critically yet productively reflect on what we have been doing and how we can improve further. The why question warrants more serious attention as it should provide some explanations for what is happening and therefore one can find out what to do about it. Often, a lot of resources are put into a description (what works/what does not work), mistaking the results as an explanation for the reasons for it to work/ not to work. This is certainly misleading. Only when one has understood what the reasons are can one make meaningful decisions what to do to change it or to prevent the same issue to re-occur.

1.2 Problem Statement

There are many distinct purposes of assessment; each assessment tool serves a different purpose that suits a specific objective. Scholars, however, have long advocated that the most important function of assessment must be to promote greater learning for students (Black & Wiliam, 1998a, 2006, 2010; Gardner, 2012; Hayward, 2015; Travis, 1996; Wiliam, 2013). It is acknowledged that assessment has the potential to be an effective mechanism that teachers can use effectively for instructional improvement and this, in turn, brings promising progress in student learning experience and learning outcomes. The problem, however, seems to be that assessment has not fulfilled its potential in the Malaysian context. In simple terms, if the utilisation of assessment information is properly and practically executed, Malaysia would expect a trend in which the number of

low-performing schools decreases. As the body of evidence shows (see details in Chapter 2), this is not the case.

As mentioned above, education in Malaysia is dominated by policies that place significant importance on high-stakes tests. These tests are now used to evaluate the quality of students, teachers, schools and the whole education system. This has slowly directed the thinking of students, teachers, parents, and society at large to perceive performance in these high-stakes tests as a predictive indication of one's future (Hashim et al., 2012; Mohamad Ali & Talib, 2013; Ong, 2010). Guided by this mentality, public examinations are highly regarded in Malaysia; the people put great faith in tests, believing that having good grades is desirable in today's competitive world. This position regarding assessment, of course, is understandable given the fact that its outcomes may lead to better educational and higher employability opportunities.

It should be of concern, though, when this belief is shared by teachers, as they could end up with potentially misguided decisions about students' actual potentials. Perceiving test scores as the reflection of students' "true ability" could trigger self-fulfilling prophecies which evoke the notion that poor performance in a test is interpreted as an indication of an "intellectual deficiency" or lack of potential. This could be alarmingly damaging to the learning growth of students, especially to those who tend to underperform. In many cases, these students become the victims of teachers' prophecies and lowered expectations. Attributing the struggle to succeed in high-stakes tests as low (intellectual) ability could be harmful to students' motivation. Research has shown that this expectation or stereotype by teachers can negatively affect students' academic performances (see Ferguson, 2003; Hughes, Gleason, & Zhang, 2005; Jussim & Harber, 2005; McKown & Weinstein, 2008; Rubie-Davies, Hattie, & Hamilton, 2006).

More importantly, the heavy reliance on test scores as the main indicator of students' ability may trigger choices of pedagogical practices that are almost certain to be off target for academically disadvantaged students. Test scores are product-oriented outcome measures; they give answers to the "what" questions, e.g., what performance level can be reached by a particular student? Answers to this kind of question do not help in decisions regarding "what do I, as a teacher, need to do to improve this student's performance?" or "why can this student not do what he is expected to do?" Failure to understand the process

and the underlying reasons for the problem may lead to the subsequent failure by teachers to employ appropriate intervention strategies for these students. Consequently, this might hinder students to fully realise their true potential.

Furthermore, the policy of the band system in Malaysian schools tends to put teachers into a dilemma. It is the aspiration of the MOE to minimise and ultimately to eradicate the number of underperforming schools by the year 2020 (Ministry of Education Malaysia, 2013). This has created pressure on teachers and school leaders to increase student performance in public examinations. In turn, this seems to exert considerable constraints on teachers, school leaders and even the students themselves to link or associate assessment to its core purpose in supporting teaching and fostering learning. Teachers are confronted with substantial pressure to decide between what is best for students and what is deemed necessary for school accountability. Not surprisingly, this leads teachers to incorporate examination-taking techniques into their pedagogical plan (also known as "teaching towards the test").

Additionally, under the increased pressure to improve Malaysia's rankings in the PISA and TIMSS tests, the purposes of assessment as a tool for monitoring students' progress and for improving instruction could be brushed aside and ignored. Benchmarking to conform to the international standard seems straight forward. The MOE certainly needs to do this so that people may know how globally competitive Malaysians are in comparison to others. However, it becomes a cause of concern when the freedom and the professional expertise as teachers to decide what and how to teach increasingly diminishes – teachers need to streamline their curriculum to "mirror" the knowledge and skills assessed in PISA, TIMSS and other high-stakes tests. Again, it is quite worrisome when assessment results are utilised by the MOE to justify the introduction of the new curriculum. Due to the influential impact of PISA and TIMSS, it is warned that the results of large-scale assessments could be easily misinterpreted, as they may lead to inaccuracies of the degree of confidence about individual students' academic ability (Elliott, Stankov, Lee, & Beckmann, 2019; Wu, 2009).

Apparently, the increasing demands to raise students' performances as an indicator of the quality of education has shifted the purpose of assessment significantly (Urdan & Paris, 1994). Student achievement in high-stakes examinations is now widely used to assess

school quality, teacher effectiveness, determine school funding and the effectiveness of reforms of educational policies (Abdul Wahid et al., 2011; Barnes, Fives & Dacey, 2017; Black, 2015; Ong, 2010). As far as educational responsibility is concerned, a reported rise in the number of low-performing schools implies just the opposite of what the MOE hopes to achieve. What is more, the prevalence of benchmarking to conform to international standards through PISA and TIMSS is reflected by their weight given in educational policies. Evidently, there is a noticeable discrepancy between rhetoric and practice. This internal conflict may confuse teachers about whether to satisfy the expectations of "outsiders" or advocate improvements for "insiders", i.e., their students. The shift of the assessment function leans more towards the accountability-oriented, and international benchmarking seems to override the core purpose to facilitate effective teaching and learning. Black and Wiliam (2006) cautioned that using information beyond this fundamental function may run the risk of misinterpreting assessment outcomes. Looking at the phenomenon of increasing number of low-performing schools, it is apparent that information drawn from test scores is not fully utilised to make instructional decisions how best to facilitate students' progress. This scenario is worrying and necessitates a return to the original aim of education, i.e., developing the potential of individuals. The question is, why do such phenomena occur? There could be many reasons, the two main ones are: (i) teachers' lack of understanding of assessment information and/or (ii) a lack of assessment tools that, in fact, would provide the information needed.

1.3 Focus of the Study

Concerns regarding students' academic performance and the urban–rural performance gap are nothing new for Malaysian educationalists. Previous studies have paid considerable attention to these issues, attempting to identify contributing factors to underperformance and associated performance gaps. The findings of a series of studies indicated that factors such as students' self-efficacy, anxiety and motivation (e.g., Dzulkifli & Alias, 2012; Lei & Mei, 2015; Md Yasin & Dzulkifli, 2011) play a role in determining students' academic success. Apart from student factors, contextual factors such as teacher–student relationships (Md Yunus, Wan Osman & Mohd Ishak, 2011), parenting style (Ishak, Suet & Poh, 2012; Othman, Azman & M. Ali, 2008) and socio-economic status (Hanafi, 2008; Saw, 2016) have been discussed as potential reasons for students' poor academic performance.

A review of the literature shows that the potential factors of Malaysian students' low performance in international and national tests have been subjected to numerous academic inquiries. The primary focus of such studies are the contextual factors of the learner, the school, the family and health-related issues. The aim of this study, however, focuses on a different pathway from earlier research. Recognising the centrality of assessment as the nucleus of Malaysian transformational programmes, it is only appropriate and timely that the attempt to explore the assessment-related factors as a possible explanation for students' academic performance be made. As a matter of fact, despite the established significant relationship between instruction and assessment, this subject matter is relatively under-researched (Ismail, Samsudin & Md. Zain, 2014; Putih, Mohd Zin & Ismail, 2016; Yusup, 2012).This study, therefore, intends to focus on two factors that have received less attention in contemporary debates, i.e., teacher's ability to utilise assessment information and the suitability of tests employed as valid sources of needed information.

The preceding section puts forth the position that a potential misunderstanding of assessment information is one of the possible barriers to improving educational practices. The critical issue underlying Malaysia's problem of unsatisfactory academic performance, especially in low-performing and rural schools, is possibly a reflection of underutilisation of what the information assessments could do or provide. In light of the existing challenges, it is assumed that misconceptions about the purpose and the potential of assessment information are a result of teachers' lack of understanding of assessment per se. Another argument relates to a situation where the assessment tools used are not providing the information teachers would need to inform their educational practice. In other words, currently used assessments are unsuitable for the purpose of identifying students' potential. This study aims to investigate the possibility of these two issues as the constraining factors that might stand in the way to a high-quality education. The following diagram (Figure 1.1) illustrates the research agenda of this study, highlighting the role of teachers, as the users of the assessment information, and the assessment tool as the determinant for the improvement of academic performance.

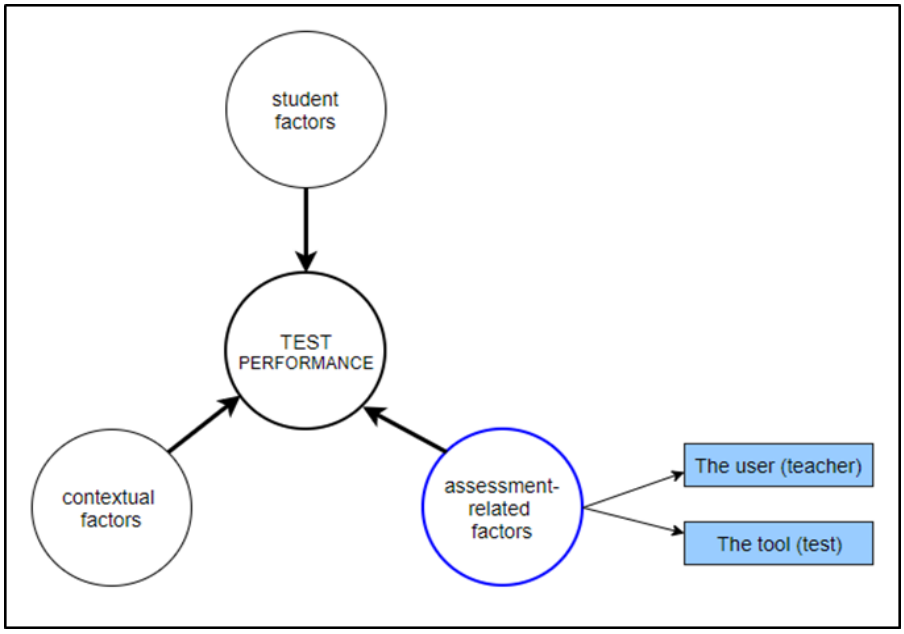


Figure 1.1: Focus of the Study

In short, apart from student and contextual factors, it is the right tool, and the sufficiently educated teacher (as a tool user) appreciating and utilising the tool appropriately that could be the answer to resolve the issue of predominantly poor academic achievement in low-performing schools (see a detailed discussion of this in the following sections).

1.3.1 Teachers' understanding of educational assessment

The notion of effectively utilising assessment information for the improvement of students' achievement, nonetheless, can only be materialised if teachers, the main user of any form of assessment, are sufficiently equipped with the necessary knowledge and understanding of how to extract value from the assessments. The MOE acknowledges that for the integral vision of quality education to take place, its success is largely dependent on the quality of teachers (Ministry of Education Malaysia, 2013). Understanding assessments, being able to choose the right tool for the task at hand and being able to utilise the information the assessment provides, is probably an under-represented dimension of a teacher's skill set. Hence, it is important for teachers to continually acquire the required knowledge and skills so that they can develop an informed criticality towards assessments to prevent misinterpretation of test scores and to use this information in their pedagogical practice.

Why should one be interested in what teachers think and do? Arguably, the ability to understand assessments can potentially be attributed to the belief system that teachers hold about assessment in general, as this will affect what they do with the information obtained by using assessment tools, and this, in turn, will have an impact on the decisions made about students and instructional approaches (Baird, 2010; Barnes, Fives & Dacey, 2015; Jussim & Harber, 2005; Pajares, 1992). Pajares (1992) advocated that beliefs are "the best indicators of the decisions individuals make throughout their lives" (p. 307). This is to say, if teachers hold misconceptions about the purpose and use of an assessment, this has the potential for consequential damage to teaching and learning. As a result, teachers may end up with an inaccurate diagnosis of students' educational needs and even inappropriate planning of interventions (Brown, Choudhry & Dhamija, 2015; Brown & Remesal, 2012; Mertler, 2009). Therefore, the notion of teachers, as the key players in education, adequately understanding the purposes and uses of educational assessments is imperative to ensure that such misconceptions do not govern the way they educate their students.

Not only are teachers' beliefs crucial, but also, what they do with the information drawn from assessment results is equally important for effective teaching and learning. Beliefs have a great influence on teachers' professional behaviour (Brown et al., 2015; Jussim & Harber, 2005). A shift in the perceived function of assessments and the meaning of test scores to influence the decision-making of school funding and programmes may have affected how teachers teach in a classroom (Urdan & Paris, 1994). The MOE initiatives to allocate selective financial awards to schools, head teachers and teachers, for example, may have led to shift towards promoting rote and superficial teaching practices to prepare students for public examinations. When school quality is tied to students' achievement in high-stakes tests, teachers are likely to pressure students to study well for the tests rather than for the purpose of learning, i.e., the acquisition of knowledge and skills. Similarly, school administrators may make some drastic decisions, such as a filtering system for student intake and intensive extra classes after school hours for the sake of raising the school's Grade Point Average (GPA). This is, of course, detrimental to the learning and development of students. If schools are allowed to continue such a practice, students will be more than likely deprived of the opportunity to develop their potential, which in turn would contradict the aim of education as outlined in the NEP. On this ground, it is

therefore equally important to examine teachers' actual use of assessment information in instructional practices.

In order to highlight teachers' understanding of the purpose and use of assessment, this study aims to draw specific attention to the existing assessment tool, the Form 1 Diagnostic Test (F1DT) which has been implemented in Malaysian schools for many years. F1DT is typically administered by teachers during the first week of the academic calendar to Form 1 students (aged 13). Before progressing to secondary level, these students sit for one of the high-stakes public examinations – the *Ujian Pencapaian Sekolah Rendah* (UPSR) or the Primary School Achievement Test – at the end of their primary education, i.e., Year 6 (comparisons of the two tests are further discussed in section 6.2.1.2.1). FIDT is designed as a school-based assessment to measure students' pre-existing knowledge, mostly on the basic knowledge of three subjects – the Malay language, the English language and mathematics. It is a curriculum-based test that aims to assess students' reading, writing and arithmetic abilities, otherwise known as *3M*, a Malay acronym for *membaca* (read), *menulis* (write) and *mengira* (count). Table 1.1 and Figure 1.2 show the examples of test constructs and test items in the English Language paper.

Table 1.1

Example of Test Construct in the English Language Paper (reproduced from (Bahagian Pembangunan Kurikulum, 2013)

Section	Construct	Weightage	Focus
A	Grammar	40%	-Identifying errors in sentences (20%) -Correcting errors (20%)
B	Vocabulary	20%	-Error identification (10%) -Cloze passage (10%)
C	Language forms and functions	20%	-Situations related to student's everyday lives
D	Information Transfer: Non-linear to linear	20%	-composite pictures/ graphs/ charts
	Total	100%	

Section D : Information Transfer

[20 marks]

[Time suggested : 20 minutes]

Questions 51 – 60

Study the pictures and the information given. Then complete the table with the correct answers.

Fish A	Fish B	Fish C
		
• Angelfish • Flat body, high fins, small head • Peaceful, but eats smaller fish • Dark green and black • RM 15 • Life expectancy of 5 years	• Goldfish • Round and short, flowing tail • Quiet and fragile, sensitive • Red and white • RM 60 • Life expectancy of 15 weeks	• Loach • Long narrow body • Hardy and adaptable, playful • Yellow and orange • RM 30 • Life expectancy of 3 months

	X	Y	Z
Type of species	51.	Goldfish	Loach
Lifespan	5 years	52.	3 months
Shape of body	Flat body, high fins, small head	53.	54.
Colour	Dark green and black	55.	56.
Price	57.	58.	RM 30
Behaviour	59.	Quiet and fragile, sensitive	60.

Figure 1.2: Examples of F1DT Test Items

Numerous attempts by the author of this thesis to find the official document justifying the implementation of the test were to no avail. In all likelihood, it does not appear to exist. The official provision for F1DT is only targeted for English language, with the purpose of identifying students' level of English language proficiency by using the score for placement in a set system[3] (Bahagian Pembangunan Kurikulum, 2013). For many years, schools have argued in the same vein for the administration of F1DT, encompassing the three subjects, to stream students according to their attainment level.

Why is it crucial to check teachers' beliefs and practices regarding this assessment tool? Ideally, diagnostic assessments should be viewed as "feedback to help inform the pupils and the teachers on the next appropriate instructional step" (Black, 1983, p.62). In other words, it is supposed to point out learners' strengths and weaknesses so that they can do something about their learning. This information is meant to be useful for teachers to modify instructions to facilitate students' educational needs. Instead, in most cases, scores of F1DT are primarily used for streaming purposes, where students are placed in certain

[3] A system that allows schools to allocate students with homogenous achievement scores to the same class, so that focused lesson planning can be carried out efficiently (effective for implementation in 2013).

classes according to the scores they attained. If teachers wrongly interpret the performance in the F1DT as a reflection of students' "true ability", this may lead to students' misplacement, and consequently teachers could end up providing inappropriate interventions for the wrong students. The issue of students' poor academic performance could be seen as one of the pieces of evidence that such misinterpretation and misdiagnosis actually occurs in schools. Hence, getting things right at the very beginning of students' first year of secondary schooling is very important to avoid the issue of wrong diagnosis, which is detrimental to their learning development.

It is, therefore, essential to reiterate that all teachers, who both generate and use assessment data, must have a solid understanding of the purpose of assessment and how the data will be utilised to make judgements about students (Stiggins, 1995; Travis, 1996). If the primary purpose of F1DT is for sorting student, it is argued that teachers could use UPSR results to sort students according to their academic achievement. Both tests are curriculum-based, and it is likely that students who obtain a good result in UPSR will perform well in F1DT. It seems that the latter does not offer new insights about students, as the information from both tests is redundant (see some examples of FIDT and UPSR test items in Appendix 1A). More importantly, using tests for testing what is already known could contribute to teachers' tiredness of using assessments. It costs time, creates pressure and diverts efforts from teaching the actual curriculum.

1.3.2 Use of appropriate assessment tools

Underutilisation of assessment information might not necessarily be an effect of lack of understanding of assessment and its related instruments. It might equally be rooted in the limited usefulness of the current assessment tool, which would make it difficult for teachers to extract appropriate value from its use. Reflecting on current challenges in Malaysian schools, doubts could be raised whether teachers are using the right tool or whether they are extracting insightful information accurately from the current instruments. At present, teachers utilise a variety of tools, yet the widening of the achievement gap is still predominant among Malaysian students. One question in this context is whether F1DT is useful in providing more meaningful information about the learning potential of a student and how he or she can be supported to improve his or her learning development.

Thus, in a situation where the existing tools seem to be functioning unsuccessfully, necessary reforms to adopt a broader range of assessment tools should be seen as the primary way to deal with the problem (Shepard, 2009). This means that schools should deploy a different instrument that allows teachers to draw meaningful and accurate inferences about students' ability to learn. If assessment tools enable the test user to derive useful insights, one could assume that the quality of subsequent interventions and instructions will improve. In turn, the learning potentials of students will be optimised.

Rooted in the reliance to stream students, information from FIDT is used to categorise students into ability groups: advance, intermediate and weak. Often, the "weak" students are labelled with various names – remedial, low performers, underachievers – because of their low scores. Yet, one wonders if a student's test score is the only thing that matters? Or is there something more important than simply knowing *what* students can do? How about knowing *how* students can do better? The point is that rather than focusing on remedial labelling, teachers should be interested in what can be done about remediation; trying to search for ways to accommodate students' educational needs to enable them to perform at the level of their true potential. Furthermore, it is feared that such labelling may reinforce feelings of hopelessness whereby many students give up. It is argued that the widespread use of traditional tests as the only viable measurement of student attainment in education is misleading, emphasising that these tests are limited to measuring knowledge and skills already acquired, not students' learning potential (Beckmann, 2006, 2014; Elliott, 2000; Elliott, Resing & Beckmann, 2018; Guthke, 1992; Guthke & Beckmann, 2000a, 2000b; Hessels, 1997).

If one were to follow this argument, the question emerges: what kind of assessment tool is needed? Obviously, what is needed is a certain type of test approach that could offer rich diagnostic insights for a better-informed decision to guide improved instruction catering to the needs of students, particularly those of underperformers. In addition to measuring academic achievement, teachers should look for a form of assessment that can provide meaningful information about (a) students' learning potential and (b) ways to help those struggling to realise their potential. These insights are crucial for teachers and also for students themselves so that they can focus on the necessary work to improve their performance (Elliott et al., 2018). Also, such a test that focuses on identifying potential could help teachers, parents and students to overcome the negative and potentially

harmful impact of fatalism and hopelessness. Like any other conventional tests, the above criteria are not covered by F1DT. In other words, F1DT might not be the right tool to provide teachers with the information they need to help their students to utilise their potential to learn.

Now, what is the solution for the above predicament? In fact, one does not need to look far. There is already an assessment approach available that is likely to fit these requirements. The solution which this study proposes is an introduction of a new assessment approach that has been widely researched and applied in educational settings for many years, called dynamic testing (DT). The proposition to use DT in this study is supported by theories and empirically established findings (see detailed discussions in Chapter 5). Theoretically, Sternberg (1998, 1999) is of the view that human abilities are a form of "developing expertise", suggesting that individuals are constantly involved in an ongoing process of acquiring skills and knowledge to develop their expertise. It is argued that traditional tests fail to identify the learning potential of individuals, which is crucial to the development of their ability. Conventional tests seem too focused on developed abilities, whilst DT aims at the potential to develop abilities. This is due to conventional tests' focus on the past or current ability, rather than on the abilities that have not been fully developed (Sternberg & Grigorenko, 2002). In support of DT, Sternberg and Grigorenko (2002) asserted:

> A major problem is not with the use of test per se, but with the kinds of tests being used. Conclusions are being drawn that go way beyond the inferences that properly should be drawn from the test scores. We believe that dynamic testing possesses the potential to make information gleaned from tests more valid and more useful. (p.ix)

But, what can DT practically offer to resolve the problem at hand? Underpinned by the contribution of Vygotsky's educational legacy of the Zone of Proximal Development (ZPD), the hallmark of DT lies in the integration of feedback and instruction into the assessment process (Elliott, 2003; Guthke, 1992; Sternberg & Grigorenko, 2002). DT seeks to examine the extent to which a person can improve his or her performance in a test after receiving feedback. It is believed that the incorporation of feedback may provoke an individual's learning process, and this eventually improves his or her performance in a test (Beckmann, 2014; Guthke & Beckmann, 2000a; 2000b). Such scaffolded assistance provides better insights of students' abilities to learn, indicating the

intensity of assistance a person needs to reach his or her potential (Elliott, 2003; Haywood & Lidz, 2007; Resing, 2013). DT, in fact, is not new; it has shown its practical uses remarkably in educational settings primarily for educationally disadvantaged children (Beckmann, 2006; Bosma & Resing, 2012; Elliott & Resing, 2015; Guthke & Beckmann, 2000a; Hessels, 2009; Hessels, Berger & Bosson, 2008; Wiedl, Mata, Waldorf, & Calero, 2014; Zaaiman, van der Flier, & Thijs, 2001). It is anticipated that DT could become a promising tool that could overcome the shortcomings of F1DT.

In summary, the impetus of this study is (a) to facilitate the development of a better understanding of the diversity and multidimensionality of educational assessments by teachers and (b) to promote the use of an assessment tool that is better suited to inform teachers, parents and students about the cognitive potential that underpins educational performance.

1.4 Research Objectives and Research Questions

The unsatisfactory achievement of Malaysian students in underperforming and rural schools has been a national concern for many years. This study aims to contribute to a better understanding of the potential reasons for this particular issue. The study focuses on the question; why does the extensive use of assessments in the Malaysian education system "fail" to bring about the intended positive changes in instructional strategies and pedagogies to foster students' academic progress? In addressing this question, the present study postulates two potential reasons: (i) teachers' lack of understanding of assessment information; and (ii) the unsuitability of the currently used assessment tools (fitness for purpose). The two elements – the teacher and the tool – are not necessarily mutually exclusive. This study is an effort to empirically investigate the interplay of these two elements and how they might contribute to an alleged underutilisation of the assessment in Malaysia's efforts to deliver on the strategic goal of world-class education.

The main objectives of this study are twofold: (i) to assess teachers' beliefs and practices concerning educational assessment and (ii) to explore the effects of introducing an alternative assessment tool on teachers' reported assessment beliefs and practices. Specifically, this study aims to:

1) Examine teachers' beliefs about the purposes and uses of an assessment tool and their practice regarding its outcome
2) Assess the alignment between teachers' assessment beliefs and practices
3) Investigate the influence of individual and school variables on teachers' assessment beliefs and practices
4) Introduce the concept of DT as an alternative method to better understand students' learning potentials
5) Analyse the potential effects of introducing DT on teachers' beliefs and practices of the current assessment tool
6) Explore teachers' responses to the application of DT and its relative potentials and barriers for the practicality of implementation in Malaysia

1.4.1 Understanding teachers' assessment beliefs and practices

The shift in the perceived function of assessment from obtaining useful information about individual students' levels of developed abilities to using their test scores as the key means to evaluate teachers and schools has an impact on how teachers view educational assessments. Teachers might even perceive test scores as a true and "objective" reflection of students' potential to learn. In the context of a perceived pressure to improve a school's performance, teachers may subsequently feel forced to only focus on high-scoring students. As a consequence, the true purpose of assessment, that is, to support effective teaching and learning, is overlooked and ignored.

Another issue worth investigating is the extent to which teachers use assessment information in their instructional practices. This is built on the assumption that teachers' behaviour and actions are influenced by their professional beliefs (Brown et al., 2015; Jussim & Harber, 2005; Urdan & Paris, 1994). If teachers hold false beliefs about assessment or have insufficient understanding of assessment and its outcomes, their contribution to working towards the country's aspiration to world-class education will be limited, to say the least.

The research reported in this thesis, therefore, (i) investigates the status quo of teachers' beliefs about the purpose and use of F1DT and (ii) assesses teachers' use of its outcomes in their educational practices. It needs to be emphasised here that this study is interested

in assessing teachers' impression of usefulness of the test rather than its actual purposes and uses. Acknowledging the influence of extraneous variables – individual as well as school characteristics – on teachers' beliefs and practices of educational assessments (Brown & Remesal, 2017; Koloi-Keaikitse, 2012; Mansour, 2013; Rubie-Davies et al., 2012), this study also aims to examine if individual characteristics and school variables could have a significant impact on teachers' reported assessment beliefs and practices.

Guided by the line of arguments outlined above, this study intends to answer the following research questions:

1. What are teachers' beliefs about the purposes and uses of F1DT?
2. What are teachers' assessment practices regarding the use of F1DT?
3. To what extent do teachers' beliefs about the purposes and uses of F1DT align with their assessment practices?
4. To what extent are individual characteristics and school variables associated with teachers' assessment beliefs and practices?

1.4.2 Introducing an alternative assessment approach

Notwithstanding its widespread use in Malaysian schools for many years, it is argued that F1DT has failed to keep its promises to help teachers and students to extract meaningful insights to improve teaching and learning effectively. Current reports indicated a relative increase in numbers of low-performing secondary schools (Ministry of Education Malaysia, 2015, 2016, 2017). Reflecting on this alarming situation, the usefulness of F1DT has been called into question. Arguably, F1DT is not suitable for identifying low-performing students' learning potential. Such information, however, is crucial to guide teachers' next steps to help and monitor students' progress to improve their academic performance.

The lack of informative assessment tools calls for alternatives that could assist teachers to better understand students' learning potential. In pursuit of a more efficient approach, this experimental study intends to introduce an alternative tool, i.e., DT, hoping it might make a difference to the current situation. In an exploration to study teachers' responses to this alternative measure, an intervention to introduce DT in the form of an educational workshop for teachers is designed to find out whether the exposure has a positive impact

on (i) teachers' attitude towards assessment and (ii) the alignment between teachers' assessment beliefs and practices.

The effects of introducing DT as a proposed alternative to currently used approaches to assessment will be the focus of the second set of research questions addressed in this study:

5. Does the introduction of DT change teachers' reported beliefs about the purposes and uses of F1DT?
6. Does the introduction of DT change teachers' reported assessment practices regarding the use of F1DT?
7. To what extent does the introduction of DT affect the alignment between teachers' reported beliefs about the purposes and uses of F1DT and their reported assessment practices?
8. What are teachers' opinions with regard to the potential and barriers to implementing DT as an alternative tool in Malaysian schools?

In sum, this study is primarily explorative, as it examines the assumption of teachers' underutilisation of assessment information and/or the unsuitableness of assessment tools that could be seen as contributing factors to the perennial problem of poor academic performance in low-performing schools in Malaysia.

1.5 Chapter Summary

This study seeks to better understand the ways to address the prevailing educational challenges of academic underperformance in Malaysian secondary schools. Two issues are put forward: (1) teachers' attitudes towards assessment and the perceived value of assessment information, and (2) the suitability of currently used assessment instruments for providing information that is useful for teachers, students and parents in their attempt to develop academic performance. This introductory chapter has presented the framework of the study that governs the scope of the investigation. It has also outlined the objectives of the study, which guided the design of the eight research questions. The next chapter describes the contextual background of the Malaysian education system, aiming to provide a better understanding of the rationales for the current study and its relevance and timeliness.

2 Contextual Background

The objective of the research reported in this thesis is to address several key issues concerning the underutilisation of assessment information in the Malaysian context. As mentioned in Chapter 1, it is proposed that two factors – lack of understanding of assessment outcomes in test users and/or the use of insufficiently informative assessment tools – could be the reasons that the assessment system put in place in Malaysia fails to facilitate effective instructions to improve students' academic performance.

This chapter provides contextual background information about the status quo regarding several elements in the education system, aiming to help understand the rationale for the study reported here. The first section describes the school and assessment systems as currently implemented in Malaysia. The subsequent sections look into two key features of the government transformational efforts: (i) the school-based assessment (SBA) as an assessment reform; and (ii) the comprehensive educational blueprint – the Malaysian Educational Blueprint (MEB) – as the pathway towards achieving national aspirations for world-class quality of education. The last section in this chapter reflects on the challenges to the implementation of an effective assessment system in Malaysia.

2.1 School and Assessment System

Under the national education policy, the MOE provides a de facto free education for all children in Malaysia, where public and private schools co-exist. The education system is divided into pre-school education, primary education, secondary education, post-secondary education and tertiary education. As secondary education is within the scope of this study, the following section exclusively focuses on details of pre-school, primary and secondary levels.

Formal education starts with pre-school education for young children aged four to six years old and is provided by several government agencies as well as non-governmental organisations. This is followed by six years of primary school, i.e., Years 1 to 6. As stipulated in the Education (Amendment) Act 2002, primary education is compulsory and the provision imposes that parent(s) will be guilty of an offence, which may result in a fine or a term of imprisonment, if such condition is violated (Ministry of Education

Malaysia, 2013). At the end of Year 6, pupils sit for the *Ujian Penilaian Sekolah Rendah* (UPSR or the Primary School Achievement Test). As automatic progression to the next educational level is applied in primary education, pupils are guaranteed an entry to secondary level regardless of their results, pass or fail, in UPSR. The information from this centralised examination is primarily used to measure the mastery of pupils in the *3M* skills of reading, writing and arithmetic abilities and to award high achievers with a scholarship and opportunity to enter boarding schools.

Secondary education is divided into two levels – lower secondary and upper secondary. The former lasts for three years, referred as Forms 1 to 3. In Form 3, students sit for the national test, the *Pentaksiran Tingkatan 3* (PT3 or the Form 3 Assessment). Based on PT3 results, students have three streams to choose from: science, arts, or technical and vocational streams. Those opting for the technical and vocational streams can enrol at vocational colleges (formerly known as technical and vocational schools). Like UPSR, failing PT3 does not prevent students from moving to the upper secondary level, which lasts for another two years – Forms 4 and 5. At the end of Form 5, students are required to take the *Sijil Pelajaran Malaysia* (SPM or the Malaysian Certificate of Education) or the *Sijil Pelajaran Malaysia (Vokasional)* (SPMV) for the technical and vocational stream. After completing their secondary education, students can choose to pursue their post-secondary education or tertiary education in various institutions or they can go directly into the labour market. The decision primarily depends on their achievement in SPM/SPMV.

As far as assessments are concerned, students' overall academic performance is assessed at the end of each level of education. Malaysia has traditionally practised a centralised assessment system. Public examinations are designed and developed by the MOE through its agency – the Malaysian Examinations Syndicate (MES). This means that the MES is responsible for the development and administration of the three major public examinations – UPSR, PT3 and SPM/SPMV. These examinations are considered as high-stakes as they are used to decide students' further educational opportunities, i.e., admission to boarding school, scholarship application and college or university entrance. Not only they are used as a yardstick for students' educational pathways, but also the information from these tests can be seen as the linchpin for teachers, parents and students themselves to assess how well students progress throughout their schooling years. Apart

from the standardised tests, other forms of assessments, both formative and summative, continue to be carried out (at least four times per academic year) throughout the schooling years. For these low-stakes tests, autonomy is given to schools and district education offices concerning test design and implementation. The above description of the school and assessment system is summarised in Figure 2.1.

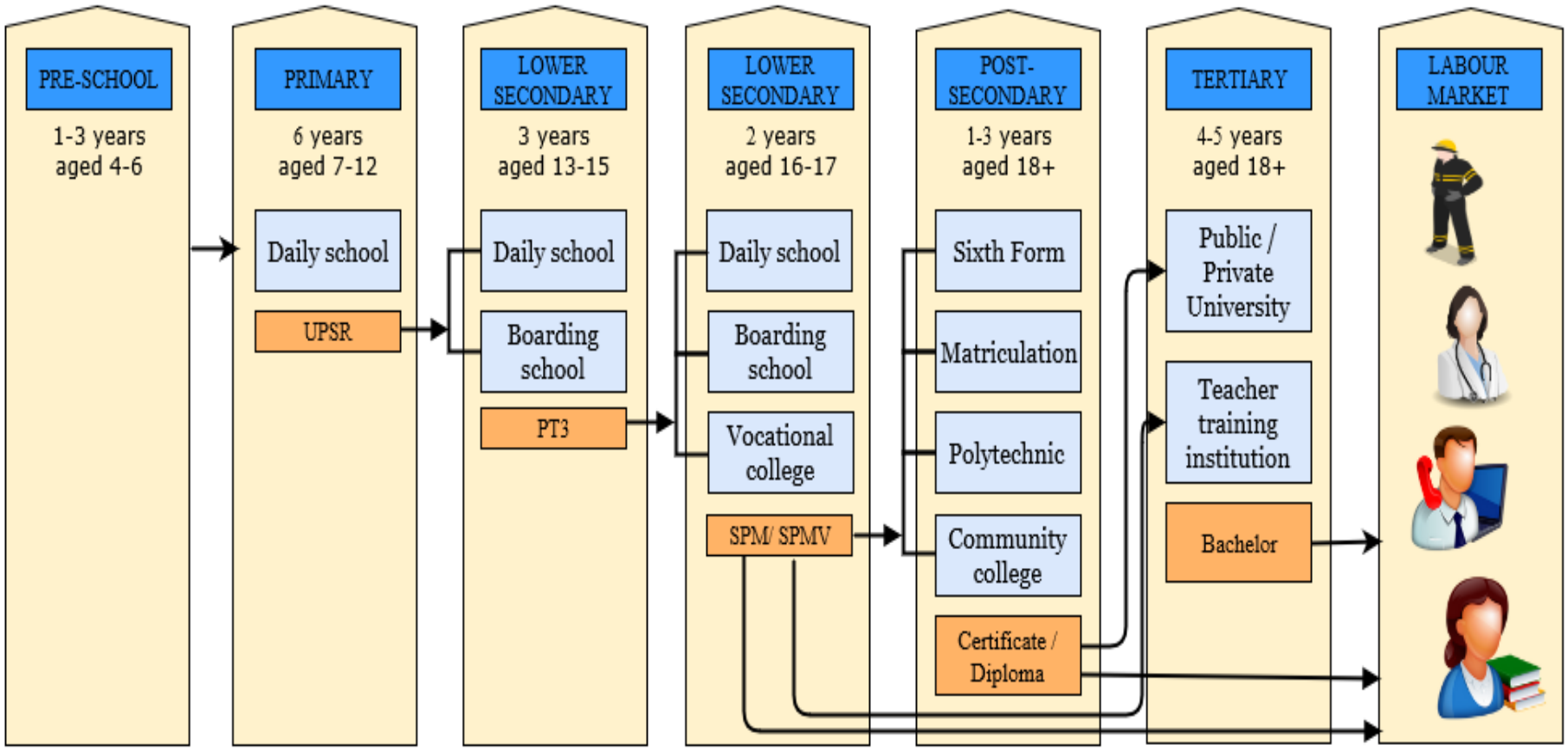


Figure 2.1: School and Assessment System in Malaysia

Another important feature of the Malaysian education system is the classification of schools into a band system which aims to monitor the performance of all government-aided schools (Ministry of Education Malaysia, 2016). At present, Malaysian schools, excluding the vocational colleges, are categorised into Bands 1 to 7; Bands 1 and 2 are the high-performing schools while Bands 6 and 7 are the low-performing ones. At secondary level, for example, the quality of a school is measured based on its composite scores, comprising 70 percent of the school's GPA and 30 percent of the Standard for Quality Education in Malaysia (SQEM) (Jemaah Nazir dan Jaminan Kualiti, 2016). The school GPA can range from 1 to 7, where a lower GPA index indicates better performance in high-stakes standardised public examinations, i.e., SPM. On the other hand, the SQEM is a self-evaluation tool that requires schools to engage in an internal quality auditing to measure five main performance dimensions: 1) school leadership; 2) organisational management; 3) management of educational programmes – curriculum, extra-curricular activities and student affairs; 4) instructional administration – teaching and facilitating; and 5) students' outcomes in public examinations (Jemaah Nazir dan Jaminan Kualiti, 2017). It is also to be noted that within the national public schools, there are a few "elite" schools. Admission to these schools is highly selective, exclusively reserved for students with excellent academic achievements in UPSR and PT3 in combination with outstanding co-curricular achievements. These schools are either daily schools or boarding/residential schools.

In its commitment to a world-class education, the Malaysian government has introduced several incentives to encourage schools in maintaining or improving their performance in public examinations. The New Deal incentive, for example, awards headmasters and principals in public schools an individual monetary incentive of RM7,500 (about GBP1,366), while teachers receive between RM900 and RM1800 (about GBP164–GBP328) based on their annual appraisal report. The reward scheme of High-Performing Schools (HPS) is another example for an examination-based assessment of school quality. The selected schools receive an annual financial allocation of RM700,000 (about GBP127,000) and an individual award of RM1000 (about GBP182) and RM700 (GBP127) for all academic and non-academic staff in secondary and primary schools, respectively (Jabatan Perdana Menteri, 2010).

As previously argued, such monetary incentives could adversely lead to a situation where teachers may engage in a distorted learning environment to prepare students for the public examination, and eventually neglect the core educational aim to develop individuals' potential holistically. In addition, such segregation of schools by students' achievement could also considerably increase the gap between low- and high-performing schools. This is because students with greater homogenous academic performance are more likely to produce similar levels of performance (OECD, 2016). This implies that schools with a larger composition of low-performing students are more likely to fall into or remain in Band 6 or 7.

2.2 Educational Assessment Reform: The School-Based Assessment (SBA)

In response to the long-standing criticism of a highly exam-oriented system, the MOE has introduced the School-based Assessment (SBA) as a new method to measure students' progress in school. The SBA has been implemented in stages, making its first emergence in 2003 when the school-based oral assessment was made compulsory for both Malay language and English language in SPM. To further strengthen its effort to make the system less exam-oriented, the MOE officially announced the National Education Assessment System (NEAS) in 2010. Under the new provision, the SBA commenced its effective use nationwide in 2011 and 2012 for primary and secondary levels respectively (Kementerian Pelajaran Malaysia, 2011).

The MOE claims that the SBA is a holistic and comprehensive educational assessment, designed in line with the aspiration of the NEP to further develop the potentials of learners intellectually, spiritually, emotionally and physically (Lembaga Peperiksaan, 2014). Mohd Yusuf (2013) asserted that this paradigm shift marks a way to a more meaningful assessment, which is characterised by its authenticity and robustness for quality education. To achieve its ambition to be a more holistic form of assessment, the SBA incorporates both academic and non-academic elements as the main components for assessing students (see Figure 2.2). PT3 is an example of the SBA-version of a central assessment, replacing the *Penilaian Menengah Rendah* (PMR or Lower Form Assessment), which was viewed as the traditional standardised test. The fundamental departure of PT3 from PMR is the autonomy of marking students' work. It is now under

the responsibility of subject teachers at the school level, with monitoring of reliability conducted at the district and state level (Lembaga Peperiksaan, 2014). Although PT3 is advocated to be a low-stakes assessment, in practice, teachers and schools continue to use its outcomes to make important decisions about students' future academic advancements.

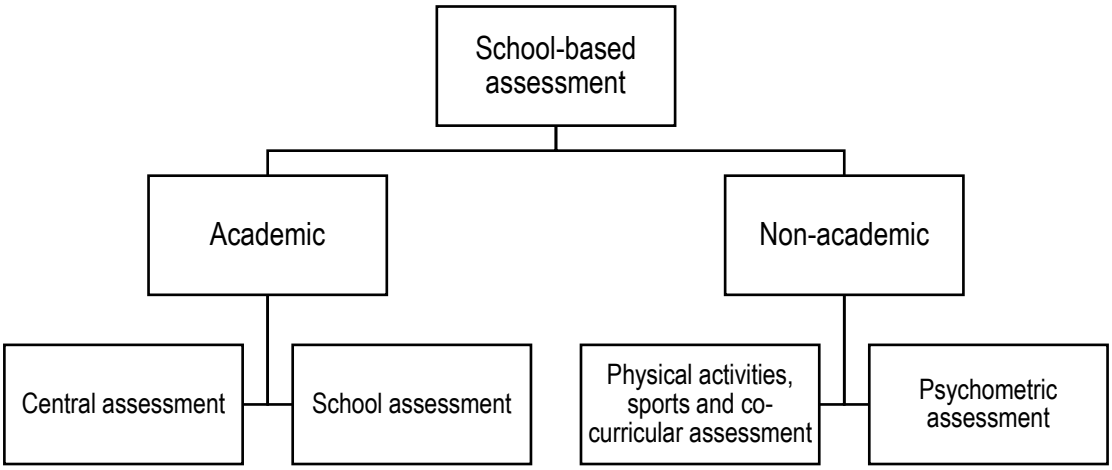


Figure 2.2: The Components of the SBA (reproduced from Lembaga Peperiksaan, 2014)

The SBA has had a mixed reception from teachers responding to the new reform. Some teachers welcomed such a move towards a more relaxed and less exam-oriented teaching and learning environment (Malakolunthu & Sim, 2010; Mansor, Ong, Rasul, Raof & Yusoff, 2013). Furthermore, teachers revealed that this contributes to a lively teaching approach and it motivates students to learn positively (Chan & Sidhu, 2006; Mansor et al., 2013; Md Omar & Sinnasamy, 2009; Sidhu, Chan & Mohamad, 2011). In contrast, others were more critical, as the introduction of the SBA was perceived as a directive from the MOE that teachers were forced to implement (Hashim, Ariffin & Muhammad Hashim, 2013). Similarly, recent studies (Abd Majid, Abd Samad, Muhamad & Vethamani, 2011; Majid, 2011; Mohd Ghazali, Yaakub & Mustam, 2012; Jaba, Hamzah, Bakar & Mat Rasid, 2013; Hashim et al., 2013) revealed that heavy workload, time constraints and lack of understanding were among the contributing factors to a negative attitude and lack of willingness to accept this new system as a shift in the educational assessment system.

The mixed responses from teachers and academicians, as well as the public, have led to a ministerial decision to review this new assessment method. Since 2014, the revised SBA has been fully implemented to revamp efforts to unleash the potential of Malaysian schools. Even after the revision, however, current studies indicated that teachers still find the implementation of the SBA a challenging task for them (e.g., Che Md Ghazali, 2016; Md-Ali et al., 2015; Md-Ali & Veloo, 2017 ; Sekharan Nair et al., 2014; Veloo & Krishnasamy, 2017).

2.3 The Malaysian Education Blueprint (MEB) 2013–2025

As articulated in the GTP, improving the quality of national education across all sectors, public and private, is the main priority for the Malaysian government. For this reason, the MOE devised its comprehensive plan, called the Malaysian Education Blueprint (MEB) 2013–2025, to steer the education system in the right direction to prepare Malaysian students for a globalised world. To achieve its objectives, the MEB sets out an ambitious vision to raise the quality standard of Malaysian education and proposes strategic and operational shifts to transform the system for improvement. The essence of the blueprint lies in its aspirational vision, which consists of two components: the system and the individual student (Ministry of Education Malaysia, 2013). The highlights of the five system aspirations and the six student attributes are illustrated in Figures 2.3 and 2.4, respectively.

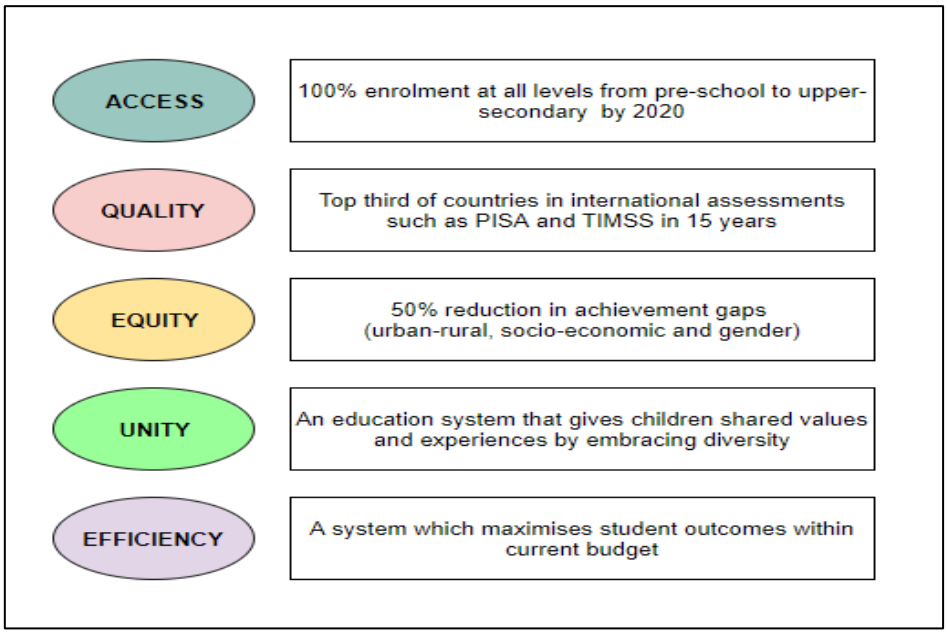


Figure 2.3: Five System Aspirations of the MEB 2013-2025

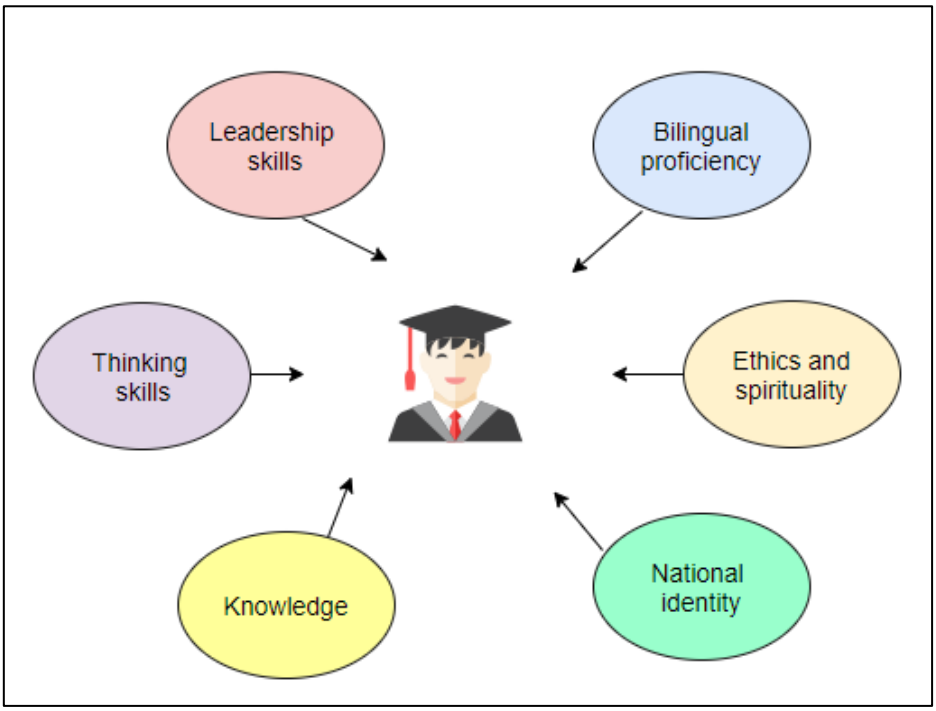


Figure 2.4: Six Student Attributes of the MEB 2013–2025

The above initiatives are envisaged to be implemented in three stages – Waves 1, 2 and 3 (see Figure 2.5).

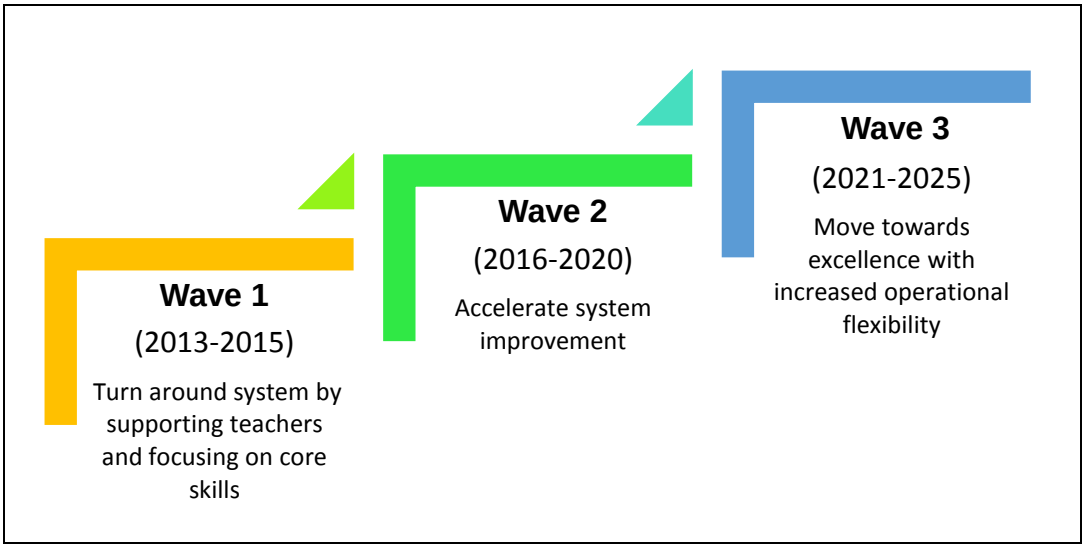


Figure 2.5: Three Stages of Implementing the MEB Transformational Initiatives

The MEB stipulates student performance as a fundamental gauge of quality education, manifested in its three elements of system aspirations – quality, equity and efficiency (Ministry of Education Malaysia, 2013). This is to say, the MOE aims to conform to the global standard, targeting the top third of high-performing countries in large-scale international assessments, such as TIMSS and PISA, within 15 years. In a national context, it is also equally important that student outcomes are equitable and budget effective. The MOE aims to reduce the urban–rural, socio-economic and gender achievement gap and to ensure the maximisation of student outcomes in return for high-cost education expenditure. In its efforts to achieve the goal of an effective world-class education system, the MEB has identified 11 shift initiatives; a cluster of collaborative plans of change with a communal involvement that seeks to engage various parties – the MOE agencies, teachers, school leaders, parents, community and private sectors.

Tracking the current performances after the launch of the MEB, annual reports (see Ministry of Education Malaysia, 2015, 2016, 2017) indicated a trajectory of encouraging outcomes, for which the MOE optimistically points out that the desired results will be achieved in years to come. On the international stage, the MEB annual report of 2016 revealed a positive sign of improvement for Malaysian students' performance in TIMSS and PISA (Ministry of Education Malaysia, 2017). In TIMSS 2015, Malaysia showed tremendous improvement in both mathematics and science, ranking among the middle third of the 39 participating countries as compared to the bottom third in 2011. As for PISA 2015, Malaysia also showed a significant increase in scores for all three literacy domains – mathematics, science and reading – which was well above the global average score of the Organisation for Economic Cooperation and Development (OECD). Comparatively, this international benchmark, however, demonstrates that Malaysia is still markedly behind other South-East Asian countries such as Singapore and Vietnam.

In the national context, the situation, however, seems to tell a different story. The annual report of the MEB in 2016 (Ministry of Education Malaysia, 2017), for example, highlighted several challenges for the Malaysian education system. In particular, academic achievement of rural students remains lower than their urban counterparts, signalling that the government effort to eradicate the urban–rural gap by 50 percent at the end of Wave 2, i.e., the year 2020, is far from being achieved. In UPSR 2016, the national achievement gap between urban and rural schools witnessed a drastic widening and the

main factor was attributed to the unsatisfactory mastery of the English language by rural students, which is, to some extent, to be expected. Although the urban–rural divide narrowed in SPM 2016, the overall results indicated that urban students performed well in all subjects. Furthermore, although it is outlined in the MEB that there will be no Band 6 or 7 schools by 2020 (Ministry of Education Malaysia, 2016), the current performances reflect the opposite. In 2016, it was reported that the number of low-performing schools (Bands 5 and 6) recorded a significant increase as compared to 2015 (see Table 2.1). This is due to the unsatisfactory student performance in SPM, which accounts for 70 percent of the overall evaluation of schools in the band system (Ministry of Education Malaysia, 2017).

Table 2.1

Secondary school band in 2013- 2016 as reported in the MEB Annual Report 2016 (reproduced from Ministry of Education Malaysia, 2017)

BAND	2013	2014	2015	2016
1	70	72	26	17
2	145	152	153	135
3	188	209	226	201
4	715	802	866	544
5	1,070	1,013	990	1,207
6	85	64	38	166

The launch of educational reforms, e.g., the SBA and the MEB, indeed signifies that the government is steadfastly committed to transforming the education system with the aim to developing young Malaysians with relevant knowledge, skills and values to thrive in the global challenges of the twenty-first century. Despite these ambitious initiatives, it seems that several concerns will continue to pose challenges for the MOE in its bid to realise a world-class education. Whilst some of the recent outcomes in TIMSS and PISA are encouraging, the evidence of internal disparities of urban–rural and high–low-performing schools continues to be worrying.

2.4 Challenges of Putting Assessment Functions into Practice

Scholars have acknowledged that assessment plays a pivotal role in facilitating effective teaching and learning (Black & Wiliam, 1998a; Gordon, 2012; Hayward, 2015; Wiliam, 2011). Putting this function of assessment into practice, however, remains a challenge for

many educational practitioners (Black & Wiliam, 1998a, 2010). Recognising this concern, the following paragraphs discuss two major challenges of putting the theoretical functions of assessment into practice in the Malaysian context.

2.4.1 The tension between accountability-oriented and learning-oriented functions of assessment

As discussed in the previous sections, the growing emphasis on increasing student performance in high-stakes tests has drastically altered the function of assessment to be more accountability-oriented (Abdul Wahid et al., 2011; Fong & Muhamad, 2017; M. Ali & Talib, 2013; Saw, 2010). The prevalence of the accountability function of assessment as a mechanism to evaluate the quality of students, teachers and schools may jeopardise the educational goal of developing the potential of the individual student. The pressure to increase student performance in high-stakes examinations may confuse teachers as to whether to prioritise the demand of the system or the learning development of students. The confusion may consequently put considerable constraints on teachers to align assessment with the improvement of teaching and to use assessment information to scaffold students' learning growth. Furthermore, the good intentions of the MOE encouraging schools to actively participate in a "band race", by offering monetary incentives, could adversely lead to a teaching-for-tests phenomenon where teachers and students engage in a rigid examination–instruction environment. This may lead students to believe that learning is merely about passing the examination and not to equip themselves with knowledge and skills to prepare for life after school.

Besides, as a result of policy enactments to increase student performance in meeting international standards in PISA and TIMSS, the government may "legitimate" the introduction of new educational reforms, particularly of the curriculum and assessment system, in its effort to achieve this national aspiration, and teachers would obviously be expected to execute the policies in their classroom. For example, the MOE has introduced the i-THINK programme with the aim to impart higher-order thinking skills (HOTS) to Malaysian students and to intentionally familiarise students with the format and style of the items tested in the PISA and TIMSS (Ministry of Education Malaysia, 2015). In this programme, teachers are given training to use the HOTS modules to complement activities in mathematics, science and reading. Teachers were reported to acknowledge

the importance of the programme, but they admitted that the modules were not fully integrated into their pedagogical practices (DeWitt, Alias & Siraj, 2016; Ganapathy, Mehar Singh, Kaur & Liew, 2017; Liew & Ganapathy, 2017). Moreover, an analysis of the baseline assessment revealed that students' competency levels in mathematics, science and reading are still weak (Ministry of Education Malaysia, 2017). This is certainly a significant concern because the efforts to promote HOTS have been actively implemented since 2013.

It is likely that assessment, as it is currently practised in Malaysian schools, may not have contributed to improving teaching and learning as much as it is expected to realise the vision of developing the potential of the individual in a holistic manner as per the aspiration of the NEP. Apparently, the tendency to focus more on high-stakes tests while ignoring other forms of assessment has caused assessment to fail to deliver its objectives to positively impact learning and teaching (Harlen, 2005; Stiggins, 2002, 2004). What teachers, school authorities and other education stakeholders need to realise is that accountability-oriented assessment is fundamentally different from the learning-oriented assessment. Arguably, assessment policies with much emphasis on the accountability function may inhibit the development of learning. Consequently, students may become objects of the assessment product, rather than active participants in learning. Moreover, substantial learning gains could be demonstrated if assessments provide meaningful interpretations that can, in turn, assist teachers to discover and develop learning potential to be an independent, effective and responsible learner in every student. This is the kind of assessment that this study will highlight.

2.4.2 Problems with the utilisation of assessment information

Student learning is likely to be enhanced if assessment information is analysed and interpreted correctly by teachers. The question is, to what extent do teachers value information derived from assessments to make instructional decisions? Looking into the dilemma of the assessment scenario in Malaysian schools, teachers have probably overlooked and disregarded the use of assessment data in pedagogical planning. What then could be the reasons for teachers disregarding the assessment data? Researchers (e.g., Chan & Sidhu, 2006; Pillay, Goddard & Wilss, 2005; Mohd Ishak, Mustapha, Mahmud & Ariffin, 2006; Mukundan & Khandehroo, 2010) identified burnout and heavy workload

as explanations for the lack of integration of assessment outcomes in the planning of instructional activities. Also, teachers cited administrative duties, heavy teaching hours and pressure to complete the syllabus for examination as the constraints on them to employ assessment information effectively in instructional decision-making (Chan and Sidhu, 2006). Furthermore, Mertler (2014) revealed that teachers are overwhelmed by the complexity of the assessment data, making it too onerous to use them. Additionally, Stiggins (1995) noted that teachers tend not to include assessment data in their instructional decision-making as a result of "fear of assessment and evaluation" (p.243). This may be true in the context of Malaysian teachers, as they may recall unpleasant assessment experiences when they were students themselves.

In addition, it is also claimed that assessment seems to have failed to ensure quality instruction simply as a result of lack of understanding among teachers on how the data should be properly interpreted (Chapman & Snyder Jr, 2000; Mertler, 2007, 2009). The findings of previous studies indicated that teachers feel less confident in accurately assessing their students (e.g., Chan & Sidhu, 2006; Koloi-Keaikitse, 2012; Malakolunthu & Sim, 2010; Md-Ali et al., 2015; Mertler, 1998, 2003, 2009; Siegel & Wissehr, 2011), claiming lack of assessment knowledge and skills as a common factor. Popham (2003, 2009) noted that this signals teachers' limited knowledge of assessment literacy. He reasoned that teachers' feeling of discomfort in interpreting and ultimately utilising assessment data accurately is largely due to the inadequate assessment training they received as pre-service and in-service teachers. It is worrying that teachers' lack of understanding of assessment may have hindered them from translating and using the information to make appropriate decisions about necessary pedagogical strategies. To a certain extent, this may have compromised students' progress in learning. Thus, it is paramount for the relevant educational agencies to provide sufficient training and supports to equip teachers with sufficient knowledge on how to use and interpret data effectively.

So, what kind of supports do teachers need? Many researchers have strongly recommended assessment literacy as a solution for the deficiency of assessment knowledge among teachers (e.g., Mertler, 2009; Popham, 2009; Siegel and Wissehr, 2011; Stiggins, 1991, 1995). Popham (2009) and Stiggins (1995) asserted that assessment literacy involves both knowledge and skills directly related to what teachers do in the

classroom, including basic assessment concepts such as validity and reliability. Another aspect to becoming an "assessment literate teacher" is the ability to interpret and transform assessment evidence to make informed and accurate decisions about teaching activities and student learning progress (Mertler, 2009; Mertler & Campbell, 2005; Stiggins, 2002). Siegel and Wissehr (2011) further expanded the framework of this professional development programme by suggesting that teachers need to possess a sound understanding of theoretical principles of assessment (knowledge of purposes and uses) as well as practical methods of use (knowledge of tools). Not only formal professional training is needed, supportive and collaborative environments, such as support from school administrators (Care & Griffin, 2009; Stiggins & Duke, 2008) and informal collegial discussion among teachers (Care & Griffin, 2009; Fullan & Watson, 2000; Majid, 2011) are also important in ensuring that teachers interpret and use assessment information accurately and appropriately. Gardner et al. (2008) referred to these formal and informal forms of support as "professional learning" (p.9), which emphasises "a change in understanding rather than merely a superficial change in teaching techniques" (p.9).

The central message here is that the focus on helping teachers to use assessment effectively should be given more emphasis in any educational policies. If teachers possess a low level of assessment literacy, they will be less likely to be able to improve teaching and learning in ways that will be advantageous to students' learning development. It is therefore pointless to advocate the fundamental function of assessment as teaching and learning improvement without steering the move for assessment awareness among teachers, who play a crucial role in the implementation and subsequent success of any educational policies and activities. Failure in supporting teachers in this regard is likely to hamper students' personal progress and development.

2.5 Chapter Summary

This chapter has given some background information about the Malaysian education system. The details of some prominent features of the system – school and assessment system, assessment reform and the MEB –were outlined. These are particularly important to provide the essential context needed to understand the statement of the research problem and to further strengthen the arguments concerning the significance of

embarking on investigating the issues of concern. The next chapter presents the literature review of the first issue of investigation about teachers' beliefs and practices of educational assessment.

3 Teachers' Beliefs and Practices about Educational Assessment

The increasing pressure for accountability of teachers and schools in the Malaysian education system arguably has an impact on how assessment is used. It tends to create the risk of misinterpreting test scores as indicators of students' potential to learn or as a reflection of teachers' quality of work. To avoid harm to students being labelled as "hopeless" and to avoid a subsequent underutilisation of their potential, better understanding of assessment and the tools employed is needed. To be able to devise effective interventions to address potential misconceptions about assessment, a thorough description of the status quo is necessary. Therefore, as part of such description, an investigation into teachers' beliefs and practices of educational assessment is patently important to inform the MOE and teachers alike on how best to adopt necessary plans to align assessment and instruction for the optimisation of student potential.

As argued earlier, teachers' potential lack of understanding of assessment may be one contributing factor to the failure to close the achievement gap between high–low and urban-rural schools in Malaysia. This study specifically aims to investigate what teachers believe about assessment and what they do with assessment-based information in their teaching. This chapter starts with a review of relevant literature about the interplay between assessment, teaching and learning for the enhancement of instruction. A review of empirical research on teachers' beliefs and practices of educational assessment is the focus of the second section. It is underpinned by the notion that the beliefs teachers hold are influential in informing their instructional activities and utilisation of assessment-based information. The last part of this chapter provides a synthesis of the reviewed literature, presenting the summary reviews and the outlines of how the present study aims to fill the gaps identified.

3.1 Interlocking Relationship: Assessment, Learning and Teaching

Assessment is a critical element of the quality of an education system. It has, therefore, also become an integral part of social life, especially when perceived as having a gatekeeping function that affects students, teachers, parents, schools and a country at large. In other words, all decisions about educational assessment or testing, whether low-

or high-stakes, influence the life course of many stakeholders. McNamara (2000) concluded that:

> Given the centrality of testing in social life, it is perhaps surprising that its practice is so little understood. In fact, as so often happens in this modern world, this process, which so much affects our lives, becomes the province of experts and we become dependent on them. The expertise of those involved in testing is seen as remote and obscure, and tests they produce are typically associated with us with feelings of anxiety and powerless. (McNamara, 2000, p.3)

Assessments are used by teachers, schools, other educational institutions and policymakers for a diverse set of reasons and with different objectives. For instance, assessments can serve the purpose of monitoring students' achievements, enhancing instruction, evaluating teachers and schools as well as evaluating national curriculum and educational policies. Educational researchers, however, have long advocated that the central purpose of assessment must be to promote effective teaching and learning (e.g., Black, 2002; Black & Wiliam, 1998a, 1998b, 2006, 2010; Brown, Irving & Keegan, 2008; Gipps, 2012; Shepard, 2000). This means that assessment outcomes are principally used to make meaningful decisions about how to best facilitate and improve learning and teaching. This calls for teachers to utilise assessment information in identifying the strengths and weaknesses of their students so that they can plan pedagogical practices to cater to the educational needs of students on an individual or group level. In support of this notion, the next section discusses the views reported in the literature on the interlocking relationship between assessment and its role in enhancing learning and teaching.

3.1.1 Assessment as monitoring and improving learning

The purposes of assessment as a tool for tracking the progress of students at school are often categorised into summative and formative functions (Harlen & James, 1997; Newton, 2007; Popham, 2009; Taras, 2005). Cizek (2010) viewed summative assessment (SA) as an assessment procedure that meets two important criteria: (a) it is collected at a specific point of an instructional period – typically at the end of a unit, semester and school year; and (b) its main purpose is to encapsulate the achievement level of a student for several purposes (e.g., grading, certification and selection). In its summative role,

assessment functions to gather the information that can be used to describe a measurement of students' learning abilities and levels of achievement (Harlen, 2007, 2012; Mertler, 2007). Evidence for this may come from formal standardised testing, particularly high-stakes tests and large-scale assessments, aiming to report and grade what has been learnt. Popham (2003) viewed tests as "inference-making enterprise" (p.4) in which teachers formally gather test-based evidence to make inferences about whether certain levels of knowledge or skills acquisition have been achieved. As SA often takes place at the end of a teaching segment, this means that an outcome or product focus is taken. Students play a passive role in this assessment procedure as no feedback on the test result is directly discussed (Sadler, 1989). As such, Harlen (2012) considered SA as mainly summarising the achievement of learning rather than supporting teaching and learning as teachers tend to "overuse" it for grading students.

In interpreting information from SA, teachers use the test scores as an indication of students' performance. At the individual level, SA yields inference about students' mastery of knowledge of and skills concerning an outlined curriculum. This indicator is widely used as a key selection criterion for students' eligibility for further educational opportunities, scholarship applications, boarding school admission, college or university entrance, etc. Also, SA performance is used as an indicator of attainment at the group level. Teachers, schools and education agencies utilise assessment data for comparative evaluation to compare and rank students' performances against that of other examinees (Popham, 2001a). Here, test results are used as a measure of students' academic performance across year groups, schools, provinces, states or countries. International large-scale assessment data from PISA, for example, serves as a benchmarking of students' achievement in mathematics, science and reading across participating countries. As shown at individual and group levels, SA, also known as assessment of learning (AoL) (Harlen, 2007; Taras, 2005), explicitly focuses on the verification and reports of evidence of students' learning, which may have profound educational and personal consequences for the student. What becomes a concern is that this is the issue where things become confused when teachers gather evidence as an indication of attainment or performance, yet inferences are being made regarding potential.

In contrast to SA, assessment functions as formative when the information is closely linked to students' learning process, providing feedback about how the assessment results

could guide teachers and students to achieve the desired learning outcome in the future. Interchangeably used as an assessment for learning (AfL) (Harlen, 2007; Taras, 2005), formative assessment (FA) is "a process in which assessment-elicited evidence is used by teachers to adjust their ongoing instructional activities, or by students to adjust the ways they are trying to learn something" (Popham, 2009, p.5). While SA is primarily product-oriented, FA aims to be process-orientated. This tends to be achieved by monitoring students' progress using low-stakes assessment procedures including classroom assessments, homework, quizzes, portfolios and group discussions. Due to its low-stakes nature, FA is flexible and informal, creating a more relaxed environment to assess students' learning progress and their understanding of particular subject matter. It is nonetheless important to emphasise that in order for FA to be meaningful and effective, it has to be planned rigorously and executed competently.

As has been advocated by many researchers, promoting learning must become the first primary purpose in designing and practising FA (e.g., Black & Wiliam, 1998a; Gardner, Harlen, Hayward & Stobart, 2008; Harlen, 2012; Sach, 2015; Sadler, 1989; Stiggins, 2007). The information gathered in assessments should allow teachers to recognise students' strengths and weaknesses and subsequently assists them to plan necessary strategies to cater for students' educational needs (Popham, 2009; Taras, 2005; Wiliam, 2013). Also, the same information is used by learners to guide them to take ownership of their learning (Black, 2015; Wiliam, 2011, 2013), developing autonomy to take actions addressing areas that require further work. When students understand their own learning progress, this can promote motivation and self-efficacy towards achieving the desired learning outcomes. In contrast to SA, FA relies on learners' active participation in the assessment process, in which they play a dual role as self-assessors and consumers of assessment information (Harlen, 2012; Stiggins, 2007).

Success in utilising the information from FA to promote learning, therefore, involves a recursive process in which both teachers and students are engaged in setting common learning targets (Harlen, 2012; Stiggins, 2007). Instrumental to this is the importance of continual feedback that aims to provide meaningful insights to regulate teaching and accelerate learning (Black & Wiliam, 1998a; Harlen, 2012; Mulliner & Tucker, 2017; Poulos & Mahony, 2008; Sadler, 1989; Stiggins, 2007; Wiliam, 2013). Crucially, this gives greater responsibility to teachers to be skilful in delivering the interpretation of the

assessment to students. Effective feedback should be "diagnostic and prescriptive" (Guskey, 2005, p.6), requiring teachers to offer detailed information about "what students [are] expected to learn, [identify] what was learned well, and [describe] what needs to be learned better" (p.6). Additionally, Evans (2013) viewed feedback as "corrective" (p.71), where teachers explicitly provide explanations of the areas that need improvement. Obviously, in disseminating the outcomes of the assessment, teachers should become facilitators guiding students to understand their own learning progress – what they can do, what they cannot yet do, where to go next and what to do next.

However, unless the feedback is understood (by learners) and is articulated well in a constructive way (by teachers), it will not greatly enhance instruction (Higgins et al., 2001; Hattie & Timperley, 2007; Orsmond et al., 2013). Reflecting on my experience as a teacher, I admit that the task of disseminating assessment-related information to students, particularly struggling learners, can be a formidable challenge. It requires teachers to really understand what kind of information they need to convey, how to ensure that the information is properly communicated and understood by students and how to encourage students to act upon the information provided. To tailor teaching to the individual needs of the student, teachers in fact would need information about learning processes instead of learning products. Conventional static testing does not provide this information.

The above distinction between summative and formative functions of assessment, however, leads to confusion when teachers struggle to distinguish the two. It is a challenge to explicitly separate the two, as they are apparently overlapping and complementary, making the distinction unclear in practice (Harlen, 2005; Harlen & James, 1997; Newton, 2007; Taras, 2005). The demarcation between the two aspects tends to be even more blurry when SA information can be used for formative purposes (Black, Harrison, Lee, Marshall & Wiliam, 2004). On occasions, SA could be formative if the same information from standardised tests, such as mid-term exams, is used to identify students' strengths and weaknesses from learning and is later used to plan for remediation or improvement for the next academic term. This is empirically proven in a study by Taras (2001), which found the use of tutor feedback and student self-assessment to be beneficial for improving SA tasks. Interestingly, such findings raise the question of

the meaningfulness of a strict distinction between SA and FA. Responding to this, Harlen (2005) argued:

> The two main purposes of assessment … are for helping learning and for summarizing learning. It is sometimes difficult to avoid referring to these as if they were different forms or types of assessment. They are not. They are discussed separately only because they have different purposes; indeed the same information, gathered in the same way, would be called formative if it were used to help learning and teaching, or summative if it were not so utilised but only employed for recording and reporting. (Harlen, 2005, p.208)

The distinction has also triggered ongoing debates among scholars. Harlen (2005, 2007, 2012) maintained his view that the distinction is still relevant; separating assessments as a dimension of purposes and uses, rather than belonging to either of the two categories. Meanwhile, others emphasise the key differences between the two relating to the purposes they serve (Black et al., 2004) and the use to which the results are put (Sadler, 1989). Following her "breakthrough" finding, Taras (2005, 2009) proposed a shift in perspective in this debate, where she asserted "both SA and FA are processes …. This SA can be implicit and the formative focus only made explicit" (Taras, 2005, p.468). Admittedly, the distinction may confuse teachers and this can affect the way they perceive and practice assessment. To rectify the confusion, Newton (2007), whose work provides a detailed discussion on this issue, concluded that it is important for those involved in the assessment system to avoid misleading categorisation of assessment purposes. He said, "to avoid getting ourselves confused, and to avoid confusing others, we need to use the language of assessment with greater precision" (Newton, 2007, p.157) and thus recommended that an explicit definition of the primary purpose is a priority in an assessment design.

The ongoing debate over the meaningfulness of distinguishing assessments into formative and summative functions is not the focus of this study. The contrast between SA and FA in the preceding paragraphs does not intend to say one assessment is better than the other. Whatever the argument, Gardner (2012) argued that "assessment of any kind should ultimately improve learning" (p.107). In a similar vein, Black and Wiliam (1998a, 2006, 2010) consistently advocated that assessment in education, first and foremost, serves the purpose of supporting learning. This view is supported by Stiggins (2007), who believed that the purpose of assessment is to turn failure into success – to avoid failure becoming

chronic and thus harm struggling students. Reflecting the same perspective, Care and Griffin (2009) put forward an enriching idea that assessment is not an identification of problems; rather, it is supposed to be used to identify the Zone of Proximal Development (Vygotsky, 2012), by providing students necessary scaffolds to reach their full potential.

To sum up, in one way or another, assessments are all meant to achieve the ultimate goal of promoting greater student learning. Obviously, both SA and FA are central to the development of students' learning. Black (1998) asserted that the labels "formative" and "summative" should be treated as "two ends of the same spectrum" (p.34). This means that both co-exist and should be closely integrated to enhance teaching and ultimately support learning.

3.1.2 Assessment as teaching improvement

While it is beneficial to monitor students' learning, assessment is also useful to guide teachers in pedagogical improvement. There has been considerable literature that has posited assessment as something that is deeply integrated into instruction, not an activity that merely evaluates learning (e.g., Care & Griffin, 2009; Gordon, 2012; Looney, Cumming, van Der Kleij & Harris, 2017; Pellegrino, 2014; Shepard, 2000). Recognising this, Care and Griffin (2009) asserted that "assessment is for teaching" (p.56) and urged teachers to collaborate in professional learning teams to use assessment data to inform their teaching. Several key roles of teachers in utilising assessment information for teaching improvement are discussed in the following paragraphs.

Firstly, for the optimal effectiveness of teaching and learning, assessment (be it formative or summative) must be a reflection of learning objectives (Boud & Falchikov, 2006; Brown, Lake & Matters, 2008; Gipps, 2012; Guskey, 2003; James & Lewis, 2012; Pellegrino, 2014; Postareff, Virtanen, Katajavuori & Lindblom-Ylänne, 2012). This suggests that the focus should be on how to incorporate assessment into the teaching and learning process to achieve the desired learning outcomes. Therefore, teachers should identify and set the learning goals prior to instruction to assist them in developing and choosing appropriate approaches to teaching and assessment accordingly (Gipps, 2012; James & Lewis, 2012). The idea of goal-setting before learning takes place will facilitate teachers to structure their pedagogical approaches for the attainment of the targeted

objectives and outcomes. Also, it will help teachers to identify suitable assessment methods and tasks that may provide evidence of how far the intended objectives and outcomes have been met. Furthermore, this will benefit students, giving them a chance to prepare for the learning in which they will be engaged (Boud & Molloy, 2013).

Secondly, it is equally important that teachers need to be cautious about the impact of judgement by assessment on the life of students, both academically and personally. In interpreting evidence from SA, for example, test scores are commonly considered as demonstrating attributes of excellence in learning. If this is misinterpreted, teachers are more likely to engage with erroneous decision-making about students' next phase of learning with an invalid interpretation of the data concerning their learning potential (Mertler, 2014; Stiggins, 2004, 2007). In the context of this study, information from F1DT may be misleading if its results are not consistent with students' performances in UPSR. This situation may confuse teachers to decide which information is reliable and valid to describe the actual capabilities that students have. If teachers insist on using the data without really understanding their meaning, inappropriate interpretations of the test information could lead to self-fulfilling prophesies, especially for struggling learners, promoting the feeling of hopelessness and thus encouraging them to stop trying. Misinterpretation of assessment data may contribute to misdiagnosis of student needs (by teachers), misunderstanding of actual ability (by the student) and miscommunication of student progress (to parents). The deleterious impacts of misinterpretation of assessment data are obvious and this affects the lives of all stakeholders. The harm of inaccurate interpretation of assessments is a powerful call for teachers to learn to interpret assessment data accurately because they are responsible to provide valid and reliable judgements about students' evidence and progress of learning (Popham, 2009; Stiggins, 1995, 2002; Volante & Fazio, 2007). Essentially, it must be emphasised that assessment literacy is an important skill for teachers if effective and sustainable assessment practices are to be established.

Thirdly, and perhaps the most important role, is teachers' utilisation of assessment feedback. As previously mentioned, feedback is the key element to effective learning (e.g., Black & Wiliam, 1998a; Boud & Molloy, 2013; Brown & Hattie, 2012; Evans, 2013; Hattie, 2015; Hattie & Timperley, 2007; Poulos & Mahony, 2008; Sadler, 1989; Wiliam, 2013). This suggests that assessment information is only useful if teachers can

extract accurate feedback to inform their own teaching and improvements to students' learning. In relation to the impact of feedback on teaching, teachers are expected to use assessment information to monitor their own teaching – pointing out students' learning achievements and later planning pedagogical strategies as interventions for enrichment and/or remediation. Often, one problem associated with the above expectation is related to teachers' making sense of assessment data (Even, 2005; Mertler, 2014; Stiggins, 2004). Thus, one way to ensure the effectiveness of feedback in teaching is through a collaborative effort among teachers (Care & Griffin, 2009; Harlen, 2005; Hayward, 2015; Majid, 2011). This approach basically allows teachers to meet, discuss and develop ideas about assessment in groups to support each other to improve their pedagogical strategies.

The above discussion about the roles of teachers in utilising the assessment information to inform instruction can be summarised in Figure 3.1.

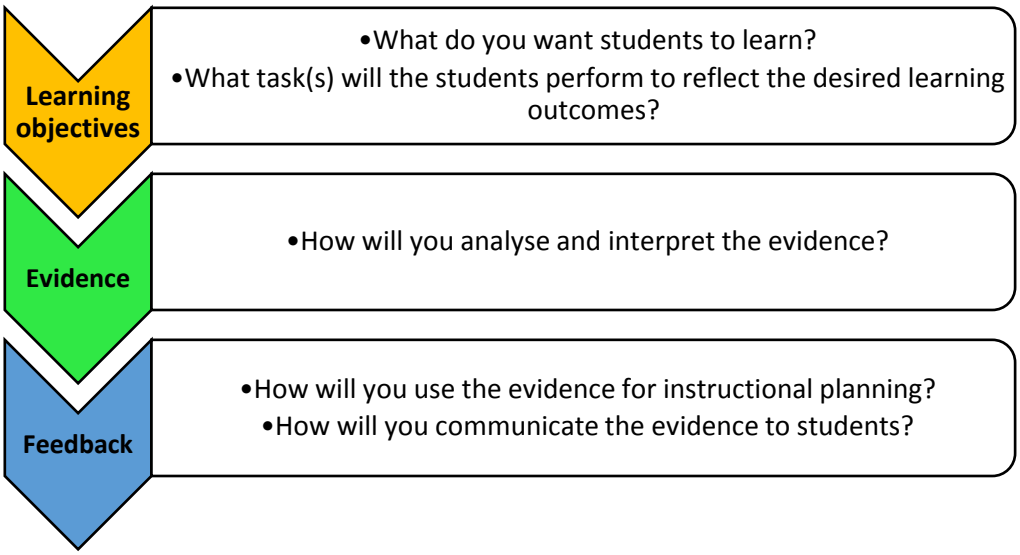


Figure 3.1: Teacher's Roles in the Utilisation of Assessment in Instruction
(Adapted from Pellegrino (2014, p.70))

The key point in this section underscores the notion that the effectiveness of assessment is highly dependent on teachers. Gardner et al. (2008) viewed teachers as a powerful agency that can foster change in education. To do this, teachers need to change their perspective and practices of assessment in order to bring the expected positive impacts in teaching and learning. This requires teachers not to view assessment as an isolated entity

from the instructional process, but rather as a guiding mechanism to inform necessary actions for improvement. Most importantly, none of the benefits of assessment, either for students or teachers, will accrue if teachers are unable to effectively transform assessment information into instructional activities. Therefore, the need for teachers to be knowledgeable users of test information is a prerequisite for the success of an assessment system.

3.1.3 Aligning the three components

As discussed above, the relationship between assessment and the process of learning and teaching is not a one-directional path. This relationship is rather a complex, reciprocal interaction. It is clear from the literature that if an assessment is tightly integrated into the instruction, it can be a valuable experience for both teachers and students. In his remark, Guskey (2003) agreed that when teachers ensure that "assessments become an integral part of the instructional process and a central ingredient in their efforts to help students learn, the benefits of assessment for both students and teachers will be boundless" (p.11), acknowledging the interrelatedness of the three domains. Arguably, it is only through assessment that teachers can be informed about the effectiveness of instructional activities that are intended to result in student learning. Its role of interplay, connecting teaching and learning, has considerably emphasised assessment as the central process of effective instruction. This also means that failure to articulate this relationship may jeopardise students' learning development, which is evidently reflected in the current widening of the achievement gap between high–low and urban–rural schools in Malaysia. It could be said that assessment brings teaching and learning together (Black & Wiliam, 2006; Brown, Irving & Keegan, 2014; Gordon, 2012; Wiliam, 2013). The interplay between assessment, teaching and learning is represented by the Venn diagram in Figure 3.2.

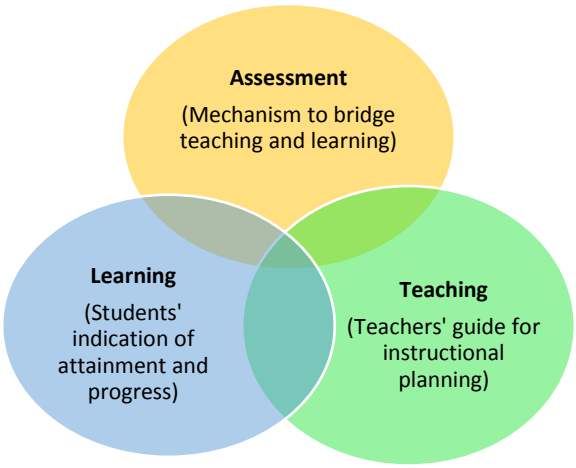


Figure 3.2: Interplay between Assessment, Teaching and Learning

A well-designed assessment that assists teachers in making decisions about more accurately informed instruction should fundamentally reflect two key features: (i) be cyclical in nature (Christoforidou, Kyriakides, Antoniou & Creemers, 2014; Mertler, 2014; Rust, O'Donovan & Price, 2005; Wiliam & Black, 1996) and (ii) be interactive (Black & Wiliam, 1998, 2010).

In regard to its cyclical nature, it is recommended that an assessment needs to be linked to all stages of instruction (Mertler, 2014). This suggests that each process of assessment and instruction is linked to and dependent on others. Wiliam and Black (1996) considered assessment as a cycle of three main phases – eliciting evidence, interpreting evidence and taking action. Using a social constructivist approach, Rust et al. (2005) expanded the cyclical nature of assessment into a two-tier assessment process model, proposing two parallel cycles of the assessment process: one for students and the other for teachers. The model involves four main essentials – engaging with assessment criteria, creating assessment criteria, engaging with feedback and enabling the feedback to work. In another work, Christoforidou et al. (2014) put forward a cycle of four phases that are interrelated– planning and construction of tools, assessment administration, recording, and reporting. Arising from the discussion in the previous sections and drawing from these three major works, an assessment can be viewed as an ongoing and complementary process, embedded in five stages: setting learning objectives, implementing the

intervention (overlapping stage), collection of evidence, interpretation of evidence, communicating feedback and planning of the intervention (see Figure 3.3).

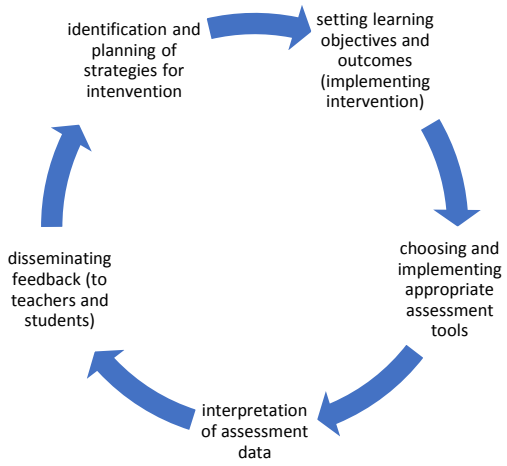


Figure 3.3: Cyclical Nature of Assessment
(Adapted from Christoforidou et al., 2014; Mertler, 2014; Rust et al., 2005; Wiliam & Black, 1996)

Besides,, it is equally important to note that assessment should be interactive, in which feedback is the instrumental element (Black & Wiliam, 1998, 2010; Harlen, 2012; Sadler, 1989; Stiggins, 2007; Wiliam, 2013). To achieve a greater harmony of integrating assessment in instruction, teachers and students should work together towards better performance in instruction and assessment outcome (Mulliner & Tucker, 2017). Teachers, in the first place, are required to understand assessment information in order to identify the progress or difficulties of their students (Popham, 2009; Taras, 2005; Wiliam, 2013), so that they can plan pedagogical practices to scaffold an effective and supportive teaching and learning environment and to check students' readiness for progression. Not only teachers should be assessment-literate, but students must also be responsive to the outcome of the assessment (Harlen, 2012; Stiggins, 2007). This requires students to use feedback from teachers to plan the necessary steps to improve their learning progress and achievement (Black, 2015; Evans, 2013; Hattie, 2015; Wiliam, 2011, 2013). To achieve this, teachers would need a form of assessment that can help them to point out not only what students have been doing but also how they can improve themselves.

Theoretically, it is argued that assessment in its essence is an ongoing process that encompasses teaching and learning. Relating to the interaction between the three components, Gordon (2012) concluded that:

> Teaching, learning, and assessment are increasingly viewed as functioning in symbiotic relationships one to the other. Although each has an independent history and a separate traditional constituency, they are, perhaps, best viewed as parts of a whole cloth, where parts are differentially emphasised at various times and for different purposes. (Gordon, 2012, p.1)

For the above notion to succeed, the cyclical and interactive nature of assessment requires teachers and students alike to mutually engage in assessment throughout learning and teaching activities. Applying these ambitions in practise remains a challenge. The phenomenon of the widening gap in high–low-performing schools and urban–rural schools in Malaysia may raise doubts about the way teachers use assessment information in their educational practices. Adversely, the complexities of teacher–student collaboration, as well as the integration of assessment into the teaching and learning process, appear to deter teachers from utilising assessment effectively to plan appropriate strategies that could really work for their students, especially struggling learners.

The above discussion points show that assessment is a complex undertaking as it involves multiple aspects that every teacher needs to deal with. This complexity demands teachers to better understand and articulate the various forms, purposes, functions, expectations and needs of assessment in order to enhance teaching and to improve learning. If we are to advocate that assessment should be embedded in teaching and learning, then simply describing it conventionally as formative and summative contributes nothing to the understanding of assessment in promoting the improvement of student learning and enhancement of pedagogical activities. The division may confuse teachers to the extent that they struggle to distinguish which assessment to use for what purpose. It is feared that the confusion may impinge on the whole process of teaching and learning, which could be detrimental to students, teachers and the education system at large. Thus, what matters most is to focus on the utilisation of information gleaned from the assessment to fit its intended purposes. For this to happen, well-qualified and sufficiently assessment-literate teachers are needed. This is why addressing teachers' understanding of assessment, particularly looking at their beliefs and practices, is an essential first step for

identifying necessary actions to be taken for improving education. The following section aims to provide detailed descriptions of elements underpinning the topic concerning teachers' beliefs about assessment and their actual practices regarding its information in instructional designs.

3.2 Studies on Teachers' Beliefs and Practices of Educational Assessment

Many scholars have established that the beliefs teachers hold about their professional work (e.g., teaching, learning and assessment) are instrumental to the way in which they engage in it (see Barnes et al., 2015; Borg, 2001; Jussim & Harber, 2005; Opre, 2015; Pajares, 1992; Skott, 2015). This means that teachers act upon what they believe. Acknowledging this bi-directional relationship, Nespor (1987) proposed that efforts to better understand a phenomenon in an educational setting should begin with the understanding of teachers' beliefs, which define and reflect their work.

In the literature, a plethora of research regarding teachers' beliefs about assessments also indicates their influence in shaping what teachers do in teaching and assessment practices (e.g., Barnes, Fives & Dacey, 2017; Brown et al., 2008; Segers & Tillema, 2011; Stiggins, 2004; Urdan & Paris, 1994). According to this perspective, teachers' assessment practices – collecting, interpreting and communicating assessment results and using the information to make decisions – are guided by the way they conceive assessment. Following this standpoint, it is argued that because teachers' beliefs have been demonstrated to potentially influence their educational practices, the actions of teachers may also have a significant impact on students' learning improvement. The following sections review empirical studies that specifically focus on teachers' beliefs about assessment and how they put their beliefs into practice.

3.2.1 Definition of terminologies

Before going further into the review, this section looks at the definition of several terminologies that are commonly used to address this topic. It is important to clarify these keywords because "the difficulty in studying teachers' beliefs has been caused by definitional problems, poor conceptualizations and differing understandings of beliefs and belief structures" (Pajares, 1992, p.307). Within the literature of teachers' beliefs,

there are a number of terminologies and definitions that are used to describe this construct. The following paragraphs highlight four frequently used terms – belief, conception, perception and knowledge.

Perhaps the most commonly appearing term is "belief". Nespor (1987) delineated belief as a system which encompasses four distinctive features: (a) existential presumption (i.e., personal truths that are immutable and unknown), (b) alternativity (i.e., the ideal situations, not the present realities), (c) affective and evaluative components (i.e., influenced by personal preferences rather than rationality) and (d) episodic storage (i.e., derived from personal experiences or events). Nespor's idea was taken up by Pajares (1992), who referred to beliefs as "messy constructs" (p.307) and viewed them as incontrovertible personal "truths" held by individuals, deriving from experiences, that are known and unknown, with strong affective and evaluative aspects. Similarly, beliefs are "psychologically held understandings, premises, or propositions about the world that are thought to be true" and are seen as "lenses that affect one's view of some aspect of the world or as dispositions toward action" (Philipp, 2007, p.259). The elements of beliefs as evaluative and emotive and their value of truth are also shared by Borg (2001), who said that belief was "a proposition which may be consciously or unconsciously held, is evaluative in that it is accepted as true by the individual, and is therefore imbued with emotive commitment; further, it serves as a guide to thought and behaviour" (p.186). In studying teachers' beliefs, she explicitly defined them as a term that describes any beliefs of relevance to the teaching environment. Furthermore, Mansour (2010) also put forward his proposition that the concept of belief is used to explain "a teacher's idiosyncratic unity of thought about objects, people, and events, and their characteristic relationships that affect his/her planning and interactive thoughts and decisions" (p.514).

Some researchers preferred to use the term "conceptions" rather than beliefs (e.g., Azis, 2015; Brown, 2004; Calveric, 2010; Dayal & Lingam, 2015; Harris & Brown, 2009; Leong, 2014; Opre, 2015; Thompson, 1992). Thompson (1992) viewed conceptions as "a more general mental structure, encompassing beliefs, meanings, concepts, propositions, rules, mental images, preferences, and the like" (p. 130). This is to say, Thompson claimed that beliefs are a subcategory of conceptions. Her sentiment is shared by Brown (2004), who described conceptions as "multifaceted and interconnected" (p.302), serving as a framework for teachers to understand, respond and interact with the teaching

environment. Reflecting a similar view, Ponte (1994) used conceptions as a "conceptual substratum" (p.170) that is closely linked to individuals' attitudes, expectations and understandings of a given situation. This means that the use of "conceptions" by these authors allows for beliefs and conceptions to be embedded within a single construct. However, there are also authors making efforts to differentiate between the two terms. For example, Ponte (1994) asserted that "they [beliefs] state that something is either true or false, thus having a prepositional nature. Conceptions are cognitive constructs that may be viewed as the underlying organising frames of concepts. They are essentially metaphorical" (p.169). Perhaps, because the nature of educational activities is complex and dynamic, "the distinction [between beliefs and conceptions] may not be a terribly important one" (Thompson, 1992, p.130).

In contrast to the two key terminologies, other researchers (e.g., Cheng, 1999; Maclellan, 2001; Mat Hassan & Talib, 2013; Mertler, 2009; Sach, 2012; Sahinkarakas, 2012; Urdan & Paris, 1994) preferred to use the term "perception". It is more common to find this term in the field of psychology (e.g., Bruner, 1957; Bruner & Postman, 1949; Hayes, 2000), philosophy (e.g., Brandom, 1981; Crane, 2009; Foster, 2000) and science (e.g., VanRullen & Koch, 2003). According to Bruner and Postman (1949), perception is "powerfully determined by expectations built upon past commerce with the environment. When such expectations are violated by the environment, the perceiver's behaviour can be described as resistance to the recognition of the unexpected" (p.222). Expanding on his previous definition, Bruner (1957) believed that perception is built upon the construction of a set of categories. He claimed that an inference is made about a phenomenon in which stimulus inputs are sorted systematically. In another work, Neisser (1976), as cited in Hayes (2000), defined perception as a dynamic cycle of cognitive construction aiming to make sense of experience. He added that perception is cyclic because it is guided by what one expects to encounter as well as what has been already encountered. In explaining his own view of perception, Hayes (2000) claimed that one tends to be selective in what one perceives as relevant, implying the idea that one will ignore what is viewed as unimportant. In the field of examining teachers' beliefs about assessment, Sach (2012) favoured this term instead of beliefs or conceptions. She asserted that teachers might differ in their perceptions about assessment due to the dynamic interaction of the teaching environment, including teachers' individual characteristics and school factors.

"Knowledge" is another term that is often linked to the literature of teachers' beliefs. Apparently, this term has not been used in a uniform way. Some scholars viewed knowledge as a subset of conception (e.g., Philipp, 2007; Thompson, 1992) while some considered beliefs as a part of knowledge (e.g., Furinghetti, 1996; Kagan, 1992; Mansour 2010; Pehkonen & Pietilä, 2003; Ponte, 1994). Kagan (1992), for instance, defined beliefs as "a particularly provocative form of knowledge" (p.65) and argued that teachers' professional knowledge (fact) is a more accurate belief than mere opinions. In discussing the relationship between the two constructs, Mansour (2010) asserted that teachers' existing beliefs act as a filter and thus control the gaining of knowledge. His claim of beliefs as a form of knowledge seems contradictory when he offered the distinction between the two by suggesting "knowledge often changes, beliefs are "static". Furthermore, he added that "whereas knowledge can be evaluated or judged, such is not the case with beliefs since there is usually a lack of consensus about how they are to be evaluated" (Mansour, 2010, p.514) Despite the efforts to associate knowledge with beliefs or conceptions, some researchers, however, understand the terms beliefs and knowledge as non-overlapping categories and that each has their own delimitations (Ernest, 1989; Dayal & Lingam, 2015; Nespor, 1987; Pajares, 1992). Ernest (1989) distinguished knowledge from beliefs by suggesting that knowledge is the cognitive outcome of education, while beliefs are the affective outcomes. He further explained that knowledge of something is fundamentally different from feeling about something. In another work, Nespor (1987) proposed that beliefs are "disputable, more inflexible and less dynamic than knowledge" (p.311

Generally, it is apparent in the literature that there is no commonly agreed on definition or term that can be used exclusively regarding the study of teachers' beliefs. There have been numerous efforts to provide definitions, yet the common core concepts have not been easily defined (Pajares, 1992; Skott, 2015). Nespor (1987) called this an "entangled domain" (p.325). Possibly this is as a result of the complexity of teachers' beliefs, which involve deeply entangled and dynamic interactions of the contexts surrounding the setting in which teachers work (Brown, 2004; Muijs & Reynolds, 2002; Nespor 1987; Opre, 2015). Pajares (1992) further suggested that the choice to adopt terminologies suitable for a study is determined by the agendas of the researchers. The literature also reported that researchers have been using at least two terms interchangeably (e.g., Barnes et al., 2015; Sahinkarakas, 2012). Furthermore, the given definitions of terms, as discussed above, are

complementary in nature. It seems that they are intertwined in an intricate way because of the complexity of the world around us. For these reasons, I am not convinced that there is need to use only one term to describe the multiple facets of this research involving the combination of several variables related to the function and use of assessment in teaching and learning. Thus, I decided that three main terms – beliefs, perceptions and conceptions – will be used interchangeably throughout the thesis, but apparently the first is the most frequently used. I opted to leave out the term knowledge as it is different from beliefs, perception and conceptions as the above discussion was referring to.

Within the framework of the present research, teachers' beliefs are used to characterise teachers' personal thoughts about the purposes and uses of assessment tools (i.e., F1DT and DT), taking into account the relationship of teachers' individual characteristics (e.g., teaching experience) and school factors (e.g., location) as possible predictors influencing their beliefs and practices. It is also important to mention here that the term "perception" is to be used in the questionnaire as described later on as it is the most equivalent to the Malay word (*persepsi*) that connotes the meaning of beliefs within the scope of the study. The translation of the other two terms is considered to be incongruous with the agenda of this study. The Malay translation of beliefs, *kepercayaan*, denotes religious and cultural beliefs, while conception is translated as *konsepsi*, which means ideas or concepts.

3.2.2 Previous empirical studies

This section documents the review of empirical works on the beliefs and practices of assessment involving the participation of practising teachers from different countries, underscoring the notion that this is a global experience in the education community, not only confined to the local setting of the current research. The review of empirical research is structured into four main areas in relation to the research questions at hand.

3.2.2.1 Teachers' beliefs about assessment

In the literature of teachers' beliefs about assessment, scholars placed much emphasis on investigating teachers' views about SA and FA. Presumably, the widespread use of standardised tests as the key indicator of students' academic achievement captured the attention of the academics to examine what teachers think about this conventional assessment approach. Urdan and Paris (1994) asserted that it is essential to understand

teachers' opinions about standardised tests, as these views may determine the way they prepare, administer and use the test information. The 153 K-8 teachers participated in their study perceived standardised testing in a negative way, implying it is educationally beneficial for neither students nor teachers. Shockingly, the respondents also avowed the practice of falsifying test scores because of the pressure to ensure their students perform well in the tests. Dissatisfaction with the high-stakes tests is further reflected by teachers who participated in Barksdale-Ladd and Thomas' (2000) qualitative study. They regarded tests as disempowering them from deciding appropriate strategies for instructional improvement because of the tendency of teaching-for-test instruction. Accordingly, the same concern is seen in other studies (see Abrams, Pedulla & Madaus, 2003; Lai & Waltman, 2008; McNamara, 2010), reporting teachers' claim of feeling that they have less control over their instruction. They admittedly spent most of the teaching activities mirroring the content and format of the tests. Apparently, these results suggest that teachers are unconvinced that standardised tests are useful for improving pedagogical practices.

Teachers' negative perceptions about the consequential effects of the standardised tests probably stemmed from their understanding of the purposes of assessments. Therefore, Delandshere and Jones (1999) conducted a study to examine teachers' perception of the purposes of assessment. The data from 14 interview sessions demonstrated that teachers utilised assessment for three main functions: student placement, grading and preparation for high-stakes exams. Summarising teachers' responses, Delandshere and Jones (1999) pointed out that assessment was "a required means of conveying information to external audiences (parents, district, state, other teachers) and rarely as a way to understand learning and inform teaching" (p.229). Other studies in the Chinese context (see Brown & Gao, 2015; Brown, Hui, Yu & Kennedy, 2011; Brown, Kennedy, Fok, Chan & Yu, 2009) also reported the influence of the accountability function on teachers' use of assessment for test preparations. Teachers in Hong Kong and Guangzhou, for example, tended to perceive assessment primarily as a function to make teachers accountable for students' performance (Brown et al., 2011). Teachers claimed that the government policy of using public examinations to benchmark teaching and school quality had specifically led them to prepare students for the examination. Similarly, the dominance use of high-stakes examinations to evaluate teachers and schools is hardly less upsetting in other countries such as the USA (Barnes et al., 2017; Lai & Waltman, 2008), Ecuador (Brown

& Remesal, 2017), Egypt (Gebril & Brown, 2014) and Iran (Pishghadam, Adamson, Shayesteh Sadafian & Kan, 2013). Evidence from these studies implies that teachers' beliefs are very much influenced by the endorsement of testing demands to ensure students perform well in the tests. As a result, teachers predominantly deployed SA to find out what students had learnt and this eventually reduced significant use of FA in instruction.

In contrast, teachers from a context where the accountability function is of lower importance tended to view the purposes of assessment differently. Countries with few compulsory national examinations tend to encourage teachers to engage more in classroom assessments and FA procedures are used to a greater extent than SA (Barnes et al., 2015). In 2004, Brown invented a 50-item inventory called the Teachers' Conception of Assessment (TCoA) in which he proposed four conceptions of assessment: (i) assessment improves teaching and learning; (ii) assessment makes students accountable for learning; (iii) assessment makes teachers accountable; or (iv) assessment is irrelevant and negatively affects teachers, students and the curriculum. He used the TCoA to examine assessment beliefs of 525 New Zealand primary school teachers and managers. The respondents demonstrated their strong agreement to utilising assessment information mainly for teaching and learning improvement. A parallel sentiment is also shared by teachers from Finland (Shalberg, 2011), Spain (Brown & Remesal, 2012), the Netherlands (Segers & Tillema, 2011) and Fiji (Dayal & Lingam, 2015), indicating that assessment should be embedded in the teaching and learning process; their work is devoted to improving classroom practices. Obviously, it is noted here that in the contexts where assessment policy is predominantly focused on FA, teachers tend to view instructional improvement as the main purpose of educational assessments.

The critics over negative consequences of high-stakes tests have resulted in researchers introducing the idea of AfL, which is claimed to focus more on improvement-oriented functions (Black & Wiliam, 1998b, 2010; The Assessment Reform Group, 1999; Stiggins, 2002). Central to this is Black and Wiliam's (1998b) seminal work, *Inside the Black Box*: *Raising standards through classroom assessment*, which presented an extensive meta-analysis of 250 studies on FA. This literature review championed the potential use of FA and demonstrated that FA can result in significant academic gains for students. Motivated by this, Leighton, Gokiert, Cor and Heffernan (2010) administered a

survey to 272 grade 7–12 teachers in Canada, aiming to compare teachers' beliefs about classroom tests and large-scale tests. The research was initiated based on the authors' arguments that most assessments, particularly traditional tests and large-scale assessments, are not "cognitively diagnostic" (p.7) because of the lack of integration between assessments and empirically derived models of learning. The survey reported teachers' appreciation of the potential of classroom assessments. They viewed FA procedures as providing more valuable diagnostic information concerning students' learning process and pedagogical strategies for performance improvement. Similarly, responses to the questionnaire from 67 lower and middle school teachers in the UK identified a wide range of perceptions about this assessment approach (Sach, 2012). The key finding reported teachers' strong agreement about the promising impacts of FA to support teaching and learning.

Alongside the global winds of change, the education landscape in Malaysia has also experienced assessment reforms with the introduction of the SBA. An investigation of teachers' perception of and readiness for the SBA has become the most explored topic since its implementation in 2003. Generally, the findings showed mixed reactions from teachers. Several studies revealed that teachers responded positively to this assessment approach; acknowledging its positive effects in enhancing teaching and learning (Chan & Sidhu, 2006; Che Md Ghazali, 2016; Md Omar & Sinnasamy, 2009; Sidhu et al., 2011; Mohammad Radzi & Md Sawari, 2016). Additionally, teachers appreciated it as a way forward to move on from the intensity of examination-oriented culture (Malakolunthu & Sim, 2010; Mansor et al., 2013). However, there are teachers who expressed unfavourable opinions over the implementation of the SBA in schools. Studies indicated that teachers were not ready to accept this new shift away from the traditional testing system (Abdul Wahid et al., 2011; Jaba et al., 2013; Tuah, 2006). In a similar tone, some teachers disparagingly claimed that the SBA was "fundamentally idealistic" (Hashim et al., 2013, p.4) and was "just another test" (Abdul Majid et al., 2011, p.119). The inconsistency of the findings is seemingly caused by a series of challenges faced by teachers within their respective contexts.

The preceding findings generally suggest that teachers' beliefs about assessment are conditional on the present cultural and educational policy within a particular society. This means that the way teachers view assessment tends to align with the dominant uses and

purposes assigned to assessment in their respective country. Not only that, these empirical studies demonstrate the evidence that the dichotomous division of SA and FA is obviously contentious in practice.

3.2.2.2 Teachers' assessment practices

Assessment involves a wide spectrum of activities, ranging from the process of information gathering to the way the information is used for decision making about students' learning progress. A considerable amount of research has covered multifaceted issues concerning teachers' assessment practices, some of which are presented in the following section.

Teachers utilise a variety of assessment methods to collect information about students' learning progress. To examine this, Mertler (1998) conducted a survey for 625 elementary, middle and high school teachers in Ohio, asking their preferences regarding the use of traditional versus alternative assessment approaches. Mertler described multiple-choice, true/false, short answer, completion and essay as traditional assessments, while informal observation, performance assessment, portfolio and exhibition/recitals as alternative assessments. The finding revealed that teachers in middle and high schools used traditional assessment techniques more frequently than their counterparts in elementary schools. In a more recent study, Vlachou (2018) detailed findings from interviews and class observations of five science teachers that indicated the leading role of SA in Greek classroom practices. She claimed that the participants predominantly applied a teacher-directed approach, giving no room for students to play active roles in the assessment activities. Box, Skoog and Dabbs (2015) also reached the same conclusion, attributing the complexity of internally constructed beliefs concerning the accountability function and externally imposed constraints as mitigating factors for teachers to use FA in their instruction. Despite the claim that FA is effective in improving student learning, it seems that the practices of its use have not been fully embraced by teachers.

One of the most critical issues regarding assessment practice is the ability of teachers to effectively communicate assessment feedback to students. This is because the potential impact of feedback is crucial to scaffold students to be independent learners who are able

to monitor and take the necessary actions for the development of their own learning (Black, 2015; Evans, 2013; Hattie & Timperley, 2007; Wiliam, 2011). Therefore, teachers should be able to communicate clear and specific information about what students have achieved and what should be done to improve or remediate their performance (Stiggins, 1991). To examine this, Carless (2006) conducted a survey and interviews to elicit teachers' view on their practices of giving feedback to students. Teachers felt a lack of feedback on examination is normal because of its nature for student grading. They further commented that students were only interested in their marks and grades. As for the other forms of assessment, they were not able to provide detailed feedback due to large class size and lack of time. Another challenge surrounding the delivery of feedback is the mismatch of expectation between students and teachers about what entails "good" feedback (Carless, Salter, Yang & Lam, 2011; O'Donovan et al., 2016; Orsmond et al., 2013; Mulliner & Tucker, 2017). Mulliner and Tucker (2017), for example, reported that teachers were frustrated over students' unresponsiveness to the comments they received. Teachers felt that their feedback was fair, constructive and detailed. Contrastingly, students tended to find teachers' feedback ambiguous and lacking details on what could be done for improvement. While there is considerable evidence that feedback is potentially the most powerful factor for improvement of student achievement (Black & Wiliam, 1998b; Evans, 2013; Hattie & Timperley, 2007), the results of the above studies showed that the task of providing effective feedback remains a challenge for teachers.

The key issue underlying the problem of ineffective communication of feedback is closely associated with teachers' knowledge and skills about assessment (Even, 2005; Mertler, 2014; Stiggins, 2004). Arguably, accurate interpretation of assessment is certainly important to ensure the information is used effectively. As such, it is reasonable to advocate that assessment competency is a prerequisite skill that teachers should possess to be able to make informed decisions about students and instruction (Popham, 2009; Stiggins, 1995; Volante & Fazio, 2007). In several studies, however, teachers expressed their feeling of unpreparedness to adequately assess their students (e.g., Koloi-Keaikitse, 2012; Mertler, 1998, 1999; Sach, 2012; Siegel & Wissehr, 2011) and to implement assessment activities according to the requirements of educational agencies (e.g., Abdul Majid et al., 2011; Hashim et al., 2013; Md-Ali et al., 2015; Md Omar & Sinnasamy, 2009). Also, teachers admitted that they did not have an adequate understanding of basic

assessment concepts such as validity and reliability. For example, teachers participating in studies by Sekharan Nair et al. (2014) and Sidhu et al. (2011) revealed that they had difficulty in interpreting the SBA test score, particularly the use of the band category to describe student performance. Following this, they tended to engage in unreasonable leniency in grading and this could raise the issue of validity and reliability because of the varying interpretation of scoring criteria among teachers. Often, teachers attributed this deficit of assessment knowledge to their lack of professional training during pre-service programmes (see Mertler, 2003; Mertler & Campbell, 2005; Volante & Fazio, 2007) as well as in-service professional development programmes (see Mertler, 2005, 2009; Sidhu et al., 2011; Veloo & Krishnasamy, 2017). Aiming to examine the effectiveness of a two-week workshop on classroom assessments, Mertler (2009) elicited teachers' responses from the administration of pre-test and post-test questionnaire and teachers' reflective journal. The study demonstrated that training was highly beneficial for teachers in guiding them to engage in better assessment practice. Similarly, the finding from Zhang and Burry-Stock (2003) showed that teachers felt that the training they received was helpful and practical for use in improving their assessment activities in the classroom. Collectively, these studies signal that adequate supports for assessment literacy should be given greater attention to ascertain teachers' deep understanding of assessment so that effective assessment practices can be achieved.

The above challenges faced by teachers in their assessment practices have called for an increased effort among relevant education stakeholders, particularly school leaders and policymakers, to remediate the problematic areas. If these issues remain unresolved, this may exasperate teachers and they may find it difficult to effectively integrate assessment in instruction. Consequently, the ultimate aim of teaching and learning improvement may not be established.

3.2.2.3 Belief–practice relationship

The above-mentioned studies have often examined assessment beliefs and assessment practices separately. Acknowledging the potential impact of teachers' beliefs on their actions and educational practices (Barnes et al., 2015; Delandshere & Jones, 1999; Jussim & Harber, 2005), several projects were undertaken to examine the link between teachers' conceptions of assessment and practices of assessment.

The associative link between teachers' beliefs and their assessment practices was addressed in an interview-based study conducted by Postareff, Virtanen, Katajavuori and Lindblom-Ylänne (2012). The findings from 28 interviews with teachers suggest that most of the respondents perceived assessments as "reproductive" (p.87). in the sense that it is used to measure how well students can repeat, describe or apply the knowledge gained from the content of study module. This conception was significantly reflected in their practices with the use of traditional methods, particularly the paper-and-pencil examination, at the end of each module.

Furthermore, the correlation between perception and practice was also demonstrated by teachers in Hong Kong. Brown et al. (2009) employed the TCoA and Practice of Assessment Inventory (PrAI) to examine the linkages between teachers' self-reported conception of assessment to their self-reported assessment practices. The respondents, 300 primary and secondary teachers, agreed with the notion that an assessment functions to improve teaching and learning and these could be achieved by making students accountable to perform well in the examination. An analysis of their reported practices showed a strong use of pedagogical strategies for coaching students to prepare them for examination. Correspondingly, teachers in a study by Varatharaj, Abdullah and Ismail (2015) also indicated that their beliefs about assessment were congruent with their practices. The study aimed to explore perceptions about teacher autonomy and its effects in implementing the SBA. In the study, teachers who positively viewed the notion of autonomy in implementing the SBA appeared to have a positive attitude in their assessment practices. In contrast, teachers who felt they were not highly autonomous seemed to implement assessment in their instruction differently. They commented that the centralised curriculum limited freedom in teaching and learning process and they had to finish the syllabus for examination purposes. Based on the consistency of results demonstrating the direct link between beliefs and practices, it is concluded that teachers' practices of assessment are contingent on how they perceive assessment. This is in line with the findings of other studies (Dixon, Hawe & Parr, 2011; Jaba et al., 2013; Md Omar & Sinnasamy, 2009; Reimann & Sadler, 2017).

Conversely, the findings of the above studies are different from what is experienced by teachers in Indonesia (Azis, 2015). Using mixed-method data from a questionnaire and

semi-structured interviews, Azis (2015) found that teachers agreed with the assertion that assessment information should be utilised primarily for teaching and learning improvement. However, teachers admitted that the country's education policy, which emphasises the nationwide examination, is seen to be a hindrance to their efforts to fulfil the perceived function of assessment. Furthermore, the lack of correspondence between assessment beliefs and practices is also shared by teachers in England (James & Pedder, 2006) and in Turkey (Büyükkarcı, 2014). In James and Pedder's (2006) study of 558 English teachers, it is reported that teachers acknowledged the benefits of classroom assessment to facilitate learning development and to promote student autonomy. Their reported practices of assessment activities nevertheless indicated misalignment with the perceptions that they have. Büyükkarcı's (2014) research also documented a similar gap between what is perceived and what is practised. In this study, it seemed that Turkish primary teachers endorsed formative assessment as a sound method of promoting improvement in teaching and learning. Despite the positive view, teachers avowed to use assessment mostly for summative purposes. In effect, these findings challenge the notion of the direct influence of beliefs on one's actions. It is apparent here that teachers' beliefs about assessment are disconnected from their practices.

The above studies, to some extent, support the prior claim that the relationship of belief–practice is complex in nature. The interaction between beliefs and reported practices vary across individuals and contexts. Arguably, this potential disparity between teachers' assessment conceptions and practices seems to emerge in the context where the government adopts a more performance-oriented or learning-oriented education policy (Brown & Remesal, 2012; Büyükkarcı, 2014; James & Pedder, 2006). Moreover, some of the studies imply that the presumption of a direct cause between beliefs and teachers' educational practices is disputable and non-conclusive.

3.2.2.4 Variables influencing beliefs and practices of assessment

Research on teachers' assessment perception as well as practice of assessment suggests that the two domains are very much impacted by external factors. This is to say, there are several variables, either related to teachers' individual characteristics or other external factors, which may mediate how teachers perceive assessment and how they use it in their pedagogical planning.

Teachers' beliefs differ inter-individually. Some research has focused on how teachers' instructional beliefs and practices including assessment activities are related to their individual characteristics. The most prominent teacher variable in this context is teaching experience (see Alkharusi, Aldhafri, Alnabhani & Alkalbani, 2012; Fong & Muhamad, 2017; Koloi-Keaikitse, 2012; McMullen, 1999; Mertler, 1998, 2003; Rosas, 2014; Sach, 2012; Sahinkarakas, 2012; Vidacovich, 2015). For example, in a study of 691 primary and secondary school teachers in Botswana, Koloi-Keaikitse (2012) found that more experienced teachers had a more positive attitude towards assessments when compared with novice teachers. This is explained by higher levels of confidence in their ability to engage in assessment activities such as grading and constructing assessment tasks by more experienced teachers. Similarly, Fong and Muhamad (2017) inferred that teaching experience also involves teachers' preparedness in implementing the SBA. However, Mertler's (1998) study showed no significant relationship between teachers' years of teaching and their implementation of assessment activities.

It was also found that teachers' educational level may have an impact on how they perceive assessments and how they put assessment information into use. Studies by McMullen (1999) and Koloi-Keaikitse (2012) demonstrated that teachers with a higher educational background were more likely to favour the learning-oriented function of assessment. The finding also highlighted that the less educated teachers felt less confident about their assessment skills and practices, presumably because they had not received adequate professional training in assessment. In contrast, other studies reported no significant link between teachers' academic qualifications and their assessment beliefs and practices (see Mehrgan, Hayati & Alavi, 2017; Rosas, 2014).

Although this sub-topic is largely under-researched in the literature, associations have been found between the position teachers hold in school and their beliefs and practices about assessments. Rosas (2014) conducted a study to examine the difference between teachers and principals regarding their assessment literacy. The finding indicated that the administrators, on average, exhibited a higher level of knowledge to administer, grade and interpret the assessment information than teachers. This is in line with the previous research by Impara, Plake and Fager (1993a, 1993b) and Perry (2013), implying that teachers were less knowledgeable and skilful in their assessment practices in comparison

to their superiors. It was reported that teachers' poor assessment knowledge and skills is due to a lack of assessment support and training.

Apart from the individual characteristics of teachers themselves, the school context in which they operate might also have an impact on teachers' beliefs and practices regarding assessments. To examine this, Mertler (1998) administered a questionnaire to 625 teachers in urban, suburban and rural schools in Ohio, asking about their assessment practices, particularly the use of assessment techniques. With respect to the use of traditional assessment techniques (such as matching, true/false and multiple choices items), there was no difference between teachers in urban, suburban and rural schools. Teachers in suburban schools, however, utilised alternative assessments more often than their rural counterparts. Despite the claim that school factors may determine teachers' instructional beliefs and practices, investigations about the influence of this element on the beliefs and practices of teachers regarding assessment are very limited. The focus of inquiry has centred on the relationship between school factors and other facets such as teachers' classroom management (Rahimi & Asadollahi, 2012) teachers' efficacy and goal orientation (Rubie-Davies et al., 2012) and student performance (Hassan & Rasiah, 2011).

The research cited above seem to suggest that there are several external variables that are seen as being influential in terms of assessment beliefs and practices of teachers. As teachers have different individual characteristics, e.g., teaching experience, educational background and position, they are more likely to differ in their beliefs and practices regarding assessments. Furthermore, school contextual variables also appear to correlate with the differences in teachers' perceptions about assessment and their actual practices of assessment in instruction. The above conclusions suggest that diverse research outcomes often also have to do with the contexts that are likely to differ considerably from one country to another.

3.3 Summary and Gaps of Previous Studies

The above review supports the notion that a greater understanding of teachers' assessment beliefs and practices may lead to greater educational success. It is suggested that the beliefs teachers hold may affect their instructional practices (e.g., decision-making),

which in turn may influence the learning experience for their students. Acknowledging the important nexus between teachers' belief and practice, it is argued that this topic should be on the research agenda (Borg, 2001; Deneen & Brown, 2016; Nespor, 1987; Opre, 2015). In line with this argument, the outline of the research scopes that highlight teachers' assessment beliefs and practices is discussed in the following section. In doing so, I start with the summary of what is known in the literature based on the findings of the previous studies. Later, I specify what this study is intended to investigate to address gaps in our understanding as indicated in the literature.

Based on the review of previous studies, four main themes emerge: (a) the use of SA and FA; (b) the challenges of assessment practices; (c) the alignment of beliefs and practice; and (d) the influence of contextual factors on teachers' beliefs and practices.

Firstly, the focus of previous studies on teachers' assessment beliefs and practices has mainly centred around examining the use of SA and FA. It is indicated that teachers' beliefs and practices of assessment are contextually defined, highlighting the nature of cross-sample differences. In an environment where FA is primarily used (e.g., New Zealand and Finland), teachers believe in instructional improvements as the key function of assessment, while in SA-dominant countries like China and Egypt, teachers regarded assessments to be more about accountability and evaluation of teachers and schools. It is also noted that the findings of the reviewed studies have unintentionally created a harmful portrayal of summative assessment as "bad" and formative as "good". Interestingly, this reflects the view that the distinction between SA and FA is contentious in theory as well as in practice (Harlen, 2005; Hayward, 2015; Stiggins, 2002; Taras, 2005). Secondly, many studies have also delved into a series of challenges faced by teachers in their assessment practices. This includes miscommunication of assessment feedback, lack of assessment knowledge, time constraints and heavy workloads. Many teachers admitted that these "flaws" require serious attention from the school administrators as well as relevant educational agencies.

Thirdly, the review has further revealed that the studies of teachers' beliefs about assessment and teachers' actual use of assessment are often investigated as separate items. Notwithstanding the established notion about the influential impact of teachers' beliefs in shaping their educational practices, this topic, i.e., the link between belief and practice,

remains relatively under-researched (Jaba et al., 2013; James & Pedder, 2006; Postareff et al., 2012). Moreover, the reviewed articles attempting to study the relationship between the two domains revealed inconsistent results, which is compelling evidence that teachers' beliefs do not necessarily have a direct causal impact on their actions. Corresponding to the idea that beliefs and practices are contextually defined, the last theme of the reviewed papers revealed that teachers' beliefs about assessment and the ways in which they put them into practice are likely to be related and moderated by the contextual diversity of the school and personal characteristics (Brown & Remesal, 2017; Koloi-Keaikitse, 2012; Rubie-Davies et al., 2012).

What is lacking in the literature? The present study aims to uncover two main issues that have not been fully explored in the literature. First, research into teachers' beliefs and practices about assessment is relatively new in Malaysia. To date, the focal point was predominantly on the beliefs and practices of the SBA. I acknowledge the efforts by the MOE towards assessment reforms, but the existence of other assessment tools that have been implemented over a long period and are still in use until today should not be ignored. In addressing this issue, the present study specifically aims to investigate teachers' beliefs and practices about the purposes and uses of the currently used diagnostic test, i.e., F1DT. To my knowledge, the use of F1DT has never been addressed in academic research despite its long-standing implementation in Malaysian schools. Thus, there is a pressing need to understand its relevance in the current environment, in which great emphasis is given to students' and/or school performance. Such research has to provide empirical evidence to be used by the relevant agencies in their decision to either maintain, abolish, or replace the F1DT. In addition, it is also of interest in this study to explore the impact of various teacher characteristics and school contextual variables on teachers' thoughts and actions about the purposes and uses of F1DT.

Second, this study also responds to the need for more empirical studies investigating the direct association between the beliefs and practices of teachers regarding assessment. In the current context, I found that there are limited studies in this topic. Out of 22 reviewed studies, I found only three studies (i.e., Jaba et al., 2013; Md Omar & Sinnasamy, 2009; Varatharaj et al., 2015) attempting to directly link the relationship between teachers' beliefs and practices regarding assessment. Furthermore, previous research fails to provide clarity with regard to an alignment of belief and practice. So, it is safe to assume

that teachers' actions are not necessarily preceded by the conceptions they have. In light of this study, it is of interest to establish whether Malaysian teachers' beliefs about assessment are aligned with the actual utilisation of assessment information when making decisions for instruction. Also, can we be confident that Malaysian teachers will replicate the findings of studies in other countries (e.g., China, Indonesia and Turkey) with similar contexts in which the accountability function of assessment is prevalent?

Altogether, this study aims to explore (a) teachers' beliefs about the purposes and uses of F1DT; (b) their assessment practices; (c) the relationship between these beliefs and practices; and (d) the relationship between contextual variables and beliefs and practices.

3.4 Chapter Summary

In sum, this chapter has provided a review of relevant literature about critical issues related to teachers' assessment beliefs and practices. The first part of the chapter outlined the conceptual views on the significant interplay between assessment and the process of teaching and learning. The review indicated that scholars have reached an agreement that the fundamental purpose of assessment is to enhance teaching and to promote greater learning. The second part documented empirical works undertaken to investigate teachers' understanding of assessment, particularly looking into two important topics – teachers' beliefs and practices. The findings pointed out that assessment is a complex and dynamic process. This is reflected in the ways in which teachers perceive assessment and how they put them into action. Coupled with other variables influencing the beliefs and practices of teachers, it is concluded that the interaction of the two domains does not occur in a vacuum. In the last section, I summarised the findings of the reviewed articles, outlining what is lacking in the literature and how this study focuses on addressing the identified gaps. Most importantly, this literature exploration offered me background insights that are crucial for the development of an instrument to measure the issues of interest. The descriptive reporting of the questionnaire development is presented in the next chapter.

4 Pilot Studies: Development and Validation of the Survey of Educational Assessment (SEA)

The main argument brought forward is that teachers' understanding of educational assessment is crucial to ensure its outcome is fully utilised to promote informed and effective decision-making about instruction and students' learning. Underpinned by the notion that the ability to understand assessment is closely linked to teachers' belief systems (e.g., Barnes et al., 2017; Brown et al., 2008; Brown et al., 2015; Segers & Tillema, 2011; Stiggins, 2004), the first step of this research is an attempt to examine teachers' perceptions about the purposes and uses of F1DT, a diagnostic test used in Malaysian schools. Furthermore, scholars have also advocated that teachers' actual actions in a classroom are seen as a reflection of their beliefs (Barnes et al., 2015; Borg, 2001; Nespor, 1987; Opre, 2015; Pajares, 1992; Skott, 2015). Following this argument, it is also fundamental to investigate teachers' actual use of the information from this test in the teaching and learning process. As there is no existing instrument that fits the context of the study, this necessitates the development of a questionnaire to measure the issues of investigation.

This chapter presents a detailed description of two pilot studies, conducted between August 2015 and August 2016, to develop an instrument suitable for use in the main study. The chapter begins with the rationales of undertaking the pilot studies; emphasising the primary objectives why these small-scale projects are as vital as the main project. In the subsequent sections, I recount the experience of conducting the two pilot studies – denoted as Pilot Study 1 (PS1) and Pilot Study 2 (PS2). In doing so, I report about the sampling strategy and the study participants, the process of developing the new questionnaire, the data collection procedures and data analysis as well as questionnaire revisions and modifications. At the end of the chapter, I relate my contemplative notes as a novice researcher, considering the experiences gained and the insights acquired throughout this one-year preliminary study.

4.1 The Rationale for the Pilot Studies

A pilot study was the first step for the current research. It is a scientific tool that allows researchers to conduct a preliminary small-scale investigation to learn what might or

might not work in a subsequent larger project. Many experts acknowledge that a pilot study is an essential precursor for many research projects (Gillham, 2008; Thabane et al., 2010; van Teijlingen & Hundley, 2001; van Teijlingen, Rennie, Hundley & Graham, 2001). Arguably, the findings of pilot studies are equally important, as they can inform researchers about the detailed successes and improvements to the process and design of the full study (van Teijlingen & Hundley, 2001; van Teijlingen et al., 2001). Furthermore, a pilot study may also be able to identify potential problems following the implementation of the research procedures (LaGasse, 2013; Thabane et al., 2010; van Teijlingen & Hundley, 2001).

Notwithstanding the wealth of insightful information that can be obtained, reports from pilot studies are generally under-reported and underappreciated (Beebe, 2007; LaGasse, 2013; van Teijlingen & Hundley, 2001; van Teijlingen et al., 2001). In fact, van Teijlingen et al. (2001) advocated that sharing such information was important in research, as "researchers have an ethical and a scientific obligation to attempt publishing the results of every research endeavour" (p. 6). Conventionally, a report of the pilot study is a part of the methodology chapter in the main study. However, I argue it deserves a separate chapter to demonstrate how the two pilot studies are instrumental to the success of the main study.

For the purpose of this research, the pilot studies primarily aimed to develop a research instrument that allows me to measure teachers' beliefs and practices about the purposes and uses of F1DT as a diagnostic measure in Malaysian schools. Additionally, the two pilot studies intended to test the feasibility of the questionnaire items, to see if there are particular technical problems or misunderstandings of the items before using them for the main study. Fraenkel and Wallen (2008) asserted that pilot studies provide the opportunity to reveal ambiguities pertaining to rubrics, unclear and poorly worded questions and the format of the questions. Most importantly, the two studies were conducted to test the validity and reliability of the research instruments.

Other than investigating the suitability of the instrument and the feasibility of conducting the main study, these studies were also planned to explore other practical implications that need improvement for the overall process of data collection in the main study. Thabane et al. (2010) emphasised that pilot studies are routinely performed to identify

potential problems that may affect the whole research effort. It is, therefore, essential to uncover the success or failure of the proposed research procedure, research schedule, data analysis methods and sampling techniques before embarking on the main fieldwork.

In short, the pilot studies were conducted to achieve the following objectives:

a) To develop an instrument that enables the capture of teachers' beliefs and practices about educational assessment
b) To conduct a psychometric analysis of the questionnaire to ensure the quality of the items
c) To elicit insights about the operational feasibility of the questionnaire
d) To identify amendments and improvements for the questionnaire and the proposed research processes of data collection

4.2 Pilot Study 1: Survey of Educational Assessment (SEA I)

The Pilot Study 1 (PS1) was launched in August 2015 as the first attempt to develop a questionnaire that aimed to measure teachers' perceptions and practices about educational assessment. As a starting point, I reviewed articles of relevant studies, which provided noteworthy insights that I used to design the questionnaire items and to adjust the items according to the specific context, i.e., culture, language and constraints.

Within the literature of teachers' beliefs, a plethora of studies has identified their beliefs about assessment and the practices of assessment information as one of the major focuses of investigation. From the literature exploration, as presented in the preceding chapter, there are several instruments designed to measure these issues quantitatively and/or qualitatively. Brown (2004), for example, invented his questionnaire called the Teacher Conception of Assessment (TCoA). He also extended the TCoA to a related inventory to measure teachers' practice about assessment, entitled the Teacher Assessment Practices Inventory (T-APrI) (Brown et al., 2009). Both inventories have been extensively used in empirical studies covering multiple countries that focus on teachers' beliefs and practices about assessment (e.g., Brown & Gao, 2015; Brown & Remesal, 2012; Brown et al., 2015; Gebril & Brown, 2013; Harris & Brown, 2009; Kyaruzi, Strijbos, Ufer, & Brown, 2018).

Another set of related works from which I generated questionnaire items came from different sources, mainly the works of Mertler (Mertler, 2003, 2005, 2009; Mertler & Campbell, 2005), and other researchers (Popham, 2009; Shin, 2015; Siegel & Wissehr, 2011; Volante & Fazio, 2007). For example, the Classroom Assessment Literacy Inventory (CALI) (Mertler, 2003) and Assessment Literacy Survey (Volante & Fazio, 2007) were used to examine the assessment literacy of in-service and pre-service teachers. What is common in these studies is that assessment literacy (knowledge of assessment) profoundly affects teachers in making an assessment-related decision. Noting the importance of this dimension, I included some of the items asking teachers' general perception about assessment literacy in education.

In PS1, the focus of the questionnaire was on perceptions of educational assessments in general. In the questionnaire, it was explicitly defined that an assessment is "a general term that refers to teacher-made classroom tests (i.e., quizzes and topical or monthly tests) and standardised tests (i.e., mid-term, final and trial examinations)". In many Malaysian schools, quizzes and topical/ monthly tests are often individually designed by subject teachers to assess students' progress on a particular topic(s). The term "standardised tests" in this questionnaire, on the other hand, specifically refers to low-stakes tests that are usually conducted at the end of each school term (mid-year (May) and end-year (October) terms) as summative tests. Typically, such tests are designed at school, district or state level by a group of subject teachers.

4.2.1 Method

4.2.1.1 Participants

Regarding the sample size of the pilot study, it appears that there is no consensus in the literature of how to determine an appropriate number of respondents (Hertzog, 2008; Johanson & Brooks, 2010; Julious, 2005; Thabane et al., 2010). Nevertheless, Hertzog (2008) pointed out that when the objective of the pilot test is to assess the internal consistency of an instrument, small sample size is certainly inadequate.

The questionnaire was administered in two versions, an online and paper-and-pencil version. Participants were recruited by ways of convenience sampling as a cost-effective approach to conduct the pilot study (Fraenkel & Wallen, 2008). The sample comprised

primary, secondary and pre-university teachers of rural (105 teachers) as well as urban schools (25 teachers). The paper-and-pencil version was distributed to two secondary schools in Sipitang, one of the districts in Sabah. The schools were chosen due to easy access and convenience to my fieldwork schedule. The online questionnaire was made available to primary and secondary teachers from Sabah and other states in Malaysia. After the termination of the study, responses of 130 teachers with an average teaching experience of 3.4 years (SD = 1.22) were collected. Most of the participants were female, accounting for 73% of the total sample. This is a reflection of the whole population where the teaching profession in Malaysia is dominated by females. Sixty-three teachers responded to the online version of the questionnaire while 67 completed the paper-and-pencil version of it. The respondents were largely from Sipitang and there was no traceable information about the location of those who responded online. A summary of the demographic characteristics of the respondents is presented in Table 4.1 below:

Table 4.1

Summary of Respondents' Demographic Information

Demographic Information	Characteristics	*N*
Gender	Male	35
	Female	95
Educational level	Diploma	3
	Bachelor	97
	Master	29
	PhD	1
Years of teaching experience	Less than 2 years	11
	Between 2 to 5 years	19
	Between 6 to 10 years	35
	Between 11 to 15 years	34
	More than 15 years	31
Teaching level	Primary	19
	Secondary	96
	Pre-University	15
School band	Band 1	2
	Band 2	7
	Band 3	14
	Band 4	14
	Band 5	86
	Band 6	2
Location of school	Rural	105
	Urban	25

4.2.1.2 Instrument

As a result of literature explorations, I developed the Survey of Educational Assessment (SEA) – a prototype set of 57 items which I adapted from various existing instruments. The initial version of the items was divided into two major parts – Perception of Assessment (PoA) and Practice of Assessment (ProA). Each contained items addressing four main underlying dimensions:

- Assessment usefulness for improvement of teaching and learning (Instructional Improvement)
- Assessment for teacher and school accountability (Accountability)
- Assessment as irrelevant to the works and life of teachers and students (Irrelevance)
- Assessment literacy

The questionnaire was divided into three sections – demographic information, PoA and ProA. The first section included items asking about respondents' background (i.e., gender, level of education, years of teaching experience, grade level of teaching assignment, position in the school, and assessment training) and school information (i.e., type of school, school location and school band). The second section consisted of 33 items and was designed to measure teachers' perception of educational assessment. It used a seven-point rating scale, ordered from agree to the left and disagree to the right. The third section comprised 24 items corresponding to the statements in the PoA section. This section was intended to examine how frequently teachers use assessment in their instructional decision-making using ordinal frequency rating scales – never, rarely, seldom, often and always. Table 4.2 summarises the content of SEA I prototype (see the full list of items in Appendix 4A).

Table 4.2

Content of SEA I

Dimensions:	Perception of Assessment (PoA)	Practice of Assessment (ProA)
Instructional Improvement	13 items	9 items
Accountability	5 items	4 items
Irrelevance	9 items	7 items
Assessment Literacy	6 items	4 items
Total number of items	33 items	24 items
Rating scale	Level of agreement Seven-point	Ordinal frequency Five-point

Another central topic to be discussed here is that of questionnaire translation. The context in which this instrument was intended to be used made it necessary to consider the language issue carefully. Therefore, an instrument translation is a crucial element in item development for non-English respondents. It is essential to avoid potential errors in the accuracy of the meaning of the source language (Maneesriwongul & Dixon, 2004) and to eliminate contextually sensitive language barriers (Esposito, 2001). Acknowledging the process of translation is not a simple task; a Malaysia-based certified translation agency, was appointed to translate all the items into the Malay language, the target language of the respondents. After that, the translated version was reviewed by a bilingual lecturer who works at one of the public universities in Malaysia. The revision by an academician helped to ensure that the technical terms used in the questionnaire were translated appropriately. Also, it allowed for the detection of any grammatical and syntactical errors in the piloted items. This collaborative work, technically called multi-forward translation, between the language expertise and the researcher, may produce a more informed decision about the suitability of the language for the instrument (Erkut, 2010).

In addressing the issue of translation equivalence, the term "perception" was found to be the most appropriate for use in the questionnaire. As pointed out in the previous chapter, the terms "conception" and "belief" do not accurately reflect the meaning of teachers' belief which frames the scope of this research. To eliminate confusion, the term "perception" is used throughout this chapter and at any point referring to the questionnaire.

4.2.1.3 Procedure of data collection

It is obligatory for research students to obtain an ethics clearance from the Ethics Committee of School of Education, Durham University. Bryman (2012) highlighted that ethics approval is fundamental before conducting research to avoid harm to the participants and lack of consent and to protect respondents from deception and invasion of privacy. For this study, I submitted the ethics application containing research objectives, research design, research instrument, recruitment strategy of sample and data management information. The application was approved by the School of Education Ethics Committee on 8 July 2015 (see Appendix 4B). As a part of the ethical procedures, I also wrote emails to the authors asking for their consent to use the inventories in my study (see Appendix 4C for one of the email correspondences).

In PS1, I utilised two methods of data collection – paper-and-pencil and online survey. Using different strategies of collecting data allowed me to evaluate the operational feasibility of the questionnaire, i.e., to find out if one was more effective than the other. The former was deployed to teachers in two rural schools in Sipitang. It was a drop-off survey; a representative for each school was appointed to collect the questionnaire. Prior to undertaking the pilot study, the principals of the selected schools were contacted to get their permission to conduct the study at their respective schools.

The second method was seen as a practical solution, taking into account the geographical factors of Malaysia, especially of Sabah as the main research site of the actual project. Because of the limitations of travel cost and time, this approach was suitable and convenient, as it enabled me to reach teachers in a wider radius of schools, particularly those in other districts outside Sipitang and other states outside Sabah. Not only did the online questionnaire offer greater coverage of location, but it also allowed me to capture the heterogeneity of the sample within and across districts. For this purpose, the questionnaire was made available online via Bristol Online Survey (BOS)[4].

With regard to the protection of the anonymity and confidentiality of the respondents, information containing their identities (e.g., name, address and name of school) were not included in the questionnaire (see Appendix 4D for a participant information sheet and consent form). As Crow and Wiles (2008) pointed out, the researcher should not by any means make the respondents identifiable; they need to be pseudonymously described in the questionnaire and the project report . The data collected was stored securely in a password-protected file, ensuring that no one outside this project had access to the information provided by the respondents.

4.2.1.4 Data analysis

After the completion of the online and paper-and-pencil questionnaires, the data were analysed using Statistical Package for the Social Sciences (SPSS) version 20. I conducted an item analysis to ensure the quality of the items and to explore the dimensional/factorial structure of the questionnaire. Furthermore, this psychometric analysis allowed me to

[4] A web-based survey tool that allows researchers to develop, deploy and analyse data. The tool, developed by the University of Bristol, is offered free to Durham University students and staff for research purposes.

identify items that appeared to be ambiguous. The process of this item analysis helped me to revise and refine the questionnaire to be ready for use in the main study.

As SEA is a new instrument, it is therefore important to conduct an analysis of internal structure (dimensionality) and internal consistency (reliability) to assess the quality of the questionnaire. Creswell (2013) emphasised the need to re-establish a good internal structure and internal reliability for any modification of the existing instruments, as the original analyses may not hold for the new questionnaire. To ascertain whether responses to items compiled in SEA I can be meaningfully combined to scores reflecting teachers' perception of use and assessment-related practices, respectively, two separate Principal Component Analysis (PCA) were conducted.

In the present study, the use of PCA was more appropriate than its counterpart Factor Analysis (FA). PCA is often mistakenly considered as a type of FA when in fact they are different statistical methods conducted to achieve different objectives (Brown, 2009; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Park, Dailey, & Lemus, 2002; Preacher & MacCallum, 2002; Tabachnick & Fidell, 2013). FA is often employed to reveal latent constructs (i.e., factors) of observed variables, and thus it is associated with theory testing (Brown, 2009; Park et al., 2002; Tabachnick & Fidell, 2013). PCA, on the other hand, produces components that are simply aggregates of correlated variables with no prior underlying theory to be associated with each component (Brown, 2009; Park et al., 2002; Tabachnick & Fidell, 2013). If the objective of the analysis is to explore the relationship of the variables and to reduce them into smaller groups of meaningful components, it makes more sense to perform PCA (Brown, 2009; Park et al., 2002; Rattray & Jones, 2007; Tabachnick & Fidell, 2013). In this questionnaire development, although I had four dimensions in mind, I only planned to explore the observed variables that emerge from all the analyses instead of verifying prior theoretical assumptions about the underlying latent constructs of the variables being investigated.

Reliability is another essential psychometric property in the development of questionnaire items. This refers to the overall consistency of an instrument – the extent to which items of the instrument tend to load together to measure the same construct (Bryman, 2012; Cohen, Manion, & Morrison, 2011; Rattray & Jones, 2007). In this study, the internal

consistency index Cronbach's Alpha coefficient was utilised as an indication of reliability.

4.2.2 Results

Following the procedures used by Brown and his colleagues (e.g., Brown, 2004; Brown & Gao, 2015; Brown & Remesal, 2012; Gebril & Brown, 2014) I analysed the inventories – PoA and ProA – separately.

4.2.2.1 Perception of Assessment (PoA)

Prior to conducting PCA an assessment of the suitability of the data for this technique was conducted, looking at two statistical measures – the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity. It is expected that the KMO value should be ≥.5 to verify that the sample used for the analysis is adequate (Field, 2013; Tabachnick & Fidell, 2013). As shown in Table 4.3, the KMO value for PoA was .834, which is well above the minimum rule of thumb, and Bartlett's test was significant ($p < .05$, $df = 528$). Given these overall indicators, PCA was conducted with all 33 items.

Table 4.3

KMO and Bartlett's Test of Sphericity of 33-item Dataset (n=130)

KMO and Bartlett's Test		
Kaiser–Meyer–Olkin Measure of Sampling Adequacy		.834
Bartlett's Test of Sphericity	Approx. Chi-Square	2177.343
	df	528
	Sig.	.000

In line with the assumption that the underlying constructs, i.e., the four underlying dimensions, are independent (Brown et al., 2009), Varimax rotation was used to extract the maximum variance from the dataset and to reduce it into smaller groups of related components. According to Kline (1994), items are accepted as belonging to their intended component only when their loadings are ≥.30, and other loadings that are less than the recommended index should be ignored. The output in Table 4.4 lists the eigenvalues associated with each component, revealing seven components exceeding the value of 1. The first three components explained the relatively large amount of variance, collectively amounting to 50.55% of the total variance. The remaining four components, however,

only accounted for small variances, with 4.9%, 4.3%, 4.0% and 3.5% of the variance respectively.

Table 4.4

Total Variance Explained for 33-item Dataset (n=130)

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.75	32.601	32.601	10.75	32.601	32.601	7.434	22.52	22.527
2	3.733	11.311	43.912	3.733	11.311	43.912	4.174	12.64	35.176
3	2.188	6.631	50.544	2.188	6.631	50.544	2.557	7.748	42.923
4	1.601	4.851	55.395	1.601	4.851	55.395	2.350	7.122	50.046
5	1.410	4.274	59.669	1.410	4.274	59.669	2.127	6.446	56.491
6	1.315	3.986	63.655	1.315	3.986	63.655	1.987	6.020	62.512
7	1.146	3.472	67.127	1.146	3.472	67.127	1.523	4.616	67.127

Extraction Method: Principal Component Analysis

The detail of the initial item loadings is displayed in Table 4.5. All components, with the exception of Component 7, managed to extract ≥3 items with a loading index above .3. It is noticeable that Component 1 had the most loading with 13 items, followed by Component 2 (6 items) and Component 3 (4 items). This is why these three components accounted for most of the variance. Note that one variable loaded on Component 7. Q21 was therefore excluded from the subsequent analyses (Field, 2013; Pallant, 2013). It is also clear that there were many items cross-loaded into more than one component, particularly items Q38, Q40 and Q36. These items are recommended to be discarded as they may be an indication of problematic items due to poor or confusing wording (Field, 2013; Pallant, 2013).

Table 4.5

Component Loadings of 33-item Dataset (n=130)

Items	Components						
	1	2	3	4	5	6	7
Q11	.816	.339					
Q10	.781	.317					
Q19	.754						
Q31	.742	.409					
Q27	.730	.355					
Q25	.718						
Q22	.710						
Q13	.692	.510					
Q14	.685	.326					
Q26	.646	.389					
Q24	.593						
Q23	.552	.301					
Q38	−.480		.458				.338
Q12		.789					
Q17		.752		.301			
Q30		.630					
Q16		.625			.345		
Q15	.537	.557					
Q20	.451	.543					
Q37			.784				
Q35			.746				
Q29			.635				
Q32			.474				.452
Q34				.816			
Q33	.418			.644			
Q28			.386	.595			
Q39					.843		
Q42				.364	.599		
Q40	.335			.330	.517		
Q36		.383			.511	-.485	
Q18						.764	
Q41						.642	
Q21							.837

Note: Factor loadings <.3 were suppressed.

The same procedures were repeated as an attempt to look for a more defined component structure, which resulted in the omission of a total number of nine items – Q21, Q38, Q40, Q36, Q39, Q42, Q18, Q41 and Q20. This is because these items did not contribute to a simple component structure and/or had three cross-loading values.

At the final stage, the analysis was run for the remaining 24 items and four components were identified. The values of KMO and Bartlett's test (KMO = .869, p = .000) guaranteed factorability for the PoA. A scree plot in Figure 4.1 illustrates the number of

components. As shown, the line of the graph starts to smooth up after the point of inflexion, suggesting the existence of the four components. From the fourth component on, it is noted that the line is becoming flatter, which means that each successive component explained smaller amounts of the total variance.

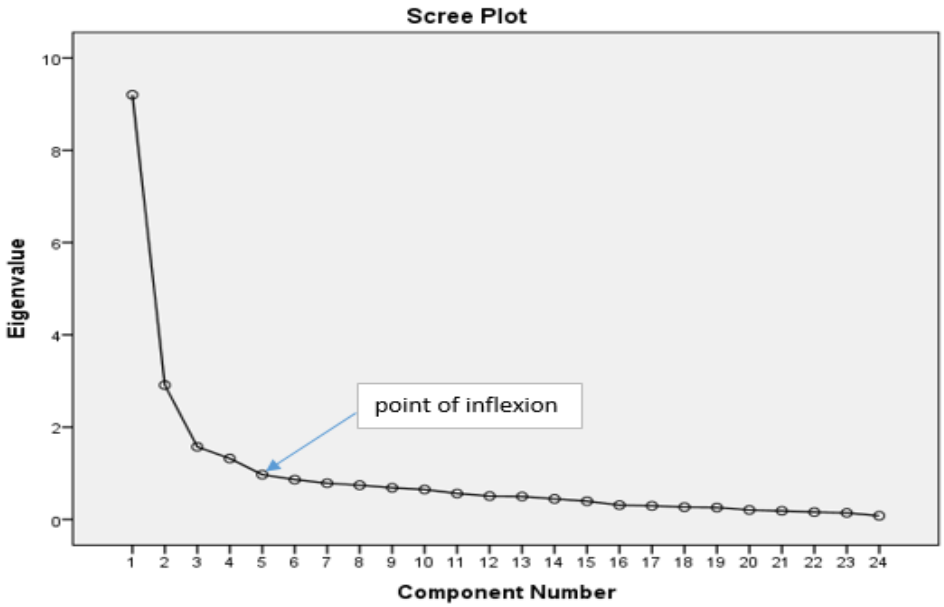


Figure 4.1: Scree Plot for 24-item Dataset

As seen in Table 4.6, the final solution yielded a better four-component structure and each component was labelled using a thematic approach (Briggs & Cheek, 1986; Ford, MacCullum & Tait, 1986; Kline, 1994). Many items were highly loaded to a particular component with a loading index range between .481 and .786. Cross-loading items decreased and none of the items belonged to three components. Nonetheless, a considerable cross-loading is apparent for items in Component 1 and Component 2, making it difficult to label these two components. Eight items from Component 1 also tapped into Component 2. This is hardly surprising because many of these items could be described as the benefits of assessment for students. This is exemplified in Q10, Q11 and Q14.

Table 4.6

Component Loadings of 24-item Dataset (n=130)

Items	Components			
	1 (11 items)	2 (6 items)	3 (4 items)	4 (3 items)
Benefit of assessment for instruction and teachers' assessment literacy				
Q25 – assessment knowledge and skills	.786			
Q19 – accurate interpretation of assessment information	.785			
Q11 – improve students' learning	.747	.457		
Q22 – cautious use of assessment results	.735			
Q10 – assessment is beneficial to students	.701	.466		
Q27 – plan strategies for teaching and learning	.685	.439		
Q26 – integrate assessment into teaching practices	.658	.402		
Q31 – how much students have learnt from teaching	.655	.559		
Q24 – teachers determine assessment quality	.602	.315		
Q14 – benefits high-achieving students	.525	.524		
Q23 – accommodate students' need individually	.481	.322		
Benefit of assessment for students				
Q16 – enjoyable experience for students		.782		
Q17 – measure students' actual mastery level		.776		.333
Q12 – fair to underperforming students		.735		
Q15 – motivate students to learn	.376	.676		
Q13 – identify students' strengths and weaknesses	.547	.667		
Q30 – focus more on students' learning progress		.653		
Irrelevant use of assessment				
Q35 – teaching for test is detrimental to learning			.743	
Q37 – focus more on students' scores			.724	
Q29 – teaching for test limits teachers' creativity			.692	
Q32 – teach against teachers' beliefs			.653	
Assessment for testing preparation				
Q34 – performance assurance for public examinations				.838
Q33 – obtain good results in public examinations	.350			.667
Q28 – controls the way teachers teach			.437	.556
% of Eigenvalues	38.3	12.1	6.6	5.5
Cronbach's Alpha (α)	.936	.840	.696	.725

Note: Factor loadings <.3 were suppressed.

Notwithstanding the above results, an examination of the reliability of this dataset revealed that Cronbach's alpha ranged between .69 and .94. This suggests that the estimation of internal consistency of the items in their respective components was adequate. However, the interpretation of Cronbach's alpha should be treated cautiously here, taking into account the small sample size of this study.

4.2.2.2 Practice of Assessment (ProA)

Initially, the appropriateness of the 24-item of ProA was examined through the KMO and Bartlett's test of sphericity. However, SPSS output of the analysis warned that "this matrix

is not positive definite", implying that a prerequisite condition for proceeding with PCA has not been met. Field (2013) explained that the most likely explanation for this situation is probably due to imbalanced numbers between the cases (sample) and the variable, causing instability in the correlation matrix. Collecting more data is suggested as a solution to rectify the problem (Field, 2013). Alternatively, it is recommended to consult a correlation matrix to check which items are correlating poorly with other items (Pallant, 2013). These items need to be excluded to allow for the rotation to converge. I chose the latter as that is the most feasible option.

Using the factor correlation matrix, I ran the analysis for few times and this led to the elimination of 16 items from the dataset. In the example shown below (Table 4.7), it is seen that Q55a and Q61a had low correlations with other items and were therefore discarded from the analysis.

Table 4.7

Factor Correlation Matrix of ProA (n=130)

Correlation Matrix										
		Q49a	Q50a	Q53a	Q54a	Q55a	Q56a	Q61a	Q62a	Q63a	Q65a
Correlation	Q49a	1.000	.522	.459	.367	−.109	.442	.116	.419	.363	.332
	Q50a	.522	1.000	.495	.458	−.056	.398	.224	.479	.265	.334
	Q53a	.459	.495	1.000	.609	.049	.442	.155	.558	.311	.334
	Q54a	.367	.458	.609	1.000	.065	.372	.256	.612	.319	.374
	Q55a	−.109	−.056	.049	.065	1.000	.189	.245	.109	.039	.189
	Q56a	.442	.398	.442	.372	.189	1.000	.177	.412	.425	.238
	Q61a	.116	.224	.155	.256	.245	.177	1.000	.400	.159	.216
	Q62a	.419	.479	.558	.612	.109	.412	.400	1.000	.416	.371
	Q63a	.363	.265	.311	.319	.039	.425	.159	.416	1.000	.335
	Q65a	.332	.334	.334	.374	.189	.238	.216	.371	.335	1.000

At the final stage of analysis, the KMO and Bartlett's tests significantly estimated that only eight items were eligible for factorability (KMO = .876, $p < .001$). A scree plot of the component loading is illustrated in Figure 4.2 below, showing that only one component was extracted from the analysis.

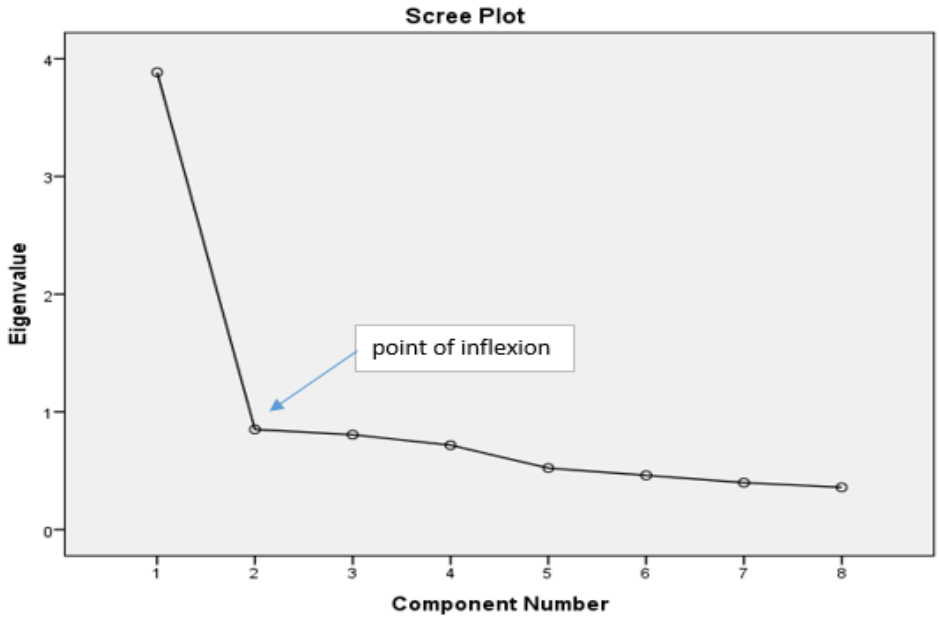


Figure 4.2: Scree Plot for 8-item Dataset

Table 4.8 details the item loading of this one component, which amounted to approximately 48.6% of all the variance. A reasonably high Cronbach's alpha statistic of .844 seems to imply that these eight items were internally consistent measuring the same construct. It is noted that items tapping on this one-component solution were similar to Component 1 of PoA.

Table 4.8

Component Loadings of 8-item Dataset (n=130)

Items	Component 1 (8 items)
Benefit of assessment for instruction and teachers' assessment literacy	
Q62a – identify students' strengths and weaknesses	.776
Q53a – improve students' learning	.773
Q54a – how much students have learnt from teaching	.754
Q50a – plan strategies for teaching and learning	.718
Q49a – integrate assessment into teaching practices	.696
Q56a – cautious use of assessment results	.661
Q63a – accurate interpretation of assessment information	.594
Q65a – predict students' future academic performance	.573
KMO	.876
Bartlett's test of sphericity	.000
% of Eigenvalues	48.6
Cronbach's alpha (α)	.844

Based on the above findings, it can be summarised that this dataset failed to demonstrate the expected component structures that were conceptualised prior to undertaking the data collection.

4.2.3 Reflective conclusion of PS1

The above analyses indicated that the research instrument was in need of item and procedural revisions before finalising it for the main study. In the following paragraphs, several specific alterations of SEA I are highlighted.

From the findings of the analyses, it is indicated that the items of SEA I seemed to behave differently from its underlying conceptual constructs. Presumably, this situation could be attributed to several "flaws". Firstly, the core point that requires a great deal of attention is the definition of assessment. SEA I operationalised assessment is the general term referring to teacher-made classroom tests and low-stakes standardised tests, directed at both formative and summative assessments. The definition perhaps created misunderstandings for the respondents. Presumably, there was a likelihood that teachers might be indecisive on how to respond to the statements in the questionnaire, as "assessment" was defined to denote two types of assessment that were apparently designed with different purposes. Thus, identifying a specific assessment battery is necessary to offer a clear direction for the respondents to provide more transparent answers.

Secondly, a closer look at the item level revealed that there were several concerns regarding clarity of the items. Table 4.9 lists a few examples of the items identified to be "problematic".

Table 4.9

Examples of Problematic Items

Issue	Explanation	Item
Item redundancy	The items have similar meaning	Q66 I conduct assessments but make little use of the result (Pr) Q59 I put little effort to use assessment information in my instructional decision-making (Pr)
Item ambiguity	The wordings of the items are more perception statements rather than practice statements	Q55 Assessment controls the way I teach (Pr) Q60 Assessment forces me to teach against my beliefs (Pr)
Item wording inconsistency	There is no wording consistency in these two corresponding items of perception and practice	Q21 I think assessment outcome is difficult to understand (P) Q48 I cannot understand assessment outcomes easily (Pr)
Inappropriate target respondent	These items are meant to be answered by students	Q16 I think assessment is an enjoyable experience for students (P)

With regard to the operational feasibilities of the data collection process, I identified that the size of the sample and the method of data collection are two crucial issues that needed further improvements for the next study. Firstly, an adequate number of samples is required as small samples tend to produce unstable solutions (as reflected in the above analyses). Secondly, only the utilisation of a paper-and-pencil questionnaire seemed practical as a method of data collection. Conducting an online survey could turn out challenging, as many potential respondents might not have easy access to the internet.

Taken the above noteworthy insights altogether, it is concluded that SEA I needed a further thorough and major modification to create a working version in which the instrument could yield a more meaningful and coherent underlying construct of the issues under investigation. Development of new dimensions and the inclusion of new items are required. Also, proper planning of methodological procedures is another concern that should be taken into due consideration for the manageability of the next stage of data collection. The subsequent section, therefore, moves on to account the details of the item reconstruction process and validation for a new version of SEA.

4.3 Pilot Study 2: Survey of Educational Assessment (SEA II)

In reflection of the item revision in PS1, it was decided that a development of new items was necessary to address the issues under investigation appropriately. As the primary objective of this research is to investigate the potential effects of familiarising teachers with the concept of DT on their perception and practice of educational assessments, it was crucial that the research instrument was able to capture the changes, if any. Thus, an explicit reference to an assessment tool is necessary to give direction for teachers to make a comparison with the information about DT. The use of F1DT appeared to serve this purpose.

Based on the findings of PS1, the emphasis for PS2 was not simply to refine the item pool, but rather to start with a refinement of the conceptualisation. Departing from Brown's (2004) conceptualisation of the conception of assessment and its practice, new sources of conceptual and empirical works about educational assessments (e.g., Black & Wiliam, 1998; Corretjer, 2016; Hayward, 2015; James & Pedder, 2006; Leighton, Gokiert, Cor & Heffernan, 2010; Newton, 2007) and in particular about diagnostic assessment (e.g., Appleby, Samuels & Treasure-Jones, 1997; Black, 1983; Gorin, 2007; Simpson & Arnold, 1983; Treagust, 1988, 2001; Van der Kleij, Vermeulen, Schildkamp & Eggen, 2015) were consulted. As a result, a new conceptual framework, as illustrated in Figure 4.3, would be used to measure teachers' perception about the purposes and uses of F1DT. It consisted of two main dimensions – perception and practice.

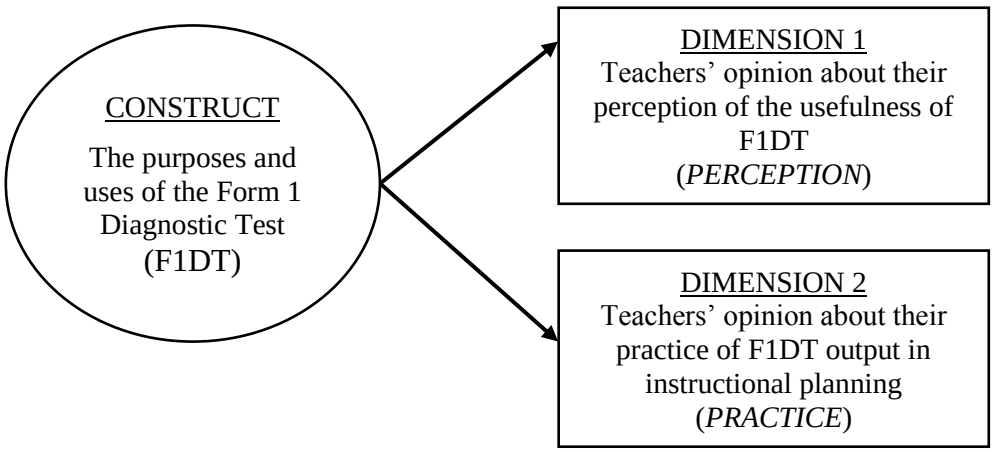


Figure 4.3: Underlying Dimensions of SEA II

4.3.1 Method

4.3.1.1 Participants

The decision concerning the sample size was determined by the sample requirements to conduct PCA. Adequate minimum sample size is important to yield meaningful results in achieving precise and interpretable estimates of component loadings (MacCallum, Widaman, Zhang & Hong, 1999; Preacher & MacCallum, 2002). Several guidelines were proposed to select participants either based on the minimum sample size *n*, or the minimum ratio of the subject *n* to the number of variables analysed *p*. Specifically, a commonly used suggestion about the absolute minimum *n* can range between at least 100 and 300 respondents (Comrey & Lee, 1992; Fabrigar, Wegener, MacCallum & Strahan, 1999; Gorsuch, 1983; Guadagnoli & Velicer, 1988). Alternatively, others have proposed the *n:p* ratio as a determinant component for a sample size decision, ranging between a ratio of 5:1 (Comrey & Lee, 1992; Gorsuch, 1983) and ten cases per variable (Everitt, 1975; Nunnally, 1978). Following the above methodological requirements, 472 teachers were approached to participate in PS2. By the end of the study, 260 completed questionnaires were returned, resulting in a *n:p* ratio of nine subjects per variable (N = 260, p = 29).

The study was conveniently deployed to five secondary schools in Sabah with a GPA of ≥ 5 based on SPM result of 2015. The response rate was relatively low (55.1%) with teachers from rural schools having a higher willingness to participate in the study as compared with their counterparts (see Table 4.10 for details of participating schools).

Table 4.10

Summary of Participating Schools and Response Rate in PS2

Name of School	Location	GPA (SPM 2015)	Enrolment of Teachers	Returned Questionnaires	Percentage
SMK PENGIRAN OMAR	Rural	5.80	102	61	59.8
SMK PADANG BERAMPAH	Rural	6.37	57	46	80.7
SMK BEAUFORT	Rural	5.63	95	62	65.3
SMK LIKAS	Urban	6.03	84	33	39.3
SMK BANDARAYA KOTA KINABALU	Urban	6.15	134	58	43.3
Total			472	260	55.1

Table 4.11 below shows detailed demographic information about the respondents participating in PS2. The total sample (169 rural and 91 urban) appeared to be almost the reflection of secondary teacher population in Sabah, which consists of 61% rural teachers (9,543) and 39% urban teachers (6,181). The participants were generally experienced teachers, with 35.8% having more than 16 years in teaching service, while only 16.4% had experience of less than five years. The majority of the respondents had bachelor degrees and were academic teachers.

Table 4.11

Summary of Respondents' Demographic Information

Demographic Information	Characteristics	*N*	Percentage
Gender	Male	92	35.4
	Female	168	64.6
Educational level	STPM	2	0.8
	Diploma	3	1.2
	Bachelor	230	88.5
	Master	24	9.2
	PhD	1	0.4
Years of teaching experience	Less than 5 years	43	16.4
	6 to 10 years	77	29.6
	11 to 15 years	47	18.1
	More than 16 years	93	35.8
Position	Academic teacher	235	90.5
	Head of panel	2	0.8
	Head of department	3	1.2
	Pre-university teacher	13	5.0
	Senior assistant	5	1.0
	Principal	2	0.8
Location of school	Rural	169	65
	Urban	91	35

4.3.1.2 Instrument

SEA II was not entirely new as it retained and refined a few relevant items from SEA I. This questionnaire was developed to explicitly measure two underlying dimensions – perception and practice about the purposes and uses of F1DT. The first dimension, *PERCEPTION*, is operationalised as the degree to which teachers agree (or disagree) about the usefulness of F1DT. This dimension is made up of "what-I-think" items pertaining to teachers' perception about the implementation of F1DT in schools. The second dimension, *PRACTICE*, is a set of "what-I-do" statements. These items are

designed to measure teachers' behaviour in relation to their assessment practices rather than their actual assessment practices, which need to be validated separately. Validation of the reported assessment practices is not within the scope of the main study.

SEA II started with respondents' demographic information – gender, education level, teaching experience and position. The main element of the questionnaire was divided into two sections. The first section contained 16 items requiring respondents to give their perception of the purposes and uses of F1DT. In the second section, teachers were asked to rate another 13 corresponding items about the extent to which they used the information from F1DT in their instructional planning. Furthermore, the questionnaire also provided a few line spaces, on the last page, for the respondents to write their comments about the issues of investigation or the features of the questionnaire. This free-text space may have given an opportunity for the participants to expand more-in-depth responses (Rattray & Jones, 2007). Table 4.12 summaries the content of the questionnaire.

Table 4.12

Content of SEA II

	Demographic information	*PERCEPTION*	*PRACTICE*
Description	Respondents' personal details	Respondents' opinion about their perception of the usefulness of F1DT	Respondents' opinion about their practice concerning F1DT output for instruction
Rating scale		Five-point level of agreement	Five-point level of agreement
Number of items	6	16	13

After finalising the items (see examples of SEA II items in Appendix 4F), the questionnaire was translated into the Malay language, using the translation in PS1 as a guideline. To check the accuracy of translation for this version, I sent it to be certified by the same lecturer who worked with me in PS1. Another stage of review involved a group of peer reviewers, consisting of eight experienced teachers who were bilingual. They were asked to provide feedback on the clarity and coherence of the items and the layout of the questionnaire. This review by peer teachers also serves as a face validity process, providing inputs about how the questionnaire is likely to appear to the potential participants of the main study.

4.3.1.3 Procedure of data collection

For this study, I followed the same ethical protocols in PS1 and submitted relevant documents to the School of Education Ethics Committee for review and approval. I received a notification of clearance on 3 December 2015 (see Appendix 4E). Prior to undertaking the study, I sent letters to the selected schools to seek consent from the principals for their participation in the research.

Again, a drop-off survey was utilised in this study. SEA II was administered via paper-and-pencil mode because this represents – in terms of logistical constraints such as internet access – the lowest common denominator across rural and urban zones in Sabah. Furthermore, this self-administered technique was intended to reduce social desirability bias, increasing willingness to provide transparent answers and to disclose sensitive information.

Assurance of confidentiality and anonymity as enclosed in the questionnaire was another measure taken to encourage respondents to take part in the study voluntarily. I designed a specific protocol to protect the identity of the participants. They were required to write down a unique anonymous code consisting of a combination of unidentifiable information – the first two letters of their mother's name, their date of birth and the first two letters of their father's name. (e.g., JA28YU).

4.3.1.4 Data analysis

As in PS1 a PCA was conducted. The guiding assumption for this analysis was the questionnaire should reflect two distinct dimensions – (i) perception and (ii) practice.

With the assumption that the items were correlated, utilisation of oblique rotation, i.e., Promax, was sought to extract SEA II into a set of meaningful components. Tabachnick and Fidell (2013) advised researchers to begin with this extraction technique before deciding to use other techniques that might seem appropriate to a particular research study.

Also, an examination of the reliability of the questionnaire was conducted using the Cronbach alpha coefficient. This was to check the extent to which all the items in a scale consistently measured the same underlying constructs.

4.3.2 Results

As an effort to test the quality of SEA II, its psychometric properties were scrutinised for systematic item analysis. The same protocols from PS1 were followed – the use of PCA for the exploration of the components and the Cronbach alpha coefficient for reliability, respectively.

An initial analysis was run to ensure the dataset was appropriate for PCA by examining the KMO value and Bartlett's test of sphericity. The SPSS output in Table 4.13 shows that the KMO value was .953, exceeding the minimum rule of thumb value of .6, and Bartlett's test achieved statistical significance ($p < .05$), indicating that the SEA II items were reasonably factorable.

Table 4.13

KMO and Bartlett's Test Sphericity of 29-item Dataset (n=260)

KMO and Bartlett's Test		
Kaiser–Meyer–Olkin Measure of Sampling Adequacy		.953
Bartlett's Test of Sphericity	Approx. Chi-Square	6389.965
	df	406
	Sig.	.000

The default setting of PCA with oblique rotation, Promax, was deployed to extract the maximum variance from the dataset and to reduce it into several smaller components. The component loading result is presented in Table 4.14. Items were tagged as "P" which denotes perception and "Pr" for practice. The initial result identified the presence of four components, with eigenvalues exceeding 1, accounting for 68.9% of the total variance. Additionally, the communalities were all above .3, further confirming that each item shared some common variance with each other. Note that there were seven items with cross-loading components, apparently belonging to more than one solid component. The first three components managed to pull together more than three items onto their respective component. The fourth component, however, had only three items with loading indices of >.7. This component consisted of strongly-worded statements that appear to be negative/unfavourable towards the application of F1DT.

Table 4.14

Component Loadings of 29-item Dataset (n=260)

All items	Component				Communalities
(29 items)	1	2	3	4	
P pre-existing knowledge		.796			.560
P learning potential		.782			.630
P strengths and weaknesses		.763			.597
P info accuracy		.758			.494
P innate abilities		.755			.492
P future performance		.702			.535
P learning improvement		.613			.591
P identification of LD		.495			.595
P streaming students		.459	.419		.610
P pedagogical decision	.320		.762		.631
P learning needs	.333		.688		.615
P intervention for LD	.305		.644		.697
P low-performing students			.635		.603
P students' comparison		.371	.621		.482
P referral to SEP			.620		.560
P quality of teaching			.508	.428	.466
Pr pedagogical decision	.886				.680
Pr learning needs	.876				.694
Pr identification of LD	.765				.619
Pr learning potential	.758				.716
Pr pre-existing knowledge	.756				.661
Pr strengths and weaknesses	.723				.707
Pr intervention for LD	.717				.635
Pr quality of teaching	.700				.442
Pr innate abilities	.679				.661
Pr future performance	.416	.387			.521
Pr little use				.813	.628
Pr directive order				.809	.663
Pr preference of UPSR				.783	.604
% of Eigenvalues (68.9%)	51.8	8.1	5.2	3.8	

Note: Factor loadings <0.3 were suppressed

Further inspection of the fourth component using the factor correlation matrix output as shown in Table 4.15 indicated that it had a low correlation value with the other components. This justified the decision to omit these items at the next stage. More importantly, the result also warranted the utilisation of oblique rotation, revealing that the items of the three components were highly correlated to each other. The correlation indices were greater than the commonly accepted threshold of .32 (Field, 2013; Tabachnick & Fidell, 2013).

Table 4.15

Factor Correlation Matrix of 29-item Dataset (n=260)

Factor Correlation Matrix				
Component	1	2	3	4
1	1.000	.686	.649	.008
2	.686	1.000	.653	.031
3	.649	.653	1.000	.111
4	.008	.031	.111	1.000

For the next stage, I retained 26 items and analysed them for factorability. An inspection of the KMO and Bartlett's test of sphericity revealed that this 26-item dataset was valid for factorability (KMO = .961, p = .000).

The results in Table 4.16 demonstrate that two components could be robustly identified. They were finally labelled as *PERCEPTION* and *PRACTICE*. Together, they accounted for 63.5% of the total variance, with *PRACTICE* contributing 57.7% and *PERCEPTION* contributing 5.8%. Here, many items had strong loading values as high as .981. Compared with the first analysis, the number of items with cross-loading components decreased. Note that the communality value of item "P quality of teaching" was .289, not even exceeding the recommended value of .3. Nonetheless, its reasonably high component loading of .604 warranted its use in the final version of SEA for the main study. It is noticeable that most of the items tapped into their respective conceptual dimension. Nonetheless, the three items (highlighted in blue) were initially conceptualised as belonging to perception, but empirically they belonged to practice. For the final version, however, they remained as perception items – justifying that there are corresponding items asking the same questions in practice.

Table 4.16

Component Loadings of 26-item Dataset (n=260)

Items	Components		Communalities
	1 (13 items)	2 (13 items)	
P future performance		.920	.660
P strengths and weaknesses		.822	.677
P innate abilities		.799	.580
P students' comparison		.767	.558
P learning improvement		.724	.642
P learning potential		.708	.659
P pre-existing knowledge		.697	.589
P info accuracy		.681	.543
P low-performing students		.673	.632
P streaming students		.658	.622
P identification of LD		.606	.617
P quality of teaching		.604	.289
P referral to SEP	.315	.484	.565
Pr learning needs	.981		.823
Pr pedagogical decision	.980		.812
Pr identification of LD	.868		.708
Pr intervention for LD	.835		.706
Pr quality of teaching	.814		.534
Pr learning potential	.803		.741
Pr pre-existing knowledge	.796		.697
Pr strengths and weaknesses	.769		.744
Pr innate abilities	.702		.691
P learning needs	.574		.623
P pedagogical decision	.574		.638
P intervention for LD	.527	.361	.697
Pr future performance	.365	.362	.466
% of Eigenvalues (63.5%)	57.7	5.8	

Note: Component loadings <.3 were suppressed

An examination of the correlation matrix for this two-component solution with Promax rotation justified the decision to use oblique, not orthogonal rotation, validating the assumptions that these 26 items are correlated (see Table 4.17).

Table 4.17

Factor Correlation Matrix of 26-item Dataset (n=260)

Factor Correlation Matrix		
Component	1	2
1	1.000	.760
2	.760	1.000

The internal consistency of the questionnaire was examined using Cronbach's alpha. The result in Table 4.18 showed that both the main components were highly internally consistent with the values of .937 (*PERCEPTION*) and .958 (*PRACTICE*). The third component, consisting of the negatively worded items, also had an acceptable reliability index of .762.

Table 4.18

Reliability Indices of SEA II

Component	Number of items	Cronbach's alpha
PERCEPTION	13	.937
PRACTICE	13	.958
NEGATIVITY	3	.762

The above analyses indicated that two distinct components were underlying teachers' responses to this SEA II prototype. The two-component structure was apparently more parsimonious and easier to analyse with highly acceptable statistical criteria – high loading value and high reliability index. It is important to mention here that a third component was added to the final item of the questionnaire. The component was tagged as "*NEGATIVITY*", consisting of the negatively worded items about F1DT. These items were included in the final version as they might be useful to understand teachers' view about the directive use of F1DT, little use of F1DT and preference for UPSR.

4.3.3 Reflective conclusion of PS2

Overall, the decision to undertake PS2 was worthy and rewarding. It ascertained that the research instrument to measure the issues of investigation was justifiable psychometrically and conceptually. It is obvious that SEA II yielded more stable and coherent component structures. It is tentatively valid for use to measure teachers' perception and practice of the targeted assessment tool. Deletion of items is unnecessary because all items are psychometrically good and conceptually coherent. Most importantly, SEA II's clear definition of the conceptual dimensions is seen as a noteworthy outcome of this study.

Additionally, it assisted me to refine methodological protocols and identified which methods would potentially work for the main study. There were no crucial issues emerging from PS2. This was attributed to a more systematic planning stage before the

data collection started. Sample size, in particular, was taken great care over to ensure a smooth running of PCA. The participants were also selected to mirror the target groups for the main study. Furthermore, one method of data collection technique, paper-and-pencil self-administration, proved to be feasible.

4.4 Lessons Learnt from the Pilot Studies

The following are several pertinent reflective notes of what I have learnt from the two pilot studies. The lessons learnt, which include item refinement, access and recruitment of participants, data collection method, as well as statistical procedures, are discussed in the subsequent sections. The final changes for the main study are also highlighted.

4.4.1 Development and modification of questionnaire items

It is worth noting that item development is the most salient feature that experienced a significant change in these preliminary studies. Here are several major improvements for the construction of the working version of SEA.

4.4.1.1 Operationalisation of educational assessment

One of the crucial modifications made from PS1 to PS2 was the operationalisation of educational assessments. The former defined educational assessments as teacher-made tests and low-stakes standardised tests, while the latter explicitly targeted a currently used diagnostic test, i.e., F1DT. This specific reference to an existing tool was deliberate to initiate a clear direction for the research participants to respond to the questionnaire. Following this, new conceptual dimensions of the underlying constructs of the questionnaire were generated – perception and practices of the purposes and uses of F1DT.

4.4.1.2 Item clarity and accuracy

As a result of the pilot studies, SEA was extensively scrutinised at the item level. Many items were revised for improvement. The new version of SEA emphasises the consistency of wording for corresponding items in both *PERCEPTION* and *PRACTICE* to ensure clarity and accuracy of the items. As shown in Table 4.19, the beginning of each statement – "I think F1DT" and "I use/used F1DT" – clarifies that each statement belongs to its respective dimension. Note that the wordings of the correlated items are consistent. If

matching items are written in a different manner, as in PS1, the respondents are likely to understand them differently. The consistency of the wording aims to direct respondents to a uniform interpretation of the meaning for both items.

Table 4.19

Examples of Consistency in Wordings of the Items

PERCEPTION I think F1DT …	*PRACTICE* I use/used F1DT …
…enables teachers to predict students' future academic performance accurately	…to predict students' future academic performance accurately
…provides teachers insights to identify students with learning difficulties (LD)	…to identify students with learning difficulties (LD)
…helps teachers to stream students according to their current attainment level	…to stream students according to their current attainment level
…helps teachers to differentiate among low-performing students	…to differentiate among low-performing students

4.4.1.3 Primary and additional items

The final version of SEA is made up of 36 items. *PERCEPTION* consists of 19 items and *PRACTICE* has 17 items. Each item has its corresponding counterpart. Two extra items in *PERCEPTION* ask the participants about general statements that are not concerned with the practice of F1DT (see the full version of the questionnaire in Appendix 4G).

In addition to the main items, two sets of additional items were included in the final version of SEA for the main study. The first set, i.e., criteria of an ideal assessment tool, was administered at the pre-intervention stage to elicit information about teachers' opinions regarding the criteria for an ideal assessment tool. The second set, i.e., reflective feedback of DT, was targeted for the intervention group, asking the respondents about their feedback after the introduction of DT. A summary of the working version of SEA is presented in Table 4.20.

Table 4.20

Final Version of SEA

	Primary dimensions		Secondary dimensions	
	PERCEPTION	*PRACTICE*	Criteria of ideal assessment tool	Reflective feedback of DT
Number of items	19	17	1	10
Target respondent	Intervention and control groups			Intervention group
Administration stage	Pre-test and post-test		Pre-test	Post-test

4.4.2 Refinement of methodological procedures

Apart from item construction and validation, this one-year preliminary study was also intended to assess the practicality of the research procedures to avoid potentially unfavourable outcomes of the whole research. The following paragraphs outline several alterations of the research protocol to improve its application in the main study.

4.4.2.1 Sampling and recruitment of participants

Due to time and logistical constraints, both pilot studies deployed a convenient sampling, recruiting the respondents based on the accessibility of my proximity and research timeline. In PS1, specific traits of the research participants were not defined prior to the study. This resulted in the random participation of primary and secondary teachers of schools in Bands 1 to 6 from various states in Malaysia. Noting the possible consequence of sampling recruitment for external validity, several efforts were made to mirror the characteristics of the target population in the main study. This is because if the respondents in the pilot studies did not possess the same characteristics to those in the main study, there was a likelihood that other participant variables might affect the validity of the questionnaire, e.g., the way they responded to the items (Bryman, 2012: Cohen et al., 2011; Fraenkel & Wallen, 2008). Hence, I recruited participating schools for PS2 based on pre-outlined criteria – location of the school, i.e., urban and rural, and a school GPA of ≥5. Representativeness, nevertheless, was not crucial at this stage because it was rather exploratory, concerning the feasibility of the instrument and research processes, than confirmatory of the research issues under investigation.

Clearly, the utilisation of two-group experimental design in the main study requires meticulous planning of the sampling strategy that particularly tailors specific criteria of the target population. Therefore, a practical design of participant comparability had been formulated for the main study (see details in Chapter 7).

4.4.2.2 Delivery mode of the questionnaire

The necessity of the online method in PS1 was justified for its practicality to reach a wider range of research subjects and its flexibility allowing the completion of the questionnaire at a time convenient to the respondent. However, several limitations of the online survey were identified and they might have risked the success of the study. Thus, I opted to utilise a paper-and-pencil method only for PS2 and the main study.

Another important consideration in choosing a single method of data collection is the accessibility issue. Most of the schools of the target population, 65 out of 85 schools, were located in rural areas. It was likely that the online approach might suffer from the inability to cover the target population adequately as a result of the lack of internet coverage.

4.4.2.3 Statistical procedures

Another key lesson from the pilot studies concerns the use of statistical procedures for data analyses. I learnt that prior planning was essential because requirements for different statistical techniques vary from one to another. The issue of sample size, in particular, was a major concern in these pilot studies. The small sample size in PS1 apparently contributed to the instability of the component loadings. This was because I did not meet the minimum sample requirement for undertaking PCA. Taking this valuable insight into due consideration, a practical plan was deployed in PS2 to collect as many samples as possible, exceeding the proposed rule of thumb for an adequate sample in PCA.

In short, it is concluded that piloting the questionnaire was essential to assess the feasibility of the instrument as well as the practicality of the proposed research procedures. A summary of what has been discussed in the preceding sections is presented in Table 4.21.

Table 4.21

Comparison between SEA I and SEA II

	SEA I		**SEA II**	
	PERCEPTION	*PRACTICE*	*PERCEPTION*	*PRACTICE*
Operationalisation of assessment	General use of any educational assessment tool		Specific use of F1DT	
Underlying dimensions	Four conceptions of educational assessment	Four conceptions of educational assessment	Perception about the purposes and uses of F1DT	Practice of F1DT output in instructional plan
Number of items	33	24	16	13
Item response format	Seven-point level of agreement	Five-point ordinal frequency	Five-point level of agreement	
Assumption of item correlation	Uncorrelated		Correlated	
Statistical procedure	Orthogonal rotation – Varimax		Oblique rotation – Promax	
Sampling strategy	Convenience sampling No prior selection criteria		Convenience sampling Pre-determined selection criteria	
Sample size	130		260	
Data collection method	Paper-and-pencil Online		Paper-and-pencil	

4.5 Chapter Summary

This chapter has documented the report of two pilot studies conducted to develop and validate the questionnaire for use in the main study. Clearly, they provided noteworthy insights that are greatly beneficial for the improvement of questionnaire items and the planning of research procedures. Furthermore, conducting these two small-scale research projects delivered a plethora of learning opportunities for me as a novice researcher. Admittedly, self-designing the questionnaire was an arduous process, but at the same time rewarding, in the sense that I discovered several unforeseen mistakes that could have affected the outcomes of the study. Most importantly, I learnt not to discount the crucial issues that emerged from the pilot studies; rather, I tried to look for practical solutions to ascertain that they would not impede the successful manageability of the actual project. The next chapter will discuss the conceptual and empirical reviews of what is in the literature about dynamic testing, the alternative assessment approach that is the focus of this study.

5 Dynamic Testing as an Alternative Assessment Approach

This study suggests that ineffective use of assessments (be it through ignoring the information they provide or through the use of inadequate assessment tools, or both) is one of the potential hindrances to the delivery of effective education. Arguably, the notion of the assessment function for teaching and learning improvement can be realised when teachers collect quality information that fairly represents students' real potential. Such information is crucial for teachers in attempting to devise instructional adjustments to cater to students' learning needs. In the context of this study, perhaps the currently used assessment battery, specifically referring to F1DT, is unable to provide such information due to its focus on measuring students' pre-existing knowledge. It is assumed that this may be the barrier to effective instruction, as F1DT may not be reasonably suitable for use, especially for disadvantaged learners. To achieve the overarching national agenda for a world-class education, it is timely and important to direct our energies to look for an alternative approach to assessment that could redeem the deficits of F1DT. As mentioned in the introductory chapter, it is argued that DT could possibly bridge the gap in the current situation. Thus, an introduction of the concept of DT to Malaysian teachers is necessary and this may offer them a new perspective on educational assessments. The research in this thesis aims to explore the effects of introducing the notion of a new assessment concept on teachers' assessment beliefs and practices. The investigation of the potential effects is elicited using the intervention-control group experimental design. This chapter outlines a conceptual and empirical review of the literature about DT, aiming to provide a basic understanding of it. As the concept of DT is entirely new in Malaysia, an overview of this assessment approach is presented in the first section. The theoretical background that underpins the idea of DT is discussed. Following that, the discussion further describes the detail of DT based on three essential questions: (i) What is DT? (ii) How different is DT from conventional testing? and (iii) How successful is DT? The focus of the second section is the review of previous studies documenting empirical evidence of DT potentials in educational settings. The last part is the summary of the review and the outline of empirical exploration in this study.

5.1 An Overview of Dynamic Testing

For a very long time, standardised tests have been widely used to gauge students' current attainment, to predict future performance and to suggest pedagogical strategies. Standardised testing has remained an important assessment tool due to its predictive validity (Elliott & Resing, 2015; Popham, 2001; Riffert, 2005; Roediger III, Putnam & Smith, 2011; Travis, 1996). Policymakers and educationalists have come to consider test scores as powerful indicators to evaluate the performance of students and the quality of teachers, schools, or the whole education system of countries. Another reason for the widespread use of psychometric tests in educational contexts is because of its perceived use in providing descriptive information about one's cognitive functioning (Black, 2002; Elliott & Resing, 2015; Phelps, 2005). The utility of cut off scores or passing scores,[5] for instance, is interpreted to reflect students' performance levels (e.g., basic, proficient or advanced). Teachers use such explicit or quantifiable outputs to identify students' relative strengths and weaknesses. This information is useful as a resource for a diagnosis towards catering students' learning needs. However, while it is acknowledged that standardised tests have delivered positive impacts in education, they have also engendered a considerable debate among scholars and educational practitioners for many years.

Many critics (e.g., Elliott, 2000; Elliott & Resing, 2015; Goslin, 1967, 1968; Guthke, 1993; Guthke & Beckmann, 2000a; Guthke, Beckmann & Dobat, 1997; Sternberg & Grigorenko, 2002; Hessels, 1997; Popham, 2001; Salmon-Cox, 1981; Tzuriel, 2001, 2005; Urdan & Paris, 1994) have identified several shortcomings of conventional tests, particularly IQ tests. The arguments against traditional tests can be summarised into four major issues:

(a) Over-emphasis on product-related information

One of the greatest criticisms of the traditional test is the assumption that knowledge and skills are the only viable yardsticks to measure student outcomes in education (Popham, 2001; Travis, 1996). The opponents of standardised tests argue that changes in student behaviour are seen as indicative of learning that has taken place (Haywood, 2006; Haywood & Lidz, 2007; Terenzini, 1989; Travis, 1996; Tzuriel, 2005). This implies that

[5] The point(s) on the test score scale that determines whether a test-taker fails or passes. Multiple cut-off scores are typically used to categorise students into levels of proficiency such as weak, intermediate or advanced.

an assessment tool should provide an accurate description of behaviour that can help students and teachers to promote learning development. The traditional tests, however, only deliver information about products of previous learning, not any information about future learning opportunities.

(b) Lack of specific description of the learning process

Because learning is a continuous process (Haywood, 2010; Grigorenko & Sternberg, 1998), it is argued that the scores of one-point-in-time tests say very little about how students learn and how they progress over time. Such limited information has substantially underestimated students' intellectual potential (Beckmann, 2006, 2014; Guthke, 1993; Guthke & Beckmann, 2000a; Elliott & Resing, 2015; Hessels, 1997; Tzuriel, 2001, 2005). Lack of description of the learning process may create challenges for teachers to provide prescriptive recommendations for a student to optimise his potentials (Haywood, Brown & Wingenfeld, 1990).

(c) Lack of sensitivity towards disadvantaged students

It is also claimed that traditional tests have been unfair to certain individuals or groups, especially students with learning disabilities (LD) and minority groups (e.g., Elliott & Resing, 2015; Guthke & Beckmann, 2000a; Hamers & Pennings, 1995; Hessels, 1997; 2009; Tzuriel, 2001, 2005). Many studies documented that traditional tests do not provide valid identification of students with LD (e.g., Gresham & Vellutino, 2010; Stuebing, Barth, Weiss, & Fletcher, 2009; Vellutino, Scanlon, & Lyon, 2000). Also, there is no strong evidence that conventional tests are educationally sensitive to the cultural, socio-economical and linguistic diversities of students from minority groups (e.g., Hessels, 1997; Pena, Iglesias & Lidz, 2001; Navarro & Mora, 2011; Resing, Tunteler, de Jong & Bosma, 2009; Stevenson, Heiser & Resing, 2016; Wiedl, Mata, Waldorf & Calero, 2014). If such biased information is interpreted misleadingly, teachers might engage in unfair assessment decisions against these students.

(d) Lack of non-intellective elements

Another criticism is that the emphasis on "purely cognitive factors" in traditional tests seems to neglect the influence of "non-intellective variables" (Haywood & Lidz, 2007, p.28) on one's intellectual status. Motivational and personality factors are no less important in determining the success in cognitive functioning (Guthke, 1993; Haywood

& Lidz, 2007; Tzuriel, 2001, 2005; Sternberg & Grigorenko, 2002). Arguably, the measurement of cognitive variables alone is not sufficient to manifest the developmental and modifiability of one's true potential.

The main concern over the deficits of traditional tests is that teachers may interpret the test scores misleadingly. This may result in self-fulfilling prophecies, which can ultimately be harmful to the learning development of students, especially those who are academically disadvantaged (Haywood et al., 1990; Jussim & Harber, 2005; McKown & Weinstein, 2008). If it is hypothesised that low test scores are the representation of actual potentials, this can demotivate students from taking further actions to improve their learning. Moreover, teachers may recommend pedagogical plans that are likely to be off-target. For these reasons, an alternative assessment that is "need sensitive" to the above criticisms is necessary. An overview of the proposed assessment approach that would possibly address the pitfalls of conventional tests is presented in the following sections.

5.1.1 Theoretical foundations underpinning the emergence of dynamic testing

The criticism that centres on the unfortunate effects of conventional testing to uncover the actual potential of learners has brought DT into the spotlight in educational settings. The presence of DT for almost five decades originated from scholarly ideas dating from as early as the 1960s. The following section presents a brief discussion of the theoretical foundations of contemporary DT.

5.1.1.1 Vygotsky's Zone of Proximal Development

In the literature, Vygotsky's (1978) major contribution on the idea of the Zone of Proximal Development (ZPD) has always been credited as a conceptual foundation of contemporary DT (Beckmann & Guthke, 1995; Dörfler, Golke & Artelt, 2009; Elliott, 2003; Elliott, Lauchlan & Stringer, 1996; Grigorenko & Sternberg, 1998; Guthke, 1992; Guthke, Beckmann & Stein, 1995; Haywood & Lidz, 2007; Lantolf & Poehner, 2007; Lidz, 1995; Lidz & Gindis, 2003; Shabani, 2012; Sternberg & Grigorenko, 2002; Tzuriel, 2000, 2001). According to Vygotsky (1978), a child's mental development can be categorised into two levels: the zone of actual development and the zone of proximal development. The first level refers to the developments of a child that have been established or developed, reflected in what a child is able to do independently. In contrast,

the ZPD is defined as "the distance between the actual developmental level as determined by independent problem-solving and the level of potential development as determined through problem-solving under adult guidance, or in collaboration with more capable peers" (Vygotsky, 1978, p.86). In contrast to the first level, Vygotsky asserted that the ZPD is a function of what has not yet been established but what is in the process of being developed. The centrality of the ZPD also points out that the interaction between a child and adults or more skilful peers can promote cognitive development. Vygotsky reasoned that the measurement of potential development is as important as assessing actual development. What a child can accomplish with the assistance of others is viewed as more indicative of a higher level of development and more informative to educational practitioners (Elliot, 2003; Elliott et al., 1996; Vygotsky, 1978).

The notion of the ZPD has captured the attention of DT advocates. The ZPD reflects a "developmental, interactive and forward-looking nature" (Sternberg & Grigorenko, 2002, p.38) of a child's cognitive function that remains undiscoverable by conventional tests. Underpinned by Vygotsky's ideas, the proponents of DT concern to combine the cognitive and non-cognitive processes within the context of assessment (Elliott, 2003; Elliot et al., 1996; Hill, 2015; Stringer, Elliott & Lauchlan, 1997). Acknowledging that a child's cognitive processes are modifiable and developmental, an assessment of learning potential becomes the nucleus of DT (Beckmann, 2014; de Beer, 2006; Guthke, 1992; Grigorenko, 2009; Grigorenko & Sternberg, 1998; Stringer, 2018). It is perceived that the measurement of learning potential provides more meaningful information about a learner's "true" ability (Beckmann, 2006; Beckmann & Guthke, 2000a; Elliott, 2000; Haywood & Lidz, 2007).

Furthermore, the idea of a child's social interaction with the experts in facilitating cognitive development has influenced the core of theoretical operationalisation of DT (Dörfler et al., 2009; Elliott et al., 1996; Grigorenko, 2009; Lidz, 1995; Shabani, 2012). Specifically, the concept of the guidance of more capable others is implemented in the form of a mediator or the prompts/cues to assist learners to complete a given task. Grigorenko (2009) implied that "others" refer to real humans or "ideal humans" (p.8), which may be books, videos or other products of human knowledge. It is apparent that the integration of assistance within the assessment procedure is the hallmark characteristic of DT (Beckmann, 2006, 2014; de Beer, 2006; Elliott, 2003; Elliott et al.,

1996; Grigorenko & Sternberg, 1998; Guthke et al., 1997; Hessels, 1997; Lidz, 1995; Resing, 2013; Sternberg & Grigorenko, 2002; Stringer et al., 1997).

5.1.1.2 The nature and development of human ability

As previously mentioned, conventional testing has become a widely debated issue due to its "failure" to be an adequate measure of learners' real ability. A critique of a test as being unsuitable to measure ability requires a priori understanding of what ability is. This basic understanding of the theoretical base of human ability may assist teachers to systematically understand how human abilities develop, how differences in abilities influence learning and what can be done to enhance ability. Thus, it is highly desirable that theoretical understanding of the nature and development of human ability precedes the description of the assessment instruments, because the instruments should be consistent with the conceptual nature of the construct, i.e., the latent abilities to be assessed (Haywood, 2006, 2010; Haywood & Lidz, 2007; Haywood, Tzuriel & Vaught, 1992; Sternberg & Grigorenko, 2002). In accordance with this view, the subsequent paragraphs discuss the views of the advocates of DT about what underlies human ability.

5.1.1.2.1 Abilities as forms of developing expertise

Sternberg (1998, 1999) discussed abilities as forms of developing expertise. According to Sternberg (1999), a fundamental assertion of this view is that "individuals are constantly in a process of developing expertise when they work within a given domain" (p.361). This means that ability is not a relatively fixed construct, but rather a continuous and developmental construct. Sternberg and Grigorenko (2002) claimed that the idea of developing expertise is conceptually consistent with the ZPD's notion of the dynamic developmental state of a child.

In conceptualising the underlying components of ability, the developing-expertise model (Sternberg, 1998, 1999; Sternberg & Grigorenko, 2002) posited six key elements that are presented below:

i. Metacognitive skills – the ability to understand, monitor and control one's cognition. Seven metacognitive skills that appear to be particularly important – problem recognition, problem definition, problem representation, strategy formulation, resource allocation, monitoring of problem-solving and evaluation

of problem-solving – are complex (Sternberg, 1998) and modifiable (Sternberg & Grigorenko, 2002).

ii. Learning skills – the acquisition of knowledge. These skills can be explicit (systematic effort to learn) and implicit (incidental learning without systematic effort).

iii. Thinking skills – the performance components. A child needs to master a set of three main thinking skills – critical (analytical) thinking skills, creative thinking skills and practical thinking skills.

iv. Knowledge. In the academic world, two types of knowledge are important: a) declarative knowledge – the "knowing that" (Sternberg & Grigorenko, 2002, p.7). which includes facts, concepts, principles and laws; and b) procedural knowledge – the "knowing how" (Sternberg & Grigorenko, 2002, p.7) that involves procedures and strategies.

v. Motivation – the driving force to success. Two kinds of motivation are necessary to succeed – achievement motivation and self-efficacy. Achievement motivation drives learners to constantly improve themselves, while self-efficacy motivates learners to believe in their ability to solve difficult tasks.

vi. Context – external factors influencing performance. All the above characteristics are affected by the context in which they operate.

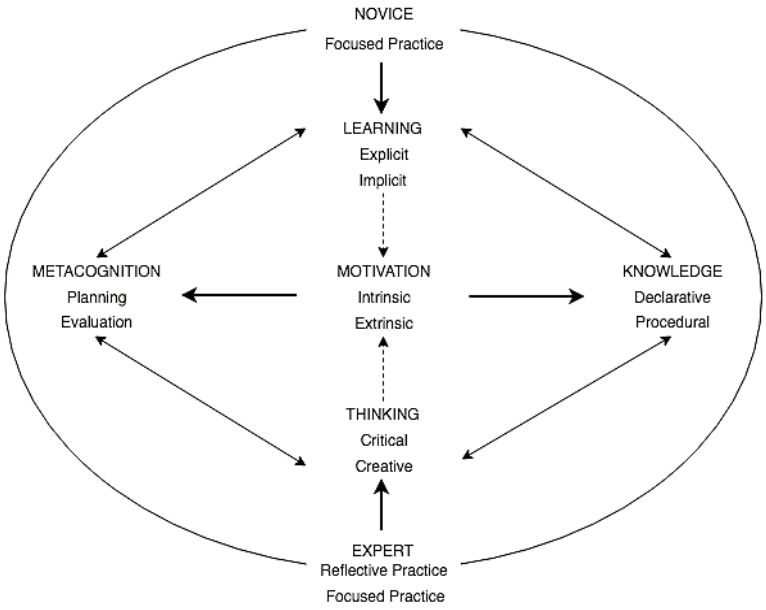


Figure 5.1: Developing-expertise Model
(Sternberg, 1998, 1999; Sternberg & Grigorenko, 2002)

As illustrated in Figure 5.1, a novice develops to become an expert through focused practice which requires the interaction of all the six elements. Central to this is motivation. The development of expertise occurs at many different levels. An individual may cycle through several elements many times to successfully reach a higher level of expertise. Although the six elements are domain-specific, they are fully interactive to the extent they influence each other directly and indirectly (Sternberg & Grigorenko, 2002).

5.1.1.2.2 The transactional perspective on human ability

Another systematic theoretical base of human ability is the transactional perspective on human ability (Haywood, 2006, 2010; Haywood & Lidz, 2007; Haywood et al., 1990; Haywood et al., 1992). This transactional perspective encompasses three core dimensions – intelligence, cognition and motivation. According to this "three-way symbiosis" (Haywood & Lidz, 2007, p.28) model, the human ability is conceptualised based on the following tenets:

i. Intelligence alone is not an adequate reflection of individual differences in ability. The various facets of ability, as presented below, are all correlated but have their unique qualities.
ii. Cognitive processes are recognised as an important element of ability. Such processes represent relatively "durable modes and strategies of logical thinking" (Haywood, 2010, p.25) and "habits of using those logic modes in the daily tasks of interpreting information from the senses, perceiving learning opportunities, approaching learning tasks, integrating new knowledge and solving problems" (Haywood & Lidz, 2007, p.23). Cognitive processes are believed to be eminently modifiable and elaborated (Haywood, 2006, 2010).
iii. Motivation, particularly task-intrinsic motivation, is also an essential dimension of human ability. It is the ability to act, think, learn, experience novelty and perform in the absence of extrinsic rewards (Haywood, 2010; Haywood & Lidz, 2007). It is both a situational variable and personality trait (Haywood & Lidz, 2007).
iv. Ability is determined by multiple factors. It is an apparent result of several mediating elements including biological and environmental/experiential influences.

The relation between the three components is not merely dynamic but rather more complex (Haywood, 2006, 2010; Haywood & Lidz, 2007). It is analogous to a "biological symbiosis" (Haywood, 2010, p.27) in the sense that these variables are not independent but are functionally connected. They appear to have a "snowball effect" (Haywood, 2010, p.28) to a greater extent so that the interaction between two variables may influence the third component. Additionally, this "tripartite conception" (Haywood, 2006, p.305) views intelligence, cognitive processes and motivation as developmental and changeable in nature (Haywood, 2006; Haywood & Lidz, 2007).

Both models responded to the conventional view that claimed that abilities are relatively stable attributes of learners determined primarily by heredity (Haywood, 2006, 2010; Lidz & Gindis, 2003; Sternberg, 1999; Sternberg & Grigorenko, 2002). Initiated by the ideas of the ZPD, these models proposed different perspectives of ability from the one that is conventionally offered. In sum, abilities are: (a) multidimensional; (b) interactive; (c) flexibly changeable and developmental; and (d) contextually defined. It is argued that traditional tests do not provide adequate manifestations of "translating" the above ideas of ability into observable and measurable performance (Haywood, 2006, 2010; Haywood et al., 1992; Haywood & Lidz, 2007; Grigorenko & Sternberg, 1998; Sternberg, 1998, 1999; Sternberg & Grigorenko, 2002). In other words, there is a disparity between the manifest variable (the test scores) and the latent variable (the proposed nature of ability).

Theoretically driven by the above notions of the ZPD and the nature of human abilities, the design of a tool that adequately reflects the conceptualisation of human cognitive development is a necessary prerequisite to obtain more insightful information about students' learning potentials. Responding to this need, DT has been suggested as a better method of measuring the complexity and modifiability of human abilities that have been fully developed as well as those that are in the process of developing (de Beer, 2006; Haywood & Lidz, 2007; Hessels, Berger & Bosson, 2008; Grigorenko & Sternberg, 1998; Lidz, 2014; Sternberg & Grigorenko, 2002; Tzuriel, 2011). Sternberg and Grigorenko (2002) advocated that:

> A major problem is not with the use of tests per se, but with the kinds of tests being used. Conclusions are being drawn that go way beyond the inferences that properly should be drawn from the test scores. We believe that dynamic testing possesses the potential to make information gleaned from tests more valid and more useful. (p.ix)

Relating this to the situation in the current context, F1DT appears to represent the failure of traditional tests in providing teachers with informative descriptions of students' real potentials. Moreover, a great emphasis over the product-oriented assessment approach seems to neglect the core idea that one's abilities are developmental, multifaceted and multi-determined. As stated in the introductory chapter, it is posited that teachers in Malaysia may benefit from assessment-related information that may provide more accurate descriptions of the complexity of human abilities. Producing such information is the unique feature of DT. The following section provides an overview of DT, conveying the ideas why DT could be a possible answer to the shortcomings of F1DT. In doing so, the following discussion is structured to answer three essential questions about DT.

5.1.2 What is dynamic testing?

So, the first question is: what is DT? In current use, DT seems to be referred to interchangeably with dynamic assessment (henceforth, DA) (Haywood & Lidz, 2007; Stringer, 2018). To avoid confusion, it is essential to clarify the conceptual differences between DA and DT (Elliott, Resing, & Beckmann, 2018; Stringer, 2018). Elliot et al. (2018) admitted that the distinctions between the two terms are often treated implicitly, rather than explicitly. To address this issue, several scholars have attempted to provide a distinction between the two terms.

Haywood and Tzuriel (2002) defined DA as "a subset of interactive assessment (IA) that includes deliberate and planned mediational teaching and the assessment of the effects of that teaching on subsequent performance" (p.40). Likewise, Resing (2013) delineated the concept of DA as a generic description of various approaches all associated with dynamically integrating feedback and instruction – fixed or individualised – in the assessment procedures. This "process of actively teaching – of an individual's perception, learning, thinking and problem-solving" (Tzuriel, 2011, p.115) aims at observing the modifiability of a child's test performance as well as informing necessary adjustment in teaching and learning towards promoting learners' competence (Hill, 2015; Lidz, 2014; Tzuriel, 2011). According to the above definitions, DA might involve, in addition to testing, portfolios, projects, interviews and other forms of information gathering that aim to provide better estimates of a child's learning ability.

In contrast, DT is one of the many procedures of the larger process of DA (Beckmann, 2006, 2014; de Beer, 2006; Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2002; Stringer, 2018) emphasising that testing and assessment are not synonymous. In his articles, Beckmann (2006, 2014) distinguished DT from a heterogeneous collection of DA procedures. Specifically, DT is explained as "a methodological approach to psychometric assessment that uses systematic variations of task characteristics and/or situational characteristics in the presentation of test items to evoke intraindividual variability in test performance" (Beckmann, 2014, p.310). The definition of DT is further restricted to "the tester-testee interaction" (Sternberg & Grigorenko, 2002, p.29) as it tends to look narrowly into the progress of a test-taker after a series of prompts is given explicitly or implicitly.

The representation of IA, DA and DT, as defined above, is illustrated in Figure 5.2 below.

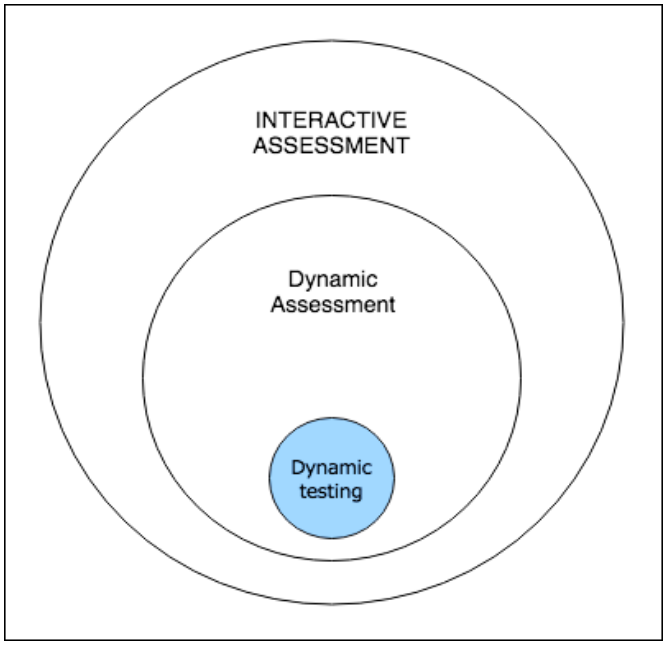


Figure 5.2: Representation of the Relationship between IA, DA, and DT

What is common in the above definitions is that an assessment procedure can be considered as "dynamic" if (a) an intervention is incorporated in the assessment process and (b) the measurement of the learners' response to the intervention (RTI) is of primary interest. Clearly, both DA and DT attempt to link assessment and intervention (Elliott et al., 2018; Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2002; Stringer, 2018;

Tzuriel, 2000); nevertheless, they differ in terms of their focus. While the goal of DA is to assess and intervene, with the primary emphasis on the intervention (Elliott et al., 2018; Grigorenko & Sternberg, 1998; Tzuriel, 2000, 2001, 2011; Sternberg & Grigorenko, 2002; Hill, 2015; Lidz, 2014; Lantolf & Poehner, 2007), DT is primarily concerned with examining the test-taker's change in test performance if prompts are provided (Elliott et al., 2018; Guthke & Beckmann, 2000a; Hessels et al., 2008; Sternberg & Grigorenko, 2002). Whilst dynamic tests are objective measures, they do not automatically represent measurements of change. They do, however, capture performance under performance-optimising conditions. In other words, an intervention is considered as an end in DA – focusing considerably on the implementation of an intervention to promote changes in students' performance. In contrast, DT uses intervention as a means to inform teachers about a better estimate of individual differences in learning and this, in turn, can be valuable for the planning of the real intervention catering to students' individual needs. In educational settings, the information from DT is arguably more practical and highly informative as it represents the dynamicity of individual differences of the learners (Beckmann, 2014; Elliott et al., 2018).

In this study, I will use DT throughout the thesis, specifically referring to the Adaptive, Computer-Based Learning Test Battery (Guthke, 1992; Guthke et al., 1995; Guthke & Beckmann, 2000a), which was adapted for use in intervention strategies for experimental groups (detailed description of this test is presented in the next chapter).

5.1.2.1 Types of dynamic testing

There are three types of DT: (i) the long-term format, (ii) the short-term format (Guthke, 1992; Beckmann & Guthke, 1995; Guthke & Beckmann, 2000a) and (iii) the train-test format that provides the student with all the information and procedural knowledge needed to be able to execute the tasks of the test (see Hessels & Vanderlinden, 2011; Hessels et al., 2008; Tiekstra, Hessels, & Minnaert, 2009; Veerbeek, Hessels, Vogelaar, & Resing, 2017) .

For simplicity, this thesis only describes the first two commonly used formats that are also referred to as "the sandwich format" and "the cake format", respectively (Sternberg & Grigorenko, 2002, p.27), or "test-train-test design" and "train-within-test design" (Dörfler et al., 2009, p.78). An illustration of the two formats is depicted in Figure 5.3.

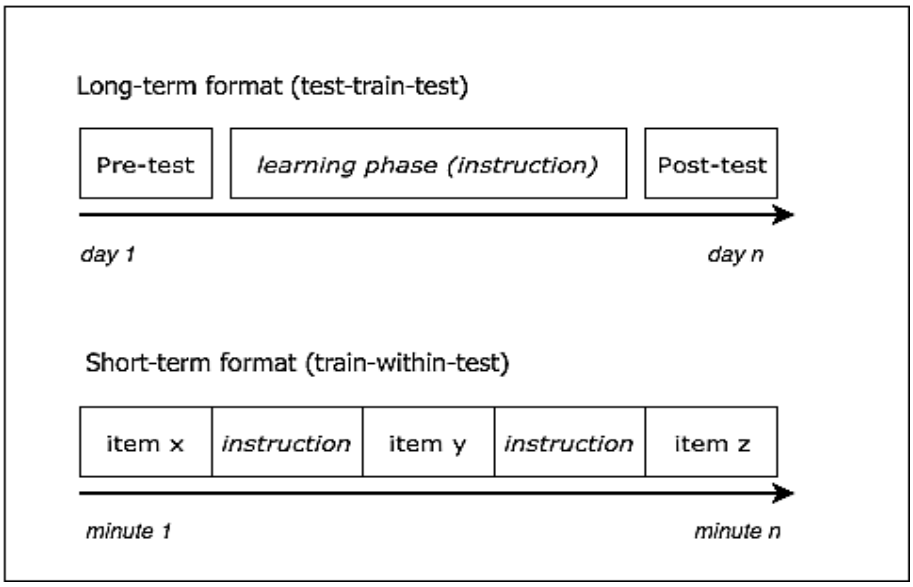


Figure 5.3: Two Common Formats of DT (adapted from Guthke (1993))

The first format is the classical learning format (Guthke, 1992, 1993), consisting of a pre-test, an intervention or a training phase and a post-test. Following the pre-test, examiners will be given instructions that are related to the skills measured in the test. This can be done individually or in groups (Guthke, 1993; Sternberg & Grigorenko, 2002). In an individual setting, the amount of training is individualised to reflect the examinee's response to the intervention, while in a group setting it is typically uniform for all students (Sternberg & Grigorenko, 2002). The post-test, which has parallel items to the pre-test, seeks to determine the extent to which the examinees' performance improves as a result of the training. The difference between the post-test and the pre-test is calculated as a reflection of students' learning potential.

The short-term format, in contrast, is often carried out in one session (Guthke, 1993; Guthke & Beckmann, 2000a; Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2002). Unlike the first format, training is integrated within the test items and consists of a series of interactive hints depending on the needs of the examinee. Here, when a test-taker answers an item correctly, the next, more complex item will be presented. But, if he/she answers incorrectly, hints are given to help him/her to successfully deal with the item (an elaborated illustration of how feedback is given can be seen in Figure 6.6). It is this assistance that plays a vital role in quantifying the test-taker's potential to learn

(Guthke, 1992, 1993; Grigorenko & Sternberg, 1998; Lidz & Gindis, 2003; Sternberg & Grigorenko, 2002).

In general, both formats embed interventions, i.e., feedbacks or prompts, to facilitate performance improvement. As indicated above, they are, however, different in terms of how assistance or training is provided. It is noted that the first format is time-consuming, thus the second format appears to be more practical for implementation (Guthke, 1992, 1993; Dörfler et al., 2009; Resing, Steijn, Xenidou-Dervou & Stevenson, 2011).

5.1.3 How different is dynamic testing from conventional testing?

As DT was developed to address the shortcomings of traditional tests, then the question worth asking is: how different it is from its counterpart? It is a point of importance to highlight the differences to shed light for practitioners on understanding the necessity of this approach to be applied in the educational context. In the subsequent paragraphs, I discuss two major "opposite characteristics" of the two assessment approaches.

5.1.3.1 Static vs. dynamic

The first principal difference pertains to the nature of the test. In the literature, the term "static" is frequently used to contrast DT with traditional testing procedures (e.g., Beckmann, 2006; 2014; Elliott et al., 2018; Elliott & Resing, 2015; Guthke, 1992; Guthke & Beckmann, 2000a, 2000b; Grigorenko & Sternberg, 1998; Lantolf & Poehner, 2007; Lidz & Gindis, 2003; Sternberg & Grigorenko, 2002; Tzuriel, 2000, 2005). Most traditional tests are static, characterised by the absence of any attempt to help a child to solve the given tasks (Elliott & Resing, 2015; Lidz & Gindis, 2003; Sternberg & Grigorenko, 2002; Tzuriel, 2000).

On the other hand, the uniqueness of DT lies in its "dynamic" feature of embedding the instruction or intervention in the assessment process (Beckmann, 2014; Elliott, 2003; Elliott & Resing, 2015; Guthke & Beckmann, 2000a; 2000b; Resing, 2013; Sternberg & Grigorenko, 2002). Beckmann (2014) further illustrated that the dynamism of DT is justified by "the systematic variation of task and situational characteristics in the item presentation" (p.310) that are incorporated in the test procedures. In many operational procedures of DT, particularly the short-term/cake format, the test-takers are provided with explicit feedback or prompts when solving cognitive tasks. Advocates of DT believe

such scaffolding may provoke an individual's learning process and may eventually initiate him/her to move to the next level of competence (Beckmann, 2006, 2014; Elliott, 2000, 2003; Haywood & Lidz, 2007; Lidz, 2014; Sternberg & Grigorenko, 2002; Tzuriel, 2000). It is noted that this supportive function originates from the notion of the more capable others of Vygotsky's ZPD.

In a conventional test, a test-taker will proceed to the next item, regardless of his/her success or failure in answering the item; whereas in DT, the test-taker receives a series of explicit feedback items if he/she gets the answer wrong. These built-in feedback items are provided to help him/her to move to the subsequent more challenging items. The goal of DT here is to examine the extent to which the test-taker can improve his/her performance after responding to the feedback given (see Figure 6.6 for more details of how prompts are provided to the examinees).

5.1.3.2 Product vs. process

Another aspect that differentiates the two is the measurement of the target construct. The focus of static tests, which is predominantly situated on students' performance of developed abilities, is often contrasted with the goal of DT, which emphasises the quantification of learning potential. This distinction is seen as a significant departure from a product-oriented assessment towards a process-oriented assessment. As noted, the construct of measurement for each assessment approach represents Vygotsky's two levels of a child's cognitive development.

In a static test, Resing and Elliott (2011) asserted that scores are an indication of the learner's zone of actual development, reflecting the level of mastery of learning that has occurred prior to the test. Its focus on the product of learning, nonetheless, has been criticised for not being able to offer an adequate picture of how a child learns, or why it might struggle to learn (Guthke & Beckmann, 2000a, 2000b; Elliott, 2003; Lidz, 2014; Resing & Elliott, 2011; Tzuriel, 2001, 2005). This has triggered a concern that a child's potential is often underestimated. Beckmann (2006) argued that the lack of learning occurs within the product-oriented assessment procedures because the test scores provide insufficient information about someone's "true" ability to learn.

Therefore, the proponents of DT postulated that the cornerstone of DT is the measurement of learning potential (Beckmann, 2014; de Beer, 2006; Elliott et al, 2018; Haywood & Wingenfeld, 1992; Grigorenko, 2009; Grigorenko & Sternberg, 1998; Lidz, 2014; Tzuriel, 2011). What is actually to be measured? In clarifying the concept of learning potential, Lidz and Gindis (2003) explained that it is a child's evolving cognitive capacity that involves both existing and projected future performance. The focus here is on the process of learning that is concerned about the cognitive modifiability of a child. Using a more quantifiable term, Beckmann (2014) described that the learning potential is operationalised as the responsiveness of the test-taker to the learning opportunities provided in the form of graduated feedback/prompts. This is similar to a definition given by Hamers and Resing (1993) – learning potential is what a child can do with proper guidance from others.

Notwithstanding the various definitions of learning potential, they are theoretically parallel with Vygotsky's ZPD. The assumption here is that cognitive processes are developmental and changeable; thus, learning potential, as manifested by the improvement in the test, can be quantifiable when some forms of help are provided. Vygotsky (1978) believed that what a child can do with assistance today will be his/her actual developmental level tomorrow. In agreement with this, Lauchlan and Elliott (2001) suggested that the estimates of learning potential are beneficial to (a) assess a child's cognitive modifiability, (b) predict future educational success, (c) make judgements about students with special educational needs, and (d) gain specific information about students' strengths and weakness for the planning of appropriate interventions.

In sum, DT has emerged as a response to the shortcomings of conventional testing, aiming to offer more informative descriptions of students' learning potential. Unique to DT, here is the summary of its fundamental assumptions:

(a) Explicit feedback and intervention are incorporated into the test procedures to provoke learning by the test-takers.
(b) The focus of the assessment is on the process of the observed learning potential (behaviour) from the learners' responsiveness to the prompts given.
(c) Assessing the learning potential (inferred from the learners' responsiveness/ modifiability) provides a more accurate description of the learners' "true" potential.

5.1.4 How successful is dynamic testing?

Having outlined the positive conceptual appraisals of DT in the preceding sections, the next important question to ask is: how successful is DT? It seems that there is a mixed reaction to the evaluation of whether DT has fulfilled its appealing "promises" in educational contexts. The following paragraphs discuss some of the conclusions suggested by the advocates of DT about this issue.

Despite the growing interest in the application of DT for educational settings, some critics have begun to point out the "flaws" of this assessment approach. Such challenges include the issues of validity and reliability (Glutting & McDermott, 1990; Lidz, 1995; Grigorenko & Sternberg, 1998; Tiekstra, Minnaert & Hessels, 2016). It is suggested that the empirical evidence of DT procedures has yet to demonstrate that they are more valid and reliable than conventional tests (Grigorenko & Sternberg, 1998; Glutting & McDermott, 1990; Sternberg & Grigorenko, 2002). It seems that these authors ignored European research that had already shown good reliability and validity.

Responding to the validity-related question, it is argued that lack of systematic empirical studies on DT validity is one of the reasons for its under-use (Beckmann, 2006; Caffrey, Fuchs & Fuchs, 2008; Sternberg & Grigorenko, 2002). Furthermore, a systematic review by Tiekstra et al. (2016) and Caffrey et al. (2008) also revealed that many of the studies did not explicitly report the reliability and validity of DT measures. To address the alleged inadequacy of this psychometric issue, Beckmann (2006, 2014) proposed that an explicit strategy should be devised to evaluate the validity of DT to convince practitioners of its values in education. He also added that validity is not about the test concept, but rather about "the appropriateness, meaningfulness and usefulness of the inferences drawn from test scores" (Beckmann, 2014, p.309). Furthermore, empirical attempts demonstrating the evidence of validity should not be put aside (e.g., Beckmann, 2006; Caffrey et al., 2008; de Beer, 2006, 2010; Hessels, 2009; Hessels-Schlatter, 2002).

While the above arguments are tendentiously concerned about the validity issues, other scholars have identified several practicality issues for its lack of widespread deployment in mainstream educational settings. Firstly, one apparent reason is lack of expertise in the field (Deutsch & Reynolds, 2000; Elliott, 2000; Lidz, 2014; Stringer et al., 1997; Tzuriel, 2011). This could be attributed to the fact that DT, or DA in general, has yet to be included

in education courses in higher education institutions and teachers are trained to use conventional tests in assessing students (Haywood & Tzuriel, 2002; Lidz, 2014; Tzuriel, 2011). Secondly, many educationalists are unfortunately still more product-oriented than process-oriented (Haywood & Tzuriel, 2002; Tzuriel, 2011). This is supported by the study of Deutsch and Reynolds (2000), in which the teachers who participated in the survey admitted they did not use DT regularly despite receiving training about it. Thirdly, it is also noted that many are still rather unfamiliar with this alternative approach (Deutsch & Reynolds, 2000; Hessels et al., 2008; Tzuriel, 2011). Even those who use DT still struggle with the interpretation of the data and the implementation of DT. Fourthly, advocates of DT acknowledged that the implementation of DT is time-consuming (Dörfler et al., 2009; Elliott, 2000; Guthke & Stein, 1996; Guthke et al., 1997; Haywood & Tzuriel, 2002; Hessels et al., 2008; Tzuriel, 2000). However, this only applies to the long-term tests and to clinical procedures.

Elliott (2000) emphasised that for DT to be effective, it requires extensive training, more experience and greater effort from the various parties involved in the assessment system. In order to make DT more compelling to researchers and practitioners, Elliott (2003) proposed several research directions to overcome its underutilisation. These include investigations about (a) the effects of DT on meaningful recommendations for intervention; (b) the employment of DT by practitioners; and (c) demonstration of successful gains resulting from the application of DT.

Probably due to the lack of use, a conclusion has been suggested that DT has not yet fully realised its potentials (Elliot, 2000; Elliott, 2003; Elliott et al., 2018). In an attempt to refute this, Beckmann (2014), however, argued that the impression of a broken promise is nurtured by a paradoxical conception of DT. It is a paradox of being too broad and being too narrow at the same time. He asserted that the assumption of the usefulness of DT centres on its validity. The problem is that the expectation that DT will always produce superior results in the context of validity studies seems unrealistic (Beckmann, 2006, 2014). This is, however, inconsistent with the conceptual understanding of how learning potential differs from what is measured in traditional static tests. Furthermore, DT has been perceived too narrowly, often equated with learning tests. Beckmann (2014) argued that DT, as a methodological approach to testing, is also applicable to the measurement of other dynamic constructs – e.g., learning ability (Guthke, 1992), learning

potential (Hamers & Pennings, 1995; Hessels, 1997; 2009), change patterns (Resing et al., 2009; Resing, Xenidou-Dervou, Steijn & Elliott, 2012), cognitive potential (Stevenson et al., 2016a), developmental trajectory (Resing, Bakker, Pronk & Elliott, 2017), intellectual change potential (Beckmann, 2001) and others.

Despite the doubts about its widespread use, proponents of DT (e.g., Beckmann, 2006, 2014; Lidz, 2014; Lidz & Gindis, 2003; Hessels, 2009) are optimistic that it will prove its worth as a relevant assessment approach if greater efforts to address its critical "drawbacks" are taken into serious consideration. Likewise, Stringer (2018), in his specific response to Elliott's (2003) argument, equates the process of realising the "promises" of DT with the concept of the ZPD. In particular, he quoted Vygotsky's (1978) saying, "the zone of proximal development defines those functions that have not yet matured but are in the process of maturation, functions that will mature tomorrow but are currently in an embryonic state" (p.86). He advocated that current evidence of DT's success is a positive sign that we are heading to a state of maturation. To fully claim that DT is superior to conventional testing, however, he concluded that "we have not yet reached 'tomorrow'" (p.26).

Taking these points together, DT largely emerges out of dissatisfaction with traditional static testing. But it is important to emphasise that DT should not be seen as superior to the static test, but rather as a complementary tool to aid teachers in improving the effectiveness of the teaching and learning process (Caffrey et al., 2008; Guthke, 1992; Guthke & Beckmann, 2000a; Haywood & Lidz, 2007; Lidz & Gindis, 2003; Sternberg & Grigorenko, 2002; Tzuriel, 2001). Although several doubts have been raised, DT has demonstrated to be more sensitive to the needs of those learners who seem to be side-lined by conventional testing. Furthermore, it has made its way to the educational setting, providing valuable insights for practitioners to understand learners in more meaningful ways than those that are revealed by static tests. This is the optimism that motivates me to embark on this research. In a situation where the currently used assessment battery seems to be less helpful, DT could possibly offer new insights to remediate the problem of the ineffectiveness of assessment in Malaysian schools. In the next section, the empirical evidence of the application of DT in the educational context is documented. It aims to demonstrate its relevance for use in enhancing teaching and learning.

5.2 Empirical Studies on Dynamic Testing

DT is conceptually appraised as an alternative assessment approach that could provide more meaningful insights about students' ability to learn. This information, in turn, is valuable for teachers in constructing appropriate interventions for scaffolding a child's learning development. Although the practicality of DT in real contexts is questionable, the impact and application of DT have been explored by researchers from various countries, particularly the USA, the Netherlands, Germany, Israel and the UK. Additionally, DT has been the subject of empirical investigation in numerous fields, e.g., special education, cognitive ability, language learning and other domain-specific situations. The robustness of the empirical literature implies that DT is not a new approach to educational assessment – it has been widely researched and fairly used for more than 50 years. The subsequent sections aim to review several studies about the application of DT in various educational settings. In this review, I include articles that employed the sandwich and cake formats; many of these studies primarily utilised standardised feedbacks or interventions in their DT procedures.

5.2.1 Validity of dynamic testing

It is emphasised that validity is one of the most fundamental psychometric properties in the evaluation of any assessment procedure (Cronbach & Meehl, 1955; Messick, 1980, 1989, 1995). There is a considerable consensus (see AERA, APA & NCME, 2014; Cronbach & Meehl, 1955; Kane, 2013; Messick, 1980, 1989, 1995; Shepard, 1993) pointing out that validity is not an inherent property of a test, but rather the interpretation and use of test information. Messick (1989) explicitly defined validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support *the adequacy and appropriateness of inferences and actions* [emphasis added] based on test scores or other modes of assessment" (p.5).

As previously mentioned, notwithstanding the level of appreciation of the promising potential of DT, it has received recurrent criticism about its validity. Allegedly, the proponents of DT have not yet succeeded in demonstrating convincing evidence for the validity of their assessment procedures (Glutting & McDermott, 1990; Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2002). According to Sternberg and Grigorenko (2002), "it is difficult to argue that this approach has proven its usefulness and has shown

distinct advantages over traditional static testing, relative to the resources that need to be expended" (p.180). Thus, continuing research on DT validation is essential to promote its practical contributions in mainstream education alongside traditional assessment methods (Caffrey et al., 2008; Guthke & Beckmann, 2000a; Lidz, 2014). In response to the validity-related doubts, the purpose of this section is to trace empirical studies conducted by advocates of DT who attempted to establish evidence of several aspects of the validity of their tools.

Because of the diversity of DT in terms of its goals, methods and theoretical perspectives on measurement, the task of addressing the issue of validity has proven to be challenging (Beckmann, 2006, 2014; Caffrey et al., 2008; Guthke et al., 1995; Lauchlan & Elliott, 2001). Thus, Beckmann (2006, 2014) proposed that a specific validation strategy should be devised to convince practitioners of the value of DT. His study (Beckmann, 2006), involving data from the administration of three learning tests of reasoning ability to 151 German eighth-grade students, is used to exemplify this strategy. His strategy focused on (a) clarification of the construct of DT; (b) explanation of construct-representative external measures; and (c) the type of validity to be measured. Firstly, he argued that "learning ability is seen as a central aspect of intelligence" (Beckmann, 2006, p.46) and thus both static and dynamic tests measure the same construct. However, traditional tests are at risk of construct under-representation because they do not provide test-takers with the opportunity to learn during the test. Apparently, there is a mismatch between the conceptualisation of the construct and the way it is measured. For this reason, the emphasis on the product-oriented information of the conventional test may be inappropriately indicative of one's ability to learn. The integration of the learning process in DT is therefore operationalised to address the construct-representation issue. Secondly, a criterion that represents an operationalisation of the target construct needs to be established before a test–criterion correlation can be concluded. Due to the process-oriented focus of DT, Beckmann (2006, 2014) argued that criteria should be predominantly qualitative rather than quantitative. In this respect, incremental validity is more uniquely attributable to DT than predictive validity (Beckmann, 2006, 2014). Thirdly, this study aimed to demonstrate predictive validity and incremental validity of DT measures. The result revealed that there was hardly any difference between DT and conventional tests in predicting participants' success in later domain-specific learning a year after the learning tests were administered. The findings, however, demonstrated

value-added information that is distinctive to DT. More importantly, incremental validity indicated that students' performance in the learning tests provided better estimates of their ability to perform in a curriculum-based test called COMBINATORICS, a computer-adaptive programme for the mathematics curriculum, which was not covered in the learning tests. In addition, incremental information also explained the presence of differential validity – the extent to which the scores correlated differently with different subgroups of the participants.

Such diagnostic gain from differential validity is also demonstrated in other studies such as Guthke and Beckmann (2000b), Hessels (2009), Hessels, Berger and Bosson (2008), Shabani (2012) and Watzke, Brieger, and Wiedl (2009). The results of Hessels et al.'s study (2008) indicated a significant training effect of DT measures in discriminating between learners with and without learning disorders. Besides, DT measures can also differentiate students in terms of their different learning strategies (see Elleman, Compton, Fuchs, Fuchs & Bouton, 2011; Resing, Xenidou-Dervou, Steijn & Elliott, 2012; Resing, Touw, Veerbeek, & Elliott, 2017; Stevenson, Heiser & Resing, 2016b). A study by Resing et al. (2017), for example, showed that DT procedures allowed for the identification of various reasoning strategies between indigenous and minority groups.

To assess the predictive validity of DT, Caffrey et al. (2008) reported a meta-analysis of 24 articles written between 1971 to 2000. In this review, the authors focused primarily on "the effects of deliberate, short-term, intervention-induced changes on student achievement" (Caffrey et al., 2008, p.257). The authors utilised mixed-method analyses to explore the data. In 15 of the studies, they used Pearson's correlation coefficients to examine the correlation between DT measures and achievement tests. For the remaining nine studies, no Pearson's correlation coefficients were reported and thus the outcomes were described narratively. Results indicated that both DT and conventional tests contributed similarly to the prediction of future achievement. Nonetheless, a very different result was obtained by (a) comparing two types of DT with contingent (individualised instruction) and non-contingent (standardised instruction) feedbacks and (b) examining the performance of sub-samples of the examinees. The review revealed that DT measures demonstrated unique predictive validity when the feedback of the assessment procedures was non-contingent, in which standardised instructions were given in response to student failure. As regards the student population, DT provided better

predictions of academic achievement of students with disabilities compared with at-risk or disadvantaged students and achieving students.

Results of the predictive power of DT in the above review are contrary to more recent studies (de Beer, 2010; Hessels, 2009; Hessels-Schlatter, 2002; Swanson & Howard, 2005; Swanson & Orosco, 2011; Watzke et al., 2009) that revealed a significant variance of DT measures in predicting later academic performance. For example, Watzke et al.'s (2009) study demonstrated that DT measures could be informative for long-term prediction. The study involved 41 schizophrenia patients who participated in a vocational rehabilitation program in Halle/Saale, Germany. The findings confirmed that estimates of learning potential from DT measures of cognitive flexibility significantly predicted work-related learning at six months into the rehabilitation programme.

In another systematic review, Tiekstra et al. (2016) analysed 31 studies to investigate the consequential validity of DT measures. The review involved studies reported from 1995 to 2011 and conducted in different continents, particularly in North America, Europe and Africa. In DT, the integration of feedback into the assessment procedures is assumed to reveal the learning potential of a child and this, in turn, may provide useful information for the design of student-tailored learning intervention (Tiekstra et al., 2016). The authors argued that it is of importance to examine the potential influence of testing procedures in practice, i.e., the extent to which students profit from the testing procedures. For the purposes of the review, consequential validity was categorised into two: (a) proximal consequential validity (the effects of testing that trigger learning during the testing procedures) and (b) distal consequential validity (the effects of testing later in the instructional environment). The former was measured from the raw or standardised scores and the latter was drawn from explicit information about the adaptation of DT outcomes in instructional practices. The outcome of the review concluded that proximal consequential validity was warranted because a learning opportunity was provided during the test. In fact, there were four studies (Elleman et al., 2011; Swanson, 2011; Swanson & Howard, 2005; Tiekstra, Hessels & Minnaert, 2009) that showed superior predictive power than those of traditional tests. Interestingly, distal consequential validity contributed to the planning of later classroom interventions to a lesser extent. This was reflected in the lack of information or guidelines for practices concerning how the estimates of learning potential are valuable for instructional decision-making. The

majority of the studies focused on reporting the raw scores and the link to the intervention following the test was not made explicitly. To unravel distal consequential validity, the authors suggested that a detailed description of learning phases and types of feedback should be highlighted in the construction of DT procedures and the reporting of the studies.

In sum, the growing increase of empirical studies is a positive sign that a successful validation of DT can be established. The outcomes of the reviewed articles provide a consolidation that DT is sufficiently valid for application in various educational contexts. More importantly, it confirms that DT is able not only to offer similar information as those of conventional tests (predictive validity) but also unique additional information (incremental validity) about students' ability to learn. This is to say, such a claim concerning the paucity of empirical evidence demonstrating the validity of DT cannot, therefore, be corroborated.

5.2.2 Applicability of dynamic testing in education

DT emerged because of constant complaints regarding the effects of traditional tests on particular groups of learners and their deficiency in providing an accurate description of these students' "true" potentials. Scholars of DT asserted that static tests seem to underestimate learning potentials of students who are academically at-risk (Guthke & Beckmann, 2000a; Elliott & Resing, 2015; Hamers & Pennings, 1995; Hessels, 1997; 2009; Sternberg & Grigorenko, 2002; Tzuriel, 2001; 2005). Perhaps one of the most appealing contributions of DT has been the works with disadvantaged children (Hessels et al., 2008; Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2002; Tiekstra et al., 2016). According to Sternberg and Grigorenko (2002), disadvantaged students are "a large class of pupils viewed as having reduced learning opportunities" (p.33). They further explained that this could be attributed to (a) lack of previous education; (b) differences in cultural and educational opportunities; or (c) explicit learning disabilities or mental deficiencies. Initiated by this definition, the works described in the subsequent sections highlight the practicality and usefulness of DT measures for the above group of learners in various educational settings.

Research results have proven that DT is worthwhile and resourceful for ethnic minority students (e.g., Hamers & Pennings, 1995; Hamers, Hessels, & Pennings, 1996; Hessels, 1997; Resing, Tunteler, de Jong, & Bosma, 2009). Hessels (1997) studied a sample of 445 Moroccan and Turkish students living in the Netherlands and 115 Dutch students, aged between five and eight years old. The author argued that the classification of ethnic minority students using the scores from traditional IQ tests may be highly inadequate in measuring students' general cognitive abilities. This may lead to inappropriate placement of students into special education classes. To address this problem, the study used two subsets of the Learning Potential Test for Ethnic Minorities (LEM) to estimate the general cognitive ability of minority group students, attempting to uncover a better estimate of their true potentials. The findings revealed that the LEM has shown to be a valid and reliable instrument not only for children from different ethnic backgrounds but also for indigenous Dutch children. Essentially, there was a strong relationship between the LEM estimates and students' later performances in achievement tests of mechanical reading, reading comprehension, spelling and arithmetic. Hessels (1997) concluded that scores in DT measures are better predictors of learning progress over time (an interval of six months) than IQ scores.

In another study, Resing et al. (2009) conducted a study aimed at examining the effects of a graduated prompt approach on students' learning strategies. Particularly, the study compared a change of strategy pattern in a seriation task, from pre-test to post-test, between second-grade students from ethnic minority backgrounds (n=55) and indigenous Dutch background (n=54). The study employed a pre-test-post-test two-group design, in which both indigenous and ethnic minority students were divided into control and treatment groups. The Seria-Think, a DT method of seriation and early mathematics skills, was administered as training for the treatment group. A short version of Raven's Standard Progressive Matrices (SPM) was given as static pre-test and post-test for all students. The hypothesis of the study confirmed that DT with graduated prompt techniques significantly contributed to the progression of strategy use – from less to more sophisticated strategies – for both Dutch and ethnic minority students. As expected, the latter group benefited the most from the training – they were able to show greater progress in using the higher-level strategy as compared with their indigenous peers. In particular, the information from graduated prompt training, i.e., the amounts and various types of hints, helped in the identification of students' strengths and weaknesses in strategy use.

More importantly, dynamically trained students performed better in more complex tasks that required superior strategy use at post-test. This further supported the assumption that static tests (as reflected from the pre-test scores) were not able to discover the "real" cognitive potential of the participants. The findings of this research correspond to more recent studies (see Resing et al., 2017; Stevenson, Heiser & Resing, 2016a; Wiedl, Mata, Waldorf & Calero, 2014), which indicated promising contributions of DT in reducing the effects of cultural differences between dominant and minority groups, and that, therefore, test bias could be minimised.

Furthermore, much of the criticism against traditional tests have stemmed from concerns that students with special learning needs or disabilities may be inappropriately placed and diagnosed (Elliott, 2000; Elliott & Resing, 2015; Lauchlan & Elliott, 2001; Tiekstra et al., 2009). To demonstrate the feasibility of DT for this group of children, Lauchlan and Elliott (2001) employed a pre-test, intervention and post-test type of DT to 30 English children with severe learning difficulties. All students were given both dynamic and static measures as the pre-test and post-test. Following the pre-tests, students were classified as high and low potential groups. 15 of the children (high: n=7, low: n= 8) received a 15-month cognitive intervention programme. The Cognitive Modifiability Battery (CMB), a series of tasks on the inductive reasoning process, was utilised for the intervention. The analyses of before-and-after intervention revealed that DT is useful in predicting the learning progress of students. It was notable that students with low potential gained the most from the intervention programme – they showed a considerable improvement from pre-test to post-test in comparison with their peers who demonstrated high potential but were not exposed to the training. In line with this finding, Lauchlan and Elliott (2001) asserted that the measurement of learning potential may only be valuable if it is accompanied by a subsequent intervention. The effect of training is also replicated in a study by Hessels et al. (2008) in which the estimates of cognitive ability of the experimental group at post-test were seen as a better predictor to students' success in later achievement tests.

Aside from much emphasis of DT in measuring cognitive ability, previous studies also herald evidence that DT can be helpful for students experiencing difficulties in curriculum-specific measures such as reading (Elleman et al., 2011; Swanson, 2011; Swanson & Howard, 2005; Swanson & Orosco, 2011), geography (M. G. P. Hessels,

2009) and chemistry (Tiekstra et al., 2009). In 2005, Swanson and Howard sampled 70 South Californian children who received special education services as an attempt to investigate whether DT can provide accurate classification of students with reading disabilities. This sample consisted of four groups of students: skilled readers (*n*=25), poor readers (*n*=14), children with reading disabilities (*n*=12) and children with reading and mathematics disabilities (*n*=19). There were two important findings that emerged from this study. Firstly, the findings supported the validity of DT measures to facilitate better classification of students with reading disabilities. Specifically, after the change scores from pre-test to post-test were analysed, it appeared that approximately 40% of students with reading disabilities and 30% of students with both reading and mathematics disabilities were inappropriately diagnosed. Secondly, the results also confirmed that DT procedures significantly predicted later performance in reading and mathematics and this is another piece of evidence that consolidates the presence of the incremental validity of DT.

In addition to the above groups of learners, DT is also regarded as a valid instrument for students who are academically at risk from intellectual disabilities. Hessels-Schlatter (2002) claimed that traditional IQ tests do not provide a reliable and valid estimate of general cognitive abilities of students who suffer from intellectual disabilities. Thus, the Analogical Reasoning Learning Test (ARLT), a dynamic test of analogical reasoning, was employed in her study with 58 special students with moderate to severe intellectual disabilities. The result yielded the validity of ARLT as a useful instrument to measure the cognitive abilities of this group of students. The predictive validity was also warranted, showing an improved performance by students who received a month's training on inductive reasoning. The result of this study coincides with Watzke et al.'s (2009) study attributing the measurement of learning potential as an informative predictor of work capability of patients with severe mental illness.

Apart from the wealth of evidence of DT in favour of academically disadvantaged learners, many studies have revealed the promising practicality of DT for other groups of students, such as gifted students (Calero, Belen & Robles, 2011; Vogelaar & Resing, 2016; Vogelaar, Bakker, Elliott & Resing, 2017), young learners (Kantor, Wagner, Torgesen, & Rashotte, 2011) and students in higher educational institutes (Embertson, 1987; Nirmalakhandan, 2007, 2013; Shabani, 2012). In a study by Calero et al. (2011),

for instance, the authors aimed to investigate the usefulness of DT in separating 127 Spanish students into gifted and non-gifted groups. It was revealed that DT not only proved to be reliable and valid for accurate identification of the sample, but it also pointed out significant intergroup differences about their performance in a diverse number of tasks. Furthermore, Shabani's (2012) study provided more empirical evidence that DT is practical for university students. His research on the feasibility of computerised DT of reading comprehension, involving 100 undergraduate students, proved to be useful for later decision-making, particularly for student placement and planning of remedial tutorials for reading modules.

As indicated by the evidence from the above-mentioned studies, it can be concluded that DT has shown rewarding potential in various educational contexts. It is also important to highlight that the assumption that DT narrowly favours specific population, e.g., ethnic minority and special students, is not conclusive. Empirically, previous studies have also documented the appropriateness of DT for other groups of learners.

5.3 Summary and Gaps of Previous Studies

In response to criticism of traditional tests, DT has been proposed as a way of uncovering information that accurately manifests students' "true" learning potentials. The above sections have suggested that the application of DT in educational settings has been positively appraised by scholars in many countries. In the following paragraphs, I will present a summary of what has been covered in previous research. I will also outline the focus of the present study, as an attempt to explore what is lacking in the literature.

The content of the reviewed articles can be summarised as having three focuses: (i) validity of DT measures; (ii) target population of DT; and (iii) DT of cognitive and domain-specific abilities.

Firstly, the literature has often focused on the issue of validation of DT procedures. As stated, this has primarily stemmed from the constant debates about "insufficient" evidence to support its validity as an appropriate assessment approach for the measurement of learning potential. Nevertheless, earlier studies, as well as the most recent works, have proven that such a claim is contradictory to the empirical evidence, which

shows that many of the DT procedures – the short-term and long-term formats – are reliable and valid for use. Overall, they showed a remarkable predictive validity (quantitative aspect) as those of the static tests. Furthermore, they have produced a consistent and unique incremental validity (qualitative aspect) in providing valuable insights that are beyond the information provided by the traditional tests. Ongoing research is however imperative to provide solid evidence in support of the validity and reliability of DT measures for widespread and practical applications (Beckmann, 2006, 2014; Caffrey et al., 2008; de Beer, 2006; Lauchlan & Elliot, 2001; Lidz, 2014; Sternberg & Grigorenko, 2002).

Secondly, many of the previous studies aimed largely at researching the feasibility of using DT with vulnerable individuals who often perform poorly in conventional tests. The findings of the reviewed studies have highlighted the value of DT as an alternative assessment approach that can potentially reduce the effect of educational inequalities for ethnic minority groups and students with special learning and intellectual disabilities. Interestingly, there is also growing evidence that application of DT can be extended to other groups of learners. Several studies have showcased encouraging results about the predictive and incremental information of DT for gifted pre-school students and university students. It is therefore disputable to say that DT only favours certain groups of the population.

Thirdly, the literature on DT has put greater attention on dynamic tests that measure a cognitive (psychological) trait (e.g., intelligence and language development) of students. Previous studies have pointed out the contribution of DT of general cognitive abilities in explaining students' improved performance in curriculum-based tests. In recent years, the proponents of this assessment method have developed DT procedures measuring performance in specific domains, e.g., language, geography and chemistry. These tests are used to measure school learning (curriculum-oriented tests) and are utilised in studies establishing the predictive utility of DT that measures intelligence. This renewed interest is still in its infancy, but it has shown a positive direction for empirical explorations and applications.

Collectively, the aforementioned review has demonstrated the applicability of DT in educational settings, with the study samples primarily focusing on the main recipient of

the assessment-related information, i.e., students. However, it is the goal of this study to obtain responses from another crucial group of users of assessment – the teachers – about the value of DT in instructional decision-making. Although there are a few studies (e.g., Bosma & Resing, 2010; Bosma, Hessels & Resing, 2012; Deutsch & Reynolds, 2000; Touw, Vogelaar, Verdel, Bakker & Resing, 2014) that embark on investigating teachers' views about the advantages of DT, there is still notable scarcity on this topic. In the context of the present study, this new assessment approach is entirely new to Malaysian teachers. Thus, the focus of previous studies on examining the direct effects of DT on the measurement of students' learning potential in the Malaysian context may not be timely. Arguably, it should be the teachers who must be well informed about this alternative tool before it can be potentially implemented for their students.

It is important to take into account that effective use of assessment requires certain conditions to be met, particularly the knowledge and skills to interpret assessment information, but, most importantly, teachers need to make changes to their thinking and continuously reflect on it in order to use assessment in a proper way (Mertler, 2003; Popham, 2009; Stiggins, 1995). It is also recommended that any attempt to change teachers' understanding of assessment must acknowledge the importance of equipping them with knowledge and skills required to appropriately interpreting assessment information (Mertler, 2003, 2005; Popham, 2009; Siegel & Wissehr, 2011; Stiggins, 1995; Volante & Fazio, 2007).

In the introductory chapter, I have proposed that the current phenomenon of poor academic performance among Malaysian students could potentially be caused by the use of unsuitable and uninformative assessment tools. Specifically, I have questioned the suitability of the currently used diagnostic test, i.e., F1DT, to provide adequate, meaningful and useful information needed to optimise students' learning performance. It is proposed that an alternative assessment method is therefore needed. As an attempt to address this issue, this research aims at introducing the concept of DT to teachers, to promote the potential contributions of DT in providing more meaningful information about students' learning ability than those offered by F1DT. Supporting the importance of assessment literacy for successful utilisation of assessment, the introduction of DT can be considered as a first practical strategy towards a more profound understanding of the

new assessment tool. It is a step to prepare teachers for the rationales for its use and to provide hands-on experience of its application in a school setting.

Taking all the above perspectives into account, the central interest of this study is to examine whether this introduction can change teachers' views about their reported perception and practices of F1DT. This is in line with the assertion that "research is necessary on the nature and process of belief change itself" (Pajares, 1992, p.329). It is also the interest of this study to examine teachers' opinions about the prospect of DT implementation in Malaysian schools.

5.4 Chapter Summary

In conclusion, literature in the field of DT has advocated refreshing paradigms and ideas as well as appealing results. This chapter has covered conceptual and empirical reviews of DT as an alternative assessment tool for conventional tests. In the conceptual review, I have outlined the rationales for the emergence of DT and the theoretical basis from which its operationalisation has originated. I have also described its features and discussed the debates about its practical applications. Correspondingly, the empirical review has traced numerous studies, conducted in different parts of the world, investigating the feasibility of DT in education. The last section has synthesised what is known in the literature and what remains to be explored. Most notably, the insights gained from this literature review are instrumental in the design of the intervention as described in the chapter that follows.

6 Intervention: Introduction and Demonstration of Dynamic Testing

As conceptually advocated and empirically researched, DT has much to offer for the effective utilisation of assessment-related information. Particularly, it is more sensitive to the needs of low-performing and disadvantaged children. Advocates of DT claim that traditional tests do not seem to do "justice" in representing an accurate description of these students' ability to learn. Relating this to the context of Malaysia, it is argued that F1DT, one of the existing traditional methods, provides information that is less valuable for teachers to make an informed decision about students' real potential. It is, therefore, fitting for schools to explore alternative assessment approaches that could offer meaningful information that is more valid and reliable in helping teachers and students to fully develop the learning potentials.

The premise of this study suggests that DT could potentially ameliorate the deficits of F1DT. Aiming to explore teachers' shift in attitude and beliefs, this study designed an intervention to familiarise teachers with the idea of DT. This intervention – with the goal of maximising the effect – comprised of two major components – experiential and educational. In this chapter, a description of these components is outlined. The first section is an account of the experiential element, in which a computer-adaptive version of DT was administered to students. This section also presents the description of the test items and the procedural administration of the test. Subsequently, the second section describes the educational component; detailing the contents of the educational talk and its implementation to teachers in participating schools.

6.1 Experiential Component

The first component of the intervention aimed at collecting "real data" that would be shared with the teachers participating in the CPD workshop. The administration of a computer-adaptive dynamic test called the Learning Test (LT) to Form 1 students was intended to be a means to an end. The main objective of this experiential component is to demonstrate a simulation of the application of DT to teachers. It is this vicarious experience of the teachers that is part of the intervention strategies (not the experience the tested students made).

Although the accuracy of the data is not a great concern at this stage, systematic and rigorous planning for this experiential component was attempted to ensure effective delivery to the teachers. The descriptions of the test and procedures for administration are presented in the following sections.

6.1.1 The Learning Test

The LT is an example of a short-term format, designed in such a way that the training phase is embedded in the test procedure. The test was an adapted version of the Adaptive Sequential Figure Learning Test (ADAFI), a subset of the three short-term learning tests from the Adaptive Computer-Assisted Intelligence Learning Test Battery (ACIL). The ACIL was used by Guthke and his colleagues (Guthke, 1992; Guthke et al., 1995; Guthke et al., 1997; Guthke & Beckmann, 2000a, 2000b; Guthke & Stein, 1996) to various groups of students in Germany.

Generally, most of the original items were retained and changes were only made to several items..After that, I appointed a professional programmer to create the test as computer-adaptive with two versions – the static and dynamic versions. The test was translated in Malay language. To ensure the accuracy of the translation, the dummy version was reviewed by several teachers.

6.1.1.1 Description of the test items

The LT consists of 32 items of figural sequences measuring reasoning ability. The items are classified into three complexity levels which are characterised by a number of varying dimensions – colour, shape and gestalt. Table 6.1 below explains each item category of the LT.

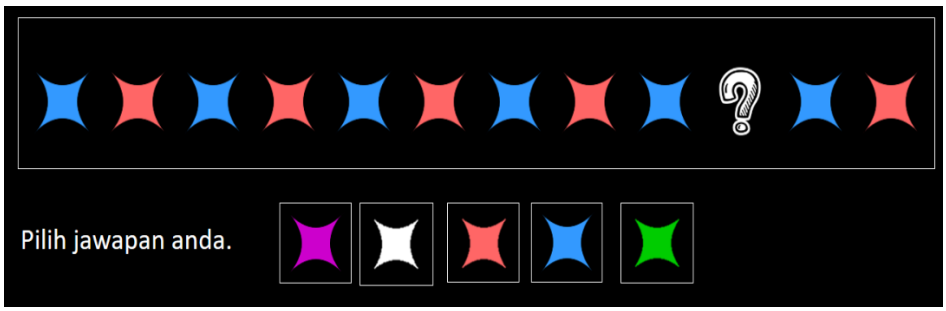
Table 6.1

Description of Classification of the LT Items

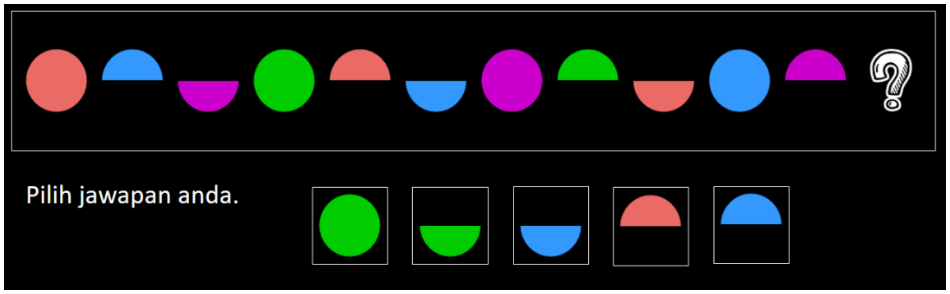
Complexity level	Description	Dimension	Number of items
Level I	One-dimensional item	Colour or shape	8 (1–8)
Level II	Two-dimensional item	Colour and shape	12 (9–20)
Level III	Three-dimensional item	Colour, shape and gestalt	12 (21–32)

The following figures illustrate the items for each level of complexity. Each item consists of 12 individual figures. In Figure 6.1, item 1 belongs to level I, characterised by its one-dimensional element – the shape. Items at level II are structured by combining two dimensions of colour and shape. As illustrated in item 14, the 12 figures contain the variations of four colours and the orientations of full and half circles. Item 23 represents a group of items at level III, the most complex level, with a combination of three dimensions. This item has four different colours, two different shapes and three variations of line border.

Level I: Item 1



Level II: Item 14



Level III: Item 23

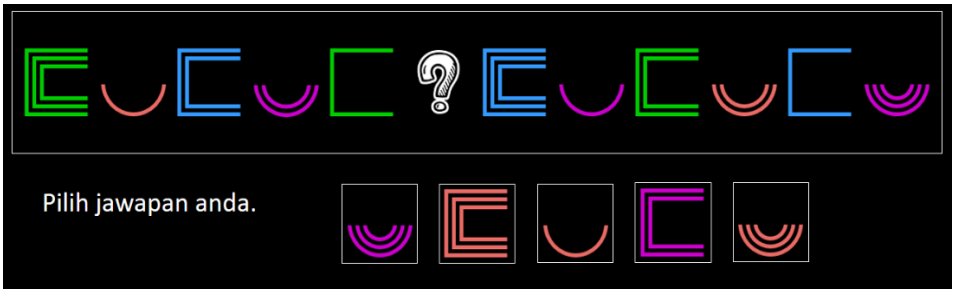
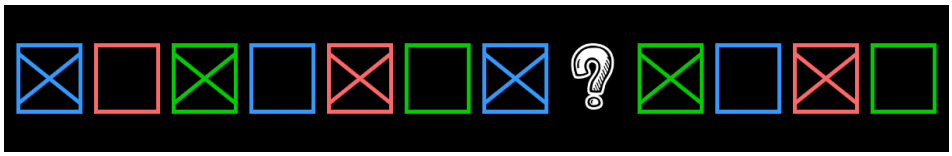


Figure 6.1: Examples of Item at Three Complexity Levels

Within each complexity level, the difficulty of the items is determined by gradually increasing the number of variations in a given dimension and by including symmetries in the sequence of the figures. This is demonstrated by the items in Figure 6.2. Items 10 and

11 are grouped in level II, characterised by the combination of two dimensions – colour and shape. Item 10 comprises figures of two different shapes with three variations of colour – blue, pink and green. Item 11 is also characterised by two shapes and three colours. However, the item difficulty is raised by adding an additional feature to one of the dimensions, i.e., the colour. In this item, note the reversal in the sequence of the third colour.

Item 10



Item 11

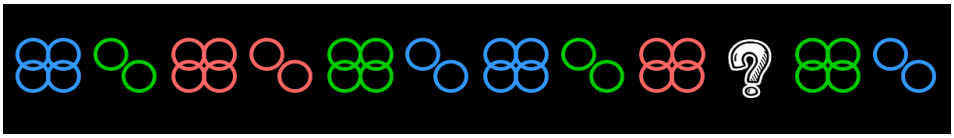


Figure 6.2: Examples of Item Difficulty

Furthermore, the item pool within a complexity level is structured in such a way to have two parallel items with the same information values (see Table 6.2). Items 9 and 10, for examples, have similar numbers of variation (two and three) for each of their two dimensions. The next group – items 11 and 12 – also has a similar pattern. It is noted that the difficulty of the task in items 11 and 12 is raised. This is indicated by the inclusion of the symmetries in one of the dimensions.

Another important element in the item pool is the presence of the obligatory items: four items for each complexity level. These items function as representatives of the respective complexity level. As seen in the above table, level II has two pairs of obligatory items. Items 13 and 14 represent a group of intermediate items, while items 19 and 20 prepare the students for another group of items of advanced complexity (further details of how these items work are explained in a later section).

Table 6.2

Information of Item Pool within a Complexity Level

	Level II											
Dimension	9	10	11	12	13	14	15	16	17	18	19	20
Colour												
1												
2												
3			S									
4							S				S	
5												
Shape												
1												
2												
3				S								
4								S				S
5												

Note: S = symmetry
Green=obligatory items
Blue=dimension(s) of the item

6.1.1.2 The procedures of test completion

The static version takes the form of a conventional computer-adaptive test (without feedback and assistance). In the dynamic version, the software is programmed to incorporate systematic predetermined prompts into the testing procedures. The presentation of how LT works as a dynamic version are explained below.

The flowchart in Figure 6.3 depicts the ramification rules to complete the test.

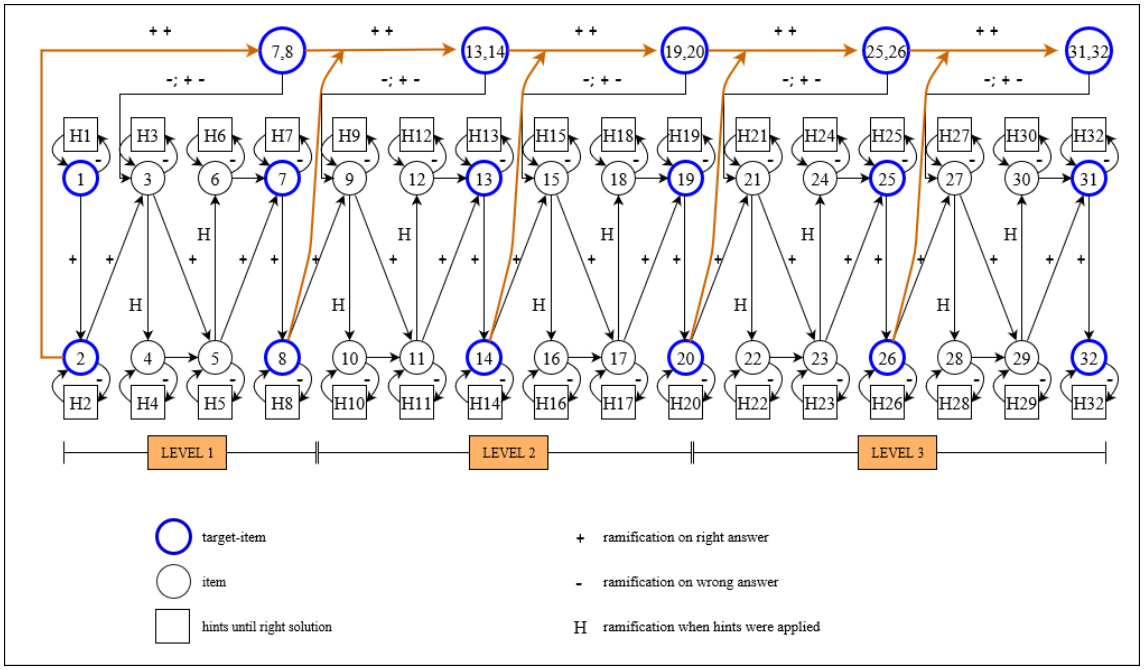


Figure 6.3: The Routes to Complete LT
(Guthke & Beckmann, 2000b; Guthke & Stein, 1996; Guthke et al., 1997)

Every student starts with items 1 and 2. These are the obligatory items in level I, functioning as a "warming up" as well as providing a basic understanding of the demands of the tasks. If both are answered correctly, the student can skip four items (items 3, 4, 5 and 6) in this complexity level. He/she can go straight to the next target items, i.e., items 7 and 8 (see the orange arrow). If the student gives correct answers to this pair, he/she can skip items 9, 10, 11 and 12 and progress to items 13 and 14 in level II. This means that when both obligatory items (in the blue circle) are solved correctly, the student can move to the next obligatory items in the subsequent levels (up to items 31 and 32). Moving from one pair of obligatory items to the next pair is only possible when the student manages to solve each pair (both items) correctly at the first attempt. If this happens, the student is only presented with 12 items in total, instead of 32 items.

In the case of answering one or both target items incorrectly, the route to answer the in-between-target items applies. For instance, if the student gets item 2 wrong at the first attempt, he/she will receive a textual prompt informing him/her that the choice is incorrect. At the second attempt, if he/she manages to get it right, this leads him/her to the next item, i.e., item 3. If he/she successfully answers item 3, he/she can progress to item 5. However, if he/she is unable to solve item 3, he/she needs to go through item 4

before moving to item 5. This “detour” aims to train the student to solve the demands of the tasks step by step (Guthke & Beckmann, 2000a). In the case of wrong answers, the level of assistance plays a crucial part in understanding and measuring the learning potential of an individual.

For each item, a student is presented with a rubric of instruction about the task. A sequence of 12 figural elements appears on the screen. The student is asked to find the missing figure (indicated by a question mark) from five alternatives shown below the item (see the screenshot of the task and example of the item in Figure 6.4).

Figure 6.4: Rubric and Example of the Task

When students respond to a question, the test automatically provides accuracy feedback (correct or incorrect). The feedback informs the test taker about the correctness of their response (see Figure 6.5). After a correct answer, the student receives the feedback “correct” and can progress to the next question. However, if he/she provides incorrect responses, the feedback “incorrect” will appear and the student is given the chance for a retry. When the student fails to answer the question at the first attempt, another prompt

appears containing an error-specific pictorial clue for assistance. Three pictorial cues in Figure 6.5 below show an example of possible prompts for items in level III, indicating its three dimensions – colour, shape and gestalt. In this example, this prompt is shown after the test taker has chosen an answer that was neither in accordance with the rule that governs the sequence of shape nor colour. Essentially, systematic feedback and clues function as training for the student to deduce the rules of the dimensional structure and sequences (Guthke & Beckmann, 2000a). They are also intended to evoke learning processes and to enable the test taker to proceed to the next complexity level.

Figure 6.5: Example of Pictorial Clues for Level III (Item 23)

As elaborated above, the student may work through various possible pathways depending on his/her ability to respond to the tasks and the systematic prompts integrated into the test procedure. The number of possible correct items and the level of assistance, therefore, vary for individual student.

6.1.1.2.1 Number of prompts

The number of prompts varies across the three complexity levels. This means that the amount of assistance for the student to reach a correct answer is determined by the number of dimensions for each level. Should the student get an item wrong in level I, he/she receives the textual feedback "incorrect" after the first attempt and a pictorial clue if he/she still gets it wrong in the second attempt. In case of an incorrect answer after the second attempt, the correct solution is shown. Thus, level I ends at the second attempt per item as all the visual clues for the five alternatives are either colour or shape. In levels II

and III, the student may have the opportunity of receiving three or four prompts, respectively. This depends on the student's susceptibility or responsiveness to the error-specific hints to solve the item correctly. Figure 6.6 depicts the possible options for students to complete an item, illustrating how textual feedback and pictorial clues are given at each attempt.

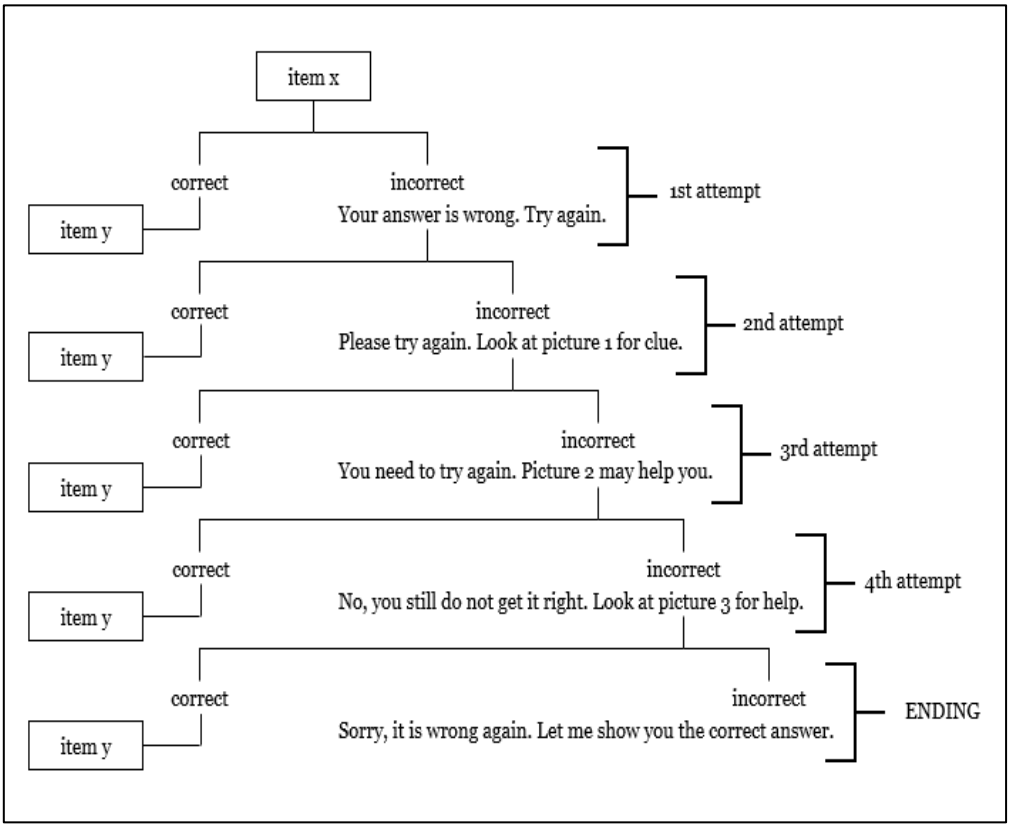


Figure 6.6: Possible Prompts to Complete an Item

6.1.1.2.2 Sequence of prompts

It is also important to mention here that the sequence of the hints is influenced by the responsiveness of the student to the alternative chosen. From item 23 above, if the student chooses the first alternative (item 23_1) after a prompt is given, a picture with a colour cue will appear to help him/her to solve the task (see Table 6.3 below). This is an indication that he/she has not grasped the notion of colour sequence in the task. If the fourth alternative is chosen (the most implausible alternative), all hints will be presented together. This suggests that the student has not understood the rules about the sequences of all the dimensions.

Table 6.3

Examples of Predetermined Prompts of LT Items

item	Sequence of prompts				
23	item23_1	item23_2	item23_3	item23_4	item23_5
	colour	shape	gestalt	all hints	
24	item24_1	item24_2	item24_3	item24_4	item24_5
	gestalt		all hints	shape	colour
25	item25_1	item25_2	item25_3	item25_4	item25_5
	gestalt		colour	all hints	shape
26	item26_1	item26_2	item26_3	item26_4	item26_5
	shape	colour	all hints		gestalt

6.1.1.2.3 The scoring procedure

In the static version, the scoring score is simplified into 1 for a correct answer and 0 for an incorrect answer. For the dynamic test, the scores were calculated depending on the number of correct items and the amount of assistance provided to arrive at the correct answer. The total score varies for each complexity level. At level I, two marks are given to each item if the student gets it right at the first attempt. One mark is subtracted for each of the prompts given if the answer is incorrect. At levels II and III, each item carries three and four marks respectively. This is characterised by the number of possible attempts for each level and each prompt will result in a deduction of one mark from the maximum mark. In the end, the total mark of all the levels is calculated. The assumption is that the higher the score, the higher the learning potential of the student.

6.1.2 Administration of the test

The test was administered to Form 1 students of SMK Padang Berampah, Sipitang. The school was chosen due to its easy access and convenience to the research schedule. In the subsequent paragraphs, the procedures of student recruitment and test administration are presented.

6.1.2.1 Recruitment of the students

The test administration involved 18 students in Form 1 (aged 13) of different ability groups. Assistance was sought from class teachers to select students from three different groups – good, intermediate and weak – to participate in this study. The classification of the groups was determined based on the students' UPSR results.

As stated, LT was administered in two versions – static and dynamic. Each version was presented to students from three ability groups (as illustrated in Figure 6.7).

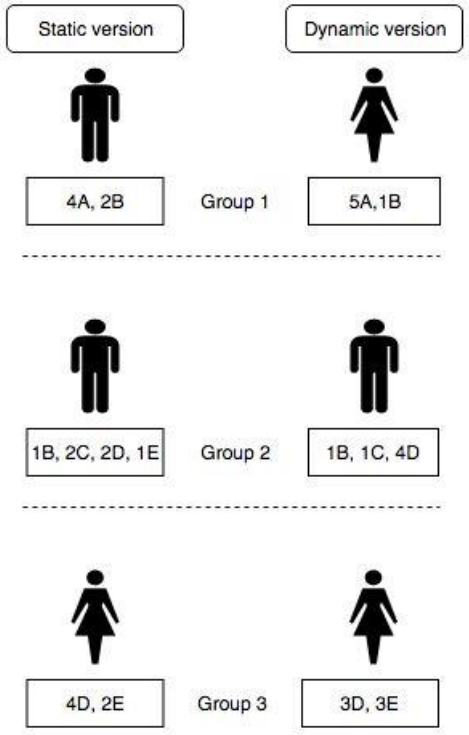


Figure 6.7: Ability Groups for Static and Dynamic Versions of LT

The above strategy of student recruitment for both versions aimed at (a) explaining the differences between students' performances in the static test (without assistance) and the dynamic test (with assistance); and (b) demonstrating students' performances across different abilities when assistance is offered. It would be expected that the differences between ability groups were greater when looking at the performance in the static test in contrast to the dynamic version. Also, one would expect that the difference between the non-dynamic and dynamic performance would be greatest for the low ability group. Such information is crucial to show to teachers the potentials of DT in providing more meaningful insights about students' potential to learn.

6.1.2.2 Administration of the test

Before the test, a parental consent form was distributed to parents or guardians of the students to ask permission for their child's participation in the research (see

Appendix 6A). In preparing for the test, I received assistance from a computer technician who helped me to book the computer lab and install LT on the laptops.

The test was administered in the computer lab. It was conducted in three different sessions because the students were from different classes. For each session, the static and dynamic versions were administered simultaneously.

At the beginning of the session, a short instruction was conducted to inform students about what they needed to do. Then, each student was given a laptop containing a static or dynamic version of LT. The assignment of the version was predetermined based on their UPSR results. The students were allowed to leave the computer lab when they had completed answering all the questions. After the test, students' scores were recorded.

Figure 6.8: The Set-up of the Test Administration

The preceding pictures show how students were set up for the test (permission to take pictures was explicitly mentioned in the consent form in Appendix 6A). The data collected serve the purpose of providing "real life" examples of the application of DT in a school setting. This information is instrumental in the design of the educational component in which teachers would be familiarised with the concept of DT.

6.2 Educational Component

This section describes the content and implementation of the educational component of the intervention. The school principals agreed to include this educational talk as one of the staff continuous professional development (CPD) programmes.

This professional workshop was designed to have a dual purpose of introducing the concept of DT and, more importantly, to demonstrate the contribution of DT to better understand students' learning potentials. Essentially, what teachers gained from this talk was expected to be reflected in their responses to the SEA questionnaire. As the objective of this study is to examine the effect of familiarising teachers with a new concept of assessment approach on their perception and practice of F1DT, it was critically important that the content concerning DT was systematically organised. Furthermore, it was also intended that this talk could elicit teachers' reflective feedback about the practicality of DT in Malaysian schools.

As an attempt for effective delivery, I designed the content of the talk and its implementation as follows.

6.2.1 Content of the educational talk

Like most professional training, the educational component was structured into three main stages – an introduction, a body and a conclusion.

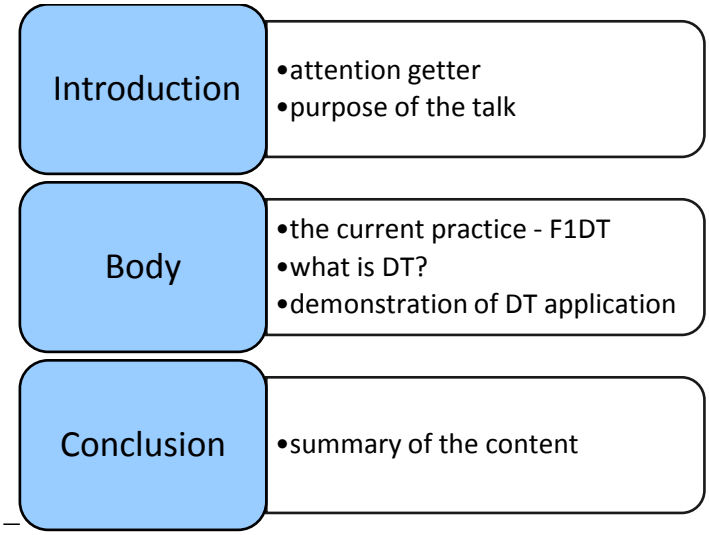


Figure 6.9: The Structure of the Educational Talk

Figure 6.9 above previews the key points of the three stages. In describing the content of each stage, several PowerPoint slides that I used for the workshop will be presented in the following sections.

6.2.1.1 The introduction

Overall, the introduction primarily served to gain the interest of the audience and to inform them of the purpose of the talk. This means that it is necessary to make a persuasive and appealing start, which was certainly challenging. To achieve this, the flow of this introduction was designed as follows.

The session started with a greeting to teachers to express my appreciation for their attendance. I briefly introduced myself as a researcher, informing the audience that this educational talk was a part of my doctoral research (slide 1 in Figure 6.10) and the content of the talk was a result of rigorous plans involving research about the conceptual and empirical reviews of the topic of interest. This starter was intended to establish the credibility of the speaker who would talk about the topic professionally.

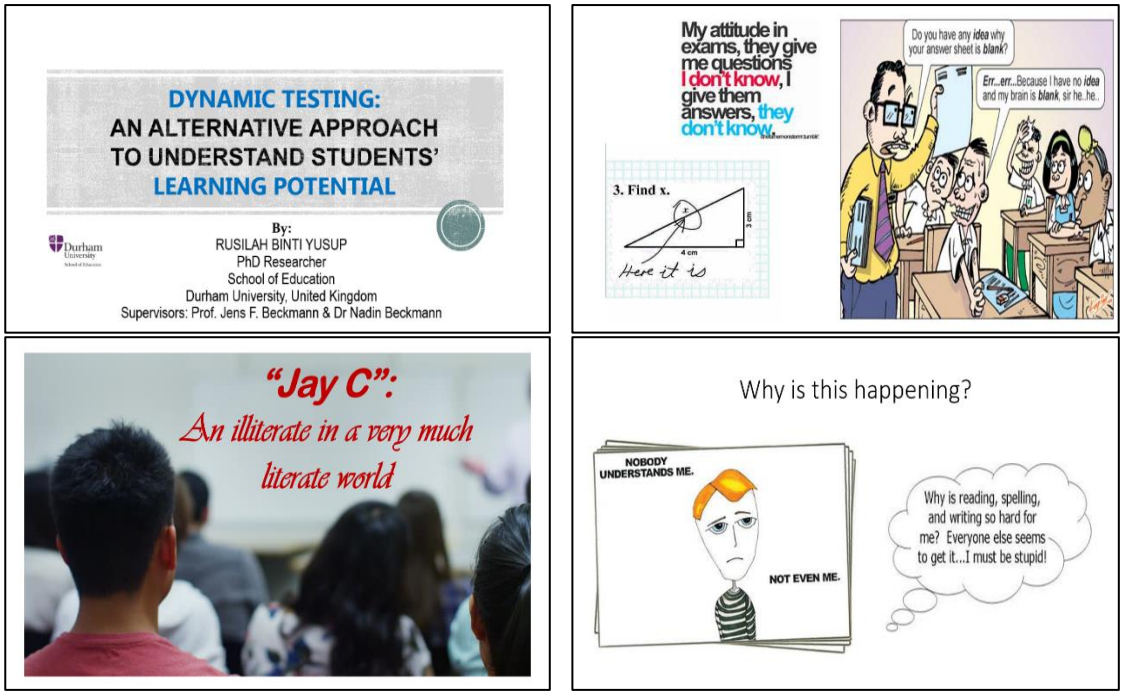


Figure 6.10: A Starter of the Introduction Stage

I also shared a little bit of information about my teaching career (e.g., years of teaching, subjects taught and former schools). This was a way for me to build a rapport and connection with the audience and to make them feel at ease. Building this rapport helped me to gain their trust, to see that I am an "insider" who understood the education system and who could relate to their current situation. Here, it was also emphasised that the intention of conducting this talk was neither to teach nor to judge what teachers were doing to be wrong, but to share new ideas that might be put into a good practice for the betterment of the national education system.

Another way to bridge the connection with the audience was sharing my personal teaching experiences relating some examples of the real-life situation in schools (slides 2, 3 and 4). In slide 3, for example, I recounted an anecdote about a "special" student, aged 16, who could barely read and write. Unsurprisingly, teachers admitted they also had encountered similar experiences and they claimed it was a challenging task to deal with, as they were not trained for special education. Apparently, using this personal experience helped me to catch the audience's attention, giving them a reason to listen to this important topic.

The following slides in Figure 6.11 showed how I prepared the audience for the talk.

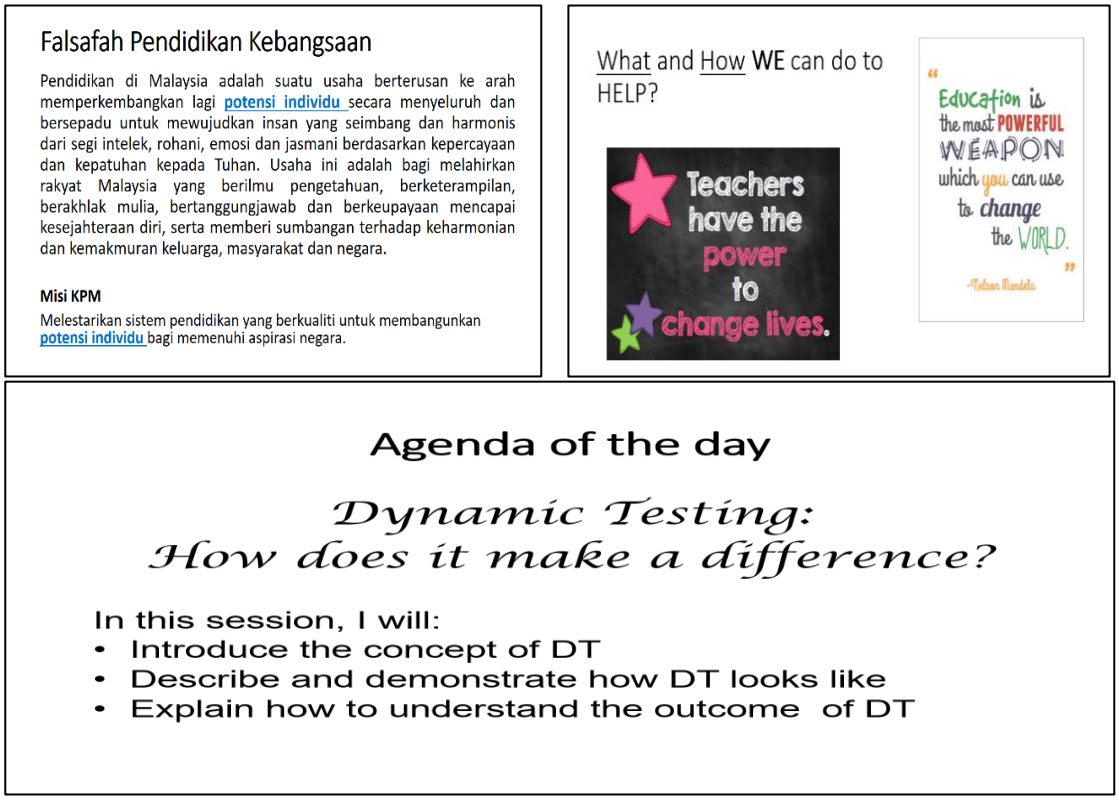


Figure 6.11: Preview of the Key Points

To bring the audience together to reflect upon this issue, I highlighted the aim of education in Malaysia is to develop individuals to their potential as aspired to in the NEP and by the MOE (slide 1). Recalling the previously mentioned problems that teachers encounter in school, I posed "thought-provoking" questions to the teachers: Is what we are doing aligned with the NEP? Are we doing things right in developing the potentials of our students? In slide 2, I urged the audience to reflect on the roles of teachers (the what and how questions) to uphold the aspiration of the NEP.

Before moving to the next stage, the purpose of the talk was revealed in slide 3 by posing another crucial question: Dynamic testing: how does it make a difference? Not only did this question help to develop the structure of the body of the talk, but it also aimed to provide an engaging flow for the audience throughout the workshop – purposefully setting them up to deduce the answers from the subsequent slides.

6.2.1.2 The body

This stage is the core element of this educational component, as it is of the whole research. It plays a crucial role in helping the audience to understand the rationales for the introduction of this alternative assessment.

To support the arguments, the main content was sequenced into three parts: (i) the current practice (UPSR and F1DT); (ii) an overview of DT; and (iii) a demonstration of the application of DT in a real-life setting. A detailed description of each part is presented in the following paragraphs.

6.2.1.2.1 The current practice

This section was designed to posit the fundamental argument for an alternative assessment tool as a solution for the previously mentioned problems. It started with a review of two existing tools that are commonly used for several purposes – measuring current achievement, boarding school entrance and class placement. In slide 1 (see Figure 6.12), a comparison of UPSR and F1DT aimed to make teachers realise the redundancy of the two tests. The focus of F1DT on measuring the existing curriculum-based knowledge of the learners is no different from the constructs assessed in UPSR. Relating to this, the significance of incremental validity was highlighted, suggesting that schools should invest time and resources for an alternative assessment approach that could offer more meaningful information beyond that provided by UPSR.

In slide 2, extracts from the study respondents' written comments about the implementation of F1DT in schools were shown to the audience (taken from the preliminary findings of the pre-intervention). Admittedly, some of the audiences were in agreement that F1DT was still relevant, while others were critical of having to use two similar tests.

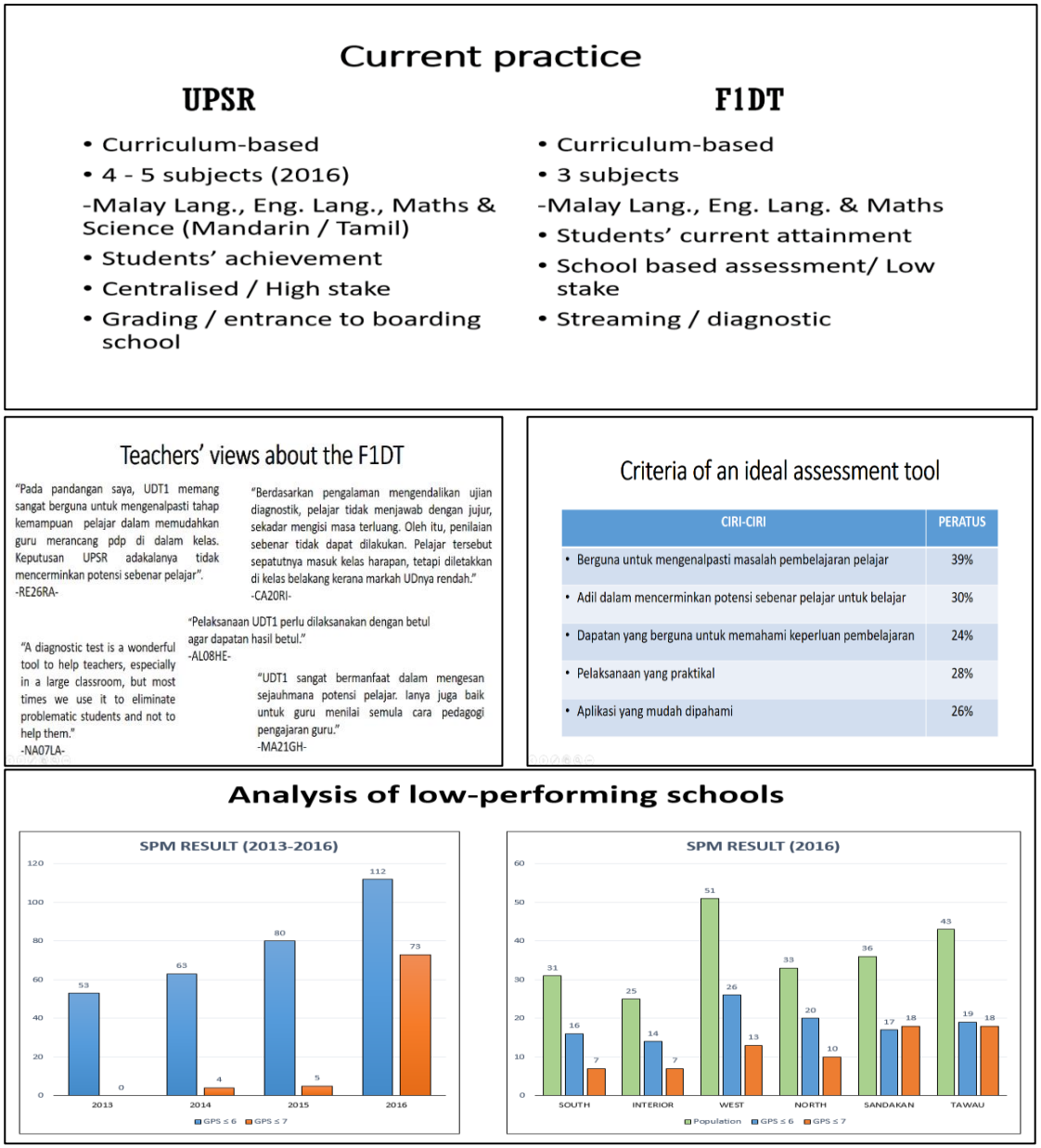


Figure 6.12: Challenges of Current Practices in Schools

It was also important to convince teachers about the necessity for a better assessment tool. Thus, the reason for this had to be framed in terms of what mattered to them – an action that served their interests and was closely linked to their day-to-day teaching activities.

Acknowledging the MOE's overarching agenda of pursuing a world-class education and improvement of students' performance in international large-scale assessments, as well as national public examinations (as explicitly mentioned in the MEB), was used to trigger teachers' interest to listen and to discuss the need for a new assessment approach. In slide 4, I showed the analysis of low-performing schools in Sabah. It was evident that the

number of schools with GPS ≥ 6 had increased for the last four years (2013–2016). The increase in the number of low-performing schools was deeply concerning, signalling something is amiss in the education system.

Reflecting on the views of teachers about the criteria of an ideal assessment tool in slide 2 (preliminary analysis of the pre-intervention), teachers were urged to reassess whether F1DT fulfilled these criteria, particularly the first two indicators: "useful for identification of students' learning problems" and "accurate indicator of students' actual learning potentials". It is argued that if it did so, then why did the numbers of low-performing schools keep increasing year by year? This question is framed to advocate the underutilisation of assessment-related information for effective instruction. It seemed that FIDT is less helpful in providing such information. That is why an alternative test is needed so that teachers can have the chance to use the information to enhance teaching and learning It was also reasoned that if teachers want to see a considerable improvement in students' learning, an assessment should not only describe students' achievements but also provide information in understanding their learning potentials. To reaffirm this point, I persuaded teachers that it would be great if there was an assessment tool that could offer information for both summative and formative purposes. This is what I intended to introduce in the next section.

6.2.1.2.2 An overview of DT

In this section, I presented the proposal for a new method, promoting DT as a practical tool that could potentially remedy the shortcomings of F1DT.

The following slides (Figure 6.13) contained the key points of the conceptual foundation of DT (as outlined in section 5.1.1). I discussed the basic conceptual explanation of why traditional tests "fail" to represent learners' "true potentials". It was emphasised that students' poor performance in a test is not necessarily an indication of limited cognitive ability. Here, I explained the key theoretical backgrounds – the nature and development of human ability (slide 1) and Vygotsky's ZPD (slide 2) – that underpin the emergence of DT.

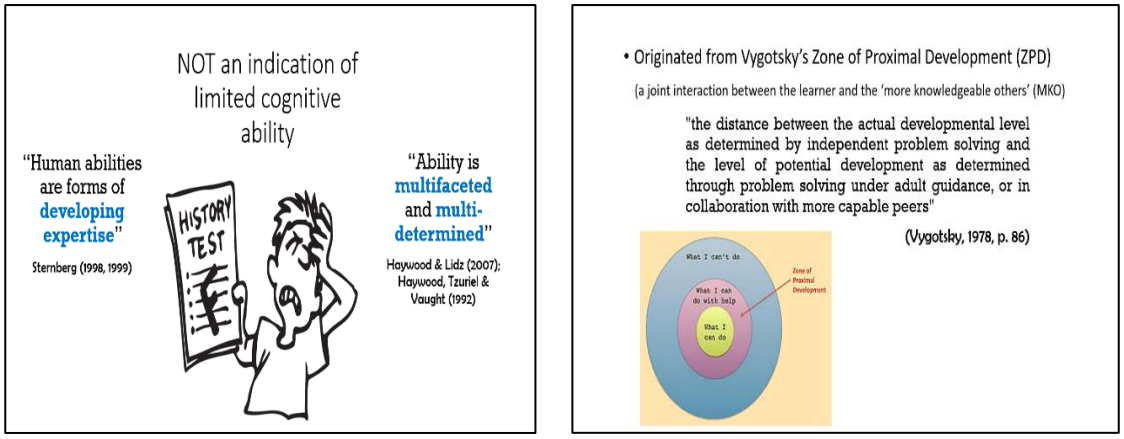


Figure 6.13: Theoretical Foundations of the Emergence of DT

I tried to make this discussion not too theoretical, but at the same time, it is essential for teachers to develop a good understanding about how these theoretical frameworks are linked to the construct they want to measure. This may assist them to understand the development of human abilities and, in turn, this can help them to plan what can be done to scaffold students' learning to achieve what they have the potential to do. This conceptual explanation was also intended to show the shortcomings of the traditional test in terms of uncovering the true potential of students.

In the following slides (Figure 6.14), the focus was to explain how DT is different from the traditional test. Slide 1 presents the definition of DT and the use of a diagram helps to show how DT is different from the traditional test. Here, I put more emphasis on detailing the central features of DT – its dynamism, the integration of feedback in the assessment procedure and the measurement of learning potentials. I also linked this operational application of DT to the conceptual background mentioned in the previous slides (Figure 6.13). This was to show that the operationalisation of DT is theoretically consistent with the nature of human ability and the notion of ZPD.

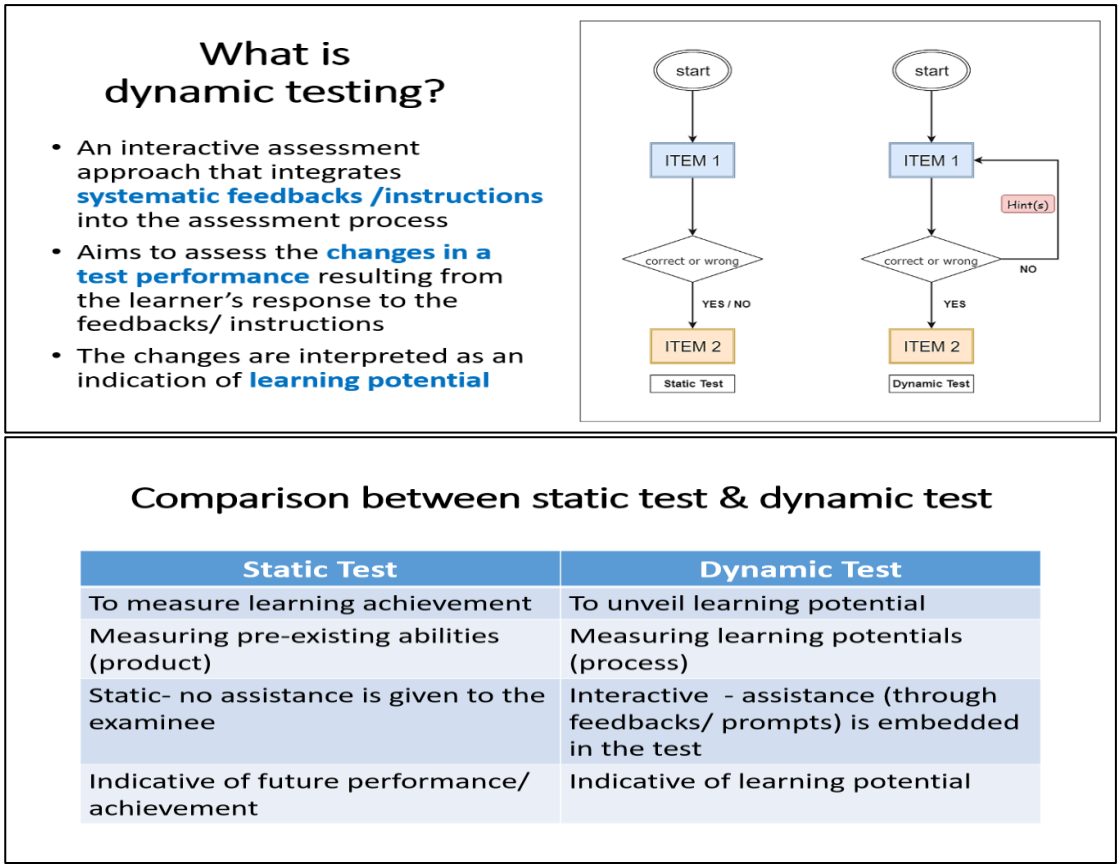


Figure 6.14: The Difference between Traditional Test and Dynamic Test

To further demonstrate the difference between the two assessment approaches, a comparison between a static test and a dynamic test was presented in slide 2. This was to reaffirm the assertion that DT is unique and potentially useful in offering meaningful information that is untapped by the traditional tests like F1DT.

6.2.1.2.3 Demonstration of application of DT in school

To avoid the risk of putting the audience off from listening to a very theoretical explanation of DT, this part was designed to demonstrate the real application of DT in school. The objective of this demonstration was to provide a hands-on experience for teachers, allowing them to see what DT looks like and how it is administered in a real-life setting. Also, it aimed to establish an impression that DT is practical and feasible for implementation in Malaysian schools.

This demonstration (see Figure 6.15) resulted from the administration of LT as described in section 6.1. As a start, a detailed description of the test items (slide 1) was presented.

Slide 2 is an example of how I explained the complexity level of the items. I also showed several screenshots of the test webpage to the audience (slides 3 and 4).

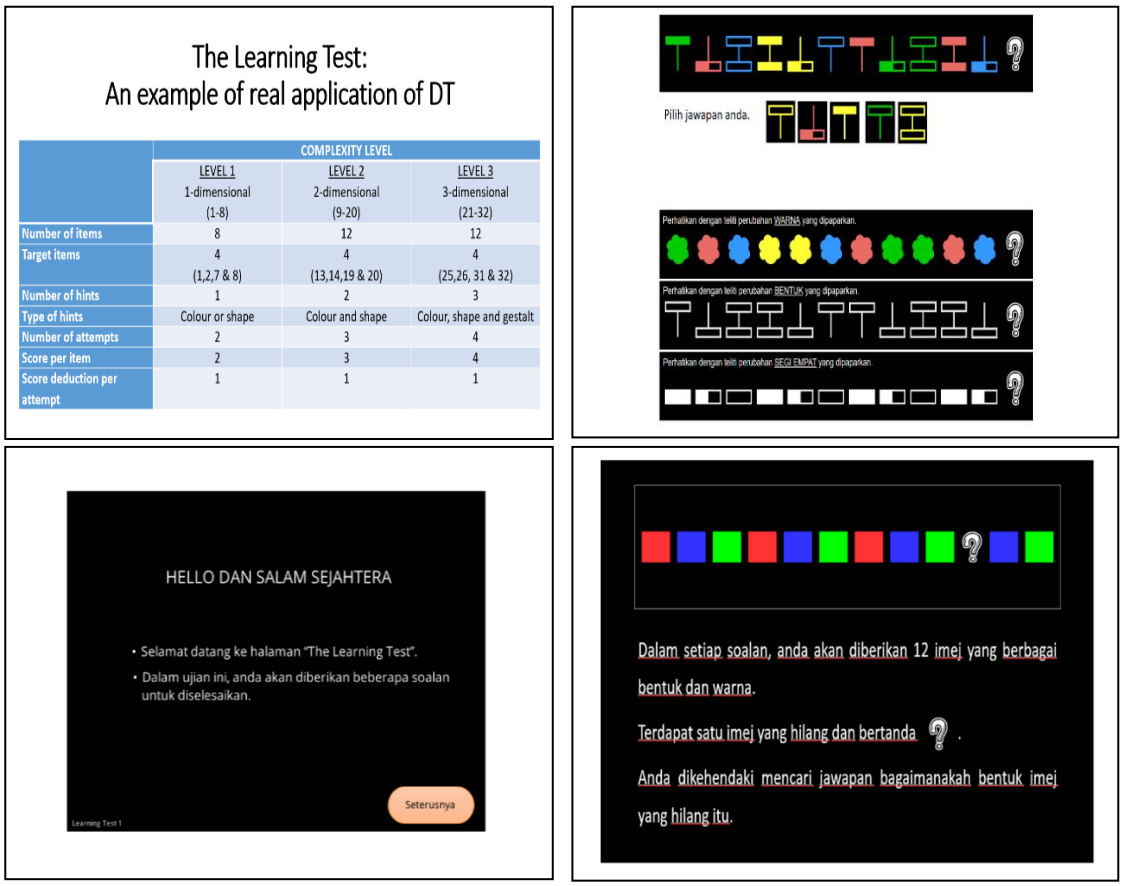
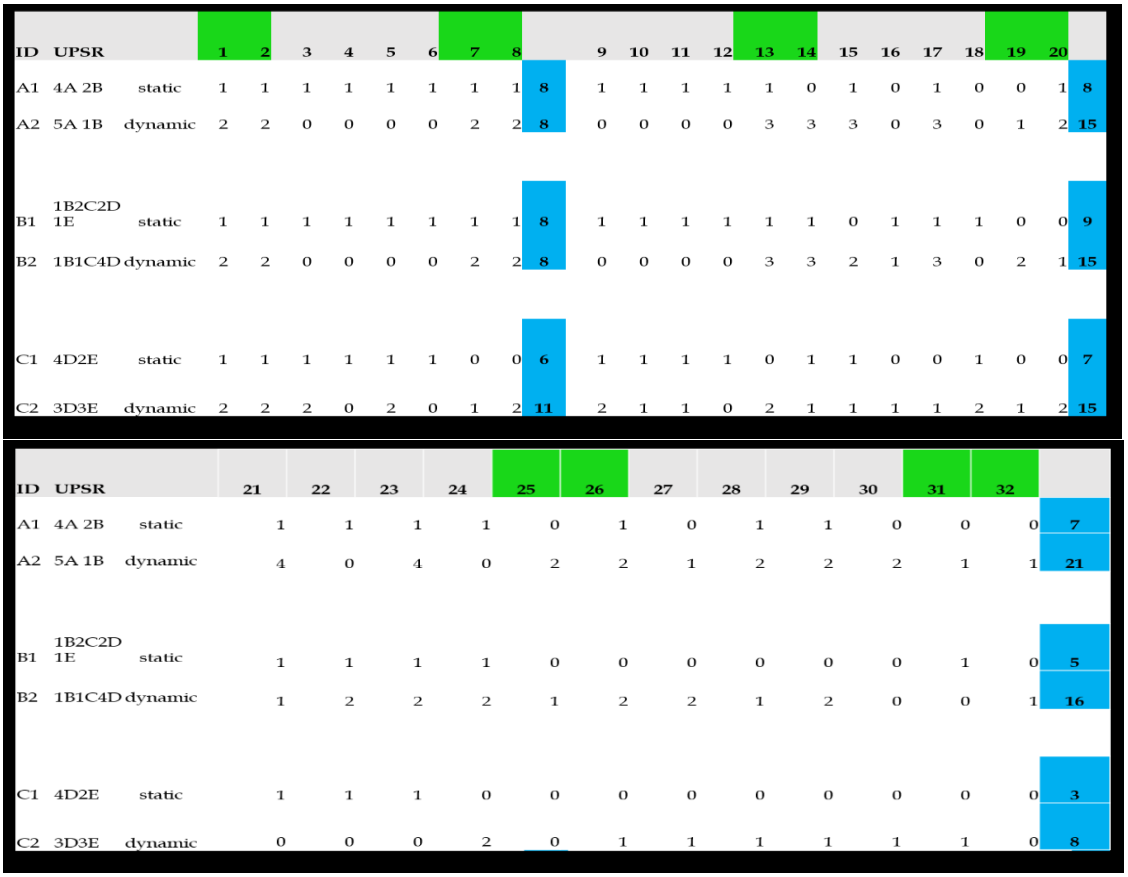


Figure 6.15: Item Description

Here, it was important that the audience was not overburdened with too much technical explanation. A complex structure might intimidate teachers and, consequently, they might lose interest in continuing to listen to the talk. Attempting to make the presentation straightforward as well as engaging, I designed a dummy version of the LT consisting of several items from each complexity level. By asking teachers to provide answers to the items, they were given the opportunity to see how the feedback and instructions work in the testing procedures.

In the following slides (Figure 6.16), the explanation of how to understand the output of DT is presented, aiming to demonstrate the promising contributions of DT in delivering more meaningful information regarding a child's potential for learning. Slides 1 and 2 are the test scores of a few students participating in the administration of the LT, representing

the static and dynamic versions as well as the ability groups. Note that the scores of the static version are indicated as 1 for a correct answer and 0 for an incorrect answer, while the scores of the dynamic version are characterised by the complexity level and the number of assistances given to the test-takers.



ID	UPSR		1	2	3	4	5	6	7	8		9	10	11	12	13	14	15	16	17	18	19	20	
A1	4A 2B	static	1	1	1	1	1	1	1	1	8	1	1	1	1	1	0	1	0	1	0	0	1	8
A2	5A 1B	dynamic	2	2	0	0	0	0	2	2	8	0	0	0	0	3	3	3	0	3	0	1	2	15
B1	1B2C2D 1E	static	1	1	1	1	1	1	1	1	8	1	1	1	1	1	1	0	1	1	1	0	0	9
B2	1B1C4D	dynamic	2	2	0	0	0	0	2	2	8	0	0	0	0	3	3	2	1	3	0	2	1	15
C1	4D2E	static	1	1	1	1	1	1	0	0	6	1	1	1	1	0	1	1	0	0	1	0	0	7
C2	3D3E	dynamic	2	2	2	0	2	0	1	2	11	2	1	1	0	2	1	1	1	1	2	1	2	15

ID	UPSR		21	22	23	24	25	26	27	28	29	30	31	32	
A1	4A 2B	static	1	1	1	1	0	1	0	1	1	0	0	0	7
A2	5A 1B	dynamic	4	0	4	0	2	2	1	2	2	2	1	1	21
B1	1B2C2D 1E	static	1	1	1	1	0	0	0	0	0	0	1	0	5
B2	1B1C4D	dynamic	1	2	2	2	1	2	2	1	2	0	0	1	16
C1	4D2E	static	1	1	1	0	0	0	0	0	0	0	0	0	3
C2	3D3E	dynamic	0	0	0	2	0	1	1	1	1	1	1	0	8

Figure 6.16: Explanation of the Test Score

The explanation of the test scores was organised into two parts. Firstly, I wanted to show the audience the difference in students' performance between the static and dynamic versions. As seen from the students' test scores, it is evident that students had a higher probability of answering the questions successfully when assistance was provided. Regardless of the complexity level, students using the dynamic version were likely to get more correct answers than students using the static version. In slide 2, for example, look at the difference of correct items between C1 (the static version) and C2 (the dynamic version). For items in level III, C2 was able to get seven correct items while C1 only answered the first three items successfully and got the remaining items inaccurate. More importantly, look at how C2 performed in level I. He answered the first pair of obligatory

items (Items 1 and 2) correctly (reflected in the maximum mark (level I= 2 marks) for both items). He then moved to the next pair of obligatory items, i.e., Items 7 and 8, but got Item 7 incorrectly at the first attempt, which led to a detour to Item 3. He got this item correct and was prompted to Item 5 without answering Item 4. Here, he required less assistance in the later part of the test process (when presented with more complex items than Item 1 and Item 2). This is an indication of learning as he has learnt how to solve the items. Although he did not manage to maintain this performance in level II and level III, teachers were given the opportunity to see what he can do if assistance is given to him to reach his potential to learn.

Secondly, I wanted to explain how DT is useful in differentiating the learning potential of students across ability group. The focus here is on the students with the dynamic version – A2, B2 and C2. It is interesting to note that C2 performed as well as A2 and B2 in level I and level II (slide 1). Although level III is the most complex, it is seen that A2 and B2 both answered ten items successfully. Impressively, C2 was not far behind them, getting seven items correct. Look at the last three items (items 30, 31 and 32). A2 received almost as many prompts as C2 to reach the correct answer. This indicates that the weakest ability group (C2) had the opportunity to reach his potential (as good as A2) at this level with the help of the feedback given. Thus, it is important to note here that getting items right is not the focus, DT is about how much support (i.e., feedback and prompts) one needs throughout the test. The less help needed over time indicates improvement in learning.

Accuracy of the above data is not a great concern (accuracy is not guaranteed as the programming was technically unstable, particularly in the calculation of the test scores). The focal point here is about the uniqueness of DT in comparison with the traditional tests. DT has much to offer to teachers and students. It offers– a better understanding of what is limiting the learners' performance and what can be helpful to facilitate higher performances. As empirically demonstrated, it is such information that makes DT more useful and reliable in describing students' ability to learn. Its central focus in measuring learning potential is crucial for inference of what students can do with the help of others, i.e., teachers and peers, in their learning development. This information is certainly valuable for teachers to better understand the potentials of their students, especially the low performers.

6.2.1.3 The conclusion

As crucial as the introduction was to grab the audience's attention, creating a memorable ending was as equally important to leave a positive and lasting impression on the audience. Although it may be short, it is often the part that the audience remembers the most.

This conclusion was an opportunity to reinforce the arguments that was built throughout the talk (see Figure 6.17). I needed to ensure that I remained consistent and steadfast about the need for an alternative assessment to support effective teaching and learning in Malaysian schools, particularly in Sabah. In recapitulating the key points of the talk, I took the audience back to the key message (DT: how does it make a difference?) which was established at the beginning of the presentation. To reflect upon this question, major potential contributions of DT in educational settings were highlighted in slide 1. It was also asserted that DT, as a supplement to traditional testing, might offer a more accurate estimate of students' potential that could benefit planning for appropriate instruction. In turn, this might lead to the likelihood of better student performances in public examinations and internationally recognised assessments.

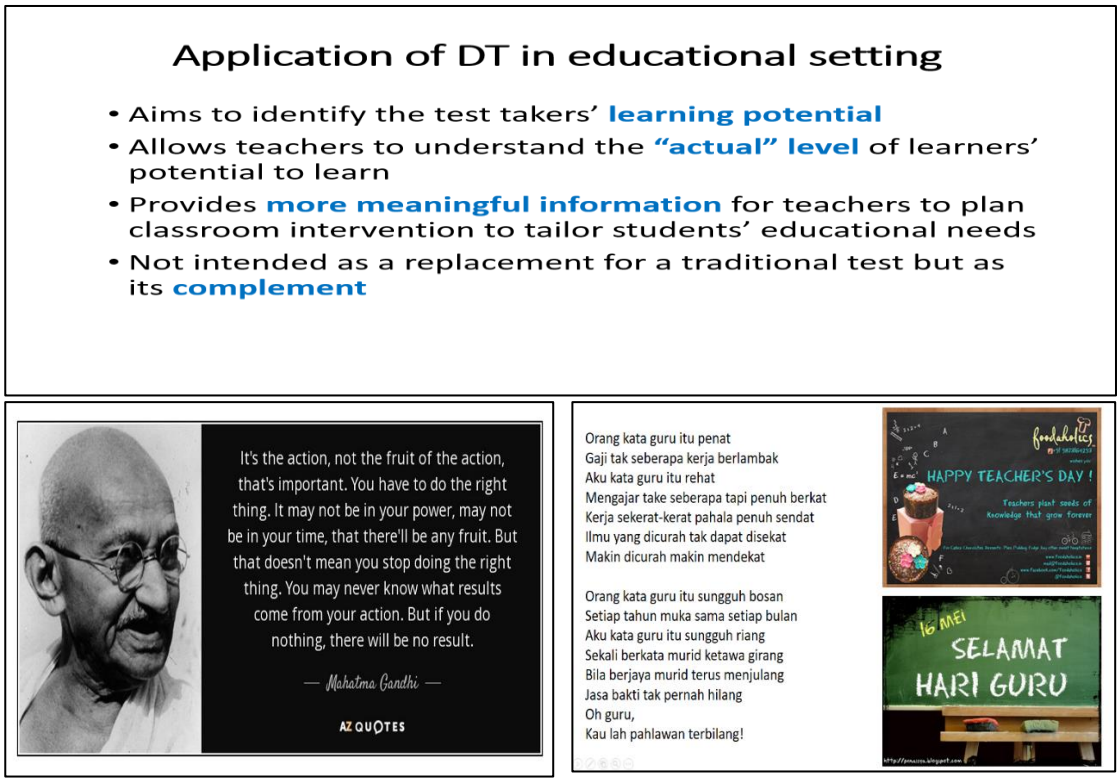


Figure 6.17: The Content of the Conclusion

Additionally, I wanted to insert a “take-away” message that might linger in the minds of the audience after the session was over. In doing this, I finished the presentation with a call for action (slide 2) and a reason to act (slide 3). In slide 2, I shared a quote by Mahatma Gandhi about the importance to do the right thing, implying that a useful tool is necessary to ensure that assessments may provide a positive impact in teaching and learning. To end the session, a motivational poem was dedicated to the audience (slide 3), championing teachers’ significant role in nurturing students’ potential to do their best and reach their goals in life.

6.2.2 Execution of the educational component

This CPD was conducted for 166 teachers across five schools. As mentioned earlier, the assignment of teachers into the intervention group was sought through voluntary participation, in agreement with the school administrators to fit in this talk as one of the school activities for the staffs.

Figure 6.18: Venues and Participants of the CPD

All participating schools were very accommodating. They arranged the date, time and venue for the talk. On average, the talk was delivered in one and a half to two hours, including the question-and-answer session after the presentation. It was conducted between 2 p.m. to 4 p.m., as this was the only time available for teachers after teaching hours ended. For this workshop, the participating schools allowed me to use the meeting room that was fully equipped with a projector and slide screen (see Figure 6.18).

In experimental design research, it is recommended to pilot the "treatment" before initiating the real intervention (Malboeuf-Hurtubise et al., 2016; Miller, Schoen, James & Schaaf, 2007). This allows the researcher to identify and to resolve limitations that may affect the implementation of the main study. In this study, however, I did not have the opportunity to test the feasibility of the educational component due to time constraints. The research permit only granted the whole process of data collection – pre-intervention, intervention and post-intervention – to be conducted within four months (February to May 2017).

Acknowledging the constraints, I strictly followed the flow of the content (as described above) in all schools to ensure the contents were delivered in a similar way for all participants. It was feared that any discrepancy in delivery might threaten the validity of this study.

6.3 Chapter Summary

This study aims to promote DT as an alternative assessment approach that can contribute greatly to better utilisation of assessment-related information in enhancing teaching and learning. The design of an intervention strategy is instrumental to fit the purpose of introducing this new method to Malaysian teachers. In this chapter, a description of the intervention strategy has covered its two components – the experiential and educational components. The first component is the simulation of a computer-adaptive DT, i.e., the LT, aiming to demonstrate the application of DT in a school setting. The focus of the second part of the chapter is the explanation of the educational component, conducted as a part of professional training for teachers in participating schools. The content of the training was designed based on the exploration of literature review about DT and the outcome of the real application of DT in the experiential component. Essentially, this

educational talk was crucial in eliciting information about (a) teachers' responses to their reported beliefs and practices about F1DT; (b) the alignment of teachers' assessment beliefs and practices; and (c) teachers' feedback about the potential implementation of DT in Malaysia. In the next chapter, the focus will be on the research design and methods used in the main study to answer the research questions relating to the issues of investigation.

7 Research Design and Methods

The main objective of this study, as outlined in the introductory chapter, was to explore changes in teachers' assessment beliefs and practices due to introducing an alternative to the existing diagnostic test (i.e., F1DT) which has been in use for many years in Malaysian schools. Specifically, there were eight research questions guiding this study:

1. What are teachers' beliefs about the purposes and uses of F1DT?
2. What are teachers' assessment practices regarding the use of F1DT?
3. To what extent do teachers' beliefs about the purposes and uses of F1DT align with their assessment practices?
4. To what extent are individual characteristics and school variables associated with teachers' assessment beliefs and practices?
5. Does the introduction of DT change teachers' reported beliefs about the purposes and uses of F1DT?
6. Does the introduction of DT change teachers' reported assessment practices regarding the use of F1DT?
7. To what extent does the introduction of DT affect the alignment between teachers' reported beliefs about the purposes and uses of F1DT and their reported assessment practices?
8. What are teachers' opinions with regard to the potentials and barriers to implementing DT as an alternative tool in Malaysian schools?

To address these research questions, this study collected data using a piloted questionnaire that was distributed to secondary teachers in Malaysia. The details of the research design, research site, sampling protocol, participants, instrument, data collection procedures and data analysis are explained in the following sections.

7.1 Research Design

This exploratory study used a survey to gather information about teachers' assessment beliefs and practices about one of the currently used assessment tools for Malaysian students. Surveys are considered a powerful and practical tool for gathering data from a group of people to describe some aspects (e.g., opinions, beliefs, attitudes, knowledge) of the population of which the group is a part of (Allred & Ross-Davis, 2010; Bartlett,

Kotrlik & Higgins, 2001; Cohen, Manion & Morrison, 2011; Fraenkel & Wallen, 2008). A ~~piloted~~ questionnaire, i.e., SEA, was utilised as a mode of data collection to answer the research questions. This data collection method was utilised over other techniques (i.e., observation and interview) due to its practical advantages; cost-effective, quick, high response rate and convenient for respondents (Brown, 1987; Bryman, 2012; Fraenkel & Wallen, 2008; Cohen et al., 2011).

The study attempted to establish a possible cause-effect relationship among variables of interests (i.e., the effect of introducing a new assessment concept on teachers' reported assessment beliefs and practices). To examine that effect, a quasi-experimental design was utilised. Fraenkel and Wallen (2008) claimed that experimental research is the most appropriate way to examine cause-effect relationships among variables; it engenders more robust and valid results about causal findings than other research designs.

A core feature of an experimental research is the implementation of an intervention or treatment to one sub-group, but not the other (Bryman, 2012; Fraenkel & Wallen, 2008; Cohen et al., 2011; Creswell, 2014). In the present study, an introduction of a new assessment approach, i.e., DT, was intended to see whether it could influence teachers' reported assessment beliefs and practices regarding the usefulness and use of the existing diagnostic measure, i.e., F1DT.

The study employed a two-group-pre-and-post design consisting of three phases of empirical investigation: pre-intervention, intervention and post-intervention. A detailed description of the stages is outlined below.

- **Pre-intervention phase**

The primary objective of the pre-intervention was to investigate the status quo of Malaysian teachers' views about the implementation of F1DT in schools. Specifically, this phase was an explorative approach focusing on (i) investigating teachers' assessment beliefs and practices about the implementation of F1DT as a diagnostic measure for Form 1 students; (ii) examining the relationship between teachers' reported assessment beliefs and practices; and (iii) studying the influence of a number of variables that might affect

teachers' assessment beliefs and practices. To seek answers for Research Questions 1–4, a self-developed questionnaire, i.e., SEA, was distributed to 21 participating schools.

The participating schools were not yet assigned to the intervention and control groups at this point. Instead, they were treated as a whole sample of participants to obtain the information to answer Research Questions 1 – 4. However, prior consent was obtained from the school principals to arrange for a suitable date and time for the educational workshop (i.e., the intervention phase).

- **Intervention phase**

After the first phase, five schools voluntarily agreed to participate in the study. Following this, teachers from these schools were exposed to the intervention (as described in Chapter 6). This was carried out to introduce DT as an alternative assessment and to demonstrate its application in an educational setting. For this purpose, the design of the intervention consisted of two components as illustrated in Figure 7.1.

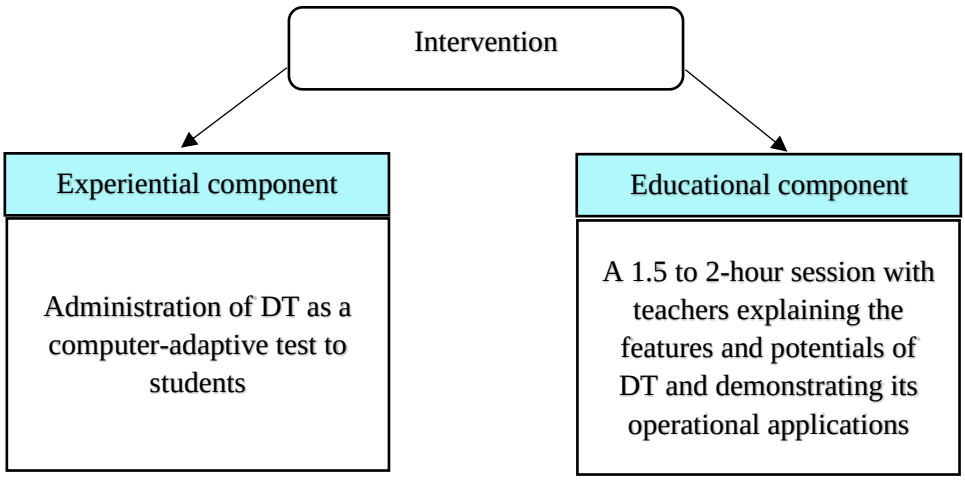


Figure 7.1: Components of the Intervention

- Experiential component

A set of DT from an existing instrument was redesigned and administered to Form 1 students in a secondary school in Sipitang district (i.e., SMK Padang Berampah) as a computer-adaptive test. This experiential component (as reported in Section 6.1) was

intended to give teachers a hands-on experience of what DT looks like and how it is feasible in a real educational setting.

- Educational component

This component was demonstrated to the five schools that agreed to carry it out as one of the CPD activities of in their respective schools. During this session, I had about one hour and 30 minutes to two hours of educational interaction with the teachers. This educational aspect was a conceptual description of DT; explaining its features, potentials and differences from the traditional test. The session also demonstrated the real-world application of DT in schools, particularly its promising potentials to provide a deeper understanding of students' actual ability to learn.

It was expected that, by the end of this intervention phase, teachers would be able to see the comparison between information provided by the current tool (i.e., FIDT) and the new tool (i.e., DT). This insight may lead them to agree more or less to the questionnaire statements on the purposes and uses of F1DT.

- **Post-intervention phase**

In the final phase, the same questionnaire used in the pre-intervention was distributed to all participating schools. The objective was to examine the extent to which teachers who were exposed to the intervention could change their responses to SEA in comparison to teachers who did not experience the same exposure. The answers to the second set of research questions (Research Questions 5–8) were deduced by comparing the responses from the intervention group and the control group to the questionnaire before and after the introduction of DT.

Figure 7.2 below depicts the descriptions of the experimental design adopted in this study.

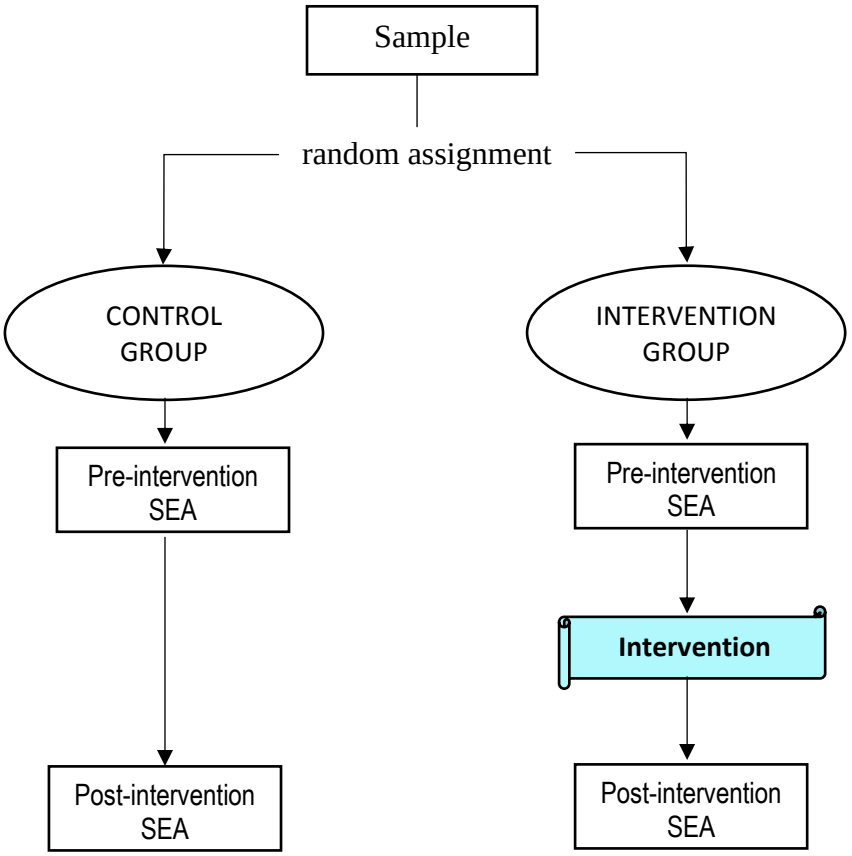


Figure 7.2: A two-group-pre-and-post-test Design

7.2 Research Site

The present study was conducted in Sabah, the second largest state in Malaysia (see a map of Malaysia in Figure 7.3). This state was chosen as the research site primarily because of two reasons: geographical and logistic factors and academic performance. First, Malaysia is geographically divided into West Malaysia in Peninsular Malaysia (11 states and three federal territories) and East Malaysia in Borneo (two states, i.e., Sabah and Sarawak, and a federal territory).

Figure 7.3: Location of Sabah in Malaysia

Sabah consists of 24 districts and are currently clustered into six educational zones (see Figure 7.4). These zones – North, West, South, Interior, Tawau and Sandakan – are primarily characterised by geographical locations and ethnic compositions. Clearly, there is variability within and across zones. It is expected that diversity across zones is greater than the differences within particular zones.

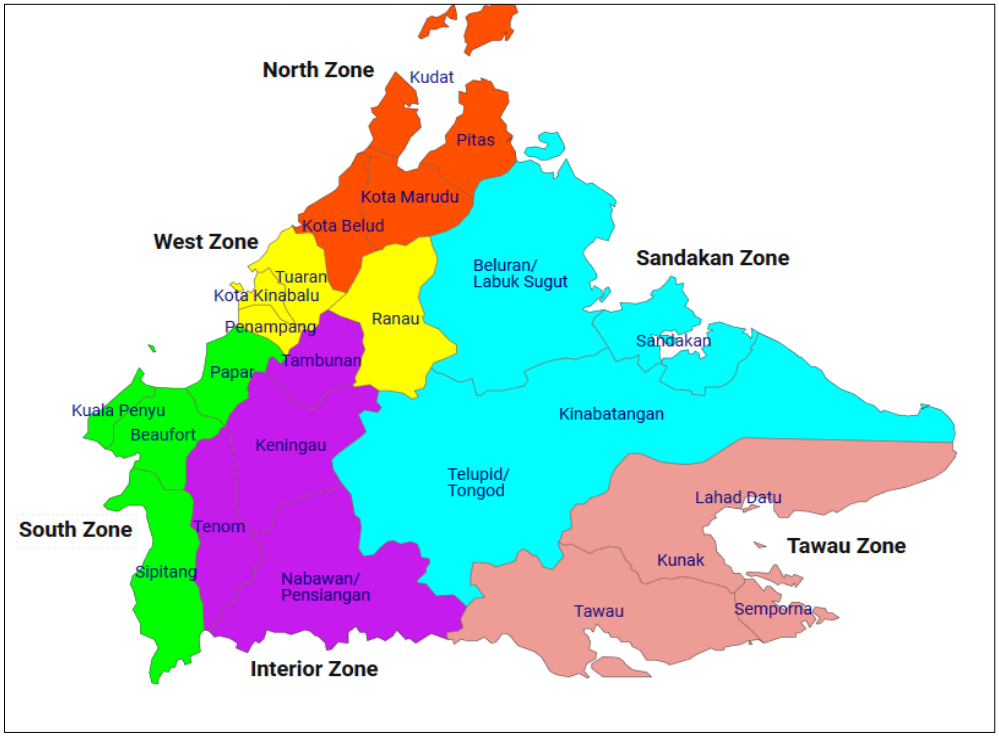


Figure 7.4: Six Educational Zones in Sabah

Due to constraints on time and financial resources, the recruitment of the sample only focused on teachers from Sabah (the state that I am originally from).

Another important reason for choosing Sabah was because of its academic performance in national high-stakes examinations. Sabah has been ranked in bottom place for overall performance in most of the public examinations for many years (Ministry of Education Malaysia, 2015, 2016, 2017). As shown in the following bar graph (Figure 7.5), there was a significant increase of low-performing schools (schools with GPA ≥6 in SPM) in Sabah.

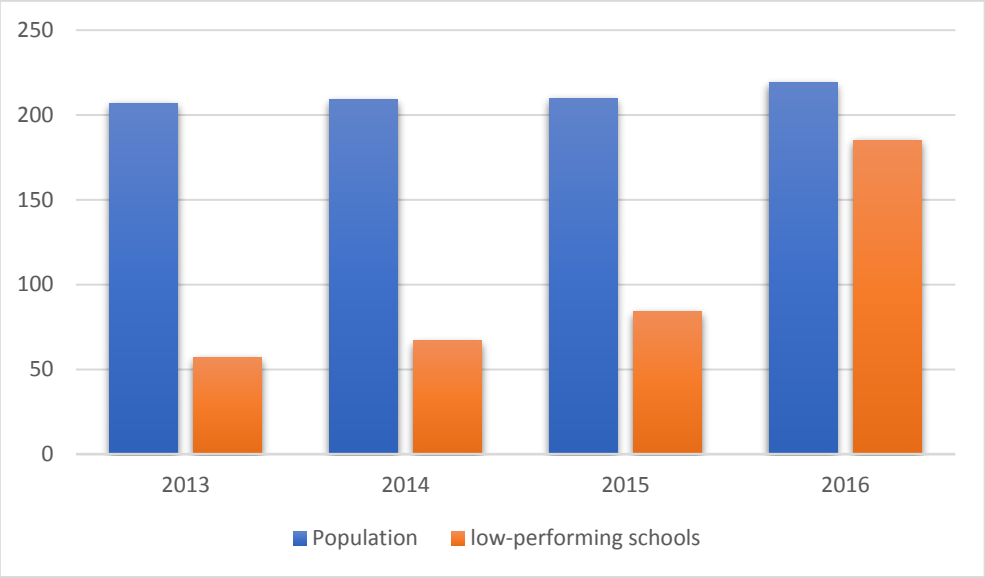


Figure 7.5: Number of Low-performing Schools in Sabah (SPM 2013–2016)
(Source: Sabah Education Department)

The above situation is a clear indication that something is amiss in the system. The big question is, if the assessment is claimed to improve teaching and learning, then why does the number of low-performing schools keep increasing year by year? Obviously, this deserves immediate attention and a systematic investigation. It is crucial and pertinent to understand what teachers in Sabah think about assessment and how they utilise assessment-related information in instructional decision-making. Such information is crucial for all the relevant educational agencies, particularly the Sabah Education Department, to find ways to improve the quality of education as envisioned in the MEB.

7.3 Methods

7.3.1 Sampling protocol

Because the population of Sabah is large and widely dispersed, a careful plan for the sampling technique was therefore crucial so that it would not jeopardise the aim of representativeness of the sample. Lack of a representative sample limits the degree of generalisability and the transferability of the findings to a wider population or setting (Bryman, 2012; Cohen et al., 2011).

Clearly, stratified sampling was more appropriate in this study as it is a way to ensure that subgroups are equally or proportionately represented within the sample, which can guard against unrepresentativeness (Bryman, 2012; Cohen et al., 2011; Marshall, 1996). Also, it can lead to a high-quality sample, as it reduces the risk of "faulty" findings in term of strata characteristics (Gorard, 2013). Stratified sampling is a probability sampling technique that allows researchers to divide the population into homogenous groups called strata (Cohen et al., 2011). A decision to adopt this sampling technique was largely influenced by the geographical factors of the potential schools, which are located in rural and urban areas in six different educational zones. To achieve a conclusive interpretation of the research findings, there is a necessity to capture the variability within and across zones. This can increase the likelihood of representativeness of the population and therefore provides an opportunity for generalisation.

The starting point in selecting a stratified sample was to determine the characteristic(s) of the strata and to randomly select samples from the subgroups. In this study, the utilisation of two-group experimental design made it deemed necessary to minimise the likelihood of threats to the internal validity of the treatment effect (Cohen et al., 2011; Fraenkel & Wallen, 2008). Therefore, when choosing prospective study participants, researchers should pay more attention to subject characteristics so as not to trigger other unintended variables that may "destroy" the estimation of any observable effects (Fraenkel & Wallen, 2008; Stuart & Rubin, 2008).

To mitigate potential threats to validity, a rigorous matching method was deployed. This method has become increasingly popular for causal inference studies in many disciplines, e.g., statistics (Diamond & Sekhon, 2006; Rosenbaum & Rubin, 1985; Sekhon, 2008),

political science (Sekhon, 2009) and economics (Dehejia & Wahba, 2002). Participant matching involves pairing the two comparison groups – the experimental and control groups – with similar observable characteristics (Cohen et al., 2011; Creswell, 2014; Rutterford, Copas, & Eldridge, 2015). It aims to minimise selection bias (Dehejia & Wahba, 2002; Sekhon, 2009), to control pre-existing differences between the groups (Sekhon, 2009; Stuart, 2010; Stuart & Rubin, 2008) and to reduce the imbalance in characteristics across groups (Rutterford et al., 2015).

While this method is expanding, there is, nonetheless, no clear-cut way in how the exact matching should be conducted (Heckman, Ichimura, Smith & Todd, 1998; King, Nielsen, Coberley, Pope & Wells, 2011; Stuart & Rubin, 2008). In this study, the potential schools were recruited from the low-performing schools with the school GPA ≥ 6 (n=85). The phenomenon of the consistent increase of low-performing secondary schools in Sabah justified the choice of the target population. Two predetermined criteria – school location (urban or rural) and educational zones (six zones) – were identified as elements of comparability between the intervention and control groups (see Figure 7.6) to ensure they were as equivalent as possible.

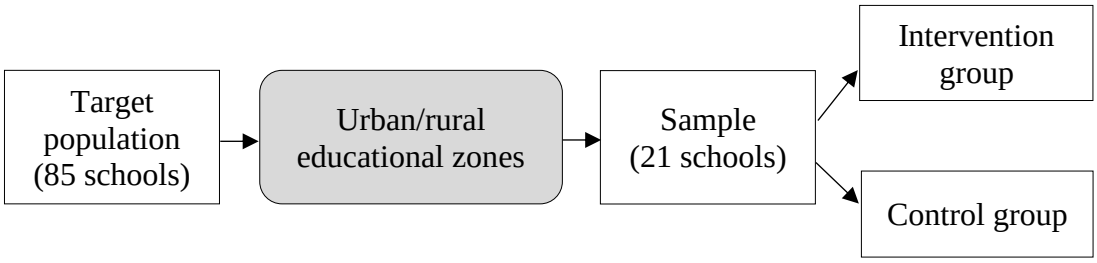


Figure 7.6: Matching and Screening Procedure for Sample Recruitment

After a random selection of 85 schools with GPA ≥ 6, 21 schools agreed to participate in this study (see Appendix 7A describing the characteristics of the schools). Schools were clustered within six educational zones as shown in Figure 7.7.

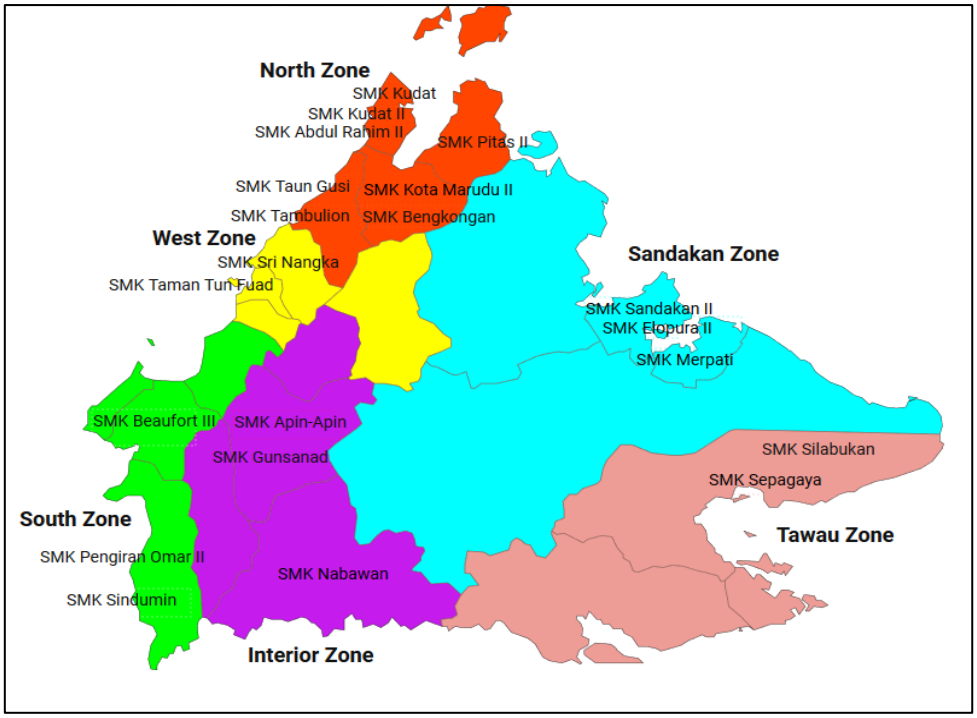


Figure 7.7: Locations of the Participating Schools

Recruiting a sample of appropriate size is instrumental to the success of a research study (e.g., Bryman, 2012; Cohen et al., 2011; Creswell, 2014; Fraenkel & Wallen, 2008). The calculation of sample size was guided primarily by the research design of this study, i.e., the two-group pre-test and post-test experimental design.

As of November 2016, the total population of secondary school teachers in Sabah was 15,724 (219 schools).[6] The target population of this study was a sub-population of secondary schools with a school GPA of ≥ 6, consisting of 6,125 teachers from 85 schools and accounting for 39% of the whole teacher population in Sabah (based on SPM results in 2015). Taking into account that DT has demonstrated promising potentials for academically-risk students, it is the interest of this study to examine the effect of introducing this new assessment approach to teachers from low-performing schools on their reported assessment beliefs and practices about the existing assessment tool, i.e., F1DT. Hence, the calculation of the sample size was determined by the expected change of questionnaire responses before and after the intervention. The effect size signifies the magnitude of effects or the strengths of relationships of the phenomenon under study (Cohen et al., 2011; Olejnik & Algina, 2000; Richardson, 2011; Stout & Ruble, 1995)

[6] Figures from Sabah Education Department at http://jpnsabah.moe.gov.my/.

This is to say, an estimate of effect size provides meaningful insights to justify the research findings, whether the observed differences or relationships between the two groups are strong enough to have a significant impact.

Thus, to determine the appropriate sample size for this two-group experimental study, I used Cohen's Table 2 (1992) calculation of sample size when an estimate of effect size is known. It is recommended that a medium effect size is desirable because it represents an effect that would likely to be "visible to the naked eyes of a careful observer" (Cohen, 1992, p.156). For an expected minimal effect size of .50 (medium effect) between two independent samples, with a significance level of .05 and a power of .80, the sample required for each group is 64 teachers or 128 teachers in total. Furthermore, caution must be exercised about potential problems such as non-response, attrition, incomplete returned questionnaires and withdrawal from the study. It is therefore advisable to overestimate the size of the required sample (Cohen et al., 2011; Gorard, 2013). It seemed that the total numbers of the intervention group (n=166) and the control group (n=215) were satisfactory to eliminate the risk of such anticipated problems.

Although the recruitment of the sample selection was screened and matched based on the characteristics of the schools, it is important to highlight that the primary unit of analysis in this study was the teachers, not the schools.

7.3.2 Participants

The questionnaire was distributed to a sample of 21 schools with an estimated enrolment of 1,521 teachers. At the end of the pre-intervention, there were 891 returned questionnaires. A total of 29 responses was removed from the analysis due to large amounts of missing information, which resulted in a 56.6% response rate (n=862) with an average number of 41 teachers per school. There were 284 male and 578 female teachers participating in this study, reflecting 33% and 76% of the whole teacher population respectively. The number of 146 urban teachers (38%) and 253 rural teachers (62%) also appeared to be similar to the whole population of secondary teachers in Sabah. Of the participants, 28% indicated that they had less than 5 years of teaching experience; 32% had 6–10 years, 14.6% had 11–15 years, 13.2% had 16–20 years and 12.2% had more than 20 years' experience. In terms of educational background, the majority of the

participants were holders of bachelor's degrees (89.4%), while teachers with master's degrees, diplomas, the STPM and SPM accounted for 9%, 0.8%, 0.6% and 0.1% respectively. A total of 86.8% of the respondents were ordinary teachers, while 13.2% were teachers with administrative duties such as principal, senior assistant and head of a department. As this study is concerned about the implementation of F1DT, all participants were asked to provide information about their familiarity with the test. A total of 546 teachers (63.3%) responded that they had experience designing, implementing and using F1DT, while 316 teachers (36.7%) answered they had no direct experience with it.

For the intervention, the intervention group was selected based on the availability of the school after negotiating with the already pre-scheduled activities of each school. Initially, it was envisioned that this study could recruit an equal number of schools with a ratio of 10:10 for each intervention group and control group. Prior to the intervention, nine schools agreed to take part in the CPD workshop to introduce the idea of DT. Four of them, however, withdrew their participation due to the administration of mid-term examinations and unadjusted pre-scheduled school activities. A similar situation occurred with the control groups, resulting in only eight schools continuing for the second phase. As a result, the response rate dropped to 59% with only 509 teachers completing the questionnaire. However, there were only 381 responses that could be used for comparison before and after the intervention. A summary of the participants' demographic information is presented in Appendix 7B.

Figure 7.8 explains the total number of samples participating in the pre-intervention and post-intervention.

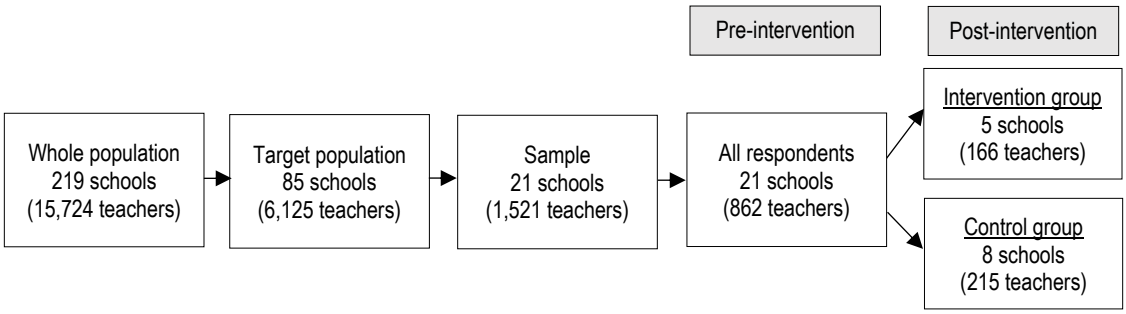


Figure 7.8: Total Number of Participants in the Study

7.3.3 Research instrument

As detailed in Chapter 4 (see section 4.4.1.3), the final version of SEA had gone through exhaustive revisions in the pilot studies, particularly PS2. As a result, SEA contained 36 items, initially covering two main dimensions – assessment beliefs and assessment practices regarding the purposes and uses of F1DT in Malaysian schools. Upon completion of data collection, an analysis of the internal structure (dimensionality) and internal consistency of this questionnaire with a new dataset (n=862) was conducted.

PCA was performed to examine the internal structure of this questionnaire. The output of KMO with a value of .940 and a significant Bartlett's Test (p = .000) verified that the questionnaire items were reasonably factorable. Deploying the Promax rotation, all 36 items were computed with the predetermined numbers of components (three components) to see whether the respective items come together on the components that they were expected to load into. The results showed a stable solution of three components. It was evident that 14 items of what-I-do (*PRACTICE*) were solidly clustered into Component 1, 16 items of what-I-think (*PERCEPTION*) into Component 2 and six negative items (*NEGATIVITY*) into Component 3 (see a table of component loadings in Appendix 7C).

The internal reliability of each component was further assessed using Cronbach's alpha. The components, *PERCEPTION*, *PRACTICE* and *NEGATIVITY*, were found to have a very high reliability (α = .923, α = .924 and α = .817 respectively).

Obviously, PCA results and reliability statistics confirmed that SEA was psychometrically good as a research instrument in this study.

Apart from the main research instrument, this study also designed a computer-adaptive test called LT (as detailed in Chapter 6) for the intervention. The test was adapted from ADAFI, a diagnostic programme used for students and adolescents in Germany (Guthke, 1992; Guthke et al., 1995; Guthke et al., 1997; Guthke & Beckmann, 2000a, 2000b; Guthke & Stein, 1996). In this study, LT utilised a series of figural sequences to measure the reasoning ability of the test-takers. This test was administered to Form 1 students as a demonstration of how DT could be practical in a real school setting. Though some outputs of the tests were used in the experiential component, the data collected in this test were not part of the main analysis to answer the research questions.

7.3.4 Procedures of data collection and ethical considerations

Prior to the data collection processes, I followed the same procedures as in the pilot studies to obtain the ethics clearance from the university. In addition to the university's obligatory regulations on undertaking a research study, an elaborate research application was also submitted to several gatekeepers before the main data collection took place. The first step to was to obtain a research permit from the Economic Planning Unit (EPU), which is a policy division of the Prime Minister's Office. This unit then issued a research pass following approval from the referral agencies concerned, i.e., Ministry of Education Malaysia and Sabah Education Department. Upon receiving the approval from the relevant agencies, initial contacts were established with the District Education Offices and the principals of the identified schools to inform them about the research and to obtain consent for participation in this study. Figure 7.9 below outlines the entire process of this ethics clearance.

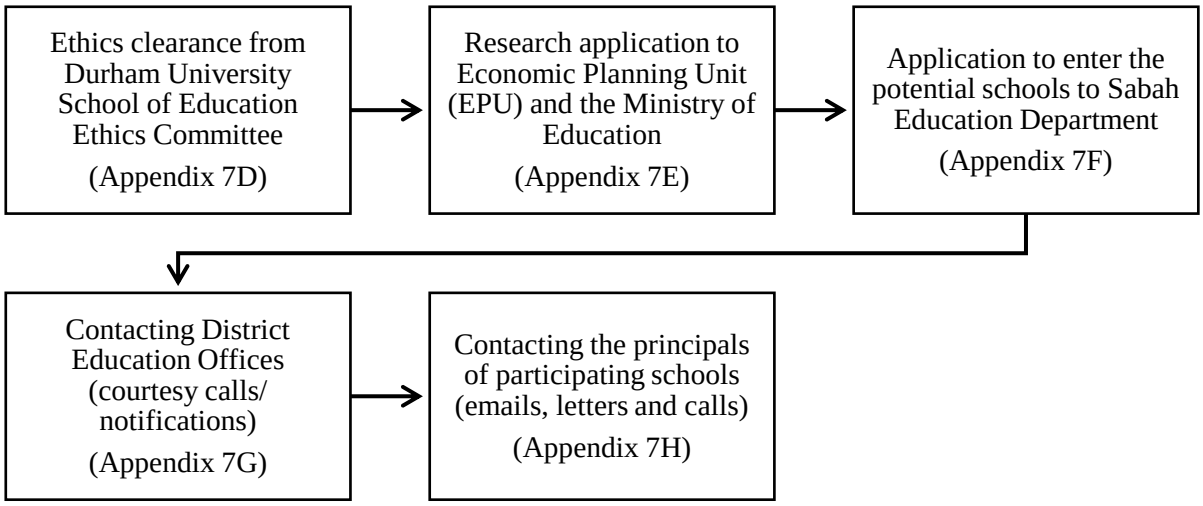


Figure 7.9: Flowchart of the Process for Ethics and Research Approval

The questionnaire was administered in paper-and-pencil form. An online survey was not practical as many schools in rural areas have poor internet access. Again, a drop-off survey was adopted here. The questionnaire was distributed to the selected schools and teachers were required to return the completed questionnaire in a box or an envelope. The appointed representative for each school then returned the box or envelope to me. A higher response rate is identified as a major advantage of the drop-off method (Allred & Ross-Davis, 2010; Bowling, 2005; Brown, 1987; Nulty, 2008) as it provides flexibility

for the respondents to complete the questionnaire at their own pace and at a time of their convenience.

The protocol to ensure anonymity and confidentiality was maintained as in the pilot studies; participants were pseudonymously identified. The respondents, however, were required to write down a unique anonymous combination of their identify – the first two letters of mother's name, day of birth, the first two letters of father's name and M/F for gender (e.g., JA28YUF). This code was used to trace which group they belonged to for the comparison of responses before and after the intervention. Teachers were also provided with an informed consent form (see Appendix 4D) which explicitly informed them that their identities would be kept anonymous and the information obtained from the questionnaire would only be used for research purposes.

7.3.5 Data analysis

This study collected data from 21 randomly selected schools in six educational zones in Sabah, making it multilevel. Its multilevel structure begins with teachers who are nested within schools, schools within districts, districts within educational zones. When teachers work within a school, it is assumed that they tend to be more homogeneous with each other than teachers randomly sampled from different schools. In other words, there is less variance within schools but greater variance between schools and across districts and educational zones. The reason for this is because teachers from the same schools have certain characteristics in common, observations of these teachers are not entirely independent. When the data are dependent, this violates the independence of observations assumption, a pre-requisite for the conventional regression model to analyse variance (Raudenbush & Bryk, 2002).

Because of the above reasons, the main statistical technique deployed in this study was hierarchical linear modelling (HLM), also termed as multilevel modelling (MLM). HLM has become a highly useful and powerful statistical technique for analysing complex hierarchical relationships (Guo, 2005; Hofmann, 1997; Osborne, 2000; Raudenbush, 1993). Scholars (e.g., Hofmann, 1997; Luke, 2004; Osborne, 2000; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; Warne et al., 2012; Woltman, Feldstain, MacKay & Rocchi, 2012) argued that conventional statistical tools, particularly ANOVA and

regression, are inappropriate when analysing multilevel data because they ignore the importance of group effects and dependence. Ignoring the nested structure of data may increase the likelihood of Type 1 error (Hoffman & Rovine, 2007; Luke, 2004; Raudenbush, 1988; Snijders & Bosker, 2012; Warne et al., 2012; Woltman et al., 2012), in which the researchers erroneously conclude the results of the study to be false, when in reality they are true. Also, Luke (2004) asserted that it is deemed necessary to analyse hierarchical data using HLM to avoid fallacies of inference of group-level information where inferences about groups are incorrectly drawn from the information obtained at the individual level. Likewise, he further cautioned that making inferences about a higher-level member using information from lower-level members is certainly inaccurate.

Further, HLM also permits researchers to examine individual- and group-level variables and interactions across levels simultaneously (Guo, 2005; Luke, 2004; Raudenbush, 1988; Warne et al., 2012; West, Ryu, Kwok & Cham, 2011; Woltman et al., 2012). In the context of this study, HLM is well suited for use to examine the influences of teachers' characteristics (e.g., educational background, years of teaching experience) and school characteristics (e.g., location) on teachers' assessment beliefs and practices concerning the purposes and uses of F1DT (see descriptions of variables of interests in Appendix 7I). It is also the interest of the study to assess the presence of cross-level interactions in affecting teachers' responses to the questionnaire. More importantly, the utilisation of a two-group pre-test-and-post-test experimental design involved data collection at different times and under different conditions for each study participant. Not only was the intervention strategy aimed at introducing DT and its potentials in educational settings, but also at observing the effect of this exposure on teachers' reported assessment beliefs and practices of F1DT. Researchers argued that when a study intends to investigate changes within subjects, data analysis is best performed using HLM that accounts for such hierarchy (e.g., Bryk & Raudenbush, 1987; Hoffman & Rovine, 2007; Osborne, 2000; Raudenbush, 1993; Raudenbush & Bryk, 2002).

An illustration of the multilevel structure of the variables of interest is depicted in Figure 7.10.

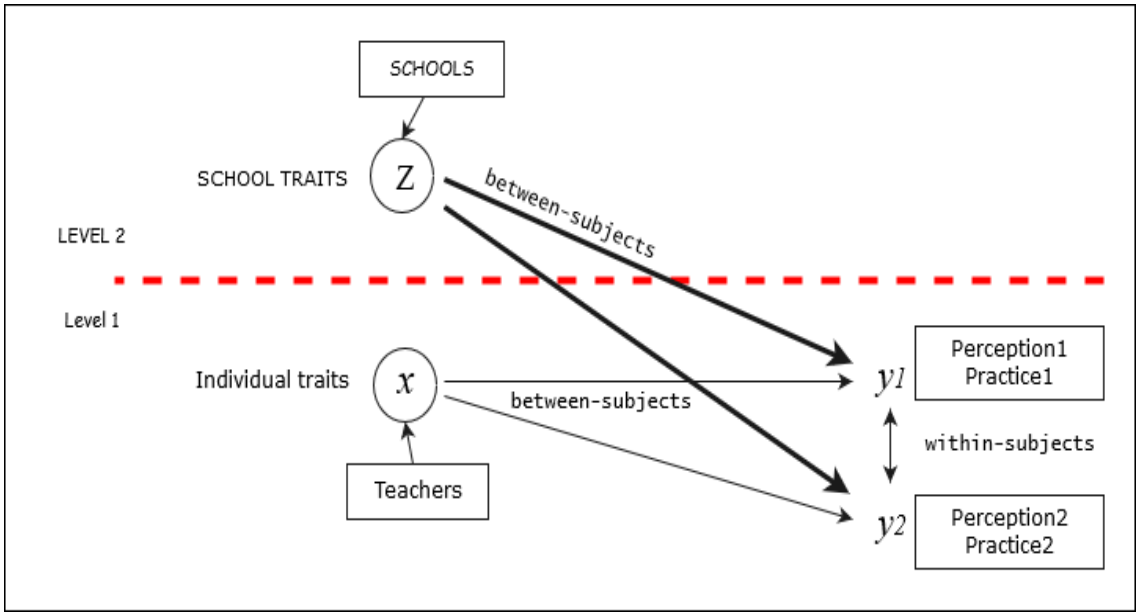


Figure 7.10: Multilevel Relationships of the Variables of Interest
(Adapted from Snijders & Bosker (2012))

As appropriate to the research questions (Research Questions 1 – 8), this study primarily focused on the basic two-level model. The sample included a total of 862 teachers (interchangeably referred to as individual level or level 1) nested within 21 schools (interchangeably referred to as group level or level 2). The largest level 1 sample size was 78 and the smallest was 19, with an average of 41.05 per school.

7.3.5.1 Pre-intervention data analysis

The pre-intervention analysis aimed to seek answers to the first set of research questions (Research Questions 1–4). It is worth reiterating that the analyses at this stage were not a hypothesis-testing approach to examine the causal relationship between the predictors of interest and the outcome variables. Rather, they were an explorative strategy to study (a) the status quo of Malaysian teachers' assessment beliefs and practices of the implementation of F1DT; (b) the relationship between assessment beliefs and practices; and (c) the influences of several variables on teachers' assessment beliefs and practices. Answers to the above issues were elicited from two different sources – the survey and the written responses. Primarily, a series of HLM analyses were conducted to analyse teachers' responses to the questionnaire. A detailed explanation of the specified analyses is presented in Chapter 8.

The data collected from teachers' written responses were analysed through thematic analysis. The main aim of this analysis was to arrange what was written by teachers into manageable data that are relevant in addressing the research question. This approach involves the process of identifying, analysing and reporting patterns or themes within data (Braun & Clarke, 2006; Maguire & Delahunt, 2017). Bryman (2012) defined a theme as an identified category of the data that is related to the research focus, specifically to the research questions at hand. This theme identification technique requires the researcher not only to summarise the data but also to interpret and make sense of them meaningfully and systematically (Maguire & Delahunt, 2017).

Data analysis began by translating all written comments into English; those very few that were written in English were kept as original as possible. To ensure the accuracy of the translation, I appointed several reviewers, including teachers and academicians, who were bilingual to verify the translated transcripts. After familiarising with the transcripts, coding was used to organise and reduce the data (containing 60 written comments) into some small chunks of information. A code is "the ascription of a category label to a piece of data, decided in advance or in response to the data that have been collected" (Cohen et al., 2011, p.668). Essentially, the method of coding adopted in this study was *a posteriori* coding – a procedure in which the researcher develops the coding categories after the collection of the data (Braun & Clarke, 2006; Cohen et al., 2011). At this stage, several initial codes from teachers' responses were generated by using a descriptive word or phrases at the left margin of each piece of data.

Later, the coded data were collated for analytic analyses to look for potential themes. As reflected in the coding procedure, the identification of emerging themes in the written comments was more data-driven and inductive, in which the themes were linked to the data themselves instead of a particular underlying theoretical framework (Braun & Clarke, 2006; Cohen et al., 2011). I organised and interpreted the themes without systematic pre-ordinated conceptions of the data, treating the responses of all teachers as they were and then relating them to the issues under investigation.

7.3.5.2 Post-intervention data analysis

The main focus of the post-intervention analyses was on how the intervention affects (a) teachers' assessment beliefs and practices about the currently used diagnostic test (F1DT)

and (b) the alignment between assessment beliefs and practices concerning F1DT. Additionally, the analysis also aimed at examining teachers' reflective feedback about the newly introduced assessment tool (DT). Specifically, the analyses aimed to answer another set of four research questions (Research Questions 5–8) in which each was analysed individually using a series of complementary statistical techniques.

Prior to the main analysis, several preliminary analyses were conducted. The first analysis aimed at examining potential differences that might exist between various groups in this study. To check for potential group differences, I conducted two analyses of differences between the groups concerned: (a) the main sample (n=862), i.e., participants who carried on and those who dropped out and (b) the comparison groups (n=381), i.e., the intervention group and the control group. Essentially, the analyses aimed at mitigating potential threats to the external and internal validity of the findings, so that the conclusion of the study could be fairly generalised. The second analysis involved the post-intervention data (n=381) to check the suitability of HLM application in this two-level dataset.

A detailed description of the specified analyses is provided in Chapter 9.

7.4 Chapter Summary

This chapter has outlined an overview of the research design and detailed explanations of the methods of this study. A two-group pre-and-post-test experimental design was employed to address the research questions. A self-developed and piloted questionnaire was distributed to 21 schools from six different educational zones at two times, before and after part of the sample was presented with information about DT. Most importantly, the intervention, consisting of the experiential and educational components, was designed to introduce DT to Malaysian teachers and to demonstrate its potential in better understanding students' learning potentials. Additionally, this chapter has described the procedures for the recruitment of the sample, demographic information of the participants, the research instruments and the protocols of data collection and ethical considerations. The last section has detailed the use of several complementary statistical techniques, particularly HLM, to answer the research questions. The following chapter will report the results obtained from the data analyses.

8 Results: Understanding Teachers' Assessment Beliefs and Practices about Form 1 Diagnostic Test

It is argued that the ability to understand assessment is a prerequisite for optimal use of assessment-related information to support effective teaching and learning (Mertler, 2009; Popham, 2009; Stiggins, 2002; Stiggins & Duke, 2008; Volante & Fazio, 2007). In relation to this, scholars (e.g., Barnes et al., 2015; Brown, 2004; Brown et al., 2009; Remesal, 2011; Sach, 2012) advocated that the ability to understand assessment is often associated with teachers' belief system. In the literature, it is highlighted that the beliefs teachers hold about their professional work interact with their practices, i.e., teachers' actions are determined by what they believe (Barnes et al., 2015; Borg, 2001; Jussim & Harber, 2005; Nespor, 1987; Opre, 2015; Pajares, 1992; Skott, 2015).

Acknowledging the influential impact of teachers' beliefs on their educational practices, this study aims to explore in-service teachers' beliefs and practices about an existing assessment tool, i.e., F1DT, that has been used in Malaysian schools for years. This chapter presents the findings of the study, which were primarily obtained from the survey data analysed from the responses of 862 teachers asking their opinions on their assessment beliefs and practices about F1DT. In the following sections, the reported findings focus on the investigation of three correlated issues: (a) teachers' assessment beliefs and practices about the implementation of F1DT; (b) the relationship between assessment beliefs and practices; and (c) variables influencing teachers' assessment beliefs and practices. Specifically, this chapter is organised to answer the first four research questions as stated below:

1) What are teachers' beliefs about the purposes and uses of F1DT?
2) What are teachers' assessment practices regarding the use of F1DT?
3) To what extent do teachers' beliefs about the purposes and uses of F1DT align with their assessment practices?
4) To what extent are individual characteristics and school variables associated with teachers' assessment beliefs and practices?

8.1 Teachers' Assessment Beliefs and Practices about the Form 1 Diagnostic Test

Research Questions 1 and 2 intended to study what teachers think and do about the information from F1DT. The answers were obtained from the survey, i.e., SEA, and teachers' elaborated comments written on the last page of the questionnaire. Both the survey and the written responses were analysed separately and reported accordingly in the following paragraphs.

8.1.1 Survey findings

Descriptive analyses – means (*M*) and standard deviations (*SD*) – were calculated to examine teachers' responses to the questionnaire items. Also, HLM analyses were performed to further explore the variances of teachers' responses to their perceptions and practices concerning F1DT.

8.1.1.1 Descriptive analyses

As a starting point, teachers' responses to the survey were analysed using descriptive statistics. A useful property of the mean (*M*) and standard deviation (*SD*) was utilised to explain the tendency (positive or negative) of the responses and to measure the amount of variability of the participants' responses, respectively. Scores on the questionnaire were firstly averaged according to each outcome variable –*PERCEPTION*, *PRACTICE* and *NEGATIVITY*. The level of agreement ranges from 0 (totally disagree) to 5 (totally agree).

Table 8.1 presents the descriptive analyses of teachers' responses to the three variables. Generally, teachers in this study tended to have a positive attitude to the implementation of F1DT. This was reflected in teachers' high agreement to the purposes and uses of F1DT (M = 3.29, SD = .67) and their use of its information in instructional decision-making (M = 3.49, SD = .69). Responding to *NEGATIVITY* items, teachers were likely to disagree that F1DT is irrelevant and has little use (M = 2.34, SD = .92).

Table 8.1

Descriptive Statistics of Teachers' Responses to SEA (n=862)

Components	*M*	*SD*
PERCEPTION (16 items)	3.29	.67
PRACTICE (14 items)	3.49	.69
NEGATIVITY (6 items)	2.34	.92

It is noted that each variable has a small *SD* of less than 1, particularly *PERCEPTION* and *PRACTICE*. This explains a small variability of the scores from the mean, suggesting a collective agreement among teachers that F1DT is an appropriate assessment tool that allows them to gather relevant information about Form 1 students. In turn, this information seems essential for them to plan the next steps in the teaching and learning processes.

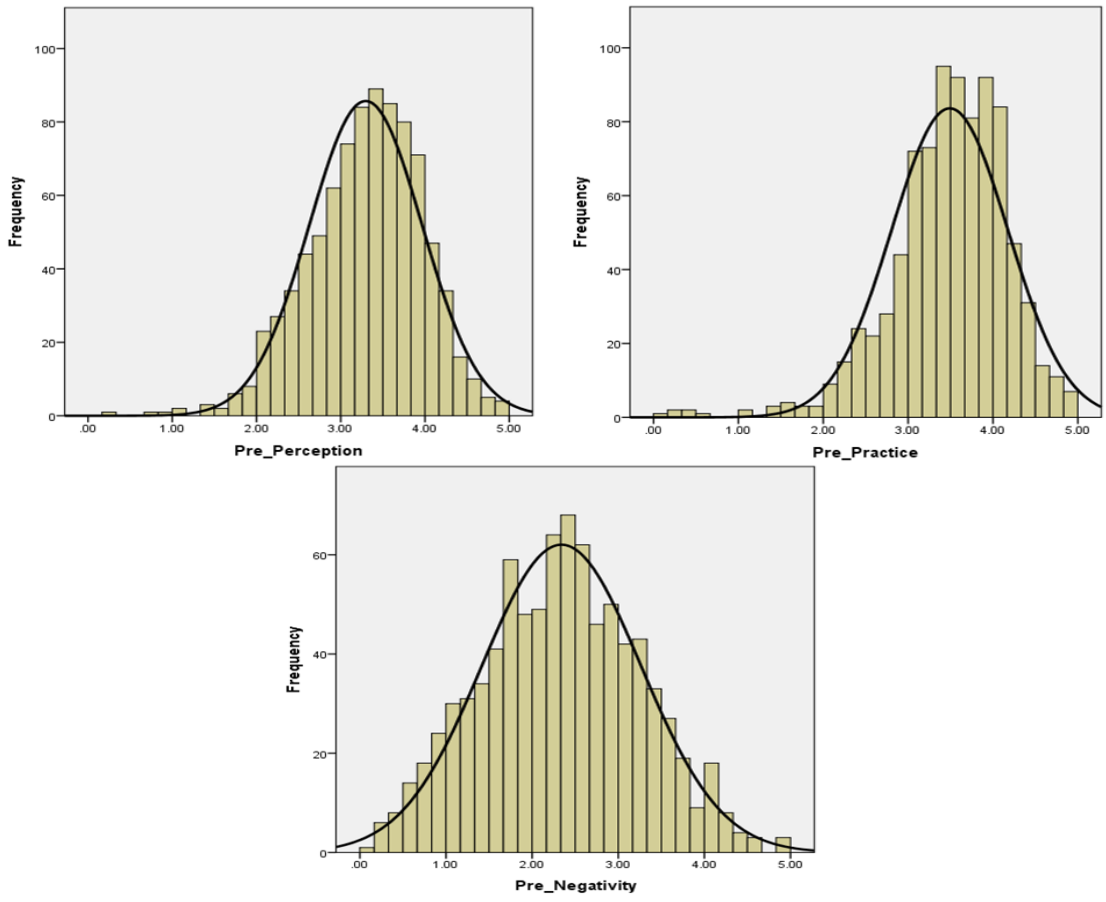


Figure 8.1: Histograms of the Mean Distributions of the Three Components

An inspection of the histograms in Figure 8.1 provides information about the distribution of the sample for each variable. As we can see, the histograms of *PERCEPTION* and

PRACTICE are right-skewed. The data were clustered on the right side, implying that the majority of teachers positively agreed that F1DT is an appropriate tool for use in schools. The histogram of the *NEGATIVITY* items, in contrast, showed a normal distribution of the means in which the highest peak of the responses is at the centre of the bell curve (*M*=2.34). Nevertheless, this also reveals that a substantial percentage of teachers had a negative view of the implementation of F1DT in schools.

For further understanding about the status quo of Malaysian teachers' perceptions and practices of F1DT, the analyses at the item level for each variable are outlined in the subsequent paragraphs.

8.1.1.1.1 Teachers' assessment beliefs about the purposes and uses of the F1DT

Table 8.2 summarises the mean scores for each item in *PERCEPTION* (*M* =3.29, *SD* =.67).

Table 8.2

Mean Scores for Individual Item in PERCEPTION (n=862)

Items	*M*	*SD*
streaming students	3.62	0.79
intervention for LD	3.58	0.83
referral to SEP	3.50	0.96
identification of LD	3.48	0.88
pedagogical decision	3.48	0.89
learning needs	3.46	0.90
learning improvement	3.45	0.86
pre-existing knowledge	3.36	0.94
learning potentials	3.30	0.98
student comparison	3.29	1.01
strengths and weaknesses	3.26	1.00
innate abilities	3.23	1.00
low-performing students	3.22	1.04
info accuracy	3.14	1.12
future performance	2.76	1.18
quality of teaching	2.62	1.20

As shown, the first nine items recorded high mean scores ranging from 3.30 to 3.62. In order of agreement, teachers in this study were likely to agree that the three most

important purposes of F1DT are: (a) to stream students according to their current attainment level (*M* =3.62, *SD* =.79); (b) to assist teachers in the planning of intervention strategies for students with learning difficulties (LD) (*M* =3.58, *SD* =.83); and (c) to use as a referral for the inclusion of students with LD in a Special Education Programme (SEP) (*M* =3.50, *SD* =.96). Furthermore, the data indicated a low *SD* (*SD* < 1) for these nine items, showing that teachers collectively believed that F1DT fits the purposes mentioned in the statements.

In contrast, the last seven items had a high *SD* greater than 1. This signalled that teachers' responses to these items were spread out over a wider range of values. Mean scores for these items were equal to or less than the mean average score (*M* = 3.29). Note that the items with least agreement were related to the use of F1DT to predict students' future performance (*M* =2.76, *SD* =1.18) and to assess teachers' quality of teaching (*M* =2.62, *SD* =1.20).

8.1.1.1.2 Teachers' assessment practices about the use of F1DT

Likewise, the findings from mean scores of individual items in *PRACTICE* supported the notion of an encouraging attitude towards the implementation of F1DT in schools (see Table 8.3).

Table 8.3

Mean Scores for Individual Item in PRACTICE (n=862)

Items	*M*	*SD*
streaming students	3.75	.83
identification of LD	3.71	.87
strengths and weaknesses	3.71	.86
learning potentials	3.66	.86
pre-existing knowledge	3.64	.87
intervention for LD	3.63	.88
learning improvement	3.58	.89
learning needs	3.53	.90
innate abilities	3.45	1.05
student comparison	3.43	1.07
pedagogical decision	3.38	0.98
quality of teaching	3.34	1.09
low-performing students	3.27	1.10
future performance	2.78	1.21

From the results in Table 8.3, the use of information from F1DT as a means for streaming students according to their attainment recorded the highest mean score (*M* =3.75, *SD* =.83), as similarly shown in *PERCEPTION*. We can also see that there was a general trend of high agreement in many of the corresponding items. With the exception of the last item, all items had a mean score above 3. This indicated that the least favourable item was the use of F1DT to predict students' future performance accurately (*M* =2.78, *SD* =1.21).

8.1.1.1.3 Teachers' responses to negative items about F1DT

In the questionnaire, six statements shown in Table 8.4 were included in *PERCEPTION* and *PRACTICE* items. After PCA analysis, however, they were extracted as a new component, tagged as *NEGATIVITY* as they consisted of negative statements about F1DT.

Table 8.4

Mean Scores for Individual Item in NEGATIVITY (n=862)

Items	*M*	*SD*
*I think the F1DT ….		
… is carried out to fulfil the administrative directives	2.84	1.24
… results are collected but largely ignored	2.20	1.23
… is largely a waste of time	1.73	1.28
**I use the F1DT….		
… but prefer to use UPSR results	2.54	1.30
… because I am told to do so	2.38	1.37
… but make little use of the results	2.35	1.24

* items included in *PERCEPTION*
** items included in *PRACTICE*

The result in Table 8.4 showed a lower agreement among teachers regarding items in this factor as compared to the items in *PERCEPTION* and *PRACTICE*. It is apparent that teachers did not consider F1DT as a directive order from the school administrators and that its results should be ignored. Of all the six items, teachers disagreed the most with the statement that F1DT is a waste of time (*M* = 1.73, *SD* = 1.28).

8.1.1.2 HLM analyses

Furthermore, HLM analyses were carried out to examine the variability of teachers' responses at different levels – individual level (level 1) and group level (level 2) – regarding their assessment beliefs and practice about F1DT.

To begin with, an unconditional (null) model, also known as the random-intercept model (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; West, Ryu, Kwok & Cham, 2011), was conducted to test whether the nested structure in this 862-dataset called for HLM application. Three null models were calculated. The model partitions the variance in the outcome variables – *PERCEPTION*, *PRACTICE* and *NEGATIVITY* – into individual (Equation 1) and group (Equation 2) variance components. In the context of this study, the equations of the null model are defined as follows:

HLM Null Model (n=862):

$$Level\text{-}1: PERCEPT1_{ij} = \beta_{0j} + r_{ij} \quad (1)$$

$$Level\text{-}2: \beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

where:

$PERCEPT1_{ij}$	= the perception of F1DT by teacher *i* in school *j*
β_{0j}	= the mean for *PERCEPTION* across schools
r_{ij}	= random errors associated with teacher *i* within school *j*
γ_{00}	= the grand mean for *PERCEPTION* across teachers and schools
u_{0j}	= the deviation (error) of school *j*'s mean from the grand mean

An essential result to examine in this model is the chi-square test (χ^2) generated in the final estimation of variance components (Hofmann, 1997; Robson & Pevalin, 2016; Woltman, Feldstain, MacKay, & Rocchi, 2012). The results of the three null models are presented in Table 8.5.

Table 8.5

HLM Output for the Null Models (n=862)

Outcome variable	Final estimation of variance components					
	Random effect	Standard deviation	Variance component	*df*	χ^2	*p*-value
PERCEPTION	INTRCPT1, u_0	0.16385	0.02685	20	72.695	<0.001
	level-1, *r*	0.64928	0.42157			
PRACTICE	INTRCPT1, u_0	0.12262	0.01504	20	47.189	<0.001
	level-1, *r*	0.67462	0.45511			
NEGATIVITY	INTRCPT1, u_0	0.16141	0.02605	20	45.859	<0.001
	level-1, *r*	0.91088	0.82970			

As shown above, the results revealed that *PERCEPTION*: χ^2 (20) = 72.70, $p < .001$, *PRACTICE*: χ^2 (20) = 47.19, $p < .001$ and *NEGATIVITY*: χ^2 (20) = 45.86, $p < .001$, indicating that there were statistically significant variances at level 2. More importantly, level 1 variabilities were also significantly different from zero ($\sigma^2 = .42$, $p < .001$, $\sigma^2 = .46$, $p < .001$ and $\sigma^2 = .83$, $p < .001$) for *PERCEPTION*, *PRACTICE* and *NEGATIVITY* respectively. From these findings, it is concluded that there is evidence of group effects for all the outcome variables and this justifies the use of HLM for this dataset.

As additional steps, the suitability of the application of HLM can be examined through the indices of the intraclass correlation coefficient (ICC) (Garson, 2013; Luke, 2004; Raudenbush, 1993; Robson & Pevalin, 2016; Snijders & Bosker, 2012) and the design effect statistics (McCoach & Adelson, 2010; Peugh, 2010; Snijders & Bosker, 2012).

ICC provides information about how similar or homogeneous individuals are within groups (McCoach & Adelson, 2010; Robson & Pevalin, 2016). It ranges from 0 to 1, where 0 means perfect data independence and 1 signifies that all individuals are completely similar within the cluster (Robson & Pevalin, 2016; Warne et al., 2012). In addition, ICC can also be interpreted as the proportion of variance in the dependent variable accounted by the level 2 units (Lorah, 2018; Luke, 2004; Snijders & Bosker, 2012; West, Ryu, Kwok & Cham, 2011). The value of ICC is computed by dividing the between-group variability (τ_{00}) by the total variability ($\tau_{00} + \sigma^2$), as the following equation shows:

$$\text{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \qquad (3)$$

Using the information from Table 8.5, where τ_{00} is level 2 variance (= .03) and σ^2 is level 1 variance (= .42), the ICC for teachers' perception about F1DT can be calculated as:

$$\text{ICC} = \frac{0.03}{0.03 + 0.42} = 0.06$$

Besides, the design effect statistics can also be used to check the presence of group variability in a two-level sample. The design effect is computed as:

$$\text{Design effect} = 1 + (n_j - 1)\,\text{ICC} \qquad (4)$$

where n_j is the average sample within each cluster. In this data, n_j is 41.05 (862/21). From the ICC of *PERCEPTION*, we can estimate the variance in teachers' responses as follows:

$$\text{Design effect} = 1 + (41.05 - 1) \times 0.06 = 3.40$$

The results of the ICC and design effect calculations for the three outcome variables are reported in Table 8.6.

Table 8.6

ICC and Design Effect of HLM Null Models (n=862)

Outcome variable	ICC	Variance component		Design effect
		Level 1	Level 2	
PERCEPTION	.06	94%	6%	3.40
PRACTICE	.03	97%	3%	2.20
NEGATIVITY	.03	97%	3%	2.20

The ICCs of all the outcome variables were less than 1 – *PERCEPTION*=.06, *PRACTICE* = .03 and *NEGATIVITY* = .03 – suggesting that the variances in teachers within a school (level 1) were greater than the variances between teachers across schools (level 2). The

result indicated that 94% of the variance in teachers' perception was at the individual level, with 6% at the school level, while for *PRACTICE* and *NEGATIVITY*, 97% of the variances were at the individual level and 3% at the school level. Although a high value of ICC is required for the application of HLM, there is no fixed guideline for the cut-off of ICC (Guo, 2005; Robson & Pevalin, 2016). Moreover, Nezlek (2008) advised researchers to focus on the structure of the nested data rather than using ICC in a decision for HLM analyses.

Further, the results of the design effects help researchers to quantify how much variation of responses is present at level 2 (McCoach & Adelson, 2010; Peugh, 2010). A high index of design effect implies a relatively large variance at between-group level (Snijders & Bosker, 2012). A design effect equal to 1 means no clustering effect, while a value greater than 1 shows the violation of the assumption of independence of observation (McCoach & Adelson, 2010). In Table 8.6, the design effects for the three outcome variables were 3.4 (*PERCEPTION*) and 2.2 (*PRACTICE* and *NEGATIVITY*). Notably, these findings satisfy the recommendation that the design effect estimates greater than 1 demonstrate a reasonable fit for HLM.

In summary, the results of the null models, the ICCs and the design effects demonstrated significant clustering effects in teachers' responses at both individual and group levels. It is therefore warranted that HLM can be employed for extended models to answer the subsequent research questions.

8.1.2 Teachers' written responses

In addition to the data from the survey, information in relation to Research Questions 1 and 2 were also gathered from teachers' reflective comments which were written on the last page of the questionnaire. These qualitative data provide additional information to gain a better insight into teachers' perception and practice about the implementation of F1DT in schools. For this purpose, I utilised a thematic approach to the analysis of the data.

Table 8.7

Emerging Themes from Teachers' Written Feedback (n=60)

Themes	Number of comments
Favourable response to F1DT	17
Unfavourable response to F1DT	17
Ideal assessment tool	11
Improvement of F1DT implementation	9
General issues about education in Malaysia	6

There were five themes emerged from the data (see Table 8.7). Examples of extracts from what was written, as original (in English) or translated (from the Malay language) for each theme are presented in the paragraphs that follow.

8.1.2.1 Favourable response to F1DT

The first theme is primarily about teachers' acknowledgement of the usefulness of F1DT in their instructional decision-making. They positively endorsed the uses of F1DT to:

a) Identify students' learning potential

> "In my view, F1DT is very useful for identifying the level of students' ability in facilitating teachers to design pedagogical strategies in the classroom. UPSR results sometimes do not reflect the true potential of a student." (RE26RA)

> "F1DT is very useful to understand the potentials of students. It is also good for teachers to help them in re-evaluating their pedagogical way of teaching." (MA21GH)

b) Identify students' weaknesses/learning problems

> "This diagnostic test needs to be implemented in helping students to identify their learning weaknesses." (PA21PE)

> "F1DT is useful for teachers to identify students' learning problems." (TA20IS)

> "F1DT is still relevant especially for subject teachers to identify the actual learning weaknesses and strengths of the students." (HO25CH)

c) Measure students' existing knowledge

> "F1DT is very important to understand students' current knowledge about the subjects tested." (LA14JA)

> "I support the implementation of F1DT as a starting point to know what students can do and what they have learnt in primary school." (NA19YA)

d) Stream students

> "F1DT is very helpful for teachers to know about students' academic achievement and this facilitates the placement of students according to their level of ability." (KA05IL)

> "F1DT is a good way to determine student's placement in a class based on their achievement level. This is to help teachers to plan for focused learning and teaching strategies to cater to students' needs." (RA07NI)

e) Facilitate the design of instructional planning

> "Information from F1DT is important for teachers in planning for teaching and learning processes." (MA08OS)

Apparently, the above positive comments reaffirm teachers' positive attitude towards F1DT, believing that it has been useful in providing information about students as well as facilitating their pedagogical plans.

8.1.2.2 Unfavourable response to F1DT

In contrast, the second theme revealed an adverse reaction by some teachers to the implementation and uses of F1DT. Teachers raised concerns about the following issues:

a) Misconception of F1DT as a diagnostic measure

> "School has a misconception about this diagnostic test. It is supposed to be a 'placement test' rather than diagnostic because we used the result to put students into set systems according to the level of reading and writing proficiency (for English subject). In my opinion, this diagnostic test is not relevant to Form 1 students at the beginning of the year as UPSR result is supposed to be reliable enough." (AM13BO – written in English)

> "A diagnostic test is a wonderful tool to help teachers, especially in a large classroom, but most times we use it to eliminate problematic students and not to help them." (NA07LA – written in English)

b) F1DT scores are not a reflection of students' actual ability

> "F1DT does not fully reflect the abilities, creativity and strengths of the students. It is just as a prerequisite for class placements." (JAYU05)

"This diagnostic test is not an accurate instrument for measuring students' abilities." (RO17AW)

"Students with specific learning problems should be given another test so that teachers can identify their learning needs appropriately." (ZA19AM)

c) UPSR is enough for the measurement of students' prior or current knowledge

"There is no need for this diagnostic test because UPSR result can be used as a measure for assessing students' prior or current knowledge." (SU17SI)

"It is preferable to use UPSR results for the purpose of streaming students. It is more valid and transparent." (RA20MA)

d) Students did not care about F1DT

"Based on my experience of conducting this diagnostic test, students did not truly care about the test, just to fulfil the requirement of the class placement. Thus, the scores are not indicative of their real performance. There were cases where good students were placed in a weak class because of their low scores in FIDT." (CA20RI)
"UPSR is good enough for the class placement. Students tended to cheat in the diagnostic test and they did not care about the scores at all." (BU18SA)

The above remarks imply that teachers may no longer be convinced that F1DT is relevant to the current situation in schools.

8.1.2.3 An ideal assessment tool

The third theme is about an ideal assessment tool. It seemed that teachers tended to prefer the following characteristics:

a) Standardised and reliable

"An assessment tool needs to be standardised to save cost and time. The standardisation may increase reliability." (NO31MO)

b) Accurate and continuous

"An ideal assessment should be the one that can provide accurate descriptions of students' progress and it should be practised continuously." (FE15AB)

c) Practical and easy

> "Good assessment tools must be practical and easy to administer." (SA03SI)

d) Does not overburden teachers

> "An ideal assessment tool is the one that does not overburden the teachers." (RO19NO)

e) Brings a positive impact

> "In my opinion, an ideal assessment tool should bring a positive impact on students' development and eventually improve the quality of education in the country." (DO02IS)

f) Improves students' achievement

> "A good assessment tool helps to improve students' academic achievement." (ME10OS)

g) Various assessment approaches

> "Evaluation of students' progress should also be conducted through other assessment tools especially the one measuring students' practical skills, not only rely on how many 'A's they have." (ID02EL)

From these views, it can be speculated that F1DT may not have truly met the criteria of an ideal assessment tool that is supposed to be beneficial for teachers and students.

8.1.2.4 Improvement of F1DT implementation

Another important theme relating to F1DT is several suggestions for improvement of its implementation. AL08HE asserted that:

> "Better implementation of F1DT is necessary for teachers to optimise its use in teaching and learning." (AL08HE)

Some of the highlighted recommendations were:

a) Combination of UPSR and F1DT results

> "Teachers should not rely heavily on the scores of F1DT. UPSR results should be taken into consideration in determining students' placement in a class." (ZA14NA)

b) Inclusion of other assessment methods

> "Various methods can be used as diagnostic assessments. Schools should include assessment of practical tasks for relevant subjects". (TI27TU)

c) Alignment of test formats with UPSR and PT3

> "The test formats of F1DT need to be aligned with UPSR and PT3 format as this may give teachers more comprehensive information about students' past and future performance." (DO17DO)

d) Inclusion of other subjects

> "This diagnostic test is important to support effective teaching and learning. Therefore, it should include all subjects, not only the three subjects – Malay language, English language and mathematics." (BA01KU)

Despite its widespread use in schools, teachers still believed that there were several issues that need to be looked into for the improvement of F1DT implementation.

8.1.2.5 General issues about education in Malaysia

Lastly, teachers also commented about a few issues in the Malaysian education system. In particular, they wrote about the followings:

a) Performance-oriented education system

> "The education system in Malaysia has more emphasis on GPA in public examinations." (ZI01LA)

b) Political interference in educational policies

> "Education in Malaysia is like "*rojak*"[7] and many political agendas interfere in the making of educational policies. This may divert the focus to develop students' learning progress" (MA03MS)

> "As long as politicians determine the direction of the education system, we are not going anywhere." (FA01AR)

c) Heavy workload for teachers

> "Teachers nowadays are over-burdened with other non-teaching tasks." (NI02MU)

[7] The term "*rojak*" is a colloquial Malay language which means eclectic mix: in this case, a mixture of differing ideas from different groups of stake-holders in education.

Notably, these comments describe the prevalent challenging issues in the Malaysian education system as discussed in Chapters 1 and 2.

Overall, it is noted that written comments from teachers not only provide valuable information about the issues of interest, i.e., F1DT, but also other relevant issues within the scope of this study.

8.2 Relationship between Teachers' Assessment Beliefs and Practices about the Form 1 Diagnostic Test

Research Question 3 aimed to examine the relationship between teachers' assessment beliefs and their assessment practices concerning F1DT. Two complementary analyses – HLM models and Pearson product-moment correlation coefficient – were used to investigate the belief-practice relationship.

Here, only the items for *PERCEPTION* and *PRACTICE* were computed, while items in *NEGATIVITY* were removed from the analyses.

Firstly, an important question in relation to the belief–practice relationship is; does belief influence practice? An extension of the null model, HLM Model 1a was computed to answer this question. Contrary to the unconditional model, the outcome variable (*PRACTICE*) is now conditional on teachers' response to the *PERCEPTION* scale. Specifically, teachers' agreement on their assessment practice of F1DT was modelled as a function of their reported perception of the purposes and uses of the F1DT. The relationship of this perception-practice is expressed in the following equations:

HLM Model 1a:

$$Level\text{-}1:\ PRACTIC1_{ij} = \beta_{0j} + \beta_{1j}*(PERCEPT1_{ij}) + r_{ij} \quad (5)$$

$$Level\text{-}2:\ \beta_{0j} = \gamma_{00} + u_{0j} \quad (6)$$
$$\beta_{1j} = \gamma_{10}$$

Table 8.8

HLM Output for Model 1a (n=862)

Final estimation of fixed effects (with robust standard errors)					
Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value
For INTRCPT1, β_0 INTRCPT2, γ_{00}	3.503101	0.034740	100.838	20	<0.001
For PERCEPT1 slope, β_1 INTRCPT2, γ_{10}	0.695719	0.040570	17.148	840	<0.001

Final estimation of variance components					
Random effect	Standard deviation	Variance component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	0.14079	0.01982	20	85.62005	<0.001
Level 1, r	0.50146	0.25146			

Table 8.8 reports the results of HLM Model 1a. The slope of *PERCEPTION* was positive (b = .69, t = 17.15, df = 840, p < .001) and this means that teachers' responses to this variable were significantly related to their responses on *PRACTICE*. Additionally, the final estimation of variance components in Table 8.8 summarises the variability of responses at individual and group levels. The estimate for level 1 variance was 0.25 and that of level 2 variance was 0.02. The ICC is then 0.02/ (0.02 + 0.25) = 0.07. This means that schools account for 7% of the variability of the responses, while 93% of the variance was primarily explained at the individual level. As expected, level 1 variance was greater than level 2 variance, further supporting the findings in the null models.

Equally important, the question should be asked whether practice shapes beliefs. To address this question, HLM Model 1b, as shown in the following equations, was carried out to estimate the outcome variable, i.e., *PERCEPTION*, as a function of *PRACTICE*.

HLM Model 1b:

$$Level\text{-}1:\ PERCEPT1_{ij} = \beta_{0j} + \beta_{1j}*(PRACTIC1_{ij}) + r_{ij} \quad (7)$$

$$Level\text{-}2:\ \beta_{0j} = \gamma_{00} + u_{0j} \quad (8)$$
$$\beta_{1j} = \gamma_{10}$$

The outputs of HLM Model 1b in Table 8.9 further confirm a strong relationship between perception and practice. Here, the finding in the slope of *PRACTICE* (b = .64, t = 11.29,

$df = 840$, $p < .001$) suggests that teachers responded positively to the items in *PERCEPTION* as a result of their agreement to *PRACTICE*-related items.

Table 8.9:

HLM Output for Model 1b (n=862)

Final estimation of fixed effects (with robust standard errors)					
Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	3.319634	0.041320	80.340	20	<0.001
For PRACTIC1 slope, β_1					
INTRCPT2, γ_{10}	0.644381	0.057082	11.289	840	<0.001
Final estimation of variance components					
Random effect	Standard deviation	Variance component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	0.17685	0.03127	20	132.09473	<0.001
Level 1, *r*	0.48260	0.23290			

An inspection of the scatterplot in Figure 8.2 further confirms that the relationship is strong, indicating a positive association between teachers' assessment beliefs and their assessment practices. It is evident that items of both variables were neatly clustered around the straight line, i.e., high agreement on items from *PERCEPTION* is associated with a high agreement on corresponding items from *PRACTICE*. In general, the result suggests there was a strong consistency between teachers' perception of the purposes of F1DT and how they used it to make decisions about students and planning for teaching and learning.

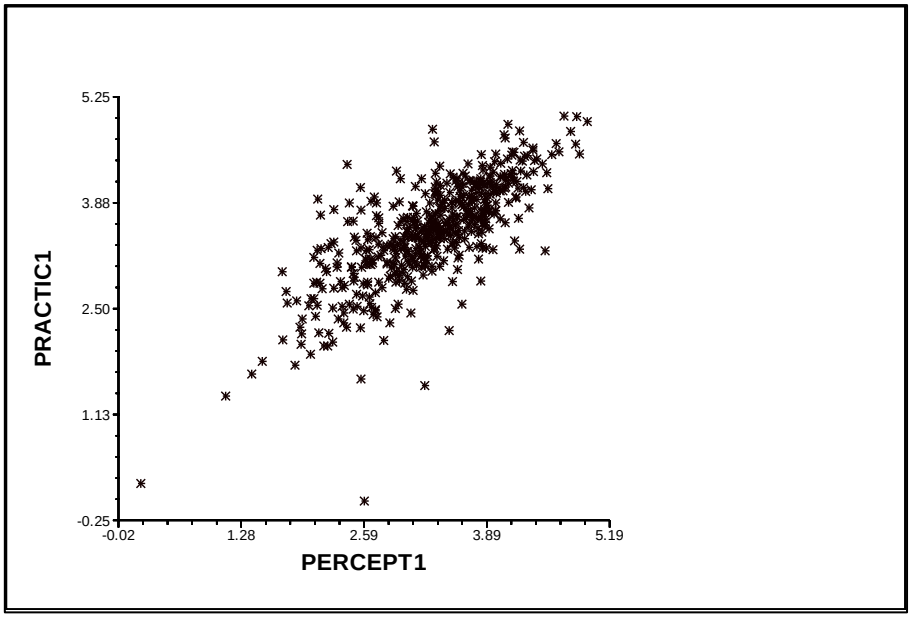


Figure 8.2: Scatterplot of the Association between *PERCEPTION a*nd *PRACTICE*

As a supplement to the above analysis, a Pearson product-moment correlation coefficient was used to examine this belief–practice relationship. A summary of the findings can be viewed in Table 8.9. Similar to the HLM analysis, the correlation between *PERCEPTION* and *PRACTICE* was found to be statistically significant (r = .685, n=862, p = .000). For this analysis, the $r \geq .50$ suggests that there was a strong correlation between the two variables (Cohen, 1988, 1992).

Overall, a series of complementary analyses, as reported above, support the notion that teachers' assessment practices are significantly influenced by their assessment beliefs (e.g., Barnes et al., 2015; Brown, 2004; Brown et al., 2009) and vice versa. In the context of this study, the beliefs teachers had about the purposes of F1DT and their practices of utilising its information to make decisions about instructional planning were pretty much aligned.

8.3 Variables Influencing Assessment Beliefs and Practices

Another objective of this study is to explore the association of several predictors – teachers' individual characteristics (level 1) and school characteristics (level 2) – as moderating factors on how teachers responded to the outcome variables – *PERCEPTION* and *PRACTICE*.

8.3.1 Influence of individual characteristics on teachers' assessment beliefs and practices

The first analysis involved an estimate of the relationship between level-1 predictors and the outcome variables. This is known as a random coefficient regression model (Hofmann, 1997; Raudenbush & Bryk, 2002; Woltman, Feldstain, MacKay & Rocchi, 2012). In this study, each outcome variable was analysed separately and denoted as HLM Model 2a (*PERCEPTION)* and HLM Model 2b (*PRACTICE)*. In these models, all level-1 predictors, i.e., years of teaching *(YEARS)*, educational background *(EDULEVEL)*, position in the school *(POSITION)* and familiarity with F1DT *(F1DT)*, were used to explain part of the variability of the outcome variables, as exemplified in the following expressions:

HLM Model 2a:

$$Level\text{-}1:\ PERCEPT1_{ij} = \beta_{0j} + \beta_{1j}*(YEARS_{ij}) + \beta_{2j}*(EDULEVEL_{ij}) + \beta_{3j}*(POSITION_{ij}) + \beta_{4j}*(F1DT_{ij}) + r_{ij} \quad (9)$$

$$Level\text{-}2:\ \beta_{0o} = \gamma_{00} + u_{0j} \quad (10)$$
$$\beta_{1o} = \gamma_{10}$$
$$\beta_{2o} = \gamma_{20}$$
$$\beta_{3o} = \gamma_{30}$$
$$\beta_{4o} = \gamma_{40}$$

There are two things to be noted here. First, all predictors were computed simultaneously. Although Tabachnick and Fidell (2013) cautioned that including all the predictors in one equation might be problematic and could likely increase the risk of achieving insufficient statistical power, it is also suggested that the use of a simple and meaningful model, which consists of uncorrelated predictors, is acceptable (Robson & Pevalin, 2016). Second, each of the level-1 predictors was entered as a group-centred variable. Kreft, de Leeuw and Aiken (1995) advised that since groups – level 1 and level 2 – are considered as separate entities, it is therefore recommended to centre within each group. Furthermore, the use of this centring procedure is preferable if the study aims to examine the effects of level-1 and level-2 predictors independently (Hofmann, 1997; Hofmann & Gavin, 1998; Kreft et al., 1995; West et al., 2011; Woltman et al., 2012). Also, this procedure yields more accurate estimates of the intercepts (Woltman et al., 2012).

8.3.1.1 The influence of level-1 predictors on assessment beliefs

The results of HLM Model 2a *(PERCEPTION)* are shown in Table 8.10. This model estimated that the fixed effects of level-1 predictors as significant influences on teachers' responses to *PERCEPTION* ($b = 3.32$, $t = 79.73$, $df = 20$, $p < .001$). As expected, the overall variability of teachers' responses was primarily due to their individual differences (ICC = .06, level 1 = 94% and level 2 = 6%).

Table 8.10

HLM Output for Model 2a (PERCEPTION) (n=862)

Predictors	**Final estimation of fixed effects (with robust standard errors)**					
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value
	For INTRCPT1, β_0					
	INTRCPT2, γ_{00}	3.316506	0.041596	79.730	20	<0.001
Years of teaching	For YEARS slope, β_1					
	INTRCPT2, γ_{10}	0.018026	0.019710	0.915	837	0.361
Education background	For EDULEVEL slope, β_2					
	INTRCPT2, γ_{20}	0.078722	0.070945	1.110	837	0.267
Position	For POSITION slope, β_3					
	INTRCPT2, γ_{30}	−0.168241	0.058702	−2.866	837	0.004
Familiarity with F1DT	For F1DT slope, β_4					
	INTRCPT2, γ_{40}	0.059704	0.050684	1.178	837	0.239
	Final estimation of variance components					
	Random effect	Standard deviation	Variance component	*df*	χ^2	*p*-value
	INTRCPT1, u_0	0.16399	0.02689	20	73.02463	<0.001
	level 1, *r*	0.64783	0.41968			

The final estimation of fixed effects showed the contribution of each predictor to the overall impact. It is revealed that teachers' position in a school ($b = -.17$, $t = -2.87$, $df = 837$, $p = .004$) made a significant unique contribution to the prediction of perceived assessment. No significant relationships were found between teachers' assessment beliefs and the other three variables; years of teaching ($b = .00$, $t = .92$, $df = 837$, $p = .361$), education background ($b = .08$, $t = 1.11$, $df = 837$, $p = .267$) and familiarity with F1DT ($b = .06$, $t = 1.18$, $df = 837$, $p = .239$). Note that the negative coefficients for *POSITION* (coded as 1 = teachers, 2 = administrators) signified a negative direction of the

relationship. This is supported by descriptive statistics in Table 8.11, implying that teachers with administrative duties ($M = 3.18$, $SD = .77$) tended to have a lower agreement in their views about the purposes and uses of F1DT as compared to teachers ($M = 3.31$, $SD = .65$). Nevertheless, the effect size was very small (Cohen's $d = 0.195$).

Table 8.11

Descriptive Statistic for POSITION (n=862)

		Descriptive statistics		
Groups		*N*	*M*	*SD*
Position	Teachers	749	3.31	.65
	Administrators	113	3.18	.77

8.3.1.2 The influence of level-1 predictors on assessment practices

Similarly, another model, HLM Model 2b, was deployed to examine the relationship between the four predictors and teachers' assessment practices regarding F1DT. A summary in Table 8.12 shows that the overall correlation between the variables of interest and the outcome variable was statistically significant ($b = 3.5$, $t = 100$, $df = 20$, $p < .001$). A closer examination at the final estimation of fixed effects found that the number of teaching years ($b = .00$, $t = .14$, $df = 837$, $p = .89$) and academic background ($b = .05$, $t = .69$, $df = 837$, $p = .492$) did not correlate with the way teachers utilised F1DT information for instructional decision-making. The findings, however, indicated that the other two teacher-characteristic variables were statistically significant, with *POSITION* b$r = -.19$, $t = -3.52$, $df = 837$, $p < .001$) recorded the strongest contribution to influence of the outcome variable and were followed by *F1DT* ($b = -.13$, $t = -2.28$, $df = 837$, $p = .023$). In regard to the variability of the responses at different levels, this model confirmed previous results that the variance of teachers' responses was largely due to their individual differences (ICC = 0.03).

Table 8.12

HLM Output for Model 2b (PRACTICE) (n=862)

Predictors	**Final estimation of fixed effects (with robust standard errors)**					
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value
	For INTRCPT1, β_0					
	INTRCPT2, γ_{00}	3.500344	0.034975	100.083	20	<0.001
Years of teaching	For YEARS slope, β_1					
	INTRCPT2, γ_{10}	0.002322	0.016857	0.138	837	0.890
Education background	For EDULEVEL slope, β_2					
	INTRCPT2, γ_{20}	0.052350	0.076163	0.687	837	0.492
Position	For POSITION slope, β_3					
	INTRCPT2, γ_{30}	−0.189666	0.053888	-3.520	837	<0.001
Familiarity with F1DT	For F1DT slope, β_4					
	INTRCPT2, γ_{40}	−0.134272	0.058928	-2.279	837	0.023
	Final estimation of variance components					
	Random effect	Standard deviation	Variance component	*df*	χ^2	*p*-value
	INTRCPT1, u_0	0.12306	0.01514	20	47.66558	<0.001
	level 1, *r*	0.67125	0.45058			

As noted above, the significant predictors (*POSITION* and *F1DT*) were negatively correlated with *PRACTICE*. As reported in Table 8.13, it is interesting to note that school administrators (*M* = 3.18, *SD* = .77), who endorse the use of F1DT as a diagnostic measure in school, tended to have a lower agreement about the practice of F1DT information in instruction. Teachers (*M* = 3.31, *SD* = .65), on the other hand, appeared to agree more on the uses of F1DT as mentioned in the questionnaire. This finding, with a relatively small effect size (Cohen's *d* = 0.203), is reasonably understandable. In practice, teachers have more active roles in designing and administrating F1DT and analysing its results than teachers with administration roles such as Principal and Senior Assistants. With regard to the familiarity of teachers with F1DT, which involves the design, administration, analysis and utilisation of F1DT, it is expected teachers who directly involve in F1DT-related activities (*M* = 3.54, *SD* = .64) tended to appreciate its uses more than teachers without such exposure (*M* = 3.40, *SD* = .75). The strength of the

relationships between teachers' familiarity with F1DT and their responses to *PRACTICE* was relatively small (Cohen's $d = 0.205$).

Table 8.13

Descriptive Statistics for POSITION and F1DT (n=862)

Predictor		Descriptive statistics		
	Groups	*N*	*M*	*SD*
Position	Teachers	749	3.51	.68
	Administrators	113	3.37	.74
Familiarity with F1DT	YES	546	3.54	.64
	NO	316	3.40	.75

8.3.2 Influence of school characteristics on teachers' assessment beliefs and practices

The next step is to investigate the significance of the relationship between level-2 predictors and the outcome variables. To test this, I employed the means-as-outcomes model (Kahn, 2011; Raudenbush & Bryk, 2002; Woltman et al., 2012), also termed intercepts-as-outcomes (Hofmann, 1997; Luke, 2004), in which *LOCATION* was grand-mean centred as the predictor of the slope in level 2. Two separate analyses, HLM Model 3a and HLM Model 3b, were conducted for *PERCEPTION* and *PRACTICE*, respectively.

HLM Model 3a:

$$Level\text{-}1:\ PERCEPT1_{ij} = \beta_{0j} + r_{ij} \quad (11)$$

$$Level\text{-}2:\ \beta_{0j} = \gamma_{00} + \gamma_{01}*(LOCATION_j) + u_{0j} \quad (12)$$

The results of HLM Model 3a are presented in Table 8.14. In this model, the output for robust standard errors is not appropriate for use due to the small number of level 2 units (n=21). Instead, I used the output from the final estimation of fixed effects without the robust standard errors, as shown below.

Table 8.14

HLM Output for Model 3a (PERCEPTION) (n=862)

Predictors	**Final estimation of fixed effects**					
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value
	For INTRCPT1, β_0					
	INTRCPT2, γ_{00}	3.317093	0.042661	77.754	19	<0.001
Location	LOCATION, γ_{01}	0.082312	0.093702	0.878	19	0.391
	Final estimation of variance components					
	Random effect	Standard deviation	Variance component	*df*	χ^2	*p*-value
	INTRCPT1, u_0	0.16397	0.02688	19	67.15321	<0.001
	level 1, *r*	0.64936	0.42167			

An examination of level 2 slope did not show a significant effect of school location on teachers' responses to their perception about F1DT (*b* = .08, *t* = .84, *df* = 19, *p* = .410). This implies that school location, i.e., urban or rural areas, did not influence teachers' overall assessment beliefs about F1DT. Additionally, the results of the variance components indicated that individual and school variabilities accounted for 94% and 6% of the total variance, respectively (ICC = 0.06), implying little variability at level 2 but still considered statistically significant (*b* = 3.32, *t* = 82.10, *df* = 19, *p* < .001).

Table 8.15

HLM Output for Model 3b (PRACTICE) (n=862)

Predictors	**Final estimation of fixed effects**					
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value
	For INTRCPT1, β_0					
	INTRCPT2, γ_{00}	3.501096	0.035735	97.975	19	<0.001
Location	LOCATION, γ_{01}	0.073893	0.078139	0.946	19	0.356
	Final estimation of variance components					
	Random effect	Standard deviation	Variance component	*df*	χ^2	*p*-value
	INTRCPT1, u_0	0.12172	0.01482	19	43.71826	0.001
	level 1, *r*	0.67472	0.45525			

Likewise, the insignificant contribution of level-2 predictors on the dependent variable was also observed in HLM Model 3b. Table 8.15 shows the results, which indicated that *LOCATION* (b = .07, t = 1.0, df = 19, p = .327) did not affect the way teachers responded to *PRACTICE*.

In general, the estimations from HLM Model 3a (ICC = 0.06) and HLM Model 3b (ICC = 0.03) further confirmed the findings of the preceding models that the differences between teachers across schools remained relatively small.

8.3.3 Interactions between level-1 and level-2 predictors on assessment beliefs and practices

The final step was to examine the cross-interactions between the two-level predictors (levels 1 and 2) in predicting the responses of teachers about their beliefs and practices concerning F1DT. To do this, I ran the slopes-as-outcomes (Hofmann, 1997; Luke, 2004) or the random intercepts and slopes model (Kahn, 2011; Raudenbush & Bryk, 2002; Woltman et al., 2012), i.e., HLM Model 5, entering level-1 predictors simultaneously as the independent variables and adding *LOCATION* as the predictor at the level 2 equation for each slope of the level-1 predictor (see equations below - HLM Model 4a (*PERCEPTION*) and HLM Model 4b (*PRACTICE*).

HLM Model 4a:

$$Level\text{-}1:\ PERCEPT1_{ij} = \beta_{0j} + \beta_{1j}*(YEARS_{ij}) + \beta_{2j}*(EDULEVEL_{ij}) + \beta_{3j}*(POSITION_{ij}) + \beta_{4j}*(F1DT_{ij}) + r_{ij} \quad (13)$$

$$Level\text{-}2:\ \beta_{0j} = \gamma_{00} + \gamma_{01}*(LOCATION_j) + u_{0j} \quad (14)$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}*(LOCATION_j)$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}*(LOCATION_j)$$
$$\beta_{3j} = \gamma_{30} + \gamma_{31}*(LOCATION_j)$$
$$\beta_{4j} = \gamma_{40} + \gamma_{41}*(LOCATION_j)$$

8.3.3.1 The influence of cross-level interaction of predictors on assessment beliefs

Generally, the results of HLM Model 4a revealed that cross-level interactions between teachers' characteristics and school characteristics on the dependent variable were insignificant. As seen in Table 8.16, it was estimated that interactions between *LOCATION* and all level-1 predictors were not significant (b = .08, t = .88, df = 19, p =

.392), which means that the interactions of all predictors at level 1 and level 2 had no influence on *PERCEPTION*.

Table 8.16

HLM Output for Model 4a (PERCEPTION) (n=862)

Cross-level interaction of predictors	**Final estimation of fixed effects**					
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value
Location* All predictors	For INTRCPT1, β_0					
	INTRCPT2, γ_{00}	3.317181	0.042680	77.723	19	<0.001
	LOCATION, γ_{01}	0.082166	0.093753	0.876	19	0.392
Location* Education background	For EDULEVEL slope, β_2					
	INTRCPT2, γ_{20}	0.081111	0.063435	1.279	833	0.201
	LOCATION, γ_{21}	0.311185	0.139584	2.229	833	0.026

Nevertheless, a closer look at the individual slope of a level-1 predictor indicated a significant effect of interaction between *LOCATION* and *EDULEVEL* on the dependent variable (b = .31, t = 2.22, df = 833, p = .026). That is, in relation to the individual's academic background, it was likely that teachers in urban (n=276) and rural (n=586) areas differ in their views about the purposes of F1DT. A plot of the mean score on *PERCEPTION* for each group of *LOCATION* and *EDULEVEL* is illustrated in a line graph below (Figure 8.3).

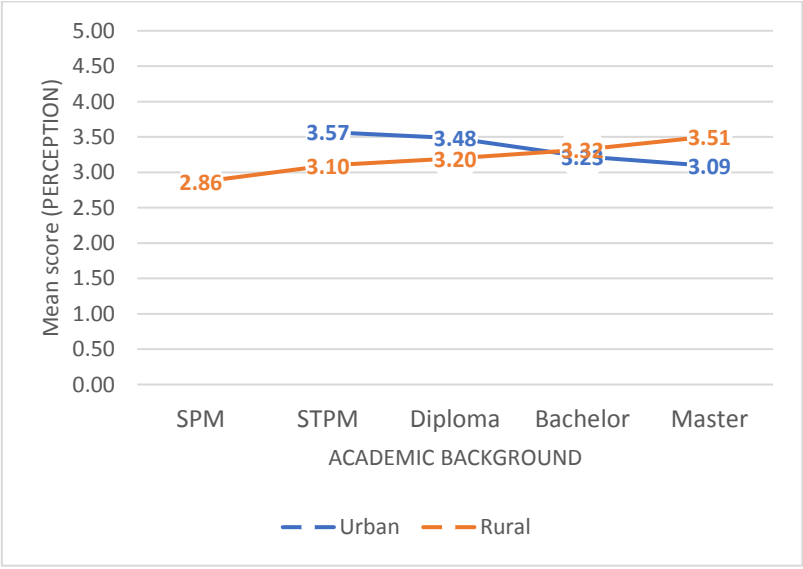


Figure 8.3: Comparison of Mean Scores of *PERCEPTION* (*LOCATION* and *EDULEVEL*)

Overall, there were differences in mean scores for most of the groups. Interestingly, there was an opposite trend in responses between teachers with pre-university academic qualifications (i.e., SPM, STPM and diploma) and university degrees (i.e., Bachelor and Master) at both urban and rural areas. It appeared that teachers in rural schools with a bachelor's degree and master's degree (n=577) tended to accept F1DT more than teachers with lower academic qualifications. On the other hand, this trend was not mirrored in urban schools where teachers with higher academic qualifications (n=272) regarded F1DT as a less preferred approach for diagnostic measures.

8.3.3.2 The influence of cross-level interaction of predictors on assessment practices

Correspondingly, the result of Model 4b (see Table 8.17) also pointed out that overall effects of level-2 predictor and all level-1 predictors on *PRACTICE* were not significant (b = .07, t = .94, df = 19, p = .358).

Table 8.17

HLM Output for Model 4b (PRACTICE) (n=862)

Cross-level interaction of predictors	Final estimation of fixed effects					
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value
Location* All predictors	For INTRCPT1, β_0					
	INTRCPT2, γ_{00}	3.501219	0.035755	97.923	19	<0.001
	LOCATION, γ_{01}	0.073640	0.078206	0.942	19	0.358
Location* Education background	For EDULEVEL slope, β_2					
	INTRCPT2, γ_{20}	0.060399	0.065621	0.920	833	0.358
	LOCATION, γ_{21}	0.352773	0.144393	2.443	833	0.015
Location* Familiarity with F1DT	For F1DT slope, β_4					
	INTRCPT2, γ_{40}	−0.099840	0.052708	−1.894	833	0.059
	LOCATION, γ_{41}	0.282123	0.106151	2.658	833	0.008

In contrast, a significant variability of an individual slope of a level-1 predictor, i.e., *EDULEVEL*, on teachers' responses to *PRACTICE* was replicated in this model. In addition, a significant effect of teachers' familiarity with F1DT was also estimated. As seen in Table 8.18, interactions of *LOCATION* with both *EDULEVEL* (b = .35, t = 2.44, df = 833, p = .015) and *F1DT* (b = .28, t = 2.67, df = 833, p = .008) contributed significantly to the prediction of teachers' responses on their assessment practices. This

suggests that urban and rural teachers appeared to be different in their assessment practices due to their educational background and experience in implementing F1DT.

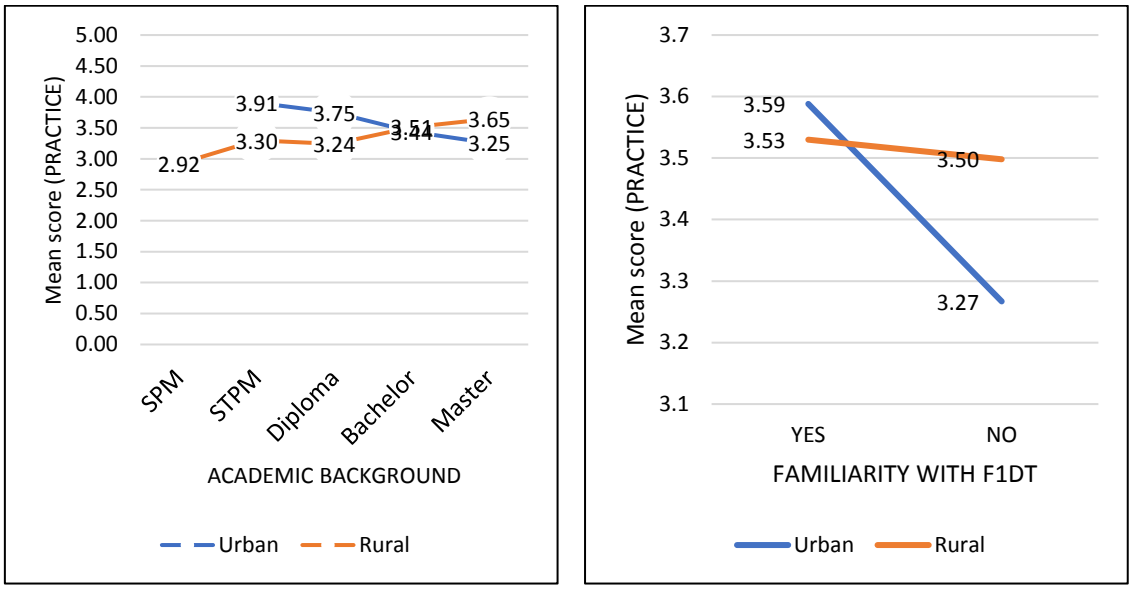


Figure 8.4: Comparison of Mean Scores of *PRACTICE* for *LOCATION* and *EDULEVEL* and *F1DT*

Figure 8.4 shows the difference in mean scores on *PRACTICE* between urban and rural teachers in terms of their education level and familiarity with F1DT. Here, a significant cross-level interaction between the location of the school and the different level of academic background of teachers indicated a similar trend to the responses on *PERCEPTION*; further supporting a strong alignment between teachers' assessment beliefs and practices. Results of HLM Model 4b estimated another salient finding, revealing a significant effect of interactions between *LOCATION* and teachers' familiarity with F1DT (see Figure 8.4). Teachers in both urban and rural schools tended to exhibit a similar pattern in which teachers who were not familiar with F1DT tended to be less favourable about its use in instructional decision-making. As stated earlier, this finding is reasonably predicted. It is interesting though; these teachers showed a noticeable disparity in their degree of agreement in using the information from F1DT in teaching and learning activities. Urban teachers tended to have a lower agreement in the use of F1DT information than teachers in rural schools.

Based on the preceding findings, it is concluded that some of the predictors studied had significant contributions on how teachers responded to the outcome variables – *PERCEPTION* and *PRACTICE*. At the individual level, a designated position held by

teachers in the school was found to be significantly related to their perceived assessment beliefs about F1DT. Similarly, there was a significant influence of teachers' position on their assessment practices, suggesting the discrepancy between teachers with administrative duties and ordinary teachers on their level of agreement about the implementation of F1DT in schools. Additionally, it was revealed that teachers' assessment practices were also influenced by their familiarity with this diagnostic test. At the school level, however, there was no indication of the school characteristics, i.e., location of the school (urban or rural), in predicting teachers' responses to their assessment beliefs and practices about F1DT. Interestingly, an investigation of the cross-level interaction of predictors indicated a significant effect of school location and teachers' educational background in predicting the way that teachers responded to items relating to assessment beliefs about F1DT. The findings also showed that urban and rural teachers were different in their views about assessment practices of F1DT due to their academic background and experience using F1DT.

All in all, it is important to keep in mind that the variability of responses should be interpreted as a sense of association, rather than as a cause-effect relationship. The predictors or variables of interest only serve as moderators of how teachers responded to the outcome variables.

8.4 Chapter Summary

This chapter has reported the findings in relation to the first four research questions. In the light of the above findings, it can be concluded that: (a) teachers generally had a positive attitude towards the implementation of F1DT; (b) the variability of teachers' responses to their perceived assessment beliefs and practices was largely contributed by their differences at the individual level; (c) there was a strong link between teachers' assessment beliefs and their practices concerning the implementation of F1DT; and (d) differences in teachers' assessment beliefs and practices were influenced by some factors related to individual and school characteristics.

The next chapter will report the findings of the post-intervention, focusing on the impact of the intervention strategy on teachers' reported assessment beliefs and practices concerning F1DT.

9 Results:The Effects of the Intervention on Teachers' Reported Assessment Beliefs and Practices

The phenomenon of ineffective utilisation of assessment in the Malaysian context, clearly indicated by the increasing number of low-performing schools, may be attributed to the use of unsuitable assessment tools in teaching and learning. Thus, it is timely and relevant for Malaysian teachers to look for a new assessment approach that could potentially remedy the drawbacks of traditional assessment tools. In an attempt to address this issue, I designed an intervention (as described in Chapter 6) aiming to familiarise teachers with the principles of DT as an alternative assessment approach with a potential to positively impact teaching and learning in schools. The rationale for this approach is based on the proposition that the currently used diagnostic test, i.e., F1DT, offers insufficient information about students' true ability to learn.

In this experimental study, the primary interest is to examine the potential effect of introducing DT as an alternative assessment tool on teachers' perceived assessment beliefs and practices concerning the purposes and uses of F1DT in making decisions about students' learning development. Insights to that effect can be gained by comparing the participants' responses to the questionnaire, i.e., SEA, before and after the intervention with those who were not exposed to the intervention. This chapter reports the results of the relevant analyses, specifically seeking to answer the following research questions:

5. Does the introduction of DT change teachers' reported beliefs about the purposes and uses of F1DT?
6. Does the introduction of DT change teachers' reported assessment practices regarding the use of F1DT?
7. To what extent does the introduction of DT affect the alignment between teachers' reported beliefs about the purposes and uses of F1DT and their reported assessment practices?
8. What are teachers' opinions with regard to the potential and barriers to implementing DT as an alternative tool in Malaysian schools?

9.1 Preliminary Analyses

As mentioned in Section 7.3.5.2, several preliminary analyses were conducted to (a) examine whether there are any pre-existing differences between the comparison groups in this experimental design and (b) check whether HLM is appropriate for this post-intervention dataset (n=381).

9.1.1 Sample analysis: Group differences

This analysis has a dual purpose to check potential differences between (a) teachers who remained in the study and teachers who withdrew from the study and (b) teachers in the intervention group and the control group. To examine this, two separate HLM analyses – HLM Model 5 (n=862) and HLM Model 6 (n=381) – were estimated.

The first analysis, HLM Model 5, was conducted for the main sample (n=862). By the end of the post-intervention, it was recorded that only 381 teachers fully participated in both phases, accounting for a 55.8% dropout rate. In order to establish the extent to which such a high attrition rate poses a threat to the generalisability of the findings (at least in terms of the teacher population in Sabah), the first step of the analysis focuses on difference (in measured attributes) between teachers who carried on with the study (n=381) and teachers who withdrew from the study (n=481).

HLM Model 5, as shown in the following equations, was carried out to estimate the outcome variables (*PERCEPTION, PRACTICE* and *NEGATIVITY*) as a function of the specified groups. Here, *G_PHASE1* (coded as 1 = carried on, 2 = dropped out) was not entered as a level-2 predictor as it refers to individual teachers who answered the questionnaire in the pre-intervention phase. They were not allocated to this group based on their respective schools.

HLM Model 5:

$$Level\text{-}1:\ PERCEPT1_{ij} = \beta_{0j} + \beta_{1j}*(G\_PHASE1_{ij}) + r_{ij} \quad (15)$$

$$Level\text{-}2:\ \beta_{0j} = \gamma_{00} + u_{0j} \quad (16)$$
$$\beta_{1j} = \gamma_{10}$$

Table 9.1 documents the result of HLM Model 5. The findings showed that *G_PHASE1*, referring to the two groups of individual teachers, did not correlate significantly to any of the outcome variables. In examining the variability of teachers' responses, a conclusion from previous findings about a large variance at the individual level was replicated in this analysis. The ICC estimates confirmed that level 2 variance had a considerably small contribution to the total variance, only accounting for less than 10%.

Table 9.1

HLM Output for Model 5 (n=862)

Outcome variable	**Final estimation of fixed effects (with robust standard errors)**						ICC
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value	
PERCEPTION	For G_PHASE1 slope, β_1						
	INTRCPT2, γ_{10}	−0.067118	0.04148	−1.618	840	0.106	.06
PRACTICE	For G_PHASE1 slope, β_1						
	INTRCPT2, γ_{10}	−0.106699	0.06574	−1.623	840	0.105	.03
NEGATIVITY	For G_PHASE1 slope, β_1						
	INTRCPT2, γ_{10}	−0.003269	0.06938	−0.047	840	0.962	.03

The above findings signal that there was no major difference between the two groups before the intervention. This means that there are no systematic differences between those who completed the study and those who dropped out, i.e., attrition is at random and not biased to certain characteristics. Thus, this allows me to continue with the analyses by comparing the intervention and control groups (*n*=381).

In the second analysis, HLM Model 6 was utilised to investigate whether there is an existing difference between the two comparison groups – the intervention group (*n*=166) and the control group (*n*=215). This analysis aimed to ascertain whether a potential difference between the two groups can be attributed to the intervention. In HLM Model 6, the outcome variables (*PERCEPT1, PRACTIC1* and *NEGATIV1)* were modelled as a function of *GROUP* (coded as 1= intervention and 2= control).

HLM Model 6:

$$Level\text{-}1:\ PERCEPT1_{ij} = \beta_{0j} + r_{ij} \quad (17)$$

$$Level\text{-}2:\ \beta_{0j} = \gamma_{00} + \gamma_{01}*(GROUP_j) + u_{0j} \quad (18)$$

Table 9.2

HLM Output for Model 6 (n=381)

Outcome variable	**Final estimation of fixed effects**						ICC
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value	
PERCEPTION	INTRCPT2, γ_{00}	3.355695	0.059240	56.646	11	<0.001	.07
	GROUP, γ_{01}	−0.045201	0.119748	−0.377	11	0.713	
PRACTICE	INTRCPT2, γ_{00}	3.564955	0.052171	68.332	11	<0.001	.04
	GROUP, γ_{01}	−0.009048	0.105186	−0.086	11	0.933	
NEGATIVITY	INTRCPT2, γ_{00}	2.357126	0.046668	50.509	11	<0.001	.00
	GROUP, γ_{01}	0.135062	0.093621	1.443	11	0.177	

The results of HLM Model 6 are summarised in Table 9.2. In this analysis, the results of the final estimation without robust standard errors are more appropriate for use as this dataset of 381 had a small number of level 2 units. From Table 9.2, it is noted that the grouping of teachers (*GROUP*) into the intervention group or the control group had no significant effect on their responses to *PERCEPTION* ($b = -.05$, $t = -.38$, $df = 11$, $p = .71$), *PRACTICE* ($b = -.01$, $t = -.09$, $df = 11$, $p = .93$) and *NEGATIVITY* ($b = .14$, $t = 1.44$, $df = 11$, $p = .18$). An inspection of the variability of the responses found that level l variance was greater than level 2 variance for *PERCEPTION* (93%) and *PRACTICE* (96%). It is, however, noted that the variance component at level 2 had essentially reached zero for *NEGATIVITY*, suggesting no variance of responses at the group level.

As there were no significant differences between the intervention group and the control group at the pre-intervention, this suggests comparability, which mitigates potential threats to the internal validity of the study. This is to say, it is reasonably valid to conclude that differences between the two groups at the post-intervention can be interpreted as a reflection of an intervention effect.

9.1.2 Suitability of HLM application

Following the procedures on group difference between several specified groups, the dataset of 381 was examined for its suitability for HLM analyses using the unconditional model, i.e., the null model. This analysis only included teachers' responses to the three outcome variables – *PERCEPTION*, *PRACTICE* and *NEGATIVITY* – in the post-intervention with no predictors.

HLM Null Model (n=381):

$$Level\text{-}1: PERCEPT_{ij} = \beta_{0j} + r_{ij} \quad (19)$$

$$Level\text{-}2: \beta_{0j} = \gamma_{00} + u_{0j} \quad (20)$$

Table 9.3 reports the results of the analysis. An assessment of the grouping effects in this dataset can be obtained from the final estimation of variance components. From Table 9.3, it is revealed that: *PERCEPTION*: χ^2 (12) = 64.71, $p < .001$; *PRACTICE*: χ^2 (12) = 63.63, $p < .001$; and *NEGATIVITY*: χ^2 (12) = 26.40, $p < .010$. This means the presence of variances at level 2 was statistically significant and this supports the use of HLM for this dataset. Essentially, this is further validated by variances at level 1, which were significantly different from zero ($\sigma^2 = .43$, $p < .001$; $\sigma^2 = .46$, $p < .001$; and $\sigma^2 = .76$, $p = .010$) for *PERCEPTION*, *PRACTICE* and *NEGATIVITY* respectively. This is another evidence that signifies the null model can be extended to several models to answer the specified research questions.

Table 9.3:

HLM Outputs for Null Models (n=381)

Outcome variable	**Final estimation of variance components**						ICC
	Random effect	Standard deviation	Variance component	*df*	χ^2	*p*-value	
PERCEPTION	INTRCPT1, u_0	0.18711	0.03501	12	64.71318	<0.001	
	level 1, r	0.65642	0.43088				.08
PRACTICE	INTRCPT1, u_0	0.19658	0.03865	12	63.63549	<0.001	
	level 1, r	0.68081	0.46350				.08
NEGATIVITY	INTRCPT1, u_0	0.12532	0.01571	12	26.40310	0.010	
	level 1, r	0.87108	0.75879				.02

To sum up, the above preliminary findings establish that (a) the attrition rate was at random, and (b) participating teachers allocated to the intervention group do not differ from those who were allocated to the control group. Also, the results of null models warranted the applicability of HLM analyses for this two-level data. In the following sections, I present the results of the analyses related to the remaining research questions.

9.2 Effects of Introducing Dynamic Testing on Teachers' Assessment Beliefs and Practices about the Form 1 Diagnostic Test

As reflected in Research Questions 5 and 6, it is the primary interest of this study to explore the effect of introducing the advantages of DT on Malaysian teachers' reported beliefs and practices concerning the existing assessment measure, i.e., F1DT. To investigate this, it is necessary to include teachers' responses at both phases in the statistical analyses. Specifically, two complementary statistical analyses –HLM analyses (HLM Model 7) and descriptive statistics – were deployed, of which the results are presented in the subsequent sections.

9.2.1 HLM analyses

To examine the potential impact of the intervention on the reported beliefs and practices about F1DT, the responses of 381 teachers, both before and after the intervention, were analysed using HLM analyses (HLM Model 7). This model was estimated to investigate whether the variance in teachers' responses to the outcome variables – *PERCEPTION*, *PRACTICE* and *NEGATIVITY* – was influenced by *TIME* (before and after intervention) and *GROUP* (the intervention group and the control group). This intercepts-and-slopes model represented a cross-level interaction where the level-1 predictor (*TIME*) and the level-2 predictor (*GROUP*) were modelled to explain the variability in the outcome variables. The equations of this model are exemplified as:

HLM Model 7:

$$Level\text{-}1:\ PERCEPT_{ij} = \beta_{0j} + \beta_{1j}*(TIME_{ij}) + r_{ij} \quad (21)$$

$$Level\text{-}2:\ \beta_{0j} = \gamma_{00} + \gamma_{01}*(GROUP_j) + u_{0j} \quad (22)$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}*(GROUP_j)$$

Table 9.4 summarises the output of HLM Model 7. An overall cross-level interaction, reflected in γ_{01}, between the predictors of interest and the outcome variables was not significant. For example, the variance of teachers across participating schools in *PERCEPTION* was insignificantly related to both level-1 and level-2 predictors (b = 0.08, t = 0.68, df = 11, p =.51). This means that group membership (the intervention group and the control group) had no influence on the strength of the relationship between *TIME* (pre- and post-intervention) and teachers' responses to items relating to their perception on the purposes of F1DT. The findings in the table below replicated the results in the sample difference (see HLM Model 6), implying that the grouping of teachers had no effect on the way they responded to the questionnaire.

Table 9.4:

HLM Output for Model 7 (n=381)

Outcome variables	**Final estimation of fixed effects**						**ICC**
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value	
PERCEPTION	For INTRCPT1, β_0						
	INTRCPT2, γ_{00}	3.275974	0.059903	54.688	11	<0.001	
	GROUP, γ_{01}	0.083254	0.121852	0.683	11	0.509	
	For TIME slope, β_1						
	INTRCPT2, γ_{10}	-0.176578	0.047180	-3.743	747	<0.001	
	GROUP, γ_{11}	0.230662	0.094650	2.437	747	0.015	.08
PRACTICE	For INTRCPT1, β_0						
	INTRCPT2, γ_{00}	3.461798	0.062125	55.723	11	<0.001	
	GROUP, γ_{01}	0.102537	0.126377	0.811	11	0.434	
	For TIME slope, β_1						
	INTRCPT2, γ_{10}	-0.216875	0.048788	-4.445	747	<0.001	
	GROUP, γ_{11}	0.211854	0.097875	2.165	747	0.031	.08
NEGATIVITY	For INTRCPT1, β_0						
	INTRCPT2, γ_{00}	2.392472	0.045633	52.429	11	<0.001	
	GROUP, γ_{01}	0.115292	0.091853	1.255	11	0.235	
	For TIME slope, β_1						
	INTRCPT2, γ_{10}	0.081257	0.063491	1.280	747	0.201	
	GROUP, γ_{11}	-0.018384	0.127371	-0.144	747	0.885	.02

A closer inspection of the individual slope of level-1 predictor, represented in γ_{11}, however, revealed different findings. Specifically, the results demonstrated that *TIME*

contributed significantly to the prediction of teachers' responses on *PERCEPTION* (b = .23, t = 2.4, df = 747, p = .02) and *PRACTICE* (b = .21, t = 2.2, df = 747, p = .03). This means, all teachers, regardless of whether being in the intervention group or in the control group, tended to be less positive about F1DT when asked to respond to the questionnaire the second time. In contrast, the result of this model estimated insignificant effect of *TIME* on *NEGATIVITY* (b = -.02, t = -0.1, df = 747, p = .89). The negative sign signifies a negative direction of effect, pointing out that teachers tended to have a higher agreement to all six negative statements about F1DT in the post-intervention.

In relation to the variance of responses across participating schools, small ICC indices for all outcome variables suggested that the variability was largely contributed by individual differences rather than group differences.

9.2.2 Descriptive analyses

As the above HLM results showed no clear indication of which group changed the most, descriptive statistics were computed to ascertain teachers' level of agreement with the three outcome variables – *PERCEPTION*, *PRACTICE* and *NEGATIVITY*. This aimed to find out whether there are any changes over time due to the fact that one group received the intervention but the other did not. To check for potential effects of the designed treatment, teachers' responses before and after the intervention were compared. The results of the analysis are shown in Table 9.5.

Table 9.5:

Comparison of Descriptive Statistics before and after Intervention (n=381)

Outcome variable	Comparison group		*M*		*SD*		Cohen's *d*
		N	Pre	Post	Pre	Post	
PERCEPTION	Intervention group	166	3.38	3.06	.67	.68	0.47
	Control group	215	3.31	3.22	.65	.68	0.14
PRACTICE	Intervention group	166	3.57	3.22	.62	.68	0.54
	Control group	215	3.54	3.40	.68	.75	0.20
NEGATIVITY	Intervention group	166	2.27	2.37	.91	.81	-0.12
	Control group	215	2.41	2.49	.90	.88	-0.09

As seen in the above table, both comparison groups – the intervention group and the control group – tended to change their answers to the statements in the questionnaire before and after the intervention. It is noted that despite their group membership, the

participants showed a similar pattern of responses before and after the intervention. The results indicated a decrease in the mean scores of two variables – *PERCEPTION* and *PRACTICE* – before and after the intervention. On the other hand, teachers seemed to agree more with the negative statements about F1DT at the post-intervention. These results corroborated the preceding findings in HLM Model 7.

Cohen's *d* was used to further investigate the observable effect of the intervention on teachers' reported beliefs and practices of the implementation of F1DT. Cohen's *d* is the most commonly used effect size when comparing two means to measure the magnitude of the experiment effect (Capraro, 2004; Fan, 2001; LeCroy & Krysik, 2007). Cohen (1988, 1992) suggested that 0.2 is a small effect, 0.5 is a medium effect, and 0.8 is a large effect.

The effect size of *PERCEPTION* (Cohen's $d = 0.47$) in the intervention group is considered to be relatively small. The effect size of *PRACTICE* (Cohen's $d = 0.54$) was found to exceed Cohen's (1988, 1992) convention for a medium effect (Cohen's $d = 0.50$), and this is a desirable effect that would likely to be "visible to the naked eyes of a careful observer" (Cohen, 1992, p.156). This is to say, teachers in the intervention group experienced a shift in perspective in the use of F1DT information for instructional planning following their exposure to the advantages of DT in better understanding the learning potentials of the student. Although the effect size of *NEGATIVITY* (Cohen's $d = -0.12$) was significantly small, the change in the intervention group from the pre-intervention ($M = 2.27$, $SD = .91$) to the post-intervention ($M = 2.37$, $SD = .81$) implies that teachers tended to agree more with the negative statements about F1DT, indicating positive feedback about the introduction of DT as an alternative assessment approach.

Results of Cohen's *d* for the control group showed a very small effect size for all the three outcome variables - *PERCEPTION* (Cohen's $d = 0.14$), *PRACTICE* (Cohen's $d = 0.2$) and *NEGATIVITY* (Cohen's $d = -0.09$). This means that differences in control group responses at two different points of time are only by chance, not deliberately affected by any extraneous variables.

Taken together, the above results from HLM Model 7 and descriptive statistics provide important insights about the differences of responses to the questionnaire before and after

the intervention. As established in the comparison of mean scores and Cohen's *d*, the greatest change was experienced by 166 teachers in the intervention group. Therefore, it can be summarised that the observed changes in teachers' reported assessment beliefs and practices concerning F1DT were attributed to their exposure to the introduction and demonstration of DT.

9.3 Effects of Introducing Dynamic Testing on the Alignment between Teachers' Assessment Beliefs and Practices about the Form 1 Diagnostic Test

To respond to Research Question 7, a Pearson product-moment correlation coefficient was estimated, analysing the impact of the intervention on the association between teachers' assessment beliefs and practices about F1DT. The objective of this analysis was to find out whether the alignment between beliefs and practices changes as a result of the intervention. Corresponding to the previous analyses on the belief–practice relationship, only items in *PERCEPTION* and *PRACTICE* were computed.

To assess the differences in the correlations between *PERCEPTION* and *PRACTICE* before and after the intervention, the data were split into two different samples – the intervention group and the control group. This was to determine which group showed a greater change in the correlation coefficients of *PERCEPTION* and *PRACTICE* at two different points in time. The results indicated that the correlations between teachers' perceptions and practices about F1DT for both groups at the pre-intervention were statistically significant and substantially strong (intervention group: $r = .775$, $p = .000$; control group: $r = .741$, $p = .000$). Similar observations were also recorded in the post-intervention, with $r = .732$ and $r = .742$ for the intervention group and the control group, respectively.

The differences between the two correlations (pre-and post-intervention) from two different samples were depicted in a line graph below. As demonstrated in Figure 9.1, significant and strong correlations were recorded for both groups before and after the intervention. Looking closely at the discrepancy in the two correlation coefficients for the intervention group (n=166), it is noted that the association between *PERCEPTION* and *PRACTICE* in the post-intervention ($r = .732$) was less strong when compared with the

recorded correlation coefficients in the pre-intervention (r = .775). This very small decrease (a difference of .043), however, suggests that there was no misalignment between assessment beliefs and practices among teachers in this group.

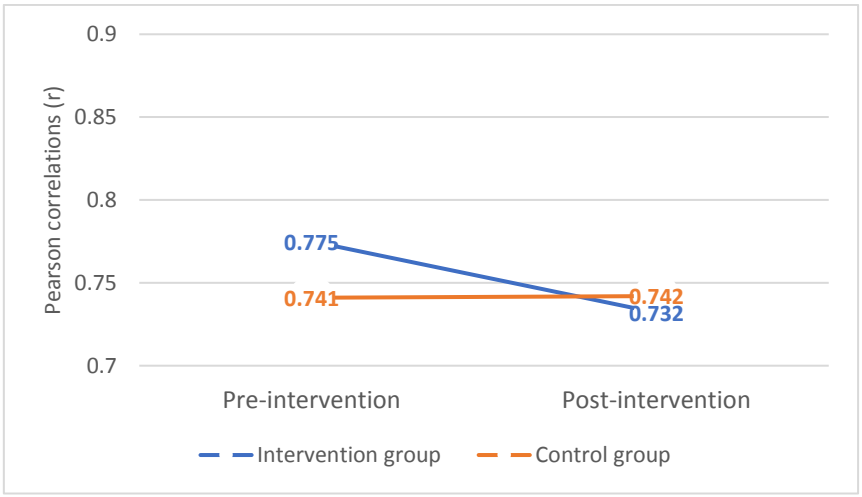


Figure 9.1: Comparison of Pearson Correlation Coefficients before and after the Intervention as an Indication of Potential Changes in the Alignment of *PERCEPTION* and *PRACTICE* (*n*=381)

An alternative way to analyse the potential change in the alignment between teachers' perception and practice regarding F1DT is to look at the changes in means in the respective scales over time (see Figure 9.2). Given that the line is parallel in both groups at the pre-and-post intervention, this confirms that there is no change in the alignment between *PERCEPTION* and *PRACTICE*.

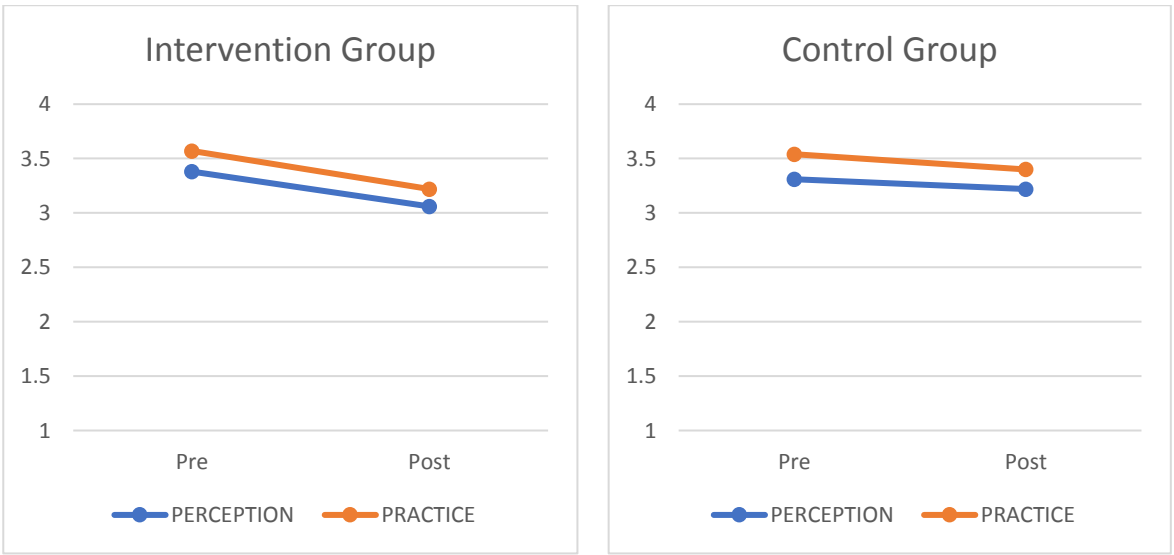


Figure 9.2: Comparison of Mean Scores before and after the Intervention as an Indication of Potential Change in Consistency between *PERCEPTION* and *PRACTICE* (*n*=381)

Overall, the above findings from Pearson correlation coefficients and the mean score comparisons concluded that the intervention, in the form of an educational workshop to introduce and demonstrate the potentials of DT, did not "disturbed" the alignment between teachers' beliefs and practices regarding F1DT.

9.4 Teachers' Reflective Feedback about Dynamic Testing

Research Question 8 seeks to study reflective feedback by teachers from the intervention group about issues related to the applicability of DT in Malaysian schools. To answer this question, the data were derived from two sources of data collection: teachers' responses to the questionnaire and their written feedback. It is important to mention that the data for the analysis were primarily responses of 166 teachers from the intervention groups who were exposed to the introduction of DT. In this section, the term Learning Test (LT) is used, specifically referring to an example of DT that was used in the intervention strategy. A separate analysis for the survey and written feedback was performed, in which the reports of the results are presented in the subsequent paragraphs.

9.4.1 Survey findings

I calculated descriptive statistics to examine the pattern and variability of how teachers in the intervention group (n=166) responded to the additional items concerning their general views about the potentials of LT in an educational setting. Table 9.6 shows the mean scores and standard deviations of the individual items.

Table 9.6:

Descriptive Statistics for DT-specific Items (n=166)

Items	*M*	*SD*
LT – new perspective	3.53	0.84
LT – useful for teachers	3.45	0.8
LT – supplementary tool	3.42	0.89
LT – easy application	3.36	0.81
LT – educational needs	3.35	0.84
LT – learning potential	3.28	0.96
LT – implementation in Malaysia	3.27	0.83
F1DT – Fairness of streaming	3.05	2.14
F1DT – redundancy with UPSR	2.83	1.07

In general, the participants of this study had a positive response to LT as an alternative assessment approach. This is reflected in the mean scores that ranged from 3.27 to 3.53 for the first seven items, relating to teachers' views about LT. Teachers generally agreed that the introduction and demonstration of LT had exposed them to a new perspective about educational assessments (M = 3.53, SD = 0.84). Additionally, teachers also viewed LT as a useful assessment tool (M = 3.45, SD = 0.80) that could be used to supplement the existing instruments (M = 3.42, SD = 0.89). The relatively small standard deviation of the seven items suggests considerable levels of agreement amongst respondents. That is to say, teachers were largely in agreement that LT has potentials for application in schools.

With respect to the two items about F1DT, the respondents thought it was appropriate to use F1DT for streaming students (M = 3.05, SD = 2.14)) but admitted that its results were redundant with the information from UPSR (M = 2.83, SD = 1.07)). This finding reflects the acknowledgement by teachers that this assessment tool is still relevant; this is possibly because F1DT is the only instrument currently available for use in Malaysian schools. However, the reported standard deviations of these items were high, suggesting a large variance in teachers' responses across the participating schools.

In addition to the above items, the survey also gathered information about the challenges of a possible implementation of LT in Malaysian schools. Teachers' answers can be viewed in Table 9.7.

Table 9.7

Challenges of Possible Implementation of LT in Malaysian Schools (n=166)

Challenges of LT implementation	Frequency	Percent
Teachers' confusion due to ever-changing assessment policies	132	81
Heavy workload	119	72
Insufficient time for teachers to familiarise with the new assessment tool	108	66
Teachers' frustration over too many changes in educational assessment	104	63
Teachers' resistance to change	66	40

The above table shows calculations of the frequency and percentage of the five listed challenges for LT implementation in Malaysia. From the table, teachers' confusion due

to many changes in assessment policies yielded the highest result (81%) while teachers' resistance to change reflected the lowest percentage (40%). In addition to the confusion over hasty changes of policies, more than half of the respondents regarded heavy workload (72%), insufficient time to understand an assessment tool (66%) and frustration over ever-changing assessment policies (63%) as obstacles to the practicality of LT for Malaysian students. This finding reflects the challenging issues faced by Malaysian teachers as previously discussed in Chapter 1.

Another important issue that is worth investigating is whether teachers in the intervention group differ significantly in their responses to these new items. To answer this question, HLM Model 8 was computed. In this model, *SLOT*, referring to five sessions at five participating schools, was estimated as a function of the variability of responses to the outcome variable (*LEARNTES*). The main effect of this level-2 predictor is depicted in the following equations:

HLM Model 8:

$$Level\text{-}1:\ LEARNTES_{ij} = \beta_{0j} + r_{ij} \quad (23)$$

$$Level\text{-}2:\ \beta_{0j} = \gamma_{00} + \gamma_{01}*(SLOT_j) + u_{0j} \quad (24)$$

Table 9.8:

HLM Output for Model 8 (n=166)

Outcome variables	**Final estimation of fixed effects**						ICC
	Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *df*	*p*-value	
LEARNTES	For INTRCPT1, β_0						
	INTRCPT2, γ_{00}	2.660000	0.542289	4.905	3	0.016	
	SLOT, γ_{01}	0.501429	0.383456	1.308	3	0.282	0

As shown in Table 9.8, the finding revealed that the main effect of level-2 predictor, i.e., *SLOT*, was not statistically significant to *LEARNTES* ($b = 0.50$, $t = 1.3$, $df = 3$, $p = 0.28$). The ICC index was zero, indicating no variance at level 2. This means that the variability of responses was primarily contributed by individual differences. Here, the clustering effect, i.e., teachers nested in five different schools, had no significant impact on the way teachers answered the items concerning their feedback about the use of DT in schools.

This finding implies that there was consensus among the participants, whose schools were located in different locations (urban and rural) and districts, that DT might be practical for use in Malaysian schools. This also can be seen as evidence that the delivery of the intervention across these schools was rather consistent.

9.4.2 Teachers' written feedback

The data gathered from teachers' written feedback were analysed thematically. As seen in Table 9.9, five themes emerged from the data. Examples of each theme are presented in the subsequent paragraphs. It is noted that teachers wrote about LT, referring to the short-term DT that I used for demonstration in the intervention strategy.

Table 9.9

Emerging Themes from Teachers' Written Feedback (n=17)

Themes	Number of comments
Favourable response to DT	8
Challenges to changes	4
Potential for DT implementation	2
Research-related feedback	2
Unfavourable response to DT	1

9.4.2.1 Favourable response to dynamic testing

The first theme shows positive responses from teachers about the idea of LT. Generally, they favourably viewed LT as a promising assessment tool to better understand students' learning potential. Some of the positive comments are:

> "The emphasis of LT on students' learning potential is an interesting concept that I'd love to apply in my classroom. I believe it is something that we tend to overlook as the current focus is more on students' achievement." (SU22KI)

> "This Learning Test is what teachers need to better measure students' potential. It offers meaningful insights for teachers to help developing students' learning potentials with appropriate pedagogical strategies." (NU10SE)

> "LT is more suitable to be used as a guide to identify students' potential so that teachers can design teaching and learning strategies to cater to students' needs." (RU07AB)

> "LT is not only useful for teachers but also allows students to identify and develop their potentials." (EL12AY)

This theme affirms Malaysian teachers' encouraging reaction to the appealing advantages of DT in the educational setting.

9.4.2.2 Challenges to change

The second theme is about challenges to change in the Malaysian assessment system including constant protest about heavy workload and many changes of assessment policies. Teachers complained:

> "LT appears to be a better way to measure students' potential and offers teachers a new perspective on assessments. However, as teachers spend more time on getting overloaded work done, they may take a longer time to understand and accept a new assessment tool." (MA21AB)

> "There are many types of assessments in the Malaysian education system and this appears to burden teachers." (RA20HA)

> "The ministry changes assessment policies many times. A new policy is introduced or the old policy is abolished in 'haste', not giving enough time for teachers to understand the current policy." (SI08EB)

> "To be honest, teachers easily accept changes in education. But, too many changes in a short time is really a nuisance, causing inconveniences to schools, teachers and students." (MA15RA)

Obviously, teachers were consistent in their views, as reflected in the survey findings in Table 9.13, that the abovementioned issues may become formidable barriers not only for the potential implementation of DT for Malaysian students but also for any assessment reforms in Malaysia.

9.4.2.3 Potential for dynamic testing implementation

In the third theme, teachers wrote about the possibility of using DT in Malaysian schools.

> "I fully support the implementation of LT for the sake of students' learning development, on condition that no extra burden is loaded to teachers' existing heavy workloads." (RE26RA)

> "LT can be potentially implemented in schools without burdening teachers and pressuring students." (HA09AH)

These two items of feedback further support teachers' positive impression about this new assessment tool.

9.4.2.4 Research-related feedback

The fourth theme is the comments written by two teachers, expressing their expectation of this study. They hoped that:

> "I hope that this study will provide meaningful insights that could contribute to the betterment of education in our country." (SH09AD)

> "Hopefully the outcome of your research can help teachers to improve the quality of teaching and learning." (KU05JU)

Here, teachers reiterated their support for any positive reforms for the betterment of education in the country.

9.4.2.5 Unfavourable response to dynamic testing

Lastly, there was only one feedback item commenting about the unsuitability of DT in Malaysia. The respondent wrote:

> "LT may not be suitable for use in Malaysian schools. This is because of the primary focus of Malaysian education policies, which place more emphasis on the products of subject-domain assessments. However, I appreciate the exposure to this assessment tool for knowledge sharing." (NO04YA)

The above comment suggests that the focus on a product-oriented assessment system is the reason why DT may not be suitable for implementation in Malaysian schools.

Altogether, it is apparent that the issues written by teachers provide supporting details to the preceding survey findings relating to Malaysian teachers' reflective feedback concerning this alternative assessment tool. In short, teachers reacted positively to LT, implying that they realised the promising advantages of this alternative tool to provide more meaningful information about students' potential to learn.

9.5 Chapter Summary

In this chapter, the results that answer the last four research questions have been presented. It is the primary objective of this study to investigate the potential effects of introducing dynamic testing on teachers' reported assessment beliefs and practices about F1DT. As reported in the preceding sections, the results of the post-intervention have revealed that: (a) teachers showed a significant change in their perceived assessment beliefs and practices about F1DT after exposure to the advantages of DT in an educational setting; (b) the introduction to DT had no effect on the alignment of the belief–practice relationship; and (c) teachers in the intervention group generally had a positive impression on the appealing potentials of DT as an alternative to the existing assessment tool.

A detailed interpretation of the findings, in relation to the specified research questions, will be further discussed in the next chapter.

10 Discussion

This study sets out to understand the potential reasons that contribute to the alleged underutilisation of assessment outcomes in Malaysian schools. Acknowledging a significant role of assessment in facilitating teaching and learning, this study started with the assumption that the phenomenon of interest could be attributed to two factors – teachers' lack of understanding of assessment information and the unsuitability of current assessment tools. The preceding chapters (Chapters 8 and 9) have reported the results of relevant analyses to answer the issues under investigation.

This chapter focuses on the discussion of the findings in response to two sets of research questions. The first cluster of research questions concerns the study of teachers' assessment beliefs about the purposes and uses of the existing assessment tool, i.e., F1DT, and the practices of its outcomes in planning pedagogical strategies for students. As this research is experimental, the second cluster of research questions aims to examine the potential effects of introducing an alternative assessment approach, i.e., DT, on teachers' reported assessment beliefs and practices concerning the use of F1DT in schools.

10.1 Understanding Teachers' Assessment Beliefs and Practices

Many researchers advocated that a better understanding of teachers' assessment beliefs and practices is instrumental to greater success in education and that this topic should be the primary focus of an inquiry by academicians (Borg, 2001; Deneen & Brown, 2016; Nespor, 1987). Essentially, it is an established notion that what teachers do in their education-related activities is said to be governed by what they believe, and these beliefs often serve as a filter through which instructional decisions are made (Barnes et al., 2015; Borg, 2001; Jussim & Harber, 2005; Nespor, 1987; Opre, 2015; Pajares, 1992; Skott, 2015). However, the notion of whether belief influences action or action influences belief is still open for debate. Poulson, Avramidis, Fox, Medwell, & Wray (2001) believed that the belief-practice nexus is complex and described it as "dialectical rather than unilateral" (p. 273). Acknowledging this bi-directional relationship is important and this will be the focus of the present study. Specifically, this study aims to understand what teachers think about assessment and how they make use of its information because they are the key players in transforming assessment information into improved teaching and learning

processes. To address this issue, the interpretations of the major findings to Research Questions 1–4 are discussed in the following sections.

10.1.1 Research Question 1: What are teachers' beliefs about the purposes and uses of F1DT?

The answer to Research Question 1 was gathered from a self-developed questionnaire, i.e., SEA, and teachers' elaborated feedback written on the last page of the questionnaire. Both approaches provided a better understanding of the status quo of teachers' assessment beliefs about the purposes and uses of F1DT.

An analysis of the means and standard deviations indicated that teachers generally perceived F1DT in a positive light. As reported at the item-level analysis, the majority of teachers were in agreement that it served its purported purpose as in the policy of many schools, i.e., mainly to stream students. This corroborates the view of teachers in Delandshere and Jones's (1999) study, who considered that assessment is more likely to be seen as serving the function of student placement and certification.

HLM analysis, i.e., the null model, produced an interesting finding on the variance of teachers' responses to their beliefs about F1DT. It was revealed that 94% of the variability in teachers' views was largely attributed to their individual differences. This is contrary to the assumption that teachers within the same school are likely to hold similar opinions about assessment-related matters. This finding, however, needs to be interpreted with caution, as there was a relatively small number of schools (n = 21) participating in this study.

The collective agreement about the purposes of F1DT in schools is further supported by elaborated feedback written by teachers, highlighting its value to measure students' pre-existing knowledge, to identify students' learning problems and to stream students. This is consistent with the findings from the survey. Some teachers, nonetheless, unfavourably viewed F1DT as irrelevant. They viewed the results of UPSR as more valid and reliable. This is probably because UPSR is rigorously designed by professionals and its quality is centrally monitored by the MES. Teachers also complained that F1DT scores did not give accurate descriptions concerning students' true learning potential. They were also in agreement with scholars (e.g., Beckmann, 2006, 2014; Elliott et al., 2018; Guthke &

Beckmann, 2000a; Haywood & Tzuriel, 2002; Hessels, 1997; Sternberg & Grigorenko, 2002) that an effective assessment tool should provide meaningful information that reflects the actual potential of a student. Essentially, teachers in the present study also believed that an assessment tool must bring a positive impact on students' learning. This is in line with Gardner's (2012) statement that "assessment of any kind should ultimately improve learning" (p.107).

Taken together, the majority of the study participants still believed that they benefitted from the implementation of F1DT. It was considered as an appropriate assessment tool that allowed them to gather information about Form 1 students before planning suitable next steps for teaching and learning.

10.1.2 Research Question 2: What are teachers' assessment practices regarding the use of F1DT?

Correspondingly, a positive attitude towards F1DT is also reflected in how the participants answered the items relating to the use of outcomes drawn from F1DT in the planning of pedagogical strategies catering to students' educational needs.

The data from the survey indicated a consistent finding with Research Question 1. In ranking the use of F1DT, there seemed to be a general agreement that F1DT was used mainly for student placement. Teachers also agreed that F1DT was inappropriate for use in predicting a child's future academic performance. This seems to be in line with the argument posited in this study, highlighting the shortcomings of this traditional test in providing a better prediction of what students can do in later stages of learning. With regard to the variability of the responses, HLM analysis also estimated a greater variance at the individual level, signalling insignificant differences across schools despite their location in a rural or urban area in six different educational zones.

Qualitative data from the written responses also provide information about teachers' practices to use the outcomes of F1DT in instructional decision making. A salient theme revolved around suggestions to improve the implementation of F1DT. In particular, teachers recommended the inclusion of other assessment approaches and subjects in providing a comprehensive description of students' learning potential. This implies that, like previous studies by Adams and Hsu (1998) and McMillan, Myran, and Workman

(2002), teachers acknowledged the necessity for a variety of assessment techniques to make informed decisions about students and instructions.

Overall, the positive attitudes of teachers about their reported F1DT practice further consolidated the predominant use of F1DT for streaming students according to their attainment level. Notably, the findings of this study seem to affirm the notion that the actions of teachers in their profession, including assessment-related activities, are a reflection of their beliefs (Barnes et al., 2015; Borg, 2001; Brown et al., 2008; Nespor, 1987; Opre, 2015; Pajares, 1992; Segers & Tillema, 2011; Stiggins, 2004; Urdan & Paris, 1994).

10.1.3 Research Question 3: To what extent do teachers' beliefs about the purposes and uses of F1DT align with their assessment practices?

While it is generally advocated that teachers' beliefs are influential in shaping their educational practices, studies attempting to examine the direct link between the two domains remain relatively under-researched (Jaba et al., 2013; James & Pedder, 2006; Postareff et al., 2012). Responding to such concern, Research Question 3 was formulated to specifically investigate whether Malaysian teachers' perceived usefulness of F1DT have a significant effect on their practice.

Previous research demonstrated that the notion that belief influences action is non-conclusive. Several empirical studies (e.g., Azis, 2015; Brown & Remesal, 2012; Büyükkarcı, 2014; Eren, 2010; James & Pedder, 2006) documented that there was a discrepancy between what teachers believed and what they reported practising.

The findings of this study, nevertheless, indicated notable insights that were different from the above. As reflected in the results of the first two research questions, a statistically significant relationship was detected between assessment beliefs and practices. The results of HLM Model 1a showed that teachers' assessment beliefs shape the way teachers responded to items asking about their practices of F1DT. These findings, to some extent, support the prior presumption that teachers' beliefs have a direct causal link on their professional practices. This is consistent with studies conducted by Malaysian researchers (e.g., Jaba et al., 2013; Md Omar & Sinnasamy, 2009; Varatharaj et al., 2015) as well as

researchers from other countries (e.g., Brown et al., 2009; Dixon et al., 2011; Postareff et al., 2012; Reimann & Sadler, 2017).

Nevertheless, results of HLM Model 1b provides compelling evidence that the "reverse" association also applies. It was revealed that teachers' positive responses to *PERCEPTION*-items were significantly influenced by their agreement to *PRACTICE*-items. This suggests that practice does not necessarily come after beliefs; sometimes, changes in beliefs may occur because of a change in practice. Notably, high values of *r* from both HLM Models signal a large effect of the relationship between the two variables. Correspondingly, a Pearson product-moment correlation coefficient also reported a strong and positive association between teachers' assessment beliefs and practices.

In conclusion, the findings of this study further confirm a strong link between beliefs and practices. In the context of the study, Malaysian teachers, particularly in Sabah, strongly endorsed the implementation of F1DT and tended to make extensive use of its information in making decisions about Form 1 students. Adversely, its widespread use in Malaysian schools for many years could have been the reason for their positive attitude towards this assessment tool.

10.1.4 Research Question 4: To what extent are individual characteristics and school variables associated with teachers' assessment beliefs and practices?

In the literature, a body of research has demonstrated that teachers' assessment beliefs and practices are largely moderated by the diversity of external factors (e.g., Alkharusi et al., 2012; Brown & Remesal, 2017; Calveric, 2010; Fong & Muhamad, 2017; Koloi-Keaikitse, 2012; Mehrgan et al., 2017; Mertler, 1998; Sach, 2012; Vidacovich, 2015). Acknowledging this issue, another purpose of this study is to explore the possible influences of teacher characteristics (e.g., years of teaching, educational background, position and familiarity with F1DT) and school characteristic (i.e., school location) on teachers' assessment beliefs and practices concerning F1DT.

In general, statistically significant correlations have been established between some of the investigated predictors and teachers' assessment beliefs and practices. Specifically, two individual characteristics (level-1 predictors), i.e., the positions of teachers in the school and their familiarity with F1DT, are found to impact and influence their

assessment beliefs and practices. In contrast, no significant impacts of years of teaching experience and educational background were noted. Similarly, an examination of the location of the school (level-2 predictor) showed that there was no significant difference between teachers in urban and rural schools regarding their reported beliefs and practices about the implementation of F1DT.

The above findings shared some consistent and inconsistent results with the previous research (see examples in section 3.2.2.4). This study, however, revealed some particularly unique findings that need to be highlighted. Firstly, the position of teachers in the school was found to be correlated with both assessment beliefs and assessment practices. This implies that a designated position that teachers hold in schools seemed to have an influence on the way they perceived the purposes of F1DT and the use of its outcome in instruction. It is argued that as teachers have different responsibilities, their perceptions and involvement and use of assessment information are more likely to differ (Adams & Hsu, 1998; Zhang & Burry-Stock, 2003).

In this study, it was noted that participants with administrative duties tended to have a relatively lower agreement on the implementation of F1DT. This finding is unexpected, as the decision on its use was determined and endorsed by school administrators. This could be the result of the small number of administrators (n = 113) participating in this study. The high level of agreement among ordinary teachers, on the other hand, was likely to be anticipated, as they are the main users of this assessment tool; they are actively involved in the process of test development, test administration and data analysis.

Secondly, the experience of teachers with F1DT was found to be insignificantly related to their reported assessment beliefs. On the other hand, familiarity with the test, i.e., exposure to the test and/or direct involvement in conducting and using the test information, seemed to have a significant impact on the use of the test results in making decisions about students' learning. The contrast between these findings is reasonably understandable. This shows that teachers may have a positive impression about any kind of assessment tool, knowing that assessment is instrumental in promoting effective teaching and learning. However, this is different when it concerns the actual use of a specific assessment tool. It is important for teachers to have adequate knowledge and skills in interpreting and translating assessment information to make informed and

accurate decisions about appropriate strategies to foster learning progress (Mertler, 2009; Popham, 2009; Stiggins, 1995, 2002; Stiggins & Duke, 2008; Volante & Fazio, 2007). Essentially, this result further supports the view that both knowledge and skills are directly related to what teachers do in the classroom, including assessment-related activities (Mertler & Campbell, 2005; Popham, 2009; Siegel & Wissehr, 2011; Stiggins, 1995).

Lastly, another interesting finding is concerned with the presence of cross-level interactions of some of the predictors on assessment beliefs and practices. Two significant interactions were observed. Firstly, it is noted that teachers in urban and rural schools were very different in their assessment beliefs due to their educational background. Teachers in urban schools with higher academic qualifications (n = 272) were less accepting of F1DT as a diagnostic measure in school. Adversely, holders of bachelor's (n = 521) and master's degrees (n = 56) in rural schools tended to be more accepting of the implementation and use of F1DT. These findings are somewhat puzzling but are potentially attributed to the relationship between teachers' academic background and their familiarity with F1DT. Presumably, teachers in urban schools with higher academic qualifications were not involved directly with the implementation and use of F1DT as the number of teachers with experience of F1DT (n = 138) and without experience (n = 138) was equal. In contrast, the participants from rural schools largely consisted of teachers with bachelor's and master's degrees, and the majority of them (n = 408) had direct experience using F1DT. Secondly, urban and rural teachers also demonstrated differences in assessment practice because of their experience in implementing F1DT. The variance between urban and rural schools was evident, particularly with teachers who had no direct involvement with F1DT. This suggests that the above explanation of the relationship between educational background and familiarity with F1DT seems to be true in explaining the disparity of responses between the participants from rural and urban schools.

All of the above scenarios reflect the complexity of assessment, which involves many contexts in which teachers work. The findings highlight that the degree of consistency (or inconsistency) between teachers' conceptions and practices is shaped, in part, by the various settings in which teachers work **(**Brown & Remesal, 2012; Nespor, 1987**;** Rubie-Davies, Flint, & McDonald, 2012) and the constraints imposed upon them (Chan &

Sidhu, 2006; Jaba et al., 2013; Harris & Brown, 2009; Poulson et al., 2001). Additionally, the above findings explain to some extent why the variability of teachers' responses on *PERCEPTION* and *PRACTICE* was largely due to individual differences. The variability of the responses, however, should be interpreted as a sense of prediction, not in a causal sense.

10.2 Effects of the Introduction of Dynamic Testing on Teachers' Reported Assessment Beliefs and Practices

To address the second argument about the unsuitability of existing assessment tools in providing accurate descriptions about students, this study aimed to introduce DT as an alternative to the current scenario where assessment seems to fail to meet expectations in facilitating effective teaching and learning. Proponents of DT (e.g., Beckmann, 2006, 2014; Elliott, 2000; Elliott et al., 2018; Guthke, 1992; Guthke & Beckmann, 2000a, 2000b; Haywood & Lidz, 2007; Hessels, 1997; Sternberg & Grigorenko, 2002) believed that DT has shown a remarkable contribution in providing teachers with far greater insights about a child's potential to learn, and that such information is helpful for teachers in pointing out the best ways to help students, especially the struggling ones, to optimise their learning potentials. In order to achieve the objective of the study, a rigorous plan to design an appropriate intervention was initiated (as explained in Chapter 6). The following sections outline a synthesis of the findings for Research Questions 5–8 in examining the effects of the introduction and demonstration of DT on teachers' reported assessment beliefs and practices.

10.2.1 Research Question 5: Does the introduction of DT change teachers' reported beliefs about the purposes and uses of F1DT?

Research Question 5 sought to examine the effect of introducing DT on the reported assessment beliefs of the teachers participating in the educational workshop. HLM analyses and descriptive statistics were examined to answer this question.

Results from HLM Model 7 confirmed the significant effect of *TIME* (i.e., before and after the intervention) in determining teachers' responses to their perceived usefulness of F1DT. Although the effect size was relatively small, it was still statistically significant.

It was also noted that, regardless of the grouping of teachers into intervention and control groups, the study participants tended to be less positive about F1DT.

Specifically comparing the responses of teachers in the intervention and control groups, the results revealed that the former demonstrated a greater change in their beliefs after the intervention than the latter. This means that teachers who participated in the educational workshop appeared to be less in agreement with the assessment-belief items (*PERCEPTION*) following the introduction of DT concept. The magnitude of the effect of the CPD was relatively small, but teachers became more aware of the promising advantages of DT to better understand the learning potential of their students. On the contrary, a very small effect size in the control group suggests that the differences in teachers' responses were random.

In short, it can be concluded that the observable changes in the intervention group were the result of the intervention that was designed to promote the benefits of DT in schools. This implies that the participants in the intervention group reacted favourably to the appealing potentials of DT in overcoming the deficit of traditional tests. This positive appraisal of DT is in line with previous studies (e.g., Bosma & Resing, 2010; Bosma et al., 2012; Deutsch & Reynolds, 2000; Touw et al., 2014) that showed a growing interest among educational practitioners about the use of DT in an educational setting.

10.2.2 Research Question 6: Does the introduction of DT change teachers' reported assessment practices regarding the use of F1DT?

With regard to Research Question 6, the results replicated a pattern similar to that shown in Research Question 5. The changes in responses to *PERCEPTION* also occurred in the way teachers responded to items in *PRACTICE*.

A closer look at the responses of the group of interest, i.e., teachers in the intervention group, showed that many items relating to the use of information from F1DT in instructional decision-making indicated a large decrease in the means after the intervention. The effect size of Cohen's $d = 0.54$ signalled a large effect of the experiential and educational components of the intervention on the impression of teachers towards the intended uses of F1DT. It was likely that teachers realised that it was inappropriate to use F1DT to make judgement about students (e.g., differentiation of low-performing students,

prediction of future academic performance and streaming). Clearly, teachers' lower level of agreement to the use of F1DT in making appropriate decisions about students signals their acknowledgement of the promising benefits of DT in an educational setting. In other words, teachers indeed recognised the pitfalls of F1DT in providing accurate insights to describe students' actual potentials to learn.

The uniformity of teachers' lower level of agreement to *PRACTICE* items in the post-intervention is an important indicator that the introduction and demonstration of DT affected the way they responded to the usefulness of F1DT information in making decisions about students' educational needs. More importantly, the impact of the intervention on teachers' reported assessment practices implies the importance of assessment training and support to equip teachers with relevant knowledge and skills on how to effectively use and interpret assessment-related data. This is in line with a strong recommendation by scholars (e.g., Care & Griffin, 2009; Chan & Sidhu, 2006; Md-Ali et al., 2015; Mertler, 2009; Mertler & Campbell, 2005; Popham, 2003, 2009; Siegel and Wissehr, 2011; Stiggins, 1991, 1995, 2002) that professional assessment literacy is mandatory for pre-service and in-service teachers in developing their ability to accurately interpret and transform assessment evidence to improve teaching and learning activities.

In brief, the answers to this research question provide further consolidation of the previous findings on *PERCEPTION* that the intervention had, in fact, influenced teachers' reported practices concerning the uses of F1DT information in making judgements and decisions about Form 1 students.

10.2.3 Research Question 7: To what extent does the introduction of DT affect the alignment between teachers' reported beliefs about the purposes and uses of F1DT and their reported assessment practices?

Another objective of this study is to check the effect of the introduction and demonstration of DT on the relationship between teachers' reported assessment beliefs and practices. The answer to Research Question 7 was derived from the results the Pearson product-moment correlation coefficient.

As reflected in Research Questions 5 and 6, it is revealed that there was a significant relationship between assessment beliefs and practices and the intervention. A closer

investigation into the responses of teachers in the post-intervention indicated a much lower agreement on *PERCEPTION* and *PRACTICE* by the intervention group than the control group. In order to verify this initial finding, it is also important to check whether the intervention had an impact on the strong alignment between assessment beliefs and practices as reported in the pre-intervention (see Research Question 3).

The results of the Pearson product-moment correlation coefficient and comparison of mean scores for *PERCEPTION* and *PRACTICE* revealed that the intervention had no significant effect on the alignment between assessment beliefs and practices. However, a comparison of correlation coefficients and mean scores at two different points in time showed a slight difference between the two groups. Looking closely at the intervention group, the association between assessment beliefs and practices tended to weaken in the post-intervention more than the control group. Although insignificant, this can be seen as a result of the exposure of teachers to the new assessment approach.

Again, this is a positive evidence that the experiential and educational components of the intervention provided teachers with a refreshing perspective about a more useful assessment tool that could be put into practice in Malaysian schools. Previous studies (e.g., Mertler, 2009; Mertler & Campbell, 2005; Sidhu et al., 2011; Veloo & Krishnasamy, 2017; Volante & Fazio, 2007; Zhang & Burry-Stock, 2003) demonstrated that assessment training is highly beneficial for teachers in guiding them to engage in better assessment activities. Collectively, the findings of this study and others suggest that adequate training on assessment literacy should be given more emphasis to ascertain that effective utilisation of assessment information in teaching and learning can be achieved.

10.2.4 Research Question 8: What are teachers' opinions with regard to the potential and barriers to implementing DT as an alternative tool in Malaysian schools?

Having introduced and demonstrated the potentials of DT, this study is also interested in eliciting reflective feedback from the group of interest, i.e., teachers in the intervention group, about their general views of this new assessment approach and relevant issues of the possibility of its implementation in Malaysian schools.

The data for this last research question were collected from two sources – the survey and teachers' written feedback. The overall response from the survey demonstrated a positive reception about this alternative assessment. Teachers largely agreed that the introduction and demonstration of DT had given them a new perspective about educational assessment. Additionally, they appeared to be optimistic that DT could be put into use in Malaysian schools as a supplement to the existing traditional measures. However, they asserted that there could be some barriers to its implementation, like any other educational reforms. As expressed by teachers in previous studies (e.g., Abdul Majid et al., 2011; Chan & Sidhu, 2006; Hashim et al., 2013; Jaba et al., 2013), confusion over many changes in assessment policies and heavy workloads were seen as the major hindrances to possible success in using DT by teachers and for students. To some extent, this finding affirms the arguments put forward in the introductory chapters about challenging issues in Malaysian education.

A similar pattern of results was derived from the thematic analysis of the written feedback. Two major themes emerged from the data: (a) favourable responses to DT and (b) challenges for change. It may be concluded that teachers perceived DT positively as a promising measure that could be utilised to better understand students' learning potential. Nonetheless, the issues of too many changes in education policies and an overload of clerical work were also echoed by teachers as formidable barriers for any reform taking place. Teachers in previous studies (see Abdul Majid et al., 2011; Chan & Sidhu, 2006; Hashim et al., 2013; Jaba et al., 2013; Mukundan & Khandehroo, 2010) also shared similar sentiments regarding challenges of putting assessment information into practice. This study indeed provides confirmation of the impression that teachers are hard-working and committed, but they are over-worked and underpaid. It seemed that these constant claims deserve immediate and systematic investigation by the MOE. If these reported issues remain unresolved, teachers may lose faith in the system and, consequently, the national agenda for world-class quality education may fail to achieve its objectives.

In short, the above findings are indicative of teachers' appreciation of the potential of DT in an educational setting, highlighting a consensus that DT could be a practical complement to or replacement of F1DT in providing meaningful information about Form 1 students. Clearly, teachers acknowledged that information gleaned from DT (as

demonstrated in the intervention) could be beneficial for them in designing appropriate strategies towards optimisation of students' learning potentials. To a greater extent, the encouraging feedback from the study participants may signal a "let-go" message for the implementation of F1DT as teachers can obtain the necessary information from a more valid and reliable assessment tool, i.e., UPSR.

10.3 Chapter Summary

This chapter has discussed a synthesis of the major findings in response to the eight research questions formulated in this study. In doing so, this chapter has been organised into two main sections.

The first section investigated the issue of alleged lack of understanding about assessment information. Based on the findings that emerged from Research Questions 1–4, it is concluded that (a) teachers are relatively content with what they do and what is expected of them in term of test use; (b) what teachers think about F1DT positively and strongly correlated with the way they used its information in instructional decision-making; (c) individual and school characteristics to some extent influenced teachers' assessment beliefs and practices; and (d) the variance in responses to assessment beliefs and practices was primarily contributed by individual differences.

In the second section, the issue of a lack of informative assessment tool was addressed. Specifically, this section synthesised the impacts of introducing an alternative assessment approach on teachers' perceived assessment beliefs and practices regarding the currently used diagnostic test. Major findings from Research Questions 5–8 can be summarised as follows: (a) teachers received DT positively and appeared to be optimistic that this assessment tool is practical and feasible for use in Malaysian schools; (b) teachers tend to demonstrate openness towards new ideas and seemed to be willing to engage in change that is expected to have positive impacts for their students; and (c) reported changes in assessment beliefs and practices showed the value of assessment training in supporting teachers to become more assessment literate.

The next chapter will extend this discussion chapter by reflecting on the implications of the findings and recommending directions for future research.

11 Conclusion

This study seeks to respond to the alleged underutilisation of assessment information in facilitating effective teaching and learning in Malaysian schools. Acknowledging the significant role of assessment in instruction, this study highlights two assessment-related factors as the potential reasons for the above phenomenon: (i) teachers' lack of understanding of assessment and (ii) lack of an informative assessment tool. In understanding the issues of interest, this study utilises a two-group pre-test and post-test experimental design to answer the formulated eight research questions.

This chapter focuses on an extended discussion of the emerging findings as outlined in Chapter 10. It highlights the contribution of the study, looking specifically at its implication for academic research and professional practices. After considering the limitations of this research, the study looks forward to the future by recommending several ideas that could be hugely productive and significant for the next research projects. This chapter ends with a reflective conclusion that goes back to the main arguments about two potential factors that have been repeatedly mentioned throughout this thesis.

11.1 Contribution of the Study

This section discusses the contributions that this study makes to the body of knowledge, particularly around the literature of assessment beliefs and practices and the application of DT. This is followed by the contributions of the study findings to several issues in professional practices.

11.1.1 Contribution to the knowledge in the field of study

The current study, on its own, provides several new contributions to the body of knowledge.

First, it offers new insights into the literature of assessment beliefs and practices and the relationship between the two in the current context. This topic is relatively new in Malaysia. It has only come to the fore after the implementation of assessment reform, i.e., the introduction of the SBA. It should be noted, however, that all previous studies

primarily centred on teachers' views and acceptance of the implementation of the SBA (e.g., Abdul Majid et al., 2011; Jaba et al., 2013; Md Omar & Sinnasamy, 2009) and only a few focus on teachers' knowledge of the SBA (e.g., Majid, 2011; Sekharan Nair et al., 2014). Apparently, none of them investigated how beliefs may or may not influence teachers' practices concerning educational assessment. Thus, the novelty of this study lies in its attempt to explore what teachers think and do with an existing assessment tool, i.e., F1DT, that has had roots in Malaysian schools for decades. To date, after an exhaustive effort, it appears that no study ever attempted or published about this traditional diagnostic test. As such, it is proper to say that this thesis is to be treated as the pioneering empirical endeavour of its kind.

Second, this study is a response to the call by Malaysian researchers (e.g., Ismail et al., 2014; Putih et al., 2016; Yusup, 2012) for critical studies of assessment-related elements that may have a potential impact on Malaysian students' academic performance. In contrast to many studies (e.g., Dzulkifli & Alias, 2012; Hanafi, 2008; Lei & Mei, 2015; Saw, 2016) that emphasise on student and contextual factors in determining academic success, this study focuses on the research assessment-related element, which is insofar a rarity. This is a very important factor that has an influential impact on the teaching and learning process, which in turn can contribute to improvement in students' academic performance. The findings in this study provide a different lens through which to observe that teachers' understanding of assessment information and the use of more useful assessment tools are instrumental in enhancing instructional activities. It is, therefore, necessary for all stakeholders like educationists, researchers and education agencies to further explore the significant contribution of assessment-related aspects in understanding the underlying reasons for academic underperformance among Malaysian students.

Third, this study is distinctively original in terms of the rigorous design of the intervention consisting of experiential and educational components. The objective was to introduce and demonstrate the appealing potentials of DT as an alternative to traditional assessment measures. Besides, this study reveals an encouraging result in which the majority of teachers in the intervention group acknowledged the information provided by DT as meaningful and relevant for instructional planning. To some extent, this finding enriches the literature on teachers' appreciation of the practical values of DT in an educational setting. To date, only a few studies (e.g., Bosma & Resing, 2010; Bosma et al., 2012;

Deutsch & Reynolds, 2000; Touw et al., 2014) examine teachers' views on DT. Previously, the main focus of DT researchers was more on the investigation of the validity of DT and its application to students. Furthermore, teachers' appraisal of DT refutes the claims that it has failed to take root in mainstream practices. In this respect, the findings in this study provide encouraging evidence that DT is feasible for use in devising more focused classroom-based interventions, particularly for students in low-performing schools. Moving forward, the findings of this study can also offer a refreshing perspective about the possibilities for DT use in Southeast Asia and beyond since unlike in Europe and the US, empirical works on DT are, to a large extent, still uncommon in Asia.

Fourth, in relation to the above, another salient finding of this study is the change of beliefs among teachers. In the pre-intervention, teachers had a high level of agreement that F1DT was suited for the purposes and uses as mentioned in the questionnaire. The responses to the same statements, however, changed in the post-intervention. Teachers in the intervention group tended to agree less with the purposes and uses of F1DT after the introduction of DT. This is contrary to Guskey's (1986) linear model, proposing "a significant change in teachers' beliefs and attitudes is likely to take place only *after* [emphasis added] changes in student learning outcomes are evidenced" (p.7). In other words, Guskey suggests that changes in beliefs are likely to occur primarily after the implementation of a new assessment and evidence of students' improvement in learning. In the context of the present study, changes in teachers' beliefs happened even before the real implementation of DT. It appears that teachers' reflection about the differences between DT and F1DT had led to the changes in their beliefs. The findings of this study support the ideas of Cooney (1999) who argued that teachers can change their beliefs after they become more reflective about their instructional practices including teaching and assessment activities. While most of the literature centres on the investigation of beliefs and practices and the association between the two, this study contributes a significant novelty in the body of knowledge regarding the implication of professional development programmes (to introduce a new assessment tool) in changing teachers' reported assessment beliefs and practices way in advance before the real implementation of the tool occurs.

Fifth, this study makes a significant contribution in terms of the substance and methodology in the study of assessment beliefs and practices and also to the field of DT

through its unique design of an intervention that incorporated the experiential and educational elements. This design is notably different from the previous studies of assessment beliefs and practices, which primarily focused on a cross-sectional study that involves collecting and looking at data at one specific point in time. Hitherto, the application of a time-and-energy-consuming repeated measurement approach is rare in this field. Perhaps, Mertler (2009) who runs a pre-test and post-test on his respondents in examining their perceptions about professional assessment training appears to be the only exception. In the field of DT, this study shares the common deployment of pre- and post-test experimental design. Nonetheless, the designed intervention of the previous studies has typically aimed at researching the feasibility and/or effectiveness of a specific DT measure on individual students. This study, on the other hand, attempted to measure the impact of the introduction of DT concept on teachers' reactions to the purposes and uses of the currently-used assessment tool.

Sixth, this study also sheds new light on the methodological aspect of academic research when dealing with a representative sample that is clustered in multilevel structures. The use of increasingly promising HLM applications makes it possible to capture the variability of responses at different levels (i.e., individual and group levels) and over time (i.e., before and after intervention). Previously, the systematicity of variability of responses tended to be analysed with less attention given to the nested structures on the outcome variables, and this, from a statistical point of view, violates the assumption of independence of observation, a fundamental requirement for analysing group differences (Hofmann, 1997; Luke, 2004; Osborne, 2000; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; Warne et al., 2012). Interestingly, this study has demonstrated the advantages of HLM in examining the presence of statistically significant cross-level interactions between the studied predictors and the outcome variables – assessment beliefs and practices.

Collectively, this study makes significant contributions to academic research by enriching the content and methodological aspects of the field of assessment beliefs and practices and the application of DT.

11.1.2 Contribution to professional practices

The finding of this study reinforces the argument that teachers' beliefs are instrumental to the understanding of their practices in assessment-related activities. It further supports the established notion about the powerful role of beliefs in influencing teachers' enactment of their educational practices (Brown et al., 2015; Jussim & Harber, 2005; Mansour, 2013; Opre, 2015; Pajares, 1992; Segers & Tillema, 2011). This study also shows that belief does not necessarily affect one's actions, it can be that practices may – over time – also shape teachers' beliefs. Acknowledging this bi-directional relationship, this study indeed makes a substantial contribution to professional practices as follow.

First, in order to ensure that teachers enact any assessment reforms in a meaningful way, effective introduction of change should address both aspects, i.e., beliefs and practices, in parallel. If DT is to be implemented, for example, an introductory workshop needs to be considered to understand teachers' existing beliefs and knowledge of assessment. Such information is essential for the education agencies to devise strategies in helping teachers to make sense of the rationales for assessment reforms. This is crucial since previous studies have demonstrated that when there was reform, the policy and implementation did not align very well (see Albert Jonglai, 2017; John, 2018; Yan & Cheng, 2015). Hence, Gardner et al. (2008) suggested the need for a continuous support and training programme, including teachers' networking and professional learning collaboration, which emphasises "a change in understanding rather than merely a superficial change in teaching techniques" (p.9). Such professional learning was exemplified in this study, where the two components of the intervention served as a practical strategy towards creating a more reflective understanding of DT as an alternative assessment tool. This is important to convince teachers of the rationales for its application in a school setting. Arguably, teachers need to have a clear idea about why DT is relevant for use in a situation where the existing measure is not helpful in bringing the intended positive impact of assessment on the teaching and learning process. This means that teachers' understanding of the change is crucial because the success of the assessment activities depends greatly on how they perceive and interpret the reform.

Second, the findings of the study also offer practical contributions to strategise the implementation of assessment training for teachers. As the data suggest, it seems that there was uniformity among teachers regarding their assessment beliefs and practices.

Despite greater individual differences, teachers from different schools across six educational zones tended to think and behave uniformly in the implementation of F1DT and also in their reaction to the introduction of DT. A possible explanation for this could be partially attributed to the centralised education system in Malaysia, in which any education-related matters are managed by the MOE through the top-down approach. The MOE and the state education departments, in particular, could take advantage of this uniformity to organise a regional or divisional (according to educational zones) professional training for assessment literacy or to develop assessment reforms in accordance with the needs of the local context. This collegial interaction is proven to be highly effective for improvement in teaching and learning (Care & Griffin, 2009; Harlen, 2005; Majid, 2011; Hayward, 2015). This platform, through networking across schools within educational zones, gives teachers opportunities to meet, discuss and reflect upon their assessment activities and later identify and use mutually agreed interventions and teaching strategies that can scaffold students' learning progress. Additionally, taking the local context into consideration, this collaborative networking encourages teachers to have a shared vision and gives them greater participation in decisions about their professional development within their respective educational zones. Moreover, it also offers valuable insights to district education offices when planning for relevant interventions to remediate the assessment-related problems faced by schools in their jurisdiction. Essentially, this study implies that proper management of training should be given much emphasis to ascertain teachers' deep understanding of assessment so that the effective execution of its information can be warranted.

Third, the study also highlights the advantages of experiential training and reflection as core components of professional development training. This suggests that training should engage teachers in both active and interactive learning – where teachers become actively involved in activities such as problem-solving, discussion, simulations and application (Garet, Porter, Desimone, Birman & Yoon, 2001; Hunzicker, 2011). As exemplified in the experiential component of the intervention strategy, teachers were shown samples of DT items and features and a demonstration of its application in a real school setting. This practical session enables teachers to digest and understand the new assessment approach and help them decide how best to adapt it to their teaching context. Additionally, effective professional development should be reflective in the sense it enables teachers to learn from their experience in a conscious and systematic manner (Korthagen, 2017). In the

study, teachers were able to review the functions and uses of F1DT as a long-standing diagnostic measure in Malaysian schools and to reflect upon the benefits and relevance of DT in the current situation. Hence, this study recommends educational authorities to place more importance on experiential and reflective aspects when designing and implementing professional development so that its activities can bring positive impacts on teachers' knowledge and skills and improvement in their educational practices.

Fourth, and most importantly, teachers' positive feedback about DT is a convincing signal for the necessity to adopt an alternative assessment measure that offers a more informative description of students' actual potential to learn. If such potential is understood accurately, teachers will be able to improve the quality of necessary interventions and consequently, the learning potentials of students can be optimised. Furthermore, the finding of this study may facilitate the referral agencies particularly Sabah Education Department and district education offices to consider adopting a new assessment approach that is more helpful to remedy the situation of the evident increase of low-performing schools in Sabah. Yet, prior to introducing a new assessment tool, the challenges to assessment reforms should not be underestimated. Reports from previous studies (e.g., Abdul Majid et al., 2011; Albert Jonglai, 2017; Jaba et al., 2013; John, 2018; Yan & Cheng, 2015) revealed that issues such as misunderstanding the rationales for reform, resistance to changes, inadequate resources and insufficient time for training are often associated with the ineffective implementation of assessment reforms. Therefore, anticipating such problems in advance allows educational authorities to formulate appropriate strategies to ensure the desired outcomes – i.e., optimisation of students' learning potential and improvement in teaching and learning – can be achieved.

Overall, this study foregrounds the importance of teacher assessment training and recommends several strategies for its effective implementation. The introduction of DT as a supplementary to the existing assessment measures is also proposed.

11.2 Limitations and Recommendations for Future Studies

No study is completely flawless and this research is no different. Nevertheless, I reflected that the limitations of this study are not "flaws", but opportunities that could lay the groundwork for more productive research in the future. While acknowledging some of its

limitations, this section puts more focus on how to turn the identified research constraints into potential avenues for academic inquiries that other researchers could explore.

First, as the introductory chapter has highlighted, relatively little prior work has paid attention to this issue when seeking to understand the underlying reasons that contribute to students' underperformance. In the context of Malaysia, the focal points of the previous investigation primarily focus on the contextual factors of the learner, the school, the family and health-related issues as the potential explanations to the problem of unsatisfactory academic performance (e.g., Dzulkifli & Alias, 2012; Ishak et al., 2012; Lei & Mei, 2015; Othman et al., 2008; Saw, 2016). The finding of this study reiterates the argument that an investigation of teachers' assessment beliefs and practices is a necessity in an educational setting as teachers are the key players for the success of assessment activities (Deneen & Brown, 2016; Nespor, 1987; Opre, 2015; Stiggins, 1995; Travis, 1996). Therefore, continued research in assessment-related topics should be attempted. Not only this is a valuable avenue to better understand the impact of teachers' beliefs on their assessment practices but also to offer insightful ideas to the education agencies and teachers themselves to take appropriate actions to improve assessment activities. This is to ensure Malaysia's aspiration for world-class quality education can be warranted.

Second, in relation to the above, the review of the literature indicates that assessment-related topics in Malaysia appear to centre on the implementation of the SBA and tend to overlook the practicality of other existing assessment tools. This study is the first empirical study conducted on F1DT. Despite its use for many years in Malaysian schools, it receives no attention for investigation from academic researchers. How valid and reliable is F1DT in providing accurate descriptions about a student's real ability to learn? How different is information drawn from F1DT and UPSR? Why do schools still need F1DT? While not included in this thesis, these are interesting research questions that might prove to be significant aspects for future studies. Seeking the answers to these questions would be helpful for schools to review the contribution and relevance of F1DT in facilitating teaching and learning processes.

Third, the result of this study demonstrates an encouraging reception of the application of DT from teachers. Nonetheless, a one-time workshop introducing DT is not conclusive

enough to say that all Malaysian teachers positively approve of this assessment approach. Due to time and logistical constraints, the finding of this study could only be used as a representation of the target population, i.e., teachers in Sabah, not as a generalisation for the other Malaysian teachers in 13 states. Therefore, more empirical studies on the nationwide application of DT are necessary to further support the initial findings of this work. Looking forward, comparative studies on the views of DT implementation and its potential to remedy the limitations of traditional tests could be explored. For example, researchers may consider examining the similarities and differences in views about the potentials of DT for under-achievers and high-achievers. It is also interesting to compare the functions and uses of F1DT and DT in extracting information that is beneficial for enhancing teaching and learning. Additional studies on DT in the Malaysian context can also be extended to empirically validate the feasibility of this alternative assessment tool to better understand students' learning potential. While DT was administered to Form 1 students, the data obtained was not analysed and included in this thesis due to the shortcomings of the technical stability of the dynamic version of the test. Analysis of such data would complement the findings of the study by revealing the appealing advantages of DT to provide meaningful information, i.e., indication of learning potential, that could be greatly helpful for teachers in designing appropriate interventions. Essentially, ongoing research is needed to provide robust evidence about the validity and reliability of DT measures to promote its widespread application in educational settings (Beckmann, 2006, 2014; Caffrey et al., 2008; de Beer, 2006; Lauchlan & Elliot, 2001; Lidz, 2014; Sternberg & Grigorenko, 2002).

Finally, teachers' positive attitude to F1DT and encouraging feedback about DT application can fairly be interpreted as a representation of the target population, i.e., teachers in low-performing schools. As teachers from high-performing schools were not represented in this study, their views should be explored in future research, as an attempt to generalise the conclusions drawn from this study. Methodologically, this study obtained data about teachers' assessment practices using the questionnaire and the validity of their responses are yet to be verified. Hence, a structured observation could be very useful to validate their reported assessment practices because it entails a thorough description of the behaviour of individuals that occurs "live" in natural contexts (Bryman, 2012; Cohen et al., 2011). In this way, researchers can directly observe what is taking

place on the site rather than relying on second-hand accounts of what teachers reporting what they have done.

As highlighted above, this study has paved the way for further research in many ways. The limitations of this study offer some refreshing ideas that could potentially be explored in future works related to the issues at hand.

11.3 Concluding Thoughts

As set out in the introductory chapter of this thesis, the primary objective of this study is to better understand the phenomenon of poor academic performance among Malaysian students. Acknowledging the seemingly significant relationship between assessment and instruction, the focus of the study lies in the question: Why do assessment outcomes "fail" to bring about the expected positive changes in instructional strategies and pedagogies to foster students' academic progress? In addressing this crucial question, this explorative study puts forward two key assessment-related elements – the user (i.e., the teacher) and the tool (i.e., the test). Specifically, it examines whether teachers' lack of understanding of assessment information or the unsuitability of the currently used assessment tool (i.e., F1DT) or a combination of both, could contribute to the alleged underutilisation of assessment information in improving teaching and learning.

Having presented and discussed the findings of the study in Chapters 8, 9 and 10, here are the main conclusions. The first major finding suggests that teachers seemed to have a good understanding of the purposes and uses of the assessment tool concerned (i.e., F1DT). Overall, the participants of the study perceived F1DT favourably and used its information mainly for the measurement of prior knowledge, student placement and identification of students' learning problems as purported in the test document. This "loyal" sentiment towards F1DT is certainly understandable, as it has been the only diagnostic measure available for use for many years. Essentially, this study concludes that teachers' attitude and behaviour towards the implementation of FIDT are seen as indirect indicators of their understanding of the intended purposes and uses of the test. Reflecting upon the second major finding, however, it appears that teachers' good understanding of the purposes and uses of assessment information is non-conclusive.

The second salient conclusion of this study indicates teachers' encouraging responses to the introduction of DT as an alternative assessment approach in an educational setting. Notwithstanding the sentiment that FIDT is still relevant and useful, teachers positively appreciated the potentials of DT to extract meaningful information about students' ability to learn. This positive feedback implies that teachers were more reflective of the limitations of F1DT to provide noteworthy insights to describe students' actual potentials to learn. In other words, the shortcoming of F1DT is likely the potential reason that contributes to teachers failing to utilise and transform its information in devising appropriate interventions that are instructionally focused to improve teaching and subsequently to scaffold student's learning development. This is to say, relative changes in teachers' reported beliefs and practices about F1DT signal a necessity for assessment reforms to adopt an alternative assessment approach. Most importantly, this study exudes the optimism that DT can be a promising approach to complement the shortcomings of conventional tests, particularly F1DT. As evident in the previous studies, DT has shown remarkable predictive validity as those of the traditional tests, and uniquely, it has produced consistent incremental validity that goes beyond the information provided by the traditional tests.

On a final note, it is hoped that this study has addressed the above arguments in a meaningful and systematic manner. It is important to emphasise that the aim of this study is not to find fault of the existing practice. Rather, it is an exploratory attempt to better understand the current situation and to propose an alternative measure that can be put to good use to reach appropriate and meaningful decisions about students, instruction and the curriculum as a whole. Most importantly, all parties – teachers, school administrators, researchers, policymakers and even the public – should work hand in hand to offer compromises towards the betterment of education in Malaysia.

> "It's the action, not the fruit of the action, that's important. You have to do the right thing. It may not be in your power, may not be in your time, that there'll be any fruit. But that doesn't mean you stop doing the right thing. You may never know what results come from your action. But if you do nothing, there will be no result."
>
> Mahatma Gandhi (1869–1948)

Appendices

Appendix 1A: Examples of FIDT and UPSR Test Items

FIDT

UJIAN DIAGNOSTIK MATEMATIK TINGKATAN 1
SMK PAYA RUMPUT
MELAKA

Nama : _____

Jawab **semua** soalan yang berikut . Untuk soalan berbentuk objektif, tulis pilihan jawapan pada ruangan yang disediakan

1 Dua puluh dua ribu dua belas' ditulis dalam angka ialah
Twenty-two thousand and twelve' written in numerals is
A 2 212
B 20 212
C 22 012
D 22 120 ☐

2 4 836 – 3 160 + 1 031 =
A 645 767 C 1 ☐
B 765 D 2 707

3 Darabkan RM13 597 dengan 14.
Multiply RM13 597 by 14.
A RM169 512
B RM175 468
C RM188 506
D RM190 358 ☐

4 9 buah kalkulator yang sama berharga RM440.10. Hitung harga bagi 15 buah kalkulator.
9 similar calculators cost RM440.10. Calculate the cost of 15 calculators.
A RM733.50
B RM745.70
C RM758.90 ☐

1

UPSR

SULIT 2 015/1

1 Antara berikut, yang manakah akan menjadi 600 apabila dibundarkan kepada ratus yang terdekat?
Which of the following will be 600 when rounded off to the nearest hundred?
A 546
B 556
C 655
D 665

2 Rajah 1 menunjukkan empat kad nombor.
Diagram 1 shows four number cards.

40 128	52 720	X	20 100

Rajah 1
Diagram 1

Jumlah nombor pada empat kad itu ialah 150 000.
Apakah nilai X?
The total number on the four cards is 150 000.
What is the value of X?
A 37 052
B 42 948
C 48 162
D 57 152

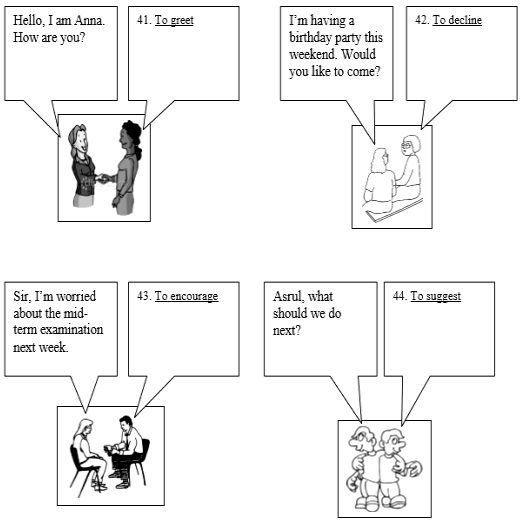
Section C : Forms and Function

[20 marks]

[*Time suggested : 30 minutes*]

Questions 41 - 50

Write the correct responses in the spaces provided to fit the situations.



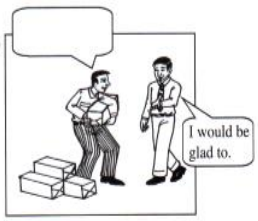
SULIT 2 013

Section B
[30 *marks*]

Question 21

Write a suitable response for each picture in the space provided.
Tulis jawapan yang sesuai bagi setiap gambar di ruang yang disediakan.

(*a*)

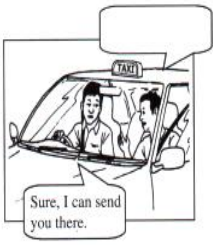


Answer

_____
_____
_____

[2 *marks*]

(*b*)



Answer

_____
_____
_____

[2 *marks*]

Appendix 4A: List Items of the SEA I

PERCEPTION	PRACTICE
Usefulness/ Improvement for Teaching and Learning	
(13 ITEMS)	(9 ITEMS)
Q10 I think assessment is beneficial to students	
Q11 I think assessment helps students to improve their learning	Q53 I use assessment to help students to improve their learning
Q13 I think assessment can identify students' strengths and weaknesses	Q62 I use assessment to identify students' strengths and weaknesses
Q15 I think assessment can motivate students to learn	Q43 I use assessment to motivate students to learn
Q16 I think assessment is an enjoyable experience for students	
Q17 I think assessment results are accurate to measure students' actual mastery level	
Q20 I think assessment results reflect students' actual learning ability/potential	
Q23 I think assessment information can accommodate students' need individually	Q44 I use assessment information to accommodate students' needs individually
Q26 I think assessment is integrated in teaching practices	Q49 I integrate assessments in my teaching practices
Q27 I think assessment is useful to plan strategies for teaching and learning	Q50 I use assessment to plan strategies for teaching and learning
Q31 I think assessment is a way to determine how much students have learned from teaching	Q54 I use assessment to know how much students have learned from teaching
Q36 I think assessment results predict student's future academic performance accurately	Q65 I use assessment results to predict students' future academic performance
Q30 I think assessment forces teachers to focus more on students' learning progress rather than their scores	Q58 I focus more on students' learning progress rather than their scores
Accountability	
(5 ITEMS)	(4 ITEMS)
Q24 I think teachers determine the quality of assessment outcomes	
Q33 I think assessment helps students to obtain good results in public examinations	Q45 I use assessment to prepare students for public examinations
Q34 I think learning is geared towards performance assurance for public examinations	Q61 I teach my students examination-taking techniques
	Q46 The priority of my work is to help students to pass public examinations
Q39 I think assessment is a good way to evaluate teachers	Q47 I am assessed for appraisal based on students' performance in public examinations
Q42 I think assessment outcome is an accurate indicator of a school's quality	

Irrelevance of Assessment	
(9 ITEMS)	(7 ITEMS)
Q12 I think assessment is fair to under-performing students	
Q14 I think assessment benefits high-achieving students	
Q18 I think assessment results are collected and filed but ignored	Q66 I conduct assessments but make little use of the result
Q28 I think assessment controls the way teachers teach	Q55 Assessment controls the way I teach
Q29 I think teaching for test/ examination preparation limits teachers' creativity to plan pedagogical strategies	Q52 I have plenty of time to design my pedagogical strategies creatively
Q32 I think assessment forces teachers to teach in a way against their beliefs	Q60 Assessment forces me to teach against my beliefs
Q35 I think teaching for test is detrimental to students' learning	
Q37 I think assessment forces teachers to focus more on students' scores rather than their learning progress	Q51 I focus more on students' scores rather than their learning progress
Q38 I think assessment is a waste of time	Q59 I put little effort to use assessment information in my instructional decision making
	Q57 I devote more time preparing and completing materials for assessment rather than designing pedagogical strategies
Assessment Literacy	
(6 ITEMS)	(4 ITEMS
Q19 I think accurate interpretation of assessment information is essential	Q63 I can interpret assessment information accurately
Q21 I think assessment outcome is difficult to understand	Q48 I cannot understand assessment outcomes easily
Q22 I think assessment results should be treated cautiously	Q56 I treat assessment results cautiously
Q25 I think assessment knowledge and skills are important for teachers	Q64 I need further training to interpret assessment information effectively
Q40 I think assessment literacy programme is essential to promote ongoing professional development for in-service teachers	
Q41 I think exposure for assessment literacy training in university / teacher training institution is not adequate	

Appendix 4B: Letter of Ethics Approval from the School of Education Ethics Committee (Pilot Study 1)

Durham University
School of Education

Shaped by the past, creating the future

8 July 2015

Rusilah Binti Yusup
PhD in Education

rusilah.yusup@durham.ac.uk

Dear Rusilah

Exploring the Potential of Dynamic Testing: A Study of under-performing students in Malaysia

I am pleased to inform you that your application for ethical approval for the above research has been approved by the School of Education Ethics Committee. May we take this opportunity to wish you good luck with your research.

Dr. J. Beckmann
Chair of School of Education Ethics Committee

Leazes Road
Durham, DH1 1TA
Telephone +44 (0)191 334 2000 Fax +44 (0)191 334 8311
www.durham.ac.uk/education

Appendix 4C: Permission to Use Questionnaire Items (email correspondences)

YR

YUSUP R.
Fri 19/06/2015, 15:36
gt.brown@auckland.ac.nz; BECKMANN J.F.; BECKMANN N.

Reply

Dear Dr Brown,

I am a first year PhD student at Durham University UK, under the supervision of Dr Jens F. Beckmann and Dr Nadin Beckmann. I am still in the early stage of my doctoral dissertation, initially titled "Exploring the Potential of Dynamic Testing: A Study of Under-performing Students in Malaysia". One of the issues addressed in my study is about teachers' perception of educational assessment and how it influences their instructional decision making.

I am writing to ask permission to use your questionnaire on Teacher Conception of Assessment (TCoA) and Teacher Practices of Assessment Inventory (PrAI). I would like to adapt your questionnaire items in my research study and I would be grateful if you could send me the full version of both inventories.

Thank you for considering my request. If you agree, I will properly cite the source in my work. Please let me know if you'd like me to follow any special instructions for acknowledging the materials.

Looking forward to your response.

Sincerely,
Rusilah Yusup
School of Education
Durham University, UK
rusilah.yusup@durham.ac.uk

GB

Gavin Brown <gt.brown@auckland.ac.nz>
Fri 19/06/2015, 15:55
YUSUP R.; BECKMANN J.F.; BECKMANN N.

Reply

C-TAPrI.docx 23 KB
CoA-IIIshort.doc 116 KB
CoA-III long inventory.d... 152 KB

3 attachments (290 KB) Download all Save all to OneDrive - Durham University

HI Yusup—thank you for asking
Yes you may use the two inventories in your research
The citation for the two inventories is as follows:
Yu, W. M., Chan, J. K. S., Fok, P. K., Kennedy, K. J., & Brown, G. T. L. (2008). *Teacher Assessment Practices* (T-APrI) *Inventory*. Unpublished test. The Hong Kong Institute of Education. Hong Kong, PRC.
Brown, G. T. L. (2001-2003). *Teachers' conceptions of assessment (TCoA) inventory* (Versions 1-3). Unpublished test. Auckland, NZ: University of Auckland.
Please note that a study done in India used items from both inventories and produced a different result to that reported in 2009
This might be important for your research if you are planning to collect data in Malaysia
Brown, G. T. L., Chaudhry, H., & Dhamija, R. (2015). The impact of an assessment policy upon teachers' self-reported assessment beliefs and practices: A quasi-experimental study of Indian teachers in private schools. *International Journal of Educational Research, 71*, 50-64. doi: 10.1016/j.ijer.2015.03.001
Best wishes in your studies

Gavin T L Brown, PhD
Director Quantitative Data Analysis and Research Unit (Quant-DARE)
School of Learning, Development & Professional Practice
Faculty of Education & Social Work, The University of Auckland
Private Bag 92019, Auckland, 1142, New Zealand
Tel: +649 623 8899 ext. 48602; Fax: +649 623 8827
http://www.education.auckland.ac.nz/uoa/gavin-brown
New available: Brown, G. T. L., Chaudhry, H., & Dhamija, R. (2015). The impact of an assessment policy upon teachers' self-reported assessment beliefs and practices: A quasi-experimental study of Indian teachers in private schools. *International Journal of Educational Research, 71*, 50-64. doi: 10.1016/j.ijer.2015.03.001

Appendix 4D: Participant Information Sheet and Consent Form

Durham University
School of Education

Shaped by the past, creating the future

31st July 2015

Participant Information Sheet

Title: **RESEARCH ON EDUCATIONAL ASSESSMENT**

You are invited to take part in a research project of "A Study of Teachers' Perception and Use of Educational Assessment: Exploring the Potentials of Dynamic Testing in Malaysian Schools". Please read this form carefully and ask any questions you may have before agreeing to be in the study.

The study is conducted as part of my PhD studies at Durham University. This research project is supervised by Professor Jens F. Beckmann (j.beckmann@durham.ac.uk) and Dr. Nadin Beckmann (nadin.beckmann@durham.ac.uk) from the School of Education at Durham University.

The purposes of this study are twofold: i) to assess teachers' perception and practice of educational assessment and ii) to explore the effects of introducing an alternative approach to assessment (the introduction of DT).

If you agree to participate in this study, you will be asked to answer several statements about (i) general demographic information and (ii) perception about Form 1 Diagnostic Test and its practices.

Your participation in this study will take approximately 20 minutes.

You are free to decide whether or not to participate. If you decide to participate, you are free to withdraw at any time without any negative consequences for you.

All your responses or other data collected will be anonymous and kept secured. All files containing any information will be stored on a password protected device. In any research report that may be published, no information will be included that will make it possible to identify you and your school individually. There will be no way to connect your name to your responses at any time during or after the study.

If you have any questions, requests or concerns regarding this research, please contact me via email at rusilah.yusup@durham.ac.uk.

This study has been reviewed and approved by the School of Education Ethics Sub-Committee at Durham University (date of approval: 08/07/2015)

RUSILAH BINTI YUSUP

Leazes Road
Durham City, DH1 1TA
Durham University, UK

Shaped by the past, creating the future

Durham University
School of Education

Declaration of Informed Consent

- I agree to participate in this study, the purpose of which are to assess teachers' perception and practice of educational assessment and to explore the effects of introducing an alternative approach to assessment (the introduction of DT).
- I have read the participant information sheet and understand the information provided.
- I have been informed that I may decline to answer any questions or withdraw from the study without penalty of any kind.
- I have been informed that all of my responses will be kept confidential and secured, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. Rusilah Yusup from the School of Education, Durham University can be contacted via email: rusilah.yusup@dur.ac.uk.
- I will be provided with a copy of this form for my records.

Any concerns about this study should be addressed to the School of Education Ethics Sub-Committee, Durham University via email to ed.ethics@durham.ac.uk.

_____

Date Participant Signature

Appendix 4E: Letter of Ethics Approval from the School of Education Ethics Committee (Pilot Study 2)

Durham University
School of Education

Shaped by the past, creating the future

3 December 2015

Rusilah Yusup
PhD

rusilah.yusup@durham.ac.uk

Dear Rusilah

STUDY OF TEACHERS' PERCEPTION AND USE OF EDUCATIONAL ASSESSMENT: EXPLORING THE POTENTIALS OF DYNAMIC TESTING IN MALAYSIAN SCHOOLS

I am pleased to inform you that your application for ethical approval for the above research has been approved by the School of Education Ethics Committee. May we take this opportunity to wish you good luck with your research.

P. M. Holmes

Dr. P. Holmes
Chair of School of Education Ethics Committee

Leazes Road
Durham, DH1 1TA
Telephone +44 (0)191 334 2000 Fax +44 (0)191 334 8311
www.durham.ac.uk/education

Appendix 4F: Examples of Questionnaire Items for SEA II

- Examples of Items for PERCEPTION

Pada pandangan saya, UDT1…….
In my opinion, the UDT1 …….

…. memberikan guru-guru maklumat yang lebih tepat berbanding UPSR *... provides teachers with more accurate information than UPSR*	**Sangat tidak setuju** 1-----2-----3-----4-----5 **Sangat setuju**
… mengukur pengetahuan sedia ada pelajar dengan tepat *... measures students' pre-existing knowledge accurately*	**Sangat tidak setuju** 1-----2-----3-----4-----5 **Sangat setuju**
… mengukur kebolehan semulajadi pelajar *... measures student's innate abilities*	**Sangat tidak setuju** 1-----2-----3-----4-----5 **Sangat setuju**
… memberikan maklumat yang mencerminkan potensi sebenar pembelajaran pelajar *... provides informative results reflecting students' actual learning potential*	**Sangat tidak setuju** 1-----2-----3-----4-----5 **Sangat setuju**

- Examples of Items for PRACTICE

Saya menggunakan UDT1….
I am using the UDT1…..

… untuk membuat penilaian mengenai pengetahuan sedia ada pelajar *... to make judgements about students' pre-existing knowledge*	**Sangat tidak setuju** 1-----2-----3-----4-----5 **Sangat setuju**
… untuk membuat penilaian mengenai kebolehan semulajadi pelajar *... to make judgements about students' innate abilities*	**Sangat tidak setuju** 1-----2-----3-----4-----5 **Sangat setuju**
… untuk membuat penilaian mengenai potensi pembelajaran pelajar *... to make judgements about students' learning potential*	**Sangat tidak setuju** 1-----2-----3-----4-----5 **Sangat setuju**
….. untuk mengetahui kekuatan dan kelemahan pelajar dalam pembelajaran *... to learn about students' strengths and weaknesses in learning*	**Sangat tidak setuju** 1-----2-----3-----4-----5 **Sangat setuju**

Appendix 4G: Research Instrument – Survey of Educational Assessment (SEA)

Kaji Selidik Ujian Diagnostik Tingkatan 1 (UDT1)

Rakan-rakan seperjuangan,

Saya ingin mempelawa tuan/puan untuk mengambil bahagian dalam kajian saya yang berfokus mengenai ujian diagnostik yang dilaksanakan di sekolah. Saya berminat untuk menyelidik pendapat tuan/puan mengenai **Ujian Diagnostik Tingkatan 1 (UDTI)** dan penggunaannya dalam amalan pengajaran. Tiada jawapan yang betul atau salah untuk pernyataan-pernyataan yang berikut. Semua maklumat yang diperolehi adalah sulit dan dirahsiakan.

Dengan mengembalikan soal selidik ini, tuan/puan bersetuju untuk mengambil bahagian dalam kajian ini secara sukarela. Sekiranya tuan/puan mempunyai sebarang pertanyaan mengenai kajian ini, sila hubungi saya melalui email rusilah.yusup@durham.ac.uk.

Kerjasama tuan/puan amatlah dihargai. Terima kasih.

****

Dear colleagues,

I would like to invite you to participate in my study which centres on the current diagnostic test that we use in schools. I am interested in your honest opinion about what you think and what you do with the ***Form 1 Diagnostic Test (F1DT)****. There are no right or wrong answers to the subsequent statements. All your responses will be kept strictly private, anonymous and confidential.*

By submitting this questionnaire, you agree to give your consent and to voluntarily participate in this research project. If you have any questions, requests or concerns regarding this research, please do not hesitate to contact me via email at rusilah.yusup@durham.ac.uk*.*

Thank you very much for your time and willingness to participate.

……..
RUSILAH BINTI YUSUP
School of Education
Durham University, United Kingdom.

Definisi kata kunci:
Menurut Akta Pendidikan 1996 (1998)

Pelajar bermasalah pembelajaran: Pelajar bermasalah kognitif yang dianggap boleh diajar dan boleh mendapat manfaat pendidikan formal. Ini merangkumi kategori Sindrom Down, Autisme ringan, Attention Deficit Hyperaktif Disorder (ADHD), terencat akal minimum, bermasalah pembelajaran spesifik (disleksia) dan lembam.

Program Pendidikan Khas: Terdapat 3 kategori program khas yang disediakan untuk pelajar berkeperluan khas – 1) sekolah pendidikan khas, 2) program pendidikan khas integrasi dan 3) program inklusif.

Definition of keywords:
According to Education Act 1996 (1998)

Student with learning difficulties (LD): *Pupils with cognitive problem- who are educable and could benefit from formal education. This includes children with Down syndrome, mild autism, attention deficit/ hyperactivity disorders (AD/HD), mild mental retardation, specific learning difficulties (such as dyslexia) and slow learners.*

Special Education Programme: *There are three special education programmes that are implemented for children with special needs- (1) the special school, (2) the integration programme and (3) the inclusive programme.*

MAKLUMAT DEMOGRAFI

DEMOGRAPHIC INFORMATION

Untuk memastikan kerahsiaan jawapan anda, sila gunakan sistem kod berikut.
*To ensure **anonymity of your responses**, please use the following **code system** that prevents tracing back your responses to your name.*

KOD/ *CODE*: _____

Contoh/ *Example*: **JA05YU**

JA: 2 huruf pertama nama ibu
First two letters of mother's name

05: hari lahir anda
Day of birth

YU: 2 huruf pertama nama bapa
First two letters of father's name

Sila nyatakan maklumat demografi yang berikut:
Please provide the following demographic information:

A) Jantina / *Gender*
□ Lelaki/ Male
□ Perempuan/ Female

B) Tahap Tertinggi Pendidikan
Highest Education Level

_____

C) Tahun Pengalaman Mengajar
Years of Teaching Experience

_____

D) Jawatan / *Position*

_____

E) Pengalaman dengan UDT1
Familiarity with the F1DT
□ Ada / *Yes*
□ Tiada / *No*

KAJI SELIDIK UJIAN DIAGNOSTIK TINGKATAN 1 (UDT1)

QUESTIONNAIRE ON FORM 1 DIAGNOSTIC TEST (F1DT)

SEKSYEN 1 / *SECTION 1*

Saya ingin mengetahui bagaimana anda menggunakan **UDT1** dan dapatannya dalam amalan pengajaran anda. Tiada jawapan yang betul atau salah untuk pernyataaan-pernyataan berikut. Sekiranya anda tidak menggunakannya pada masa ini, sila berikan jawapan sejujur yang mungkin.
Sila beri respon anda untuk pernyataan-pernyataan berikut dengan menandakan (X) pada mana-mana bahagian di garisan antara **sangat tidak setuju** (di sebelah kiri) dan **sangat setuju** (di sebelah kanan) yang mewakili pendapat peribadi anda.

*I would like to know how you use the **F1DT** and its result in your educational practice. There are no correct or incorrect answers. Even if you are currently not using it, please provide your responses as honest as possible based of your previous experience with the **F1DT**.*
Please respond to the following statements by marking (X) *at any point on the line between "**strongly disagree**" (on the left) and "**strongly agree**" (on the right) that best represents you personal opinion.*

Contoh:
Example:

Saya suka pekerjaan saya sebagai guru *I love my job as a teacher*	Sangat tidak setuju I----------X----------I Sangat setuju

Saya menggunakan UDT1….
I am using the F1DT…..

…untuk membuat penilaian mengenai pengetahuan sedia ada pelajar *…to make judgements about students' pre-existing knowledge*	Sangat tidak setuju I----------I Sangat setuju
…untuk menempatkan pelajar mengikut tahap pencapaian semasa mereka *…to stream students according to their current attainment level*	Sangat tidak setuju I----------I Sangat setuju
…untuk membuat penilaian mengenai kebolehan semulajadi pelajar *…to make judgements about students' innate abilities*	Sangat tidak setuju I----------I Sangat setuju
…untuk membuat penilaian mengenai potensi pembelajaran pelajar *…to make judgements about students' learning potential*	Sangat tidak setuju I----------I Sangat setuju
…untuk mengetahui kekuatan dan kelemahan pelajar dalam pembelajaran *…to learn about students' strengths and weaknesses in learning*	Sangat tidak setuju I----------I Sangat setuju
… untuk membuat perbandingan kalangan pelajar *…to make comparisons among students*	Sangat tidak setuju I----------I Sangat setuju

…untuk meramal pencapaian akademik masa hadapan pelajar *…to predict students' future academic performance*	**Sangat tidak setuju** I----------I **Sangat setuju**
…untuk membuat keputusan mengenai strategi pedagogi saya *…to inform decisions regarding my pedagogical strategies*	**Sangat tidak setuju** I----------I **Sangat setuju**
…untuk mengubahsuai kaedah pengajaran saya sesuai dengan keperluan pembelajaran pelajar *…to tailor my teaching according to students' learning needs*	**Sangat tidak setuju** I----------I **Sangat setuju**
…untuk mengetahui sejauh mana pelajar memerlukan bantuan dalam pembelajaran *…to find out how much help students need to improve their learning*	**Sangat tidak setuju** I----------I **Sangat setuju**
…untuk mengenalpasti pelajar yang bermasalah pembelajaran *…to identify students with learning difficulties (LD)*	**Sangat tidak setuju** I----------I **Sangat setuju**
…untuk merancang intervensi yang sesuai untuk pelajar yang bermasalah pembelajaran *…to plan appropriate interventions for students with learning disabilities (LD)*	**Sangat tidak setuju** I----------I **Sangat setuju**
…untuk membuat perbandingan kalangan pelajar berprestasi rendah *…to differentiate among low-performing students*	**Sangat tidak setuju** I----------I **Sangat setuju**
….untuk mengetahui keberkesanan pengajaran saya *…to know how effective my teaching is*	**Sangat tidak setuju** I----------I **Sangat setuju**
….kerana saya disuruh berbuat demikian *…because I am told to do so*	**Sangat tidak setuju** I----------I **Sangat setuju**
…namun kurang memanfaatkan keputusannya *…but make little use of the results*	**Sangat tidak setuju** I----------I **Sangat setuju**
…namun lebih suka menggunakan keputusan UPSR *…but prefer to use the UPSR results*	**Sangat tidak setuju** I----------I **Sangat setuju**

SEKSYEN 2 / *SECTION 2*

Sekarang saya ingin mengetahui pandangan jujur anda mengenai kegunaannya UDT1. Tiada jawapan yang betul atau salah untuk pernyataaan-pernyataan berikut.
Sila beri respon anda untuk pernyataan-pernyataan berikut dengan menandakan (X) pada mana-mana bahagian di garisan antara **sangat tidak setuju** (di sebelah kiri) dan **sangat setuju** (di sebelah kanan) yang mewakili pendapat peribadi anda.
Now I am interested in your honest view about the usefulness of the F1DT. Again, there are no correct or incorrect answers.
Please respond to the following statements by marking (X) *at any point on the line between "**strongly disagree**" (on the left) and "**strongly agree**" (on the right) that best represents your personal opinion.*

Pada pendapat saya, UDT1…….
In my opinion, the F1DT …….

…. memberikan guru-guru maklumat yang lebih tepat berbanding UPSR *… provides teachers with more accurate information than UPSR*	**Sangat tidak setuju** I----------I **Sangat setuju**
…membolehkan sekolah untuk merujuk pelajar bermasalah pembelajaran ke Program Pendidikan Khas *…allows school to refer students with learning difficulties (LD) to a Special Education Programme*	**Sangat tidak setuju** I----------I **Sangat setuju**
…mencerminkan kualiti pengajaran guru *…reflects teacher's quality of teaching*	**Sangat tidak setuju** I----------I **Sangat setuju**
...dilaksanakan untuk memenuhi arahan pihak atasan *… is carried out to fulfil the administrative directives*	**Sangat tidak setuju** I----------I **Sangat setuju**
…membantu guru-guru untuk mengubahsuai kaedah pengajaran mengikut keperluan pembelajaran pelajar *…helps teachers to adjust instructions according to students' educational needs*	**Sangat tidak setuju** I----------I **Sangat setuju**
…mengukur pengetahuan sedia ada pelajar dengan tepat *…measures students' pre-existing knowledge accurately*	**Sangat tidak setuju** I----------I **Sangat setuju**
…mengukur kebolehan semulajadi pelajar *…measures student's innate abilities*	**Sangat tidak setuju** I----------I **Sangat setuju**
…memberikan maklumat yang mencerminkan potensi sebenar pembelajaran pelajar *…provides informative results reflecting students' actual learning potential*	**Sangat tidak setuju** I----------I **Sangat setuju**
…membantu guru-guru untuk membuat perbandingan kalangan pelajar berprestasi rendah *…helps teachers to differentiate among low-performing students*	**Sangat tidak setuju** I----------I **Sangat setuju**

…keputusannya dikumpulkan namun tidak dimanfaatkan *…results are collected but ignored*	**Sangat tidak setuju** I----------I **Sangat setuju**
….memberikan guru-guru maklumat berguna mengenai sejauh mana pelajar memerlukan bantuan dalam pembelajaran *…provides teachers with useful information about how much help students need to improve their learning*	**Sangat tidak setuju** I----------I **Sangat setuju**
...membolehkan guru-guru meramal dengan tepat pencapaian akademik masa hadapan pelajar *…enables teachers to accurately predict students' future academic performance*	**Sangat tidak setuju** I----------I **Sangat setuju**
…memberikan guru- guru maklumat untuk mengenalpasti pelajar yang bermasalah pembelajaran *…provides teachers insights to identify students with learning difficulties (LD)*	**Sangat tidak setuju** I----------I **Sangat setuju**
…memberikan guru- guru maklumat yang mencukupi mengenai kekuatan dan kelemahan pelajar dalam pembelajaran *…provides teachers with adequate information of students' strengths and weaknesses in learning*	**Sangat tidak setuju** I----------I **Sangat setuju**
…membantu guru-guru untuk penempatan pelajar mengikut tahap pencapaian semasa mereka *…helps teachers to stream students according to their current attainment level*	**Sangat tidak setuju** I----------I **Sangat setuju**
…membolehkan guru-guru untuk membuat perbandingan kalangan pelajar *…enables teachers to make comparisons among students*	**Sangat tidak setuju** I----------I **Sangat setuju**
…membolehkan guru-guru untuk memperbaiki strategi pedagogi *…allows teachers to enhance pedagogical strategies*	**Sangat tidak setuju** I----------I **Sangat setuju**
…membantu guru-guru untuk merancang intervensi yang sesuai untuk pelajar yang bermasalah pembelajaran *…assists teachers to plan appropriate intervention for students with learning difficulties (LD)*	**Sangat tidak setuju** I----------I **Sangat setuju**
…adalah sesuatu yang membazir masa *…is a waste of time*	**Sangat tidak setuju** I----------I **Sangat setuju**

SEKSYEN 3 / *SECTION 3* (Additional items for Pre-intervention only)

Dalam seksyen ini, saya ingin mengetahui pandangan anda mengenai ciri-ciri sesuatu alat pentaksiran yang **ideal**. Tiada jawapan yang betul atau salah untuk pernyataan-pernyataan di bawah. Sila beri respon anda dengan menyenaraikan **LIMA** ciri-ciri terpenting untuk sesuatu alat pentaksiran yang ideal – **1** untuk yang paling penting dan **5** untuk yang kurang penting.

Sila tulis 1 hingga 5 di dalam kotak di bawah yang mewakili ciri yang terpenting hingga ciri yang kurang penting.

*In the last section, I would like to learn more about your personal opinion about the characteristics of an **ideal** assessment tool. Again, there are no correct or incorrect answers. Please respond to the following statements by ranking **FIVE** important characteristics of an ideal assessment tool – i.e., **1** for the most important and **5** for the least important.*

Please write 1 to 5 in the box below to represent the most to the least important criteria.

Pada pendapat saya, alat pentaksiran yang ideal ialah…
In my opinion, an ideal assessment tool is ….

	…berguna kepada guru-guru untuk mengenalpasti pelajar yang bermasalah pembelajaran *…useful for teachers to identify students with learning difficulties*
	…adil kepada pelajar-pelajar dalam menunjukkan potensi sebenar mereka untuk belajar *…fair to students in reflecting their actual potential to learn*
	…dapatan yang berguna untuk memahami keperluan pembelajaran pelajar *…meaningful outcomes to understand students' educational needs*
	…mudah untuk guru-guru untuk memahami aplikasinya *…easy for teachers to understand its application*
	…praktikal untuk dilaksanakan kepada pelajar *…practical to be administered to students*
	…prosedur pemarkahan yang mudah *…easy to use scoring procedures*
	…berasaskan kertas dan pensil *…paper-and-pencil based*
	…berasaskan computer *…computer-based*
	Atau ciri-ciri yang tidak disebut di atas. Sila Senaraikan di sini... *Or anything else that has not been mentioned. Please list here …*

Sila pastikan anda memberi respon untuk **semua** item.
*Please ensure you have responded to **all** the items.*

Komen (sekiranya ada)/ *Comments (if any):*

_____

_____

Jutaan terima kasih atas kerjasama anda.
Thank you very much for your participation.

SEKSYEN 3 / *SECTION 3* (Additional items for Post-intervention (intervention group only)

Dalam seksyen ini, saya ingin mengetahui pendapat anda mengenai alat pentaksiran baru "Learning Test" (LT) dan potensi pelaksanaannya dalam sekolah-sekolah di Malaysia.
Sila beri respon anda untuk pernyataan-pernyataan berikut dengan menandakan (X) pada mana-mana bahagian di garisan antara **sangat tidak setuju** (di sebelah kiri) dan **sangat setuju** (di sebelah kanan) yang mewakili pendapat peribadi anda.

In this section, I would like to know your honest opinion about the new "Learning Test" (LT) and its potential implementation in Malaysian schools.
Please respond to the following statements by marking (X) *at any point on the line between "**strongly disagree**" (on the left) and "**strongly agree**" (on the right) that best represents your personal opinion.*

Pada pendapat saya, ...
In my opinion, ...

... ia adalah adil untuk menempatkan pelajar menggunakan keputusan UDT1 *... it is fair to stream students using the F1DT result*	Sangat tidak setuju I---------I Sangat setuju
... UDT1 memberikan maklumat yang bertindan dengan maklumat UPSR *... the F1DT provides information that are redundant to the UPSR*	Sangat tidak setuju I---------I Sangat setuju
.... LT memberikan maklumat yang lebih berguna mengenai potensi pembelajaran pelajar *... the LT provides more useful information regarding students' learning potential*	Sangat tidak setuju I---------I Sangat setuju
... LT menyediakan maklumat yang lebih bermakna untuk guru-guru menyesuaikan keperluan pembelajaran pelajar mereka *... the LT offers more meaningful insights for teachers to tailor their students' educational needs*	Sangat tidak setuju I---------I Sangat setuju
... adalah sesuai untuk melaksanakan LT dalam sekolah-sekolah di Malaysia *... it is feasible to implement the LT into Malaysian schools*	Sangat tidak setuju I---------I Sangat setuju
... LT boleh menggantikan UDT1 sebagai alat pentaksiran diagnostik untuk pelajar Tingkatan 1 *... the LT can replace the F1DT as the diagnostic assessment tool for Form 1 students*	Sangat tidak setuju I---------I Sangat setuju
... konsep di sebalik LT dan aplikasinya adalah mudah untuk difahami *... the concept behind the LT and its application is easy to understand*	Sangat tidak setuju I---------I Sangat setuju
... sesi pengenalan mengenai LT ini berguna untuk saya sebagai guru *... the introductory session to the LT was useful to me as a teacher*	Sangat tidak setuju I---------I Sangat setuju

… sesi LT ini memberikan saya perspektif baru mengenai pentaksiran pendidikan *… the LT session provided me with a new perspective on educational assessment*	**Sangat tidak setuju** I----------I **Sangat setuju**

Sekiranya LT akan digunakan, apakah cabaran-cabaran untuk kejayaan pelaksanaannya dalam sekolah-sekolah di Malaysia?
If LT were to be used, what are the challenges for its successful implementation in Malaysian schools?

Tandakan mana-mana jawapan yang berkaitan.
Tick any relevant answers.

	Keengganan guru-guru untuk berubah *Teachers' resistance to change*
	Kekecewaan guru-guru terhadap perubahan yang banyak pada pentaksiran pendidikan *Teachers' frustration over too many changes of educational assessments*
	Masa yang tidak mencukupi untuk guru-guru untuk membiasakan diri dengan alat pentaksiran baru *Insufficient time for teachers to familiarise themselves with the new assessment tool*
	Kekeliruan guru-guru terhadap polisi pentaksiran yang sentiasa berubah *Teachers' confusion of ever-changing assessment policies*
	Beban kerja yang berat *Heavy workload*
	Atau faktor yang tidak disebut di atas. Sila Senaraikan di sini... *Or anything else that has not been mentioned. Please list here …*

Sila pastikan anda memberi respon untuk **semua** item.
*Please ensure you have responded to **all** the items.*

Komen (sekiranya ada)/ *Comments (if any):*

_____

_____

_____

_____

Jutaan terima kasih atas kerjasama anda.
Thank you very much for your participation.

Appendix 6A: Consent Form - Parental Permission

Durham University
School of Education

Shaped by the past, creating the future

03rd April 2017

Title: **RESEARCH ON EDUCATIONAL ASSESSMENT**

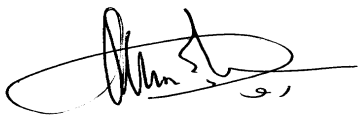
Dear parents/ guardian,
With the consent from the school, your son/daughter …….. is invited to participate in a research about educational assessments. One of the objectives of this study is to introduce an alternative assessment tool to secondary teachers. This study has been reviewed and approved by the Ministry of Education and the Sabah Education Department (date of approval: 24 January 2017).

Your child's participation will involve him/her to answer a computer-adaptive test in a computer lab. The test will take approximately 30 to 40 minutes. During the test, a lab assistant and the researcher will be there at all time to monitor the administration of the test.

There are no known risks associated with this research. The researcher will do everything to protect your child's privacy. No individual information (e.g., name, class, gender) will be revealed in any publication resulting from this study. However, the researcher may use a picture of your child (no disclosing of the face) as a pictorial documentation of the study.

Participation in this research is voluntary. You may refuse to allow your child to participate or withdraw your child from the study at any time. Your child will not be penalised in any way should you decide not to allow your child to participate or to withdraw your child from this study.

If you have any questions or concerns about this study or if any problems arise, please contact the researcher via email at rusilah.yusup@durham.ac.uk.
Thank you.



RUSILAH BINTI YUSUP

Leazes Road
Durham City, DH1 1TA
Telephone +44 (0)191 334 2000 Fax +44 (0)191 334 8311
www.durham.ac.uk
Durham University is the trading name of the University of Durham

Declaration of Informed Consent

To: Rusilah binti Yusup

Permission to be a part of a research study

I wish to inform that I

	Approve
	Do not approve

for my child ..... from Form 1 ..... to participate in the research project that will take place in the computer lab.

Name :.....
Signature :.....
Date :.....

****

Kepada: Rusilah binti Yusup

Kebenaran untuk pelajar mengambil bahagian dalam penyelidikan di sekolah

Saya ingin menyatakan bahawa saya

	Setuju
	Tidak setuju

anak saya ..... dari kelas Tingkatan 1 .....untuk mengambil bahagian dalam penyelidikan yang akan diadakan di makmal komputer sekolah.

Nama : .....
Tandatangan : .....
Tarikh : .....

Appendix 7A: Characteristics of the Participating Schools (*n*=21)

No.	Educational Zone	District	School Name	Number of samples		Element of Comparability					
						GROUP		SCHOOL GPA		LOCATION	
				PRE-TEST	POST-TEST	Intervention	Control	(GPA ≥6	(GPA ≥7	Urban	Rural
1	SOUTH	Beaufort	SMK BEAUFORT III	25				√			√
2		Sipitang	SMK. SINDUMIN	36	40	√		√			√
3		Sipitang	SMK PENGIRAN OMAR II	57	21		√	√			√
4	WEST	Kota Kinabalu	SMK TAMAN TUN FUAD	24				√		√	
5		Tuaran	SMK SRI NANGKA	36				√			√
6	TAWAU	Lahad Datu	SMK SEPAGAYA	78	67		√	√		√	
7		Lahad Datu	SMK SILABUKAN	40				√			√
8	NORTH	Kota Marudu	SMK BENGKONGAN	40	40	√		√			√
9		Kota Marudu	SMK KOTA MARUDU II	32				√			√
10		Kudat	SMK KUDAT	57	44		√	√			√
11		Kudat	SMK KUDAT II	40	36	√		√			√
12		Kudat	SMK ABDUL RAHIM II	24	13		√	√			√
13		Pitas	SMK PITAS II	56	50	√			√		√
14		Kota Belud	SMK TAUN GUSI	22				√			√
15		Kota Belud	SMK TAMBULION	42				√			√
16	SANDAKAN	Sandakan	SMK MERPATI	31				√		√	
17		Sandakan	SMK ELOPURA II	52	48		√	√		√	
18		Sandakan	SMK SANDAKAN II	59	67		√	√		√	
19	INTERIOR	Keningau	SMK GUNSANAD	32	32		√	√		√	
20		Keningau	SMK APIN APIN	19	9		√	√			√
21		Nabawan	SMK NABAWAN	60	42	√			√		√
TOTAL				862	509	5	8	19	2	6	15

Appendix 7B: Summary of Demographic Information about Respondents (Main Study)

Demographic Information	Characteristics	Pre-intervention		Post-intervention	
		n=862	%	*n=381*	%
GENDER	Male	284	32.9	121	31.8
	Female	578	67.1	260	68.2
YEARS OF TEACHING	less than 5 years	241	28.0	133	34.9
	6 to 10 years	276	32.0	116	30.4
	11 to 15 years	126	14.6	55	14.4
	16 to 20 years	114	13.2	37	9.7
	more than 20 years	105	12.2	40	10.5
EDUCATION LEVEL	SPM	1	.1	1	.3
	STPM	5	.6	0	0
	Diploma	7	.8	2	.5
	Bachelor	771	89.4	359	94.2
	Master	78	9.0	19	5.0
POSITION	Teacher	749	86.8	334	87.7
	Administrator	113	13.2	47	12.3
FAMILIARITY WITH F1DT	YES	546	63.3	242	63.5
	NO	316	36.7	139	36.5
LOCATION	Urban	276	32.0	146	38.3
	Rural	586	68.0	235	61.7

Appendix 7C: Component Loadings of SEA (862-dataset)

Pattern Matrix[a]

	Component		
	1 (14 items)	2 (16 items)	3 (6 items)
XPr learning potentials	.878		
XPr strengths and weaknesses	.871		
XPr streaming students	.860		
XPr pre-existing knowledge	.800		
XPr identification of LD	.756		
XPr learning needs	.751		
XPr Learning improvement	.750		
XPr intervention for LD	.727		
XPr pedagogical decision	.706		
XPr innate abilities	.600		
XPr students comparison	.565		
XPr Quality of teaching	.513		
XPr low-performing students	.478		
XPr future performance	.328		
XP learning potentials		.829	
XP strengths and weaknesses		.746	
XP pre-existing knowledge		.746	
XP future performance		.739	
XP innate abilities		.736	
XP referral to SEP		.676	
XP info accuracy		.666	
XP quality of teaching		.664	.308
XP identification of LD		.643	
XP intervention for LD		.612	
XP learning improvement		.610	
XP pedagogical decision		.597	
XP students comparison		.594	
XP low-performing students		.589	
XP streaming students		.568	
XP learning needs		.562	
XPr little use			.779
XPr directive order			.742
XP waste of time			.720
XPr preference of UPSR			.666
XP little use			.647
XP directive order			.630

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Appendic 7D:Letter of Ethics Approval from the School of Education Ethics Committee (Main Study)

Durham University
School of Education

Shaped by the past, creating the future

24 January 2017

Rusilah Yusup
rusilah.yusup@durham.ac.uk

Dear Rusilah

A Study of Teachers' Perception and Use of Educational Assessment: Exploring the Potential of Dynamic Testing in Malaysian Schools

I am pleased to inform you that your ethics application for the above research project has been approved by the School of Education Ethics Committee.

May we take this opportunity to wish you good luck with your research.

Yours sincerely,

Dr Nadin Beckmann
School of Education Ethics Committee Chair

Leazes Road
Durham, DH1 1TA
Telephone +44 (0)191 334 2000 Fax +44 (0)191 334 8311
www.durham.ac.uk/education

Appendix 7E: Research Approval from Economic Planning Unit (EPU)

UNIT PERANCANG EKONOMI
Economic Planning Unit
Jabatan Perdana Menteri
Prime Minister's Department
Blok B5 & B6
Pusat Pentadbiran Kerajaan Persekutuan
62502 PUTRAJAYA
MALAYSIA

Telefon : 603-8000 8000

Ruj. Tuan:
Your Ref.:

Ruj. Kami:
Our Ref.: UPE 40/200/19/3286 (31)

Tarikh:
Date: 3 November 2016

Ms. Rusilah binti Yusup
Kampung Gadong
Peti Surat 165
Beaufort, Sabah
Email : seela76@yahoo.com

APPLICATION TO CONDUCT RESEARCH IN MALAYSIA

With reference to your application, I am pleased to inform that your application to conduct research in Malaysia has been approved by the **Research Promotion and Co-ordination Committee, Economic Planning Unit, Prime Minister's Department.** The details of the approval are as follows:

Researcher's name	:	**RUSILAH BINTI YUSUP**
Passport No./ I.C No	:	**760528-12-5680**
Nationality	:	**MALAYSIAN**
Title of Research	:	**"A STUDY OF TEACHERS' PERCEPTION AND USE OF EDUCATIONAL ASSESSMENT: EXPLORING THE POTENTIALS OF DYNAMIC TESTING IN MALAYSIAN SCHOOLS"**
Period of Research Approved	:	**7 months (1.1.2017- 30.6.2017)**

2. Please take note that the study should avoid sensitive issues pertaining to local values and norms as well as political elements. At all time, please adhere to the conditions stated by the code of conduct for researchers as attached.

3. The issuance of the research pass is also subject to your agreement on the following:

a) to ensure submission of a brief summary of your research findings on completion of your research;
b) to submit three (3) copies of your final dissertation/publication; and
c) to return the research pass to the Research Promotion and Co-ordination Committee, Economic Planning Unit, Prime Minister's Department.

4. Thank you for your interest in conducting research in Malaysia and wish you all the best in your future research endeavor.

Yours sincerely,

(AZRAL IZWAN BIN MAZLAN)
for Director General
Economic Planning Unit
Prime Minister's Department
Email: azral.mazlan@epu.gov.my
Tel : 03 88725277
Fax : 03 88883798

ATTENTION

This letter is only to inform you the status of your application and **cannot be used as a research pass.**

c.c

1. Ketua Setiausaha
Kementerian Pendidikan Malaysia
Aras 1-4, Blok E8
Kompleks Kerajaan Parcel E
Pusat Pentadbiran Kerajaan Persekutuan
62604 Putrajaya
(u.p. YBhg. Dato' Sulaiman bin Wak
Pengarah BPPDP
Bahagian Perancangan dan Penyelidikan
Dasar Pendidikan

2. Pengarah
Unit Perancang Ekonomi Negeri Sabah
Jabatan Ketua Menteri
Lot 6-10, Wisma Sedia
Off Jalan Pintas, Penampang
88300 Kota Kinabalu
Sabah

Appendix 7F: Research Approval from Sabah Education Department

JABATAN PENDIDIKAN NEGERI SABAH
Sektor Pengurusan Sekolah
Aras 2, Wisma Pendidikan
Jabatan Pendidikan Negeri Sabah
Jalan Punai Tanah, 88450 Likas
KOTA KINABALU, SABAH

Tel : 088-537127/115/098/106
Fax : 088-310719
Laman Web: www.moe.gov.my/jpnsabah
E-mel : sjk.sabah@moe.gov.my

Rujukan: JP (SB)/700/07/03 Jld. 50
Tarikh : 23 Januari 2017

Rusilah Binti Yusup
11 Grove Terrace
Langley Moor, DH7 8JT
Durham, United Kingdom

Tuan,

KELULUSAN UNTUK MENJALANKAN KAJIAN DI SEKOLAH, INSTITUT PERGURUAN, JABATAN PENDIDIKAN NEGERI DAN BAHAGIAN-BAHAGIAN DI BAWAH KEMENTERIAN PENDIDIKAN MALAYSIA

Dengan segala hormatnya, saya diarah merujuk surat tuan/puan mengenai perkara di atas

2. Sukacita dimaklumkan bahawa Jabatan Pendidikan Negeri Sabah tiada halangan bagi pihak tuan menjalankan kajian **"A Study of Teachers' Perception And Use Of Educational Assessment Exploring The Potentials of Dynamic Testing in Malaysian Schools"** seperti dalam surat Kementerian Pendidikan Malaysia. Walau bagaimanapun ianya tertakluk kepada syarat-syarat berikut:

2.1 Berhubung dan berbincang dengan pentadbir sekolah tentang pelaksanaan/ perjalanan kajian tersebut.

2.2 Penyertaan warga pendidik dan murid-murid dalam kajian adalah sukarela.

2.3 Proses pengajaran dan pembelajaran atau pelaksanaan aktiviti sekolah tidak terganggu atau terjejas semasa kajian dijalankan..

2.4 Tuan tidak dibenarkan menjalankan aktiviti di kelas-kelas peperiksaan awam sekolah.

2.5 Sebarang data / maklumat serta dapatan kajian hanyalah untuk memenuhi syarat-syarat kursus pengajian sahaja.

2.6. Sila tuan kemukakan ke sektor ini senaskah laporan akhir kajian/laporan dalam bentuk elektronik berformat Pdf di dalam CD bersama naskah hardcopy setelah selesai kelak sebagai rujukan.

3. Surat kelulusan ini sah digunakan bermula dari **01 Januari 2017 hingga 30 Jun 2017**

Sekian, terima kasih.

"BERKHIDMAT UNTUK NEGARA"

Saya yang menurut perintah

MOHD ZAINI BIN YANIN
Ketua Sektor
Sektor Pengurusan Sekolah
b.p Pengarah, Jabatan Pendidikan Negeri Sabah

s.k 1. Pendaftar Institusi Pendidikan dan Guru,
Jabatan Pendidikan Negeri sabah

Appendix 7G: Letter of Notification to District Education Office

11 Grove Terrace,
Langley Moor,
DH7 8JT
Durham, United Kingdom

EN. WASLY WAHIP
Pegawai Pendidikan Daerah,
Pejabat Pendidikan Daerah Keningau,
Peti Surat 4
89007 Keningau
Sabah, MALAYSIA

10 APRIL 2017

Tuan,

PEMBERITAHUAN MENJALANKAN KAJIAN PENYELIDIKAN DI SEKOLAH MENENGAH DI DAERAH KENINGAU

Dengan segala hormatnya perkara di atas adalah dirujuk.

02. Untuk makluman tuan, saya telah mendapat kebenaran dari Jabatan Pendidikan Negeri Sabah untuk menjalankan kajian saya yang bertajuk *"A Study of Teachers' Perception and Use of Educational Assessment: Exploring the Potentials of Dynamic Testing in Malaysian Schools"*.

03. Sampel kajian saya akan melibatkan guru-guru sekolah menengah di negeri Sabah. Kajian ini memerlukan guru-guru untuk menjawab beberapa pernyataan dalam borang soal selidik mengenai ujian diagnostik yang digunakan di sekolah. Berikut adalah senarai sekolah menengah daerah Keningau yang terlibat dalam projek ini:

1. SMK Gunsanad
2. SMK Bingkor
3. SMK Apin-Apin

04. Bersama surat ini saya lampirkan dokumen-dokumen sokongan berikut untuk rujukan tuan.

i. Surat kelulusan menjalankan kajian dari EPU, Putrajaya (bagi pelajar luar negara)
ii. Surat ulasan penyelidikan EPRD, Kementerian Pendidikan Malaysia, Putrajaya
iii. Surat kelulusan menjalankan kajian dari Jabatan Pendidikan Negeri Sabah

05. Keprihatinan dan kerjasama tuan mengenai penglibatan sekolah-sekolah di daerah Keningau ini amatlah dihargai. Semoga kita bersama-sama dapat menyumbang ke arah peningkatan kualiti pendidikan negara khususnya di negeri Sabah.

Sekian, terima kasih.

Yang benar,

.....

RUSILAH BINTI YUSUP
Durham University, United Kingdom

Appendix 7H: Letter of Consent to School Principal

11 Grove Terrace,
Langley Moor,
DH7 8JT
Durham, United Kingdom

EN. MOHAMAD TAN BIN BACHO
Pengetua
SMK Sandakan II
Peti Surat 2769
90731 Sandakan, Sabah.

27 MAC 2017

Tuan,

PERMOHONAN UNTUK MENJALANKAN KAJIAN PENYELIDIKAN DI SEKOLAH

Dengan segala hormatnya perkara di atas adalah dirujuk.

02. Untuk makluman tuan, saya telah mendapat kebenaran dari Jabatan Pendidikan Negeri Sabah untuk menjalankan kajian saya untuk ijazah kedoktoran yang bertajuk *"A Study of Teachers' Perception and Use of Educational Assessment: Exploring the Potentials of Dynamic Testing in Malaysian Schools"*.

03. Sampel kajian saya akan melibatkan guru-guru sekolah menengah di negeri Sabah. Justeru, saya ingin memohon kebenaran daripada pihak tuan untuk menjalankan penyelidikan di sekolah tuan. Kajian ini memerlukan guru-guru untuk menjawab beberapa pernyataan dalam borang soal selidik mengenai ujian diagnostik yang digunakan di sekolah.

04. Bersama surat ini saya lampirkan dokumen-dokumen sokongan berikut untuk rujukan tuan.

i. Surat kelulusan menjalankan kajian dari EPU, Putrajaya (bagi pelajar luar negara)
ii. Surat ulasan penyelidikan EPRD, Kementerian Pendidikan Malaysia, Putrajaya
iii. Surat kelulusan menjalankan kajian dari Jabatan Pendidikan Negeri Sabah

05. Saya amat berharap tuan dapat mempertimbangkan permohonan saya ini sewajarnya. Semoga kita bersama-sama dapat menyumbang ke arah peningkatan kualiti pendidikan negara khususnya di negeri Sabah. Segala keprihatinan dan kerjasama tuan mengenai perkara ini amatlah dihargai.

Sekian, terima kasih.

Yang benar,

RUSILAH BINTI YUSUP
Durham University, United Kingdom

Appendix 7I: Description of Variables of Interests for HLM Analyses

Hierarchical Level	Unit of Analysis	Independent Variables (Explanatory Variables)	Coding	Outcome variables
Level-1	Teacher (*i*) (Individual)	Years of teaching (*YEARS*)	1=less than 5 years 2=6 to 10 years 3=11 to 15 years 4=16 to 20 years 5=more than 20 years	*PERCEPTION* *PRACTICE* *NEGATIVITY*
		Educational level (*EDULEVEL*)	1=SPM 2=STPM 3=Diploma 4=Bachelor 5=Master 6=PhD	
		Position in the school (*POSITION*)	1=teachers 2=administrators	
		Familiarity with F1DT (*F1DT*)	1=YES 2=NO	
		Teachers participating in the pre-intervention *(G_PHASE1)*	1=carried on 2=dropped out	
		Phase/ Time of completing the questionnaire *(TIME)*	1=pre-intervention 2=post-intervention	
Level-2	School (*j*) (Group)	Location of the school (*LOCATION*)	1=urban 2=rural	
		The comparison groups *(GROUP)*	1=intervention group 2=control group	
		Session for the CPD *(SESSION)*	1=Session 1 2=Session 2 3=Session 3 4=Session 4 5=Session 5	

References

Abdul Majid, Z., Abd Samad, A., Muhamad, M., & Vethamani, M. E. (2011). The school-based Oral English Test: Similarities and differences in opinion between teachers and students. *The English Teacher, 10*, 113–128.

Abdul Wahid, N.-A., Abdul Hamid, H., Low, Y.-M., & Mohd Ashhari, Z. (2011). Malaysian education system reform: Educationists' perspectives. In *Proceeding of the International Conference on Social Science, Economics and Art 2011* (pp. 14–15).

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice*, *42*(1), 18–29. https://doi.org/10.1207/s15430421tip4201_4

Adams, T. L., & Hsu, J.-W. Y. (1998). Classroom assessment : Teachers' conceptions and practices in Mathematics. *School Science and Mathematics*, *98*(4), 174–180.

Albert Jonglai, S. (2017). *From policy to practice : The effect of teachers ' educational beliefs and values on their interpretation of school-based assessment reform in primary schools in Malaysia*. University of Leeds.

Alkharusi, H., Aldhafri, S., Alnabhani, H., & Alkalbani, M. (2012). Educational assessment attitudes, competence, knowledge, and practices: An exploratory study of Muscat teachers in the Sultanate of Oman. *Journal of Education and Learning*, *1*(2), 217–232. https://doi.org/10.5539/jel.v1n2p217

Allred, S. B., & Ross-Davis, A. (2010). The drop-off and pick-up method: An approach to reduce nonresponse bias in natural resource surveys. *Small-Scale Forestry*, *10*(3), 305–318. https://doi.org/10.1007/s11842-010-9150-y

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.

Appleby, J., Samuels, P., & Treasure-Jones, T. (1997). Diagnosys - a knowledge-based diagnostic test of basic mathematical skills. *Computers Education*, *28*(2), 113–131.

Azis, A. (2015). Conceptions and practices of assessment: A case of teachers representing improvement conception. *TEFLIN Journal*, *26*(2), 129–154. https://doi.org/10.15639/teflinjournal.v26i2/129-154

Bahagian Pembangunan Kurikulum. (2013). *Panduan pemantapan pelaksanaan sistem set untuk Tingkatan 1 (Guidelines for the implementation of set system for Form 1)*. Putrajaya:

Bahagian Pembangunan Kurikulum.

Baird, J. (2010). Beliefs and practice in teacher assessment. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 1–5. https://doi.org/10.1080/09695940903562682

Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education*, *51*(5), 384–397. https://doi.org/10.1177/0022487100051005006

Barnes, N., Fives, H., & Dacey, C. M. (2015). Teachers' beliefs about assessment. In H. Fives & M. G. Gill (Eds.), *International handbook of research on teachers' beliefs* (pp. 284–300). New York: Routledge.

Barnes, N., Fives, H., & Dacey, C. M. (2017). U.S. teachers' conceptions of the purposes of assessment. *Teaching and Teacher Education*, *65*, 107–116. https://doi.org/10.1016/j.tate.2017.02.017

Bartlett, J. E., Kotrlik, J. W. ., & Higgins, C. C. (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, *19*(1), 43–50.

Beckmann, J. F. (2006). Superiority: always and everywhere? - On some misconceptions in the validation of dynamic testing. *Educational and Child Psychology*, *23*(3).

Beckmann, J. F. (2014). The umbrella that is too wide and yet too small: Why dynamic testing has still not delivered on the promise that was never made. *Journal of Cognitive Education and Psychology*, *13*(3), 308–323.

Beckmann, J. F., & Guthke, J. (1995). Complex problem solving, intelligence and learning ability. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 177–200). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Beckmann, N., Wood, R. E., & Minbashian, A. (2010). It depends how you look at it: On the relationship between neuroticism and conscientiousness at the within- and the between-person levels of analysis. *Journal of Research in Personality*, *44*(5), 593–601. https://doi.org/10.1016/j.jrp.2010.07.004

Beebe, L. H. (2007). What can we learn from pilot studies? *Perspectives in Psychiatric Care*, *43*(4), 213–218. https://doi.org/10.1111/j.1744-6163.2007.00136.x

Black, H. D. (1983). Introducing diagnostic assessment. *Innovations in Education & Training International*, *20*(1), 58–63. https://doi.org/10.1080/0033039830200109

Black, P. (2015). Formative assessment – an optimistic but incomplete vision. *Assessment in Education: Principles, Policy and Practice*, *22*(1), 161–177. https://doi.org/10.1080/0969594X.2014.999643

Black, P. . (2002). *Testing, friend or foe?: The theory and practice of assessment and testing*. London: Falmer Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). *Assessment for learning: Putting it into practice*. Maidenhead: Open University Press.

Black, P. J. (1998). *Testing, friend or foe?: The theory and practice of assessment and testing*. London: The Falmer Press.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (1998b). Inside the Black Box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–148. Retrieved from http://www.pdkintl.org/kappan/kbla9810.htm

Black, P., & Wiliam, D. (2006). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and learning* (pp. 9–25). London: Sage Publications.

Black, P., & Wiliam, D. (2010). Inside the black box : Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1002/hrm

Borg, M. (2001). Teachers ' Beliefs. *ELT Journal*, *55*(2), 186–188. https://doi.org/10.1097/01.chi.0000187243.17824.6c

Bosma, T., Hessels, M. G. P., & Resing, W. C. . (2012). Teachers' preferences for educational planning: Dynamic testing, teaching' experience and teachers' sense of efficacy. *Teaching and Teacher Education*, *28*(4), 560–567. https://doi.org/10.1016/j.tate.2012.01.007

Bosma, T., & Resing, W. C. . (2010). Teacher's appraisal of dynamic assessment outcomes: Recommendations for weak mathematics-performers. *Journal of Cognitive Education and Psychology*, *9*(2), 91–115. https://doi.org/10.1891/1945-8959.9.2.91

Bosma, T., & Resing, W. C. . (2012). Need for instruction: Dynamic testing in special education. *European Journal of Special Needs Education*, *27*(1), 1–19. https://doi.org/10.1080/08856257.2011.613599

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education*, *31*(4), 399–413. https://doi.org/10.1080/02602930600679050

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment and Evaluation in Higher Education*, *38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data

quality. *Journal of Public Health*, *27*(3), 281–291. https://doi.org/10.1093/pubmed/fdi031

Box, C., Skoog, G., & Dabbs, J. M. (2015). A case study of teacher personal practice assessment theories and complexities of implementing formative assessment. *American Educational Research Journal*, *52*(5), 956–983. https://doi.org/10.3102/0002831215587754

Brandom, R. (1981). Leibniz and degree of perception. *Journal of the History of Philosophy*, *19*(4), 447–479.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, *54*(1), 106–148. https://doi.org/10.1111/j.1467-6494.1986.tb00391.x

Brown, G. T. . (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy and Practice*, *11*(3), 301–318. https://doi.org/10.1080/0969594042000304609

Brown, G. T. ., Choudhry, H., & Dhamija, R. (2015). The impact of an assessment policy upon teachers' self-reported assessment beliefs and practices: A quasi-experimental study of Indian teachers in private schools. *International Journal of Educational Research*, *71*, 50–64.

Brown, G. T. ., & Gao, L. (2015). Chinese teachers' conceptions of assessment for and of learning: Six competing and complementary purposes. *Cogent Education*, *2*(1), 1–19. https://doi.org/10.1080/2331186X.2014.993836

Brown, G. T. ., & Hattie, J. (2012). The benefits of regular standardized assessment in childhood education: Guiding improved instruction and learning. In S. Suggate & E. Reese (Eds.), *Contemporary debates in child development and education* (pp. 287–292). London: Routledge. https://doi.org/10.4324/9780203115558

Brown, G. T. ., Hui, S. K. F., Yu, F. W. M., & Kennedy, K. J. (2011). Teachers' conceptions of assessment in Chinese contexts: A tripartite model of accountability, improvement, and irrelevance. *International Journal of Educational Research*, *50*(5), 307–320. https://doi.org/10.1016/j.ijer.2011.10.003

Brown, G. T. ., Irving, E., & Keegan, P. (2008). *An introduction to educational assessment, measurement and evaluation: Improving the quality of teacher-based assessment* (2nd ed.). North Shore, New Zealand: Pearson.

Brown, G. T. ., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment

for student improvement: Understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy and Practice*, *16*(3), 347–363. https://doi.org/10.1080/09695940903319737

Brown, G. T. ., Lake, R., & Matters, G. (2008). New Zealand and Queensland teachers' conceptions of learning: Transforming more than reproducing. *Australian Journal of Educational and Developmental Psychology*, *8*, 1–14.

Brown, G. T. ., & Remesal, A. (2012). Prospective teachers ' conceptions of assessment : A cross-cultural comparison. *The Spanish Journal of Psychology*, *15*(1), 75–89.

Brown, G. T. ., & Remesal, A. (2017). Teachers' conceptions of assessment: Comparing two inventories with Ecuadorian teachers. *Studies in Educational Evaluation*, *55*, 68–74. https://doi.org/10.1016/j.stueduc.2017.07.003

Brown, J. D. (2009). Principal components analysis and exploratory factor analysis — Definitions , differences , and choices. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *13*(1), 26–30.

Brown, S. (1987). Drop and collect surveys: A neglected research techniques? *Marketing Intelligence and Planning*, *5*(1), 19–23.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, *64*(2), 123–152. https://doi.org/10.1037/h0043805

Bruner, J. S., & Postman, L. (1949). On the preception of incongruity: a paradigm. *Journal of Personality*, *18*(2), 206–223. https://doi.org/10.1111/j.1467-6494.1949.tb01241.x

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(l), 147–158. https://doi.org/10.1037//0033-2909.101.1.147

Bryman, A. (2012). *Social research methods* (4th ed.). Oxford: Oxford University Press.

Büyükkarcı, K. (2014). Assessment beliefs and practices of language teachers in primary education. *International Journal of Instruction*, *7*(1), 107–121.

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *Journal of Special Education*, *41*(4), 254–270. https://doi.org/10.1177/0022466907310366

Calero, M. D., Belen, G.-M. M., & Robles, M. A. (2011). Learning potential in high IQ children : The contribution of dynamic assessment to the identification of gifted children. *Learning and Individual Differences*, *21*(2), 176–181. https://doi.org/10.1016/j.lindif.2010.11.025

Calveric, S. B. (2010). *Elementary teachers ' assessment beliefs and practices*. Virginia

Commonwealth University.

Capraro, R. M. (2004). Statistical significance , effect size reporting , and confidence intervals : Best reporting strategies. *Journal for Research in Mathematics Education*, *35*(1), 57–62.

Care, E., & Griffin, P. (2009). Assessment is for teaching. *Independence*, *34*(2), 56–59.

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education*, *36*(4), 395–407. https://doi.org/10.1080/03075071003642449

Chan, Y. F., & Sidhu, G. K. (2006). School-based assessment among ESL teachers in Malaysian secondary schools. *Jurnal Pendidikan USM*, *11*(3), 1–18. https://doi.org/10.1017/CBO9781107415324.004

Chapman, D. W., & Snyder Jr, C. W. (2000). Can high stakes national testing improve instruction: Reexamining conventional wisdom. *International Journal of Educational Development*, *20*, 457–474. https://doi.org/10.1016/S0738-0593(00)00020-1

Che Md Ghazali, N. H. (2016). The implementation of school-based assessment system in Malaysia: A study of teacher perceptions. *GEOGRAFIA Online TM Malaysian Journal of Society and Space*, *12*(9), 104–117.

Cheng, L. (1999). Changing assessment: Washback on teacher perceptions and actions. *Teaching and Teacher Education*, *15*(3), 253–271. https://doi.org/10.1016/S0742-051X(98)00046-8

Christoforidou, M., Kyriakides, L., Antoniou, P., & Creemers, B. P. M. (2014). Searching for stages of teacher's skills in assessment. *Studies in Educational Evaluation*, *40*, 1–11. https://doi.org/10.1016/j.stueduc.2013.11.006

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics and challenges. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3–17). New York: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd editio). Hillsdale, New Jersey: Lawrence Erlbaum Associates. https://doi.org/10.1234/12345678

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). Abingdon, Oxon,England: Routledge.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, New Jersey: Erlbaum.

Cooney, T. J. (1999). Conceptualizing teachers' ways of knowing. *Educational Studies in Mathematics*, *38*, 163–187.

Corretjer, O. I. (2016). *Exploring foreign language in the elementary school (FLES) teachers' attitudes and perceptions about assessment and assessment practices in the elementary world language classroom*. George Mason University.

Crane, T. (2009). Is perception a propositional attitude? *The Philosophical Quarterly*, *59*(236), 452–469.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches*. *Research design: Qualitative quantitative and mixed methods approaches* (4th editio). Thousand Oaks: Sage publications. https://doi.org/10.1007/s13398-014-0173-7.2

Cronbach, L. ., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. Retrieved from file:///C:/Users/Cinthia/Desktop/Capstone/Cronbach LJ & Meehl PE (1955) Construct validity in psychological tests.pdf

Crow, G., & Wiles, R. (2008). *Managing anonymity and confidentiality in social research: the case of visual data in community research* (NCRM Working Papers Series). Retrieved from http://eprints.soton.ac.uk/80291/

Dayal, H. C., & Lingam, G. I. (2015). Fijian teachers' conceptions of cssessment. *Australain Journal of Teacher Education*, *40*(8), 43–58. https://doi.org/10.1016/j.sbspro.2015.11.222

de Beer, M. (2006). Dynamic testing: Practical solutions to some concerns. *SA Journal of Industrial Psychology*, *32*(4), 8–14. https://doi.org/10.4102/sajip.v32i4.245

de Beer, M. (2010). Longitudinal predictive validity of a Learning Potential Test. *Journal of Psychology in Africa*, *20*(2), 225–231. https://doi.org/10.1080/14330237.2010.10820370

Dehejia, R. H., & Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics*, *84*, 151–161.

Delandshere, G., & Jones, J. H. (1999). Elementary teachers ' beliefs about assessment in mathematics : A case of assessment paralysis. *Journal of Curriculum and Supervision*, *14*(3), 216–240.

Deneen, C. C., & Brown, G. T. . (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent Education*, *3*, 1–14. https://doi.org/10.1080/2331186X.2016.1225380

Deutsch, R., & Reynolds, Y. (2000). The use of dynamic assessment by educational

psychologists in the UK. *Educational Psychology in Practice*, *16*(3), 311–331. https://doi.org/10.1080/713666083

DeWitt, D., Alias, N., & Siraj, S. (2016). Problem solving strategies of Malaysian secondary school teachers. In *Educational Technology World Conference 2016* (pp. 1–14). Bali, Indonesia.

Diamond, A., & Sekhon, J. S. (2006). *Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies*.

Dixon, H. R., Hawe, E., & Parr, J. (2011). Enacting Assessment for Learning: The beliefs practice nexus. *Assessment in Education: Principles, Policy and Practice*, *18*(4), 365–379. https://doi.org/10.1080/0969594X.2010.526587

Dörfler, T., Golke, S., & Artelt, C. (2009). Dynamic assessment and its potential for the assessment of reading competence. *Studies in Educational Evaluation*, *35*, 77–82. https://doi.org/10.1016/j.stueduc.2009.10.005

Dzulkifli, M. A., & Alias, I. A. (2012). Students of low academic achievement – their personality, mental abilities and academic performance: How counsellor can help? *International Journal of Humanities and Social Science*, *2*(23), 220–225.

Elleman, A. M., Compton, D. L., Fuchs, D., Fuchs, L. S., & Bouton, B. (2011). Exploring dynamic assessment as a means of identifying children at risk of developing comprehension difficulties. *Journal of Learning Disabilities*, *44*(4), 348–357. https://doi.org/10.1177/0022219411407865

Elliott, J. G. (2000). Dynamic assessment in educational contexts: purpose and promise. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: prevailing models and applications* (pp. 713–740). Oxford, UK: Elsevier.

Elliott, J. G. (2003). Dynamic assessment in educational settings: Realising potential. *Educational Review*, *55*, 37–41. https://doi.org/10.1080/00131910303253

Elliott, J. G., Lauchlan, F., & Stringer, P. (1996). Dynamic assessment and its potential for educational psychologists. *Educational Psychology in Practice*, *12*(3), 152–160. https://doi.org/10.1080/0266736960120303

Elliott, J. G., & Resing, W. C. . (2015). Can intelligence testing inform educational intervention for children with reading disability? *Journal of Intelligence*, *3*(4), 137–157. https://doi.org/10.3390/jintelligence3040137

Elliott, J. G., Resing, W. C. ., & Beckmann, J. F. (2018). Dynamic assessment: a case of unfulfilled potential? *Educational Review*, *70*(1), 7–17. https://doi.org/10.1080/00131911.2018.1396806

Elliott, J. G., Stankov, L., Lee, J., & Beckmann, J. F. (2019). What did PISA and TIMSS ever do for us?: The potential of large scale datasets for understanding and improving educational practice. *Comparative Education*, *55*(1), 133–155. https://doi.org/doi:10.1080/03050068.2018.1545386

Embertson, S. E. (1987). Improving the measurement of spatial aptitude by dynamic testing. *Intelligence*, *11*(4), 333–358. https://doi.org/10.1016/0160-2896(87)90016-X

Eren, A. (2010). Consonance and dissonance between Turkish prospective teachers ' values and practices : Conceptions about teaching , learning , and assessment. *Australain Journal of Teacher Education*, *35*(3), 27–48.

Erkut, S. (2010). Developing multiple language versions of instruments for intercultural research. *Child Development Perspectives*, *4*(1), 19–24. https://doi.org/10.1111/j.1750-8606.2009.00111.x

Ernest, P. (1989). The knowledge, beliefs and attitudes of the mathematics teacher: a model. *Journal of Education for Teaching*, *15*(1), 13–33. https://doi.org/10.1080/0260747890150102

Esposito, N. (2001). From meaning to meaning: The influence of translation techniques on non-English focus group research. *Qualitative Health Research*, *11*(4), 568–579. https://doi.org/10.1177/104973201129119217

Evans, C. (2013). Making sense of asssessment feedback in higher education. *Review of Educational Research*, *83*(1), 70–120. https://doi.org/10.3102/0034654312474350

Even, R. (2005). Using assessment to inform instructional decisions: How hard can it be? *Mathematics Education Research Journal*, *17*(3), 45–61. https://doi.org/10.1007/BF03217421

Everitt, B. S. (1975). Multivariate analysis: The need for data, and other problems. *The British Journal of Psychiatry*, *126*(3), 237–240.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*, *94*(5), 275–282. https://doi.org/10.1080/00220670109598763

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the black-white test score gap. *Urban Education*, *38*(4), 460–507. https://doi.org/10.1177/0042085903254970

Field, A. (2013). *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock*

"n" roll (4th ed.). London: SAGE.

Fong, P. C., & Muhamad, N. (2017). Readiness of implementation of school-based assessment among the Malay language teachers in national schools. *Advanced Science Letters*, *23*(3), 2169–2173. https://doi.org/10.1166/asl.2017.8589

Ford, J. K., MacCullum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*(2), 291–314. https://doi.org/10.1111/j.1744-6570.1986.tb00583.x

Foster, J. (2000). *The nature of perception*. Oxford: Oxford University Press. https://doi.org/10.1080/00048405885200231

Fraenkel, J. R., & Wallen, N. E. (2008). *How to design and evaluate research in education* (7th ed.). New York: McGraw-Hill.

Fullan, M., & Watson, N. (2000). School-based management: Reconceptualizing to improve learning outcomes. *School Effectiveness and School Improvement*, *11*(4), 453–473. https://doi.org/10.1076/sesi.11.4.453.3561

Furinghetti, F. (1996). A theoritical framework for teachers' conceptions. In E. Pehkonen (Ed.), *Current state of research in mathematical beliefs III: Proceedings of the MAVI-3 workshop* (pp. 19–25). Helsinki: University of Helsinki.

Ganapathy, M., Mehar Singh, M. K., Kaur, S., & Liew, W. K. (2017). Promoting higher order thinking skills via teaching practices. *The Southeast Asian Journal of English Language Studies*, *23*(1), 75–85. https://doi.org/http://doi.org/10.17576/3L-2017-2301-06

Gardner, J. (2012). Quality assessment practice. In J. Gardner (Ed.), *Assessment and learning* (2nd editio, pp. 103–121). London: SAGE publications.

Gardner, J., Harlen, W., Hayward, L., & Stobart, G. (2008). *Changing assessment practice process, principles and standards*. Retrieved from http://www.nuffieldfoundation.org/sites/default/files/JG Changing Assment Practice Final Final(1).pdf

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results rrom a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945.

Garson, G. D. (2013). Introductory guide to HLM with HLM 7 software. In G. D. Garson (Ed.), *Hierarchical linear modeling: Guide and applications* (2nd editio, pp. 55–96). Thousand Oaks, CA: Sage publications.

Gebril, A., & Brown, G. T. . (2013). The effect of high-stakes examination systems on teacher beliefs: Egyptian teachers' conceptions of assessment. *Assessment in Education:*

Principles, Policy & Practice.

Gebril, A., & Brown, G. T. . (2014). The effect of high-stakes examination systems on teacher beliefs: Egyptian teachers' conceptions of assessment. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 16–33. https://doi.org/10.1080/0969594X.2013.831030

Gillham, B. (2008). *Small-scale social survey methods*. London: Continuum International Publishing Group.

Gipps, C. V. (2012). *Beyond testing: Towards a theory of educational assessment* (Classic Ed). Abingdon: Routledge.

Glutting, J. J., & McDermott, P. A. (1990). Childhood learning potential as an alternative to traditional ability measures. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*(4), 398–403. https://doi.org/10.1037/1040-3590.2.4.398

Gorard, S. (2013). *Research design: Creating robust approaches for the social sciences*. London: Sage Publications.

Gordon, E. W. (2012). Assessment, teaching, and learning: A new vision of pedagogy. Retrieved from www.gordoncommission.org

Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: thoery and applications* (1st ed, pp. 173–201). Cambridge: Cambridge University Press.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, New Jersey: L. Erlbaum Associates.

Goslin, D. A. (1967). *Criticism of standardized tests and testing*. New York.

Goslin, D. A. (1968). Standardized ability tests and testing: Major issues and the validity of current criticisms of tests are discussed. *Science*, *159*(3817), 851–855. Retrieved from http://science.sciencemag.org/

Gresham, F. M., & Vellutino, F. R. (2010). What is the role of intelligence in the identification of specific learning disabilities? Issues and clarifications. *Learning Disabilities Research & Practice*, *25*(4), 194–206. https://doi.org/10.1111/j.1540-5826.2010.00317.x

Grigorenko, E. L. (2009). Dynamic assessment and response to intervention: Two sides of one coin. *Journal of Learning Disabilities*, *42*(2), 111–132. https://doi.org/:10.1177/0022219408326207

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, *124*(1), 75–111. https://doi.org/10.1037/0033-2909.124.1.75

Guadagnoli, E., & Velicer, W. F. (1988). The relationship of sample size to the stability of

component patterns. *Psychological Bulletin*, *103*(2), 265–275.

Guo, S. (2005). Analyzing grouped data with hierarchical linear modeling. *Children and Youth Services Review*, *27*(6), 637–652. https://doi.org/10.1016/j.childyouth.2004.11.017

Guskey, T. R. (1986). Staff development and the process of teacher change. *Educational Researcher*, *15*(4), 5–12.

Guskey, T. R. (2003). How classroom assessments improve learning. *Educational Leadership*, *60*(5), 6–11. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=9029496&site=ehost-live

Guskey, T. R. (2005). Formative classroom assessment and Benjamin S. Bloom: Theory, research, and implications. In *Annual Meeting of the American Educational Research Association* (pp. 1–11). https://doi.org/April 2005

Guthke, J. (1992). Learning tests-the concept, main research findings, problems and trends. *Learning and Individual Differences*, *4*(2), 137–151. https://doi.org/10.1016/1041-6080(92)90010-C

Guthke, J. (1993). Development in learning potential assessment. In J. H. M. Hamers, K. Sijtsma, & A. J. . Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 43–67). Lisse, The Netherlands: Swets & Zeitlinger BV.

Guthke, J., & Beckmann, J. F. (2000a). Learning test concepts and dynamic assessment. In A. Kozulin & B. Y. Rand (Eds.), *Experience of mediated learning: An impact of Feuerstein's theory in education and psychology* (pp. 175–190). Oxford, United Kingdom: Elsevier Science Inc.

Guthke, J., & Beckmann, J. F. (2000b). The learning test concept and its application in practice. In C. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 17–69). New York: JAI Elsevier.

Guthke, J., Beckmann, J. F., & Dobat, H. (1997). Dynamic testing - problems, uses, trends and evidence of validity. *Educational and Child Psychology*, *14*(4), 17–32.

Guthke, J., Beckmann, J. F., & Stein, H. (1995). Recent research evidence on the validity of learning tests. In J. S. Carlson (Ed.), *European contributions to dynamic assessment* (pp. 117–143). Greenwich, Connecticut: JAI Press Inc.

Guthke, J., & Stein, H. (1996). Are learning tests the better version of intelligence tests? *European Journal of Psychological Assessment*, *12*(1), 1–13. https://doi.org/10.1027/1015-5759.12.1.1

Hamers, J. H. M., Hessels, M. G. P., & Pennings, A. H. (1996). Learning potential in ethnic minority children. *European Journal of Psychological Assessment*, *12*(3), 183–192. https://doi.org/10.1027/1015-5759.12.3.183

Hamers, J. H. M., & Pennings, A. H. (1995). Learning potential tests for ethnic minorities. *European Journal of Special Needs Education*, *10*(1), 70–74. https://doi.org/10.1080/0885625950100107org/10.1080/0885625950100107

Hamers, J. H. M., & Resing, W. C. M. (1993). Learning potential assessment: Introduction. In J. H. M. Hamers, K. Sijtsma, & A. J. . Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues*. Amsterdam: Swets & Zeitlinger BV.

Hanafi, Z. (2008). The relationship between aspects of socio-economic factors and academic achievement. *Jurnal Pendidikan*, *33*, 95–105.

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *Curriculum Journal*, *16*(2), 207–223. https://doi.org/10.1080/09585170500136093

Harlen, W. (2007). *Assessment of learning*. London: Sage.

Harlen, W. (2012). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (2nd editio, pp. 87–102). London: Sage.

Harlen, W., & James, M. (1997). Assessment and Learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365–379. https://doi.org/10.1080/0969594970040304

Harris, L. R., & Brown, G. T. . (2009). The complexity of teachers' conceptions of assessment: tensions between the needs of schools and students. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 365–381. https://doi.org/10.1080/09695940903319745

Hashim, C. N., Ariffin, A., & Muhammad Hashim, N. (2013). Ideal vs. reality: Evidences from senior teachers' experiences on the Malaysian school-based assessment system (SBA). In *Proceedings of the Malaysian Education Deans' Council* (pp. 1–12). Retrieved from http://irep.iium.edu.my/38724/

Hashim, H. A., Freddy, G., & Rosmatunisah, A. (2012). Relationships between negative affect and academic achievement among secondary school students: The mediating effects of habituated exercise. *Journal of Physical Activity and Health*, *9*, 1012–1019. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22971880

Hassan, O. R., & Rasiah, R. (2011). Poverty and student performance in Malaysia. *International Journal of Institutions and Economies*, *3*(1), 61–76.

Hattie, J. (2015, October 27). We aren't using assessments correctly: There's a distinction between formative and summative assessments. *Education Week*, p. 23.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hayes, N. (2000). *Foundations of psychology*. London: Thomson Learning.

Hayward, L. (2015). Assessment is learning: the preposition vanishes. *Assessment in Education: Principles, Policy and Practice*, *22*(1), 27–43. https://doi.org/10.1080/0969594X.2014.984656

Haywood, H. C. (2006). A transactional perspective on mental retardation. In H. N. Switzky (Ed.), *International review of research in mental retardation* (pp. 289–314). San Diego, CA: Elsevier.

Haywood, H. C. (2010). Cognitive education: A transactional metacognitive perspective. *Journal of Cognitive Education and Psychology*, *9*(1), 21–35. https://doi.org/10.1891/1945

Haywood, H. C., Brown, A. L., & Wingenfeld, S. (1990). Dynamic approaches to psychoeducational assessment. *School Psychology Review*, *19*(4), 411–422.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York: Cambridge University Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=xQekS_oqGzoC&oi=fnd&pg=PA373&dq=Dynamic+assessment+in+practice:+Clinical+and+educational+applications.&ots=7wyEJOOMLb&sig=LpkVguMJLeCSYSSkvMW2VFXr6A0

Haywood, H. C., & Tzuriel, D. (2002). Applications and challenges in dynamic assessment. *Peabody Journal of Education*, *77*(2), 40–63.

Haywood, H. C., Tzuriel, D., & Vaught, S. (1992). Psychoeducational assessment from a transactional perspective. In H. C. Haywood (Ed.), *Interactive Assessment* (pp. 38–63). New York: Springer Science+Business Media.

Haywood, H. C., & Wingenfeld, S. (1992). Interactive assessment as a research tool. *The Journal of Special Education*, *26*(3), 253–268. https://doi.org/10.1177/002246699202600303

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). *Characterizing Selection Bias Using Experimental Data*.

Hertzog, M. . (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health*, *31*, 180–191. https://doi.org/10.1002/nur.20247

Hessels-Schlatter, C. (2002). A dynamic test to assess learning capacity in people with severe

impairments. *American Journal on Mental Retardation*, *107*(5), 340–351. https://doi.org/10.1352/0895-8017(2002)107<0340:ADTTAL>2.0.CO;2

Hessels, M. G. P. (1997). Low IQ but high learning potential: why Zeyneb and Moussa do not belong in special education. *Educational and Child Psychology*, *14*(4), 121–136.

Hessels, M. G. P. (2009). Estimation of the predictive validity of the HART by means of a dynamic test of Geography. *Journal of Cognitive Education and Psychology*, *8*(1), 5–21. https://doi.org/10.1891/1945-8959.8.1.5

Hessels, M. G. P., Berger, J.-L., & Bosson, M. (2008). Group assessment of learning potential of pupils in mainstream primary education and special education classes. *Journal of Cognitive Education and Psychology*, *7*(1), 43–69.

Hessels, M., & Vanderlinden, K. (2011). Training effects in dynamic assessment: A pilot study of eye movement as indicator of problem solving before and after training. *Educational & Child Psychology*, *28*(2), 101–113.

Higgins, R., Hartley, P., & Skelton, A. (2001). Getting the message across: The problem of communicating assessment feedback. *Teaching in Higher Education*, *6*(2), 269–274. https://doi.org/10.1080/13562510120045230

Hill, J. (2015). How useful is dynamic assessment as an approach to service delivery within educational psychology? *Educational Psychology in Practice*, *31*(2), 127–136. https://doi.org/10.1080/02667363.2014.994737

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101–117. https://doi.org/10.3758/BF03192848

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, *23*(6), 723–744. https://doi.org/10.1177/014920639702300602

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, *24*(5), 623–641. https://doi.org/10.1177/014920639802400504

Hughes, J. N., Gleason, K. A., & Zhang, D. (2005). Relationship influences on teachers' perceptions of academic competence in academically at-risk minority and majority first grade students. *Journal of School Psychology*, *43*(4), 303–320. https://doi.org/10.1016/j.jsp.2005.07.001

Hunzicker, J. (2011). Effective professional development for teachers : a checklist. *Professional Development in Education*, *37*(2), 177–179. https://doi.org/10.1080/19415257.2010.523955

Impara, J. C., Plake, B. S., & Fager, J. J. (1993a). Educational administrators' and teachers' knowledge of classroom assessment. *Journal of School Leadership*, *3*(5), 510–521.

Impara, J. C., Plake, B. S., & Fager, J. J. (1993b). Teachers' assessment background and attitudes toward testing. *Theory Into Practice*, *32*(2), 113–117. https://doi.org/10.1080/00405849309543584

Ishak, Z., Suet, F. L., & Poh, L. L. (2012). Parenting style as a moderator for students' academic achievement. *Journal of Science Education and Technology*, *21*(4), 487–493. https://doi.org/10.1007/S10956-01

Ismail, M. E., Samsudin, M. A., & Md. Zain, A. N. (2014). A multilevel study on trends in Malaysian secondary school students' science attitude: Evidence from TIMSS 2011. *International Journal of Asian Social Science*, *4*(5), 572–584. Retrieved from http://www.aessweb.com/pdf-files/ijass-2014-4(5)-572-584.pdf

Jaba, S., Hamzah, R., Bakar, A. R., & Mat Rasid, A. (2013). Acceptance towards school based assessment among agricultural integrated living skills teachers: Challenges implementing a holistic assessment. *Journal of Technical Education and Training*, *5*(1), 44–51.

Jabatan Perdana Menteri. (2010). *Government Transformation Programme: The Roadmap 1.0*. Putrajaya: Unit Pengurusan Prestasi Dan Pelaksanaan (PEMANDU).

James, M., & Lewis, J. (2012). Assessment in harmony with our understanding of learning: Problems and possibilities. In J. Gardner (Ed.), *Assessment and learning* (2nd editio, pp. 187–205). London: Sage.

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

Jemaah Nazir dan Jaminan Kualiti. (2016). *Pengiraan skor komposit bagi tujuan penentuan banding dan ranking sekolah menengah 2016 -KPM/TKPPM/ PPK/JNJK 1 (62) (Calculation of composite score for secondary school banding system 2016 - KPM/TKPPM/ PPK/JNJK 1 (62))*.

Jemaah Nazir dan Jaminan Kualiti. (2017). *Standard Kualiti Pendidikan Malaysia gelombang 2 (SKPMg2) (Standard for Quality Education in Malaysia wave 2)*. Putrajaya: Jemaah Nazir dan Jaminan Kualiti.

Johanson, G. A., & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement*, *70*(3), 394–400. https://doi.org/10.1177/0013164409355692

John, M. (2018). *Assessment reform in Malaysia : Policy into practice in primary schools*. University of Stirling, Scotland.

Julious, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics*, *4*, 287–291. https://doi.org/10.1002/pst.185

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, *9*(2), 131–155. https://doi.org/10.1207/s15327957pspr0902_3

Kagan, D. M. (1992). Implications of research on teacher belief. *Educational Psychologist*, *27*(1), 65–90. https://doi.org/10.1207/s15326985ep2701

Kahn, J. H. (2011). Multilevel modeling : Overview and applications to research in counseling psychology. *Journal of Counseling Psychology*, *58*(2), 257–271. https://doi.org/10.1037/a0022680

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12001

Kantor, P. ., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2011). Comparing two forms of dynamic assessment and traditional assessment of preschool phonological awareness. *Journal of Learning Disabilities*, *44*(4), 313–321. https://doi.org/10.1177/0022219411407861

Kementerian Pelajaran Malaysia. (2011). *Surat siaran Lembaga Peperiksaan Bil. 3 Tahun 2011: Pemakluman Pentaksiran Berasaskan Sekolah (PBS) di sekolah rendah dan menengah rendah (Circular from Examination Syndicate No.3 Year 2011: Implementation of school-based assessment in primary and lower s*.

King, G., Nielsen, R., Coberley, C., Pope, J. E., & Wells, A. (2011). Comparative effectiveness of matching methods for causal inference. *Unpublished Manuscript*, *15*.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge. https://doi.org/10.1016/0191-8869(94)90040-X

Koloi-Keaikitse, S. (2012). *Classroom assessment practices: A survey of Botswana primary and secondary school teachers*. Ball State University, Indiana.

Korthagen, F. (2017). Inconvenient truths about teacher learning : towards professional development 3 . 0. *Teachers and Teaching*, *23*(4), 387–405. https://doi.org/10.1080/13540602.2016.1211523

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*(1), 1–21. https://doi.org/10.1207/s15327906mbr3001

Kyaruzi, F., Strijbos, J. W., Ufer, S., & Brown, G. T. . (2018). Teacher AfL perceptions and feedback practices in mathematics education among secondary schools in Tanzania.

Studies in Educational Evaluation, 59, 1–9. https://doi.org/10.1016/j.stueduc.2018.01.004

LaGasse, A. B. (2013). Pilot and feasibility studies: Application in music therapy research. *Journal of Music Therapy*, *50*(4), 304–320. Retrieved from https://academic.oup.com/jmt/article-abstract/50/4/304/970685

Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues and Practice*, *27*(2), 28–45. https://doi.org/10.1111/j.1745-3992.2008.00120.x

Lantolf, J. P., & Poehner, M. E. (2007). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics*, *1*(1), 49–72. https://doi.org/10.1558/jal.v1i1.49

Lauchlan, F., & Elliott, J. G. (2001). The psychological assessment of learning potential. *British Journal of Educational Psychology*, *71*(4), 647–665. https://doi.org/10.1348/000709901158712

LeCroy, C. W., & Krysik, J. (2007). Understanding and interpreting effect size measures. *Social Work Research*, *31*(4), 243–248. https://doi.org/Article

Lei, M. T., & Mei, Y. O. (2015). Malaysian and Singaporean students' affective characteristics and mathematics performance: evidence from PISA 2012. *SpringerPlus*, *4*(563), 1–14. https://doi.org/10.1186/s40064-015-1358-z

Leighton, J. P., Gokiert, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom - versus large - scale tests : implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 7–21. https://doi.org/10.1080/09695940903565362

Lembaga Peperiksaan. (2014). *Panduan pengurusan pentaksiran berasaskan sekolah (Guidelines for the management of school-based assessment)*. Putrajaya: Lembaga Peperiksaan.

Leong, W. S. (2014). Knowing the intentions, meaning and context of classroom assessment: A case study of Singaporean teacher's conception and practice. *Studies in Educational Evaluation*, *43*, 70–78. https://doi.org/10.1016/j.stueduc.2013.12.005

Lidz, C. S. (1995). Dynamic assessment and the legacy of L.S. Vygotsky. *School Psychology International*, *16*, 143–153.

Lidz, C. S. (2014). Leaning toward a consensus about dynamic assessment: Can we? Do we want to? *Journal of Cognitive Education and Psychology*, *13*(3), 292–307. https://doi.org/10.1891/1945-8959.13.3.292

Lidz, C. S., & Gindis, B. (2003). Dynamic assessment of the evolving cognitive functions in

children. In A. Kozulin, B. Gindis, V. S. Ageyev, & S. M. Miller (Eds.), *Vygotsky's Educational Theory in Cultural Contexs* (pp. 99–118). Cambridge, UK: Cambridge University Press.

Liew, W. K., & Ganapathy, M. (2017). Promoting HOTS via ICT in ESL classrooms. In *The Seventh International Language Learning Conference*.

Looney, A., Cumming, J., van Der Kleij, F., & Harris, K. (2017). Reconceptualising the role of teachers as assessors: teacher assessment identity. *Assessment in Education: Principles, Policy and Practice*, 1–26. https://doi.org/10.1080/0969594X.2016.1268090

Lorah, J. (2018). Effect size measures for multilevel models : definition , interpretation , and TIMSS example. *Large-Scale Assessments in Education*, *6*(8), 1–11. https://doi.org/10.1186/s40536-018-0061-2

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage Publications.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Maclellan, E. (2001). Assessment for learning: the differing perceptions of tutors and students. *Assessment and Evaluation in Higher Education*, *26*(4), 307–318.

Maguire, M., & Delahunt, B. (2017). Doing a thematic analysis : A practical , step-by-step guide for learning and teaching scholars. *All Ireland of Teaching and Learning in Higher Education (AISHE-J)*, *8*(3), 3351–3364.

Majid, F. A. (2011). School-based assessment in Malaysian schools : The concerns of the English teachers. *US-China Education Review*, 393–402. Retrieved from http://education.uitm.edu.my/v1/images/stories/publication/faizah/article7.pdf

Malakolunthu, S., & Sim, K. H. (2010). Teacher perspectives of school-based assessment in a secondary school in Kuala Lumpur. *Procedia - Social and Behavioral Sciences*, *9*, 1170–1176. https://doi.org/10.1016/j.sbspro.2010.12.302

Malboeuf-Hurtubise, C., Achille, M., Muise, L., Beauregard-Lacroix, R., Vadnais, M., & Lacourse, É. (2016). A mindfulness-based meditation pilot study: Lessons learned on acceptability and feasibility in adolescents with cancer. *Journal of Child and Family Studies*, *25*(4), 1168–1177. https://doi.org/10.1007/s10826-015-0299-z

Maneesriwongul, W., & Dixon, J. K. (2004). Instrument translation process: A methods review. *Journal of Advanced Nursing*, *48*(2), 175–186. https://doi.org/10.1111/j.1365-2648.2004.03185.x

Mansor, A. N., Ong, H. L., Rasul, M. S., Raof, A. R., & Yusoff, N. (2013). The benefits of school-based assessment. *Asian Social Science*, *9*(8), 101–106.

https://doi.org/10.5539/ass.v9n8p101

Mansour, N. (2010). Impact of the knowledge and beliefs of Egyptian science teachers in integrating a STS based curriculum: A sociocultural perspective. *Journal of Science Teacher Education*, *21*(5), 513–534. https://doi.org/10.1007/s10972-010-9193-0

Mansour, N. (2013). Consistencies and Inconsistencies Between Science Teachers' Beliefs and Practices. *International Journal of Science Education*, *35*(7), 1230–1275. https://doi.org/10.1080/09500693.2012.743196

Marshall, M. N. (1996). Sampling for qualitative research. *Family Practice*, *13*(6), 522–525.

Mat Hassan, M. A., & Talib, R. (2013). Perception towards SBA implementation among teachers in Malaysian schools. In *2nd International Seminar on Quality and Affordaable Education (ISQAE 2013)* (pp. 194–200).

Mc Namara, M. (2010). *Exploring the impact of standardised assessment in the primary school classroom*. The National University of Ireland.

McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence (Part I): Understanding the effects of clustereddData. *Gifted Child Quaterly*, *54*(2), 152–155. https://doi.org/10.1177/0016986210363076

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context and the achievement gap. *Journal of School Psychology*, *46*(3), 235–261. https://doi.org/10.1016/j.jsp.2007.05.001

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers ' classroom assessment and grading practices. *The Journal of Educational Research*, *95*(4), 203–213. https://doi.org/10.1080/00220670209596593

McMullen, M. B. (1999). Characteristics of teachers who talk the DAP talk and walk the DAP walk. *Journal of Research in Childhood Education*, *13*(2), 216–230. https://doi.org/10.1080/02568549909594742

McNamara, T. (2000). *Language testing*. Oxford University Press. https://doi.org/10.1177/026553220001700103

Md-Ali, R., & Veloo, A. (2017). Teachers' autonomy and accountability in assessing students' physical education in school-based assessment. *Teacher Empowerment Toward Professional Development and Practices*, 71–83.

Md-Ali, R., Veloo, A., & Krishnasamy, H. N. (2015). Implementation of school-based assessment: The experienced teachers ' thoughts. *Australian Journal of Basic and Applied Sciences*, *9*(18), 72–78.

Md Omar, H., & Sinnasamy, P. (2009). Between the ideal and reality: Teachers' perception of

the implementation of school-based oral English assessment. *The English Teacher*, *38*, 13–29. Retrieved from http://www.melta.org.my/ET/2009/ET2009_p013-029.pdf

Md Yasin, A. S., & Dzulkifli, M. A. (2011). Differences in depression, anxiety and stress between low-and high acheiving students. *Journal of Sustainability Science and Management*, *6*(1), 169–178.

Md Yunus, M., Wan Osman, W. S., & Mohd Ishak, N. (2011). Teacher-student relationship factor affecting motivation and academic achievement in ESL classroom. *Procedia - Social and Behavioral Sciences*. https://doi.org/10.1016/j.sbspro.2011.04.161

Mehrgan, K., Hayati, A., & Alavi, S. M. (2017). Investigating the impacts of EFL teachers' age , educational background , instructional experience and gender on their beliefs about formative assessment. *International Journal of Foreign Language Teaching & Research*, *5*(18), 131–151.

Mertler, C. A. (1998). Classroom assesment practices of Ohio teachers. In *The Annual Meeting of the Mid-Western Educational Research Association*. https://doi.org/ED419696

Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education*, *120*, 285–96.

Mertler, C. A. (2003). Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference. In *The Annual Meeting of the Mid-Western Educational Research Association*. Columnus, Ohio.

Mertler, C. A. (2005). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education*, *33*(2), 76–92.

Mertler, C. A. (2007). *Interpreting standardized test scores: Strategies for data-driven instructional decision making*. Sage publications.

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools*, *12*(2), 101–113. https://doi.org/10.1177/1365480209105575

Mertler, C. A. (2014). *The data-driven classroom: How do I use student data to improve my instruction?* Alexandria, VA: ASCD. Retrieved from http://books.google.co.uk/books?id=czhcBAAAQBAJ

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

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11),

1012–1027.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037//0003-066X.50.9.741

Miller, L. J., Schoen, S. A., James, K., & Schaaf, R. C. (2007). Lessons learned: A pilot study on occupational therapy effectiveness for children with sensory modulation disorder. *American Journal of Occupational Therapy*, *61*(2), 161–169. https://doi.org/10.5014/ajot.61.2.161

Minbashian, A., Wood, R. E., & Beckmann, N. (2010). Task-contingent conscientiousness as a unit of personality at work. *Journal of Applied Psychology*, *95*(5), 793–806. https://doi.org/10.1037/a0020016

Ministry of Education Malaysia. (2001). *Falsafah Pendidikan Negara: Matlamat dan visi (National Education Philosophy: Goal and vision)*. Putrajaya: Curriculum Development Centre.

Ministry of Education Malaysia. (2013). *Malaysia Education Blueprint 2013 - 2025 (preschool to post-secondary education)*. Putrajaya: Kementerian Pendidikan Malaysia. https://doi.org/10.1016/j.tate.2010.08.007

Ministry of Education Malaysia. (2015). *Malaysia Education Blueprint: Annual report 2014*. Putrajaya: Ministry of Education Malaysia. https://doi.org/10.1017/CBO9781107415324.004

Ministry of Education Malaysia. (2016). *Annual report 2015: Malaysian Education Blueprint (2013-2015). Ministry of Education Malaysia*. Putrajaya: Ministry of Education Malaysia. https://doi.org/10.1017/CBO9781107415324.004

Ministry of Education Malaysia. (2017). *Malaysia Education Blueprint 2013-2025: Annual report 2016*. Putrajaya: Ministry of Education Malaysia.

Mohamad Ali, N. S., & Talib, R. (2013). Test anxiety in school settings: Implication on teachers. In *2nd International Seminar on Quality and Affordaable Education (ISQAE 2013)* (pp. 182–187).

Mohammad Radzi, F., & Md Sawari, S. S. (2016). Recognize teachers ' perception of the School- based Assessment (SBA) effectiveness in increasing students ' achievement in mathematics. *EDUCARE: International Journal for Educational Studies*, *8*(February),

139–146.

Mohd Ghazali, N. ., Yaakub, B., & Mustam, A. . (2012). Why do we need change?: Teachers' attitude towards school-based assessment system. In *SCR London's First International Conference on Social Sciences and Humanities in the Islamic World*.

Mohd Ishak, N., Mustapha, R., Mahmud, Z., & Ariffin, S. R. (2006). Emotional Intelligence of Malaysian Teachers : Implications on Workplace Productivity. *International Journal of Vocational Education and Training*, *14*, 7–24. Retrieved from http://hdl.voced.edu.au/10707/14289.

Mohd Yusuf, N. (2013). School-based assessment: transformation in educational assessment in Malaysia. In *Cambridge Horizons School-based Assessment: Prospects and Realities in Asian Contexts*.

Muijs, D., & Reynolds, D. (2002). Teachers' beliefs and behaviors: What really matters? *Journal of Classroom Interaction*, *37*(2), 3–15. Retrieved from https://georgefox.idm.oclc.org/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eft&AN=507792392&scope=site

Mukundan, J., & Khandehroo, K. (2010). Burnout among English language teachers in Malaysia. *Contemporary Issues in Education Research*, *3*(1), 71–76.

Mulliner, E., & Tucker, M. (2017). Feedback on feedback practice: perceptions of students and academics. *Assessment and Evaluation in Higher Education*, *42*(2), 266–288. https://doi.org/10.1080/02602938.2015.1103365

Navarro, J. J., & Mora, J. (2011). Analysis of the implementation of a dynamic assessment device of processes involved in reading with learning-disabled children. *Learning and Individual Differences*, *21*, 168–175. https://doi.org/10.1016/j.lindif.2010.11.008

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. New York: Freeman & Co.

Nespor, J. (1987). The role of beliefs in the practice of teaching The role of beliefs in the practice of teaching. *Journal of Curriculum Studies*, *19*(4), 317–328. https://doi.org/10.1080/0022027870190403

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass*, *2*(2), 842–860.

Nirmalakhandan, N. (2007). Computerized adaptive tutorials to improve and assess problem-

solving skills. *Computers and Education*, *49*(4), 1321–1329. https://doi.org/10.1016/j.compedu.2006.02.007

Nirmalakhandan, N. (2013). Improving problem-solving skills of undergraduates through computerized dynamic assessment. *Procedia - Social and Behavioral Sciences*, *83*, 615–621. https://doi.org/10.1016/j.sbspro.2013.06.117

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education*, *33*(3), 301–314. https://doi.org/10.1080/02602930701293231

Nunnally, J. (1978). *Psychometric theory*. New York: New York: McGraw-Hill.

O'Donovan, B., Rust, C., & Price, M. (2016). A scholarly approach to solving the feedback dilemma in practice. *Assessment and Evaluation in Higher Education*, *41*(6), 938–949. https://doi.org/10.1080/02602938.2015.1052774

OECD. (2016). *Low-performing students: Why they fall behind and how to help them succeed*. Paris: OECD Publishing.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*(3), 241–286. https://doi.org/10.1006/ceps.2000.1040

Ong, S. L. (2010). Assessment profile of Malaysia: high-stakes external examinations dominate. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 91–103. https://doi.org/10.1080/09695940903319752

Opre, D. (2015). Teachers' conceptions of assessment. *Procedia - Social and Behavioral Sciences*, *209*, 229–233. https://doi.org/10.1016/j.sbspro.2015.11.222

Orsmond, P., Maw, S. J., Park, J. R., Gomez, S., & Crook, A. C. (2013). Moving feedback forward: Theory to practice. *Assessment and Evaluation in Higher Education*, *38*(2), 240–252. https://doi.org/10.1080/02602938.2011.625472

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, *7*(1), 1–3. Retrieved from http://eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ638489

Othman, S. A., Azman, N., & M. Ali, M. (2008). Faktor Ibu Bapa dalam Kecemerlangan Akademik Pelajar Pekak : Kajian Kes [The parental factor in academic achievement of deaf students: A case study]. *MJLI*, *5*, 79–98.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, *62*(3), 307–332. https://doi.org/10.3102/00346543062003307

Pallant, J. (2013). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows*. Buckingham: Open University Press.

Park, H. S., Dailey, R., & Lemus, D. (2002). The use of exploratory actor analysis and principal components analysis in communication research. *Human Communication Research*, *28*(4), 562–577.

Pehkonen, E., & Pietilä, A. (2003). On relationships between beliefs and knowledge in mathematics education. *Proceedings of the CERME-3 (Bellaria) Meeting*, 1–8. Retrieved from http://www.cimm.ucr.ac.cr/ciaemIngles/articulos/universitario/concepciones/On Relationships Between Beliefs and Knowledge in Mathematics Education.*Pehkonen, Erkki; Pietil?, Anu. *Pehkonen, E. Relationships between Beliefs ... 2003.pdf

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicologia Educativa*, *20*(2), 65–77. https://doi.org/10.1016/j.pse.2014.11.002

Pena, E., Iglesias, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology*, *10*, 138–154.

Perry, M. L. (2013). *Teacher and principal assessment literacy*. University of Montana.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112. https://doi.org/10.1016/j.jsp.2009.09.002

Phelps, R. P. (2005). The rich, robust research literature on testing's achievement benefits. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 55–91). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Philipp, R. A. (2007). Mathematics teachers' beliefs and affect. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of teachers of mathematics* (pp. 257–315). Charlotte, NC: Information Age Publishing.

Pillay, H., Goddard, R., & Wilss, L. (2005). Well-being, burnout and competence: Implications for teachers. *Australian Journal of Teacher Education*, *30*(2), 22–33. https://doi.org/10.14221/ajte.2005v30n2.3

Pishghadam, R., Adamson, B., Shayesteh Sadafian, S., & Kan, F. L. F. (2013). Conceptions of assessment and teacher burnout. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 1–18. https://doi.org/10.1080/0969594X.2013.817382

Ponte, J. P. da. (1994). Knowledge, beliefs and conceptions in Mathematics teaching and learning. In *Proceeding of the Fifth International Conference on Systematic Co-operation*

between Theory and Practice in Mathematics Education (pp. 169–177).

Popham, W. J. (2001a). Teaching to the test? *Educational Leadership*, *58*(6), 16–20.

Popham, W. J. (2001b). Uses and misuses of standardized tests. *NASSP Bulletin*, *85*(24), 24–31. https://doi.org/10.1177/019263650108562204

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Virginia, USA: ASCD.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, *48*(1), 4–11. https://doi.org/10.1080/00405840802577536

Postareff, L., Virtanen, V., Katajavuori, N., & Lindblom-Ylänne, S. (2012). Academics' conceptions of assessment and their assessment practices. *Studies in Educational Evaluation*, *38*, 84–92. https://doi.org/10.1016/j.stueduc.2012.06.003

Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: The students' perspective. *Assessment and Evaluation in Higher Education*, *33*(2), 143–154. https://doi.org/10.1080/02602930601127869

Poulson, L., Avramidis, E., Fox, R., Medwell, J., & Wray, D. (2001). The theoretical beliefs of effective teachers of literacy in primary schools: An exploratory study of orientations to reading and writing. *Research Papers in Education*, *16*(3), 271–292. https://doi.org/10.1080/02671520126827

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: factor recovery with small sample sizes. *Behavior Genetics*, *32*(2), 153–161. https://doi.org/10.1023/A:1015210025234

Putih, M., Mohd Zin, Z., & Ismail, I. (2016). Reading performance of Malaysian students across gender in PISA 2012. *3L: The Southeast Asian Journal of English Language Studies*, *22*(2), 109–121.

Rahimi, M., & Asadollahi, F. (2012). EFL teachers' classroom management orientations: Investigating the role of individual differences and contextual variables. *Procedia - Social and Behavioral Sciences*, *31*, 43–48. https://doi.org/10.1016/j.sbspro.2011.12.014

Rattray, J., & Jones, M. C. (2007). Essential elements of questionnaire design and development. *Journal of Clinical Nursing*, *16*(2), 234–243. https://doi.org/10.1111/j.1365-2702.2006.01573.x

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models : A review. *Journal of Educational Statistics*, *13*(2), 85–116.

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459–496). New

York: Marcel Dekker. Retrieved from http://books.google.com/books?hl=en&lr=&id=P_1cAVvrVgoC&oi=fnd&pg=PA459&dq=Hierarchical+linear+models+and+experimental+design&ots=uL9B5s_zIR&sig=_cX2K9joSQX7l8gY9urZgcAAn9w

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis* (2nd ed). Newbury Park, CA: Sage Publications.

Reimann, N., & Sadler, I. (2017). Personal understanding of assessment and the link to assessment practice: the perspectives of higher education staff. *Assessment and Evaluation in Higher Education*, *42*(5), 724–736. https://doi.org/10.1080/02602938.2016.1184225

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education*, *27*(2), 472–482. https://doi.org/10.1016/j.tate.2010.09.017

Resing, W. C. . (2013). Dynamic testing and individualized instruction: Helpful in cognitive education? *Journal of Cognitive Education and Psychology*, *12*(1), 81–95. https://doi.org/10.1891/1945-8959.12.1.81

Resing, W. C. ., & Elliott, J. G. (2011). Dynamic testing with tangible electronics: measuring children's change in strategy use with a series completion task. *The British Journal of Educational Psychology*, *81*, 579–605. https://doi.org/10.1348/2044-8279.002006

Resing, W. C. ., Steijn, W. M. P., Xenidou-Dervou, I., & Stevenson, C. E. (2011). Computerized dynamic yesting: A study of the potential of an approach using sensor technology. *Journal of Cognitive Education and Psychology*, *10*(2), 178–194. https://doi.org/10.1891/1945

Resing, W. C. ., Touw, K. W. J., Veerbeek, J., & Elliott, J. G. (2017). Progress in the inductive strategy-use of children from different ethnic backgrounds: a study employing dynamic testing. *Educational Psychology*, *37*(2), 173–191. https://doi.org/10.1080/01443410.2016.1164300

Resing, W. C. ., Tunteler, E., de Jong, F. M., & Bosma, T. (2009). Dynamic testing in indigenous and ethnic minority children. *Learning and Individual Differences*, *19*(4), 445–450. https://doi.org/10.1016/j.lindif.2009.03.006

Resing, W. C. ., Xenidou-Dervou, I., Steijn, W. M. P., & Elliott, J. G. (2012). A "picture" of children's potential for learning: Looking into strategy changes and working memory by dynamic testing. *Learning and Individual Differences*, *22*(1), 144–150. https://doi.org/10.1016/j.lindif.2011.11.002

Resing, W. C. M., Bakker, M., Pronk, C. M. E., & Elliott, J. G. (2017). Progression paths in children's problem solving: The influence of dynamic testing, initial variability, and working memory. *Journal of Experimental Child Psychology*, *153*, 83–109. https://doi.org/10.1016/j.jecp.2016.09.004

Richardson, J. T. . (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Riffert, F. (2005). The use and misuse of standardized testing: A Whiteheadian point of view. *Interchange*, *36*(1), 231–252.

Robson, K., & Pevalin, D. (2016). *Multilevel modeling in plain language*. London: Sage Publications.

Roediger III, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *55*, 1–36. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Rosas, R. (2014). *Elementary school teachers' and principals' formative assessment beliefs, practices and assessment literacy*. California State University, Fresno.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistican*, *39*(1), 33–38.

Rubie-Davies, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology*, *76*, 429–444. https://doi.org/10.1348/000709905X53589

Rubie-Davies, C. M., Flint, A., & McDonald, L. G. (2012). Teacher beliefs, teacher characteristics, and school contextual factors: What are the relationships? *British Journal of Educational Psychology*, *82*(2), 270–288. https://doi.org/10.1111/j.2044-8279.2011.02025.x

Rust, C., O'Donovan, B., & Price, M. (2005). A social constructivist assessment process model: How the research literature shows us this could be best practice. *Assessment and Evaluation in Higher Education*, *30*(3), 231–240. https://doi.org/10.1080/02602930500063819

Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, *44*(3), 1051–1067. https://doi.org/10.1093/ije/dyv113

Sach, E. (2012). Teachers and testing: An investigation into teachers' perceptions of formative

assessment. *Educational Studies*, *38*(3), 261–276. https://doi.org/10.1080/03055698.2011.598684

Sach, E. (2015). An exploration of teachers' narratives: what are the facilitators and constraints which promote or inhibit 'good' formative assessment practices in schools? *Education 3-13*, *43*(3), 322–335. https://doi.org/10.1080/03004279.2013.813956

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(119–144).

Sahinkarakas, S. (2012). The role of teaching experience on teachers' perceptions of language assessment. *Procedia - Social and Behavioral Sciences*, *47*, 1787–1792. https://doi.org/10.1016/j.sbspro.2012.06.901

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? *The Phi Delta Kappan*, *62*(9), 631–634.

Saw, G. K. (2016). Patterns and trends in achievement gaps in Malaysian secondary schools (1999–2011): Gender, ethnicity, and socioeconomic status. *Educational Research for Policy and Practice*, *15*, 41–54. https://doi.org/10.1007/s10671-015-9175-2

Segers, M., & Tillema, H. (2011). How do Dutch secondary teachers and students conceive the purpose of assessment? *Studies in Educational Evaluation*, *37*(1), 49–54. https://doi.org/10.1016/j.stueduc.2011.03.008

Sekharan Nair, G. K., Setia, R., Abdul Samad, N. Z., Raja Zahri, R. N. H., Luqman, A., Vadeveloo, T., & Che Ngah, H. (2014). Teachers' knowledge and issues in the implementation of school-based assessment: A case of schools in Terengganu. *Asian Social Science*, *10*(3), 186–194. https://doi.org/10.5539/ass.v10n3p186

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimisation: The matching package for R. *Journal of Statistical Software*, *55*, 1–47.

Sekhon, J. S. (2009). Opiates for the matches : Matching methods for causal inference. *The Annual Review of Political Science*, *12*, 487–508. https://doi.org/10.1146/annurev.polisci.11.060606.135444

Shabani, K. (2012). Dynamic assessment of L2 learners' reading comprehension processes: A Vygotskian perspective. *Procedia - Social and Behavioral Sciences*, *32*(2010), 321–328. https://doi.org/10.1016/j.sbspro.2012.01.047

Shalberg, P. (2011). The professional educator: Lessons from Finland. https://doi.org/10.1146/annurev.genom.2.1.103

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*(1), 405–

450. https://doi.org/10.3102/0091732X019001405

Shepard, L. A. (2000). *The role of classroom assessment in teaching and learning*. *CSE Technical Report 517* (Vol. 95064). https://doi.org/10.1007/s11104-008-9783-1

Shepard, L. A. (2009). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Shin, L. W. (2015). Teachers ' assessment literacies and practices: developing a professional competency and learning framework. *Advances in Scholarship of Teaching and Learning*, *2*(2), 1–20.

Sidhu, K. G., Chan, Y. F., & Mohamad, A. (2011). Teachers' knowledge and understanding of the Malaysian school-based oral English assessment. *Malaysian Journal of Learning and Instruction.*, *8*(1), 93–115.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education*, *22*(4), 371–391. https://doi.org/10.1007/s10972-011-9231-6

Simpson, M., & Arnold, B. (1983). Diagnostic tests and criterion - referenced assessments : Their contribution to the resolution of pupil learning difficulties. *Innovations in Education & Training International*, *20*(1), 36–42. https://doi.org/10.1080/0033039830200106

Skott, J. (2015). The promises, problems, and prospects of research on teacher´beliefs. In H. Fives & M. G. Gill (Eds.), *International Handbook of Research on Teachers´Beliefs* (pp. 13–30). New York: Routledge.

Snijders, T. A. ., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd editio). London: Sage.

Sternberg, R. J. (1998). Abilities Are Forms of Developing Expertise. *Educational Researcher*, *27*(3), 11–20. https://doi.org/10.3102/0013189X027003011

Sternberg, R. J. (1999). Intelligence as Developing Expertise. *Contemporary Educational Psychology*, *24*(4), 359–375. https://doi.org/10.1006/ceps.1998.0998

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. New York: Cambridge University Press.

Stevenson, C. E., Heiser, W. J., & Resing, W. C. . (2016a). Dynamic testing: Assessing cognitive potential of children with culturally diverse backgrounds. *Learning and Individual Differences*, *47*, 27–36. https://doi.org/10.1016/j.lindif.2015.12.025

Stevenson, C. E., Heiser, W. J., & Resing, W. C. . (2016b). Dynamic testing of analogical reasoning in 5- to 6-year-olds multiple-choice versus constructed-response training items. *Journal of Psychoeducational Assessment*, *34*(6), 550–565.

https://doi.org/10.1177/0734282915622912

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan*, *86*(1), 22–27. https://doi.org/10.1177/003172170408600106

Stiggins, R. (2007). Assessment through the student's eyes. *Educational Leadership*, *64*(8), 1–38.

Stiggins, R., & Duke, D. (2008). Effective instructional leadership requires assessment leadership. *Phi Delta Kappan*, *90*(4), 285–291. Retrieved from http://www.jstor.org.cyber.usask.ca/stable/20446092

Stiggins, R. J. (1991). Assssment literacy. *Phi Delta Kappan*, *72*(7), 534–539.

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

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758–765. https://doi.org/10.2307/20440249

Stout, D. E., & Ruble, T. L. (1995). Assessing the practical significance of empirical results in accounting education research : The use of effect size information. *Journal of Accounting Education*, *13*(3), 281–298.

Stringer, P. (2018). Dynamic assessment in educational settings: is potential ever realised? *Educational Review*, *70*(1), 18–30. https://doi.org/10.1080/00131911.2018.1397900

Stringer, P., Elliott, J. G., & Lauchlan, F. (1997). Dynamic assessment and its potential for educational psychologists. *Educational Psychology in Practice*, *12*(4), 234–239. https://doi.org/10.1080/0266736960120303

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistics Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313.Matching

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155–176). Sage publications. https://doi.org/10.4135/9781412995627

Stuebing, K. K., Barth, A. E., Weiss, B., & Fletcher, J. M. (2009). IQ is not strongly related to response to reading instruction: A meta-analysis interpretation. *Exceptional Children*, *76*(1), 31–51.

Suhaili, A. (2014). *Exploring teachers' experiences on integration of higher order thinking skills (HOTS) in teaching science*. Universiti Malaysia Sarawak.

Swanson, H. L. (2011). Dynamic testing, working memory, and reading comprehension growth in children with reading disabilities. *Journal of Learning Disabilities*, *44*(4), 358–371. https://doi.org/10.1177/0022219411407866

Swanson, H. L., & Howard, C. B. (2005). Children with reading disabilities: Does dynamic assessment help in the classification? *Learning Disability Quarterly*, *28*(1), 17–34. https://doi.org/10.2307/4126971

Swanson, H. L., & Orosco, M. (2011). *Predictive validity of dynamic testing and working memory as it relates to reading growth in children with reading disabilities*. *Advances in Learning and Behavioral Disabilities* (Vol. 24). Emerald Group Publishing Limited. https://doi.org/10.1108/S0735-004X(2011)0000024003

Tabachnick, B. ., & Fidell, L. . (2013). *Using Multivariate Statistics* (6th ed.). Boston: Pearson Education.

Taras, M. (2001). The use of tutor feedback and student self-assessment in summative assessment tasks: towards transparency for students and for tutors. *Assessment & Evaluation in Higher Education*, *26*(6), 605–614.

Taras, M. (2005). Assessment - summative and formative - some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466–478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Taras, M. (2009). Summative assessment : the missing link for formative assessment. *Journal of Further and Higher*, *33*(1), 57–69. https://doi.org/10.1080/03098770802638671

Terenzini, P. T. (1989). Assessment with open eyes: Pitfalls in studying student outcomes. *The Journal of Higher Education*, *60*(6), 644–664.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., … Goldsmith, C. H. (2010). A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology*, *10*(1), 1–10. https://doi.org/10.1186/1471-2288-10-1

The Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box*. Cambridge, UK: University of Cambridge School of Education. https://doi.org/10.1017/CBO9781107415324.004

The Sport Digest. (2013). Malaysia's Deputy Prime Minister Muhyiddin Yassin: Sport molds and unifies a nation. Retrieved November 11, 2015, from thesportdigest.com

Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of research. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127–146). New York: Macmillan.

Tiekstra, M., Hessels, M. G. P., & Minnaert, A. E. M. G. (2009). Learning capacity in adolescents with mild intellectual disabilities. *Psychological Reports*, *105*, 804–814. https://doi.org/10.2466/PR0.105.3.804-814

Tiekstra, M., Minnaert, A., & Hessels, M. G. P. (2016). A review scrutinising the consequential

validity of dynamic assessment. *Educational Psychology*, *36*(1), 112–137. https://doi.org/10.1080/01443410.2014.915930

Touw, K. W. J., Vogelaar, B., Verdel, R., Bakker, M., & Resing, W. C. . (2014). Children's progress in solving figural analogies: Are outcomes of dynamic testing helpful for teachers? *Educational & Child Psychology*, *34*(1), 21–38.

Travis, J. E. (1996). Meaningful assessment. *The Clearing House*, *69*(5), 308–312. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/00098655.1996.10114327

Treagust, D. F. (2001). Diagnostic assessment in science as a means to improving teaching , learning and retention. In *UniServe Science Assessment Symposium Proceedings* (pp. 1–9).

Tuah, A. B. (2006). Improving the quality of primary education in Malaysia through curriculum innovation: Some current issues on assessment of students' performance and achievement. In *Proceedings 3rd International Conference on Measurement and Evaluation in Education (ICMEE)* (pp. 16–26).

Tzuriel, D. (2000). Dynamic assessment of young children : Educational and intervention perspectives. *Educational Psychology Review*, *12*(4), 385–435. https://doi.org/10.1023/A:1009032414088

Tzuriel, D. (2001). Dynamic assessment of learning potential. In J. Andrews, D. H. Saklofske, & H. L. Janzen (Eds.), *Handbook of Psychoeducational Assessment* (pp. 451–496). Gulf Professional Publishing. https://doi.org/10.1016/B978-012058570-0/50017-3

Tzuriel, D. (2005). Dynamic assessmet of learning potential: A new paradigm. *Transylvanian Journal of Psychology*, *Special Is*, 7–16.

Tzuriel, D. (2011). Revealing the effects of cognitive education programmes through Dynamic Assessment. *Assessment in Education: Principles, Policy & Practice*, *18*(2), 113–131. https://doi.org/10.1080/0969594X.2011.567110

Urdan, T. C., & Paris, S. G. (1994). Teachers' perceptions of standardized achievement tests. *Educational Policy*, *8*(2), 137–156. https://doi.org/0803973233

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, J. H. M. (2015). Integrating data-based decision making , assessment for Learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 324–343. https://doi.org/10.1080/0969594X.2014.999024

van Teijlingen, E. R., & Hundley, V. (2001). The importance of pilot studies. *Social Research Update*, *35*, 1–4. https://doi.org/10.7748/ns2002.06.16.40.33.c3214

van Teijlingen, E. R., Rennie, A.-M., Hundley, V., & Graham, W. (2001). The importance of

conducting and reporting pilot studies: The example of the Scottish Births Survey. *Journal of Advanced Nursing*, *34*(3), 289–295. https://doi.org/10.1046/j.1365-2648.2001.01757.x

VanRullen, R., & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences*, *7*(5), 207–213. https://doi.org/10.1016/S1364-6613(03)00095-0

Varatharaj, R., Abdullah, A. G. K., & Ismail, A. (2015). The effect of teacher autonomy on assessment practices among Malaysian cluster school teachers. *International Journal of Asian Social Science*, *5*(1), 31–36. https://doi.org/10.18488/journal.1/2015.5.1/1.1.31.36

Veerbeek, J., Hessels, M. G. P., Vogelaar, S., & Resing, W. C. (2017). Pretest versus no pretest: An investigation into the problem-solving processes in a dynamic testing context. *Journal of Cognitive Education and Psychology*, *16*(3), 260–280. https://doi.org/10.1891/1945

Vellutino, F. R., Scanlon, D. M., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers more evidence against the IQ-achievement discrepancy definition of reading disability. *Journal of Learning Disabilities*, *33*(3), 223–238.

Veloo, A., & Krishnasamy, H. N. (2017). School-based assessment in the context of secondary school physical education teachers in Malaysia. *International Journal of Social Sciences*, *3*(2), 2122–2134.

Vidacovich, C. (2015). *Measuring teachers' knowledge and use of assessment: Creating a measure as a first step toward effective professional development*. University of Denver.

Vlachou, M. A. (2018). Classroom assessment practices in middle school science lessons: A study among Greek science teachers. *Cogent Education*, *5*(1), 1–19. https://doi.org/10.1080/2331186X.2018.1455633

Vogelaar, B., Bakker, M., Elliott, J. G., & Resing, W. C. . (2017). Dynamic testing and test anxiety amongst gifted and average-ability children. *British Journal of Educational Psychology*, *87*, 75–89. https://doi.org/10.1111/bjep.12136

Vogelaar, B., & Resing, W. C. . (2016). Gifted and average-ability children's progression in analogical reasoning in a dynamic testing setting. *Journal of Cognitive Education and Psychology*, *15*(3), 349–367. https://doi.org/10.1891/1945-8959.15.3.349

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. https://doi.org/10.2307/20466661

Vygotsky, L. (2012). *Thought and language*. Cambridge, MA: MIT Press.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: MIT Press.

Warne, R. T., Li, Y., McKyer, E. L. J., Condie, R., Diep, C. S., & Murano, P. S. (2012).

Managing clustered data using hierarchical linear modeling. *Journal of Nutrition Education and Behavior*, *44*(3), 271–277. https://doi.org/10.1016/j.jneb.2011.06.013

Watzke, S., Brieger, P., & Wiedl, K. H. (2009). Prediction of vocational rehabilitation outcome in schizophrenia: Incremental prognostic validity of learning potential beyond basic cognitive performance. *Journal of Cognitive Education and Psychology*, *8*(1), 52–62. https://doi.org/10.1891/1945

West, S. G., Ryu, E., Kwok, O. M., & Cham, H. (2011). Multilevel modeling: Current and future applications in personality research. *Journal of Personality*, *79*(1), 2–50. https://doi.org/10.1111/j.1467-6494.2010.00681.x

Wiedl, K. H., Mata, S., Waldorf, M., & Calero, M. D. (2014). Dynamic testing with native and migrant preschool children in Germany and Spain, using the Application of Cognitive Functions Scale. *Learning and Individual Differences*, *35*, 34–40. https://doi.org/10.1016/j.lindif.2014.07.003

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Wiliam, D. (2013). Assessment : The bridge between teaching and learning. *Voices from the Middle*, *21*(2), 15–20.

Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, *22*(5), 537–548. https://doi.org/10.1080/0141192960220502

Woltman, H., Feldstain, A., MacKay, C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52–69. https://doi.org/10.2307/2095731

Wu, M. (2009). Issues in large-scale assessments. In *the Pacific Rim Objective Measurement Symposium*. Hong Kong. Retrieved from http://www.edmeasurement.com.au/_docs/Issues In Large-scale Assessments.pdf

Yan, Z., & Cheng, E. C. K. (2015). Primary teachers' attitudes, intentions and practices regarding formative assessment. *Teaching and Teacher Education*, *45*, 128–136. https://doi.org/10.1016/j.tate.2014.10.002

Yusup, R. (2012). *Item evaluation of the reading test of the Malaysian University English Test (MUET)*. University of Melbourne.

Zaaiman, H., van der Flier, H., & Thijs, G. D. (2001). Dynamic testing in selection for an educational programme: Assessing South African performance on the Raven Progressive Matrices. *International Journal of Selection and Assessment*, *9*(3), 258–269.

https://doi.org/10.1111/1468-2389.00178

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323–342. https://doi.org/10.1207/S15324818AME1604_4