

“Assessment for Learning” Practices Amongst the Primary School English Language Teachers: A Mixed Methods Approach

Mazidah Mohamed*, Mohd Sallehuddin Abd Aziz and Kemboja Ismail

Center for Research in Language and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Bangi, 43000, Malaysia

ABSTRACT

This study aimed to investigate the practices of the AfL strategies and the stages of assessment in pedagogy by English language teachers in primary schools. In Phase I: the Quantitative Strand, the Assessment for Learning Audit Instrument (AfLAI) was distributed to a total of 89 primary schools in the Hulu Langat district. The AfLAI results determined the cut-off points and the divergent categories to be further investigated in Phase II: Qualitative Strand. The 244 respondents were clustered into two subset participants of four higher and four lower scorers in the AfLAI who were investigated via a semi-structured interview and/or observation. The results from the AfLAI cum the cut-off points were Mean=3.7 on QCD (SD=0.74), sharing LOSC (SD=0.79) and FB (SD=0.77), and mean=3.3 on PSA (SD=0.87). From Phase I, 15 divergent categories were selected. Phase II findings on the 115 recurring categories were divided into three profiles: the higher scorers (29 categories), the lower scorers (26 categories) and the consensus (60 categories).

Consequently, the 15 divergent categories were explained in a joint display to observe the similarities and the differences of practised amongst the higher and the lower scorers in the AfLAI. The joint display affirmed that 11 categories were practiced similarly. The four differences were on the “Availability of LOs”, “Questioning strategies”, “Compliment with FB”, and “Pupils’ progress report/self-assess”. It indicated that regardless of their scores, the participants had contributed some practical

ARTICLE INFO

Article history:

Received: 29 April 2021

Accepted: 6 July 2021

Published: 21 September 2021

DOI: <https://doi.org/10.47836/pjssh.29.3.21>

E-mail addresses:

mazidah78@gmail.com (Mazidah Mohamed)

salleh@ukm.edu.my (Mohd Sallehuddin Abd Aziz)

kemboja.ismail@gmail.com (Kemboja Ismail)

*Corresponding author

approaches to the study framework, which were beneficial beyond disciplines and classroom situations.

Keywords: Assessment for learning, assessment in pedagogy, audit instrument, English language teachers, formative assessment, mixed methods, primary school

INTRODUCTION

Formative assessment (FA) practises by the English language teachers in primary schools comprised a plethora of strategies, including assessment for learning (AfL) and assessment in pedagogy. Despite being a developing theory, the concept of FA, AfL and Feedback (FB) had been ranked as one of the most effective pedagogical strategies to influence student achievement (Hattie, 2008, 2012; Hattie & Zierer, 2018, 2019).

The holistic education system in Malaysia required implementing the School-Based Assessment (SBA) policy since 2011 (Alla Baksh, 2019). The SBA policy comprised Central Assessment, School Assessment, Physical, Sports and Co-curricular Activities assessment, and Psychometric assessment (Alla Baksh et al., 2016; Ministry of Education Malaysia, 2016). The School Assessment component in SBA is the focus in this study, which combined the progress chart of individual pupils based on the FA and the Summative Assessment (SA), among the other assessments. In addition, in the primary school English language education (ELE), the revised CEFR-aligned Dokumen Standard Kurikulum dan Pentaksiran contained several suggestions

on the implementation of FA (Azman, 2016; Ministry of Education Malaysia, 2012; Mohamad Uri & Abd Aziz, 2018; Mohd Don et al., 2015).

In one of the CEFR cascades, the nine building blocks for FA were introduced, namely: sharing Learning Objectives and Success Criteria (LOSC), exemplars, starters and plenaries, deliberate practice, Questioning, Discussions (QCD), quick scans, self-assessment and peer-assessment (PSA), and Feedback (FB; Wiliam, 2018). In this study, the framework was built on the four key strategies of AfL in the Assessment for Learning Audit instrument (AfLAI; Lysaght & O'Leary, 2017) and the five stages of assessment in pedagogy (Black, 2015).

The four key strategies of AfL in the original AfLAI in this study were sharing Learning Intentions and Success Criteria, Questions and Classroom Discussions (QCD), Feedback (FB) and Peer- and Self-assessment (PSA; Lysaght & O'Leary, 2017). The original AfLAI was adapted and validated in the Hulu Langat district with modifications based on a pilot study.

The pilot study procedure was an adaptation Phase 8 in the Instrument Development and Construct Validation (Onwuegbuzie et al., 2010) procedure whereby the modified AfLAI instrument was discussed with three subject-matter experts and nine English language teachers as the validators. Then, the semi-structured AfLAI was distributed to a stratified sample among 53 English language teachers. The analysis from the 53 responses was again discussed

with the nine validators. The pilot study aimed to validate the modified AfLAI.

Among the preliminary findings, the term Sharing Learning Intentions and Success Criteria was changed to sharing Learning Objectives and Success Criteria (LOSC), and the others were maintained (Mohamed et al., 2019). In addition, the six scales in the instrument (Lysaght & O’Leary, 2013) were reduced to 5 scales (Lysaght & O’Leary, 2017) as displayed in Table 1 because none of the respondents stated that they did not know the AfL strategies. The first key strategy in the validated AfLAI for this study was how the teachers share the Learning Objectives and the Success Criteria with the pupils during the lesson, e.g., verbally or in pictures. The second key strategy was Questions and Classroom Discussions which focused on how the teachers facilitated learning using different types of questions and brainstorming. The third key strategy, feedback, allowed the teachers to assess the pupils based on their responses, whether verbally, in writing, or through gestures. Furthermore, the fourth key strategy, Peer- and Self-Assessment by the pupils, for the other pupils and themselves was dedicated to training the trait of appreciation on self-paced progress through socialisation. These four key strategies were later merged into the Implementation stage in the five stages of assessment in pedagogy (Black, 2015).

Black (2013) proposed the model of the five stages of assessment in pedagogy based on (Hallam et al., 2004), among others, in the following: (1) Clear aims, (2)

Planning activities, (3) Implementation, (4) Review and (5) Summing up. The first stage of assessment in pedagogy began with having clear aims prior to planning the activities. In the second stage, teachers were expected to plan the activities that had the potential to achieve the aims. The key strategies of AfL were mainly expected during the Implementation stage, but not limited to this stage only. The fourth stage, review of the learning, suggested using informal assessment to check the pupils’ achievement. Finally, the fifth stage, Summing up, required pupils assessment to guide the teachers in planning the next lesson. Black (2015) emphasised the perpetual interactions among the teacher and the pupils between the stages, from which the teacher could observe the assessments in the long term. The key strategies in AfL and the stages of assessment in pedagogy were to be practiced in continuous cycles in the triangle of the curriculum, pedagogy and assessment.

The definitions of FA and AfL had been debated as to whether they belong to assessment or pedagogy. Brown (2019) argued that AfL was more inclined towards pedagogy even though it contained the word ‘assessment’. The fluidity of the definition of ‘assessment’ in AfL had led to questionings on the validity of the evidence from which the pupils’ progress reports were made. This issue was previously discussed by Black (2015), whereby SA and FA were supposed to be ‘married’. In the Malaysian education system, the integration of FA, SA and other school assessments and central assessments

were prominent in the SBA assessment policy.

Current Developments in Assessment for Learning

The teachers' knowledge and understanding of FA and AfL have affected their practices to some extent (Talib et al., 2014). Many teachers preferred pencil and paper tests and high stakes examinations to FA and AfL strategies. Despite the high ranking of FA and AfL as one of the catalysts in student achievement (Hattie & Clarke, 2019), the results from high-stakes examinations were always sought after by most stakeholders. Most parents, pupils, teachers and schools preferred norm-referenced statistical results compared to the individual progress or criterion-referencing. Moreover, previous washback studies on SBA often showed that some students "were equally pessimistic about external examinations and SBA" (Alla Baksh et al., 2016, p. 1087), and some pre-service teachers indicated some barriers to the implementation of SBA (Alla Baksh et al., 2019). In other words, the examination-oriented culture could somehow affect the inclination of the stakeholders on the practices of FA or AfL in the classrooms compared to the practice of teaching to the test or other complex situations (Alderson & Wall, 1993; Tayeb et al., 2014).

In terms of implementation, workload issues were dominant in the practices of FA and AfL in SBA. The stigma of the increased workload caused by the implementation of SBA and FA was found in the UK, Brunei, Singapore, and Malaysia, among others. In the UK, some teachers named the reporting

of FA as the 'checkbox' syndrome due to the detailed requirement in keying in the pupils' progress (Marshall & Drummond, 2006). Nasri et al. (2010) found that the added paperwork in the alternative assessment was overwhelming in Brunei. In Singapore, the primary school teachers had to give lengthy and detailed written feedback to Primary 1 and Primary 2 pupils because there were no examinations for these pupils (Ratnam-Lim & Tan, 2015). Meanwhile, in Malaysia, some teachers had experienced the increased workload due to SBA while acknowledging its benefits (Mansor et al., 2013). Compared to the assessment practices in Singapore, the Malaysian teachers were more empowered to choose the assessment strategies despite having similar challenges on the increased workload (Mohamed & Abd Aziz, 2018).

The study aimed to examine the use and understanding of AfL amongst the English language teachers in primary schools using the AfLAI. It then investigated their practices in implementing AfL and assessment in pedagogy to explain any recurring similarities or differences in practices between the higher and the lower scorers in the AfLAI. The following are the research questions of this study.

Research Questions

1. What is the understanding and practices of AfL amongst the English language teachers in primary schools, based on the AfLAI?
2. What are the practices in the implementation of AfL by the higher and the lower scorers in the AfLAI?

amongst the English language teachers in primary schools?

3. What are the similarities and the differences of practices between the higher and the lower scores in the AfLAI amongst the English language teachers in primary schools?

METHODS

Phase I: Quantitative Strand

Instrument. This sequential explanatory mixed-methods study was carried out in two significant phases. In Phase I: Quantitative Strand, the AfLAI instrument was adapted

from an Irish study (Lysaght et al., 2013; Lysaght & O’Leary, 2017) with written consent via email. The modified and validated AfLAI audited four key strategies of AfL: sharing Learning Objectives and Success Criteria (LOSC=16 items), Questions and Classroom Discussions (QCD=16 items), Feedback (FB=12 items) and Peer- and Self-Assessment (PSA=14 items) using five numerical scales ascending from the labels 1=never, 2=sporadic, 3=emerging, 4=established and 5=embedded via 58 items. The following Table 1 shows the key strategies and the rating scales in the AfLAI.

Table 1
Key strategies and rating scales in the AfLAI

4 key strategies of AfL	5 rating scales	Explanation
LOSC (16 items)	5=embedded	Happens 90% of the time
QCD (16 items)	4=established	Happens 70% of the time
FB (12 items)	3=emerging	Happens 50% of the time
PSA (14 items)	2=sporadic	Happens 25% of the time
	1=never	Never happens

Procedures. The AfLAI instrument was adapted, piloted, modified and distributed to a target population of 772 English language teachers in the district of Hulu Langat. The sampling design was adapted according to the multistage cluster sampling (Babbie, 2007). The AfLAI was distributed online and in hardcopies to encourage more participation. The psychometric properties of the four scales in the AfLAI were examined via the Kaiser-Meyer-Olkin test (none statistically significant) and Bartlett’s Test of Sphericity (all statistically

significant). The cut-off points were based on descriptive statistics. Based on the results from Phase I, the cut-off point was used to identify the subset participants and the divergent categories to be investigated in Phase II.

Phase II: Qualitative Strand

Sampling. The eight subset participants among primary school English language teachers were purposively selected based on their AfLAI scores, gender, age, teaching experience and academic background. There

was only 1 participant who scored below the average mean rating in all four scales of the AfLAI. Therefore the three other lower scorers were purposively selected based on at least one lower scores in the four scales in the AfLAI.

Procedures. The nine items (four key strategies of AfL and five stages of assessment in pedagogy) in the semi-structured interview questions and the observation checklist were modified to include the 15 divergent categories from the AfLAI results. Only field notes were taken from the observations due to MoE Malaysia's audio/video recording prohibition. The participants were asked to complete a biodata and consent form prior to the interview and the observation. The investigation was halted after six interviews and 12 observations upon reaching the saturation of data. Data triangulation was made via member checking whereby the transcribed data were discussed with the participants and were only used in this study with their permission. The within-case and cross-case analyses were profiled according to the higher, lower, and general consensus.

Ethical considerations were taken by obtaining written permission from the Ministry of Education Malaysia, the Selangor State Education Department, the Hulu Langat District Education Office, the headteachers of the respective schools and the English language teachers involved in this study.

Phase I and Phase II were designed to answer RQ1 and RQ2, respectively.

RQ3 was discussed based on a tabulated joint data display from both Phases to find the similarities and the differences in the divergent categories. This research was based on the explanatory sequential mixed method procedure suggested by (Creswell, 2014a, 2014b; Creswell & Creswell, 2017; Creswell & Plano Clark, 2011; Tashakkori & Creswell, 2007). Hence RQ3 was the mixed method research question that integrated the quantitative and the qualitative phases in this study.

RESULTS AND FINDINGS

Phase I

The 244 respondents in Phase I: Quantitative Strand constituted 31.6% of the population in this study who responded to the AfLAI. Therefore, the target sample size between $n=196$ to $n=278$ out of $n=772$ could reach a 90% confidence level, 10% sampling error and .10 significance level (Cohen et al., 2007). They were predominantly women (96.3%) and bachelor degree holders (89%). Twenty-one respondents were Master's degree holders. In the preliminary study, some EL teachers complained that the AfLAI was difficult to understand. Hence it underwent back-to-back translation into Bahasa Melayu. Since the AfLAI was translated into Bahasa Melayu, language preferences were also denoted whereby 146 respondents preferred the dual-language version, 68 preferred the English language version, and 30 chose the Bahasa Melayu version. The following Table 2 displays the respondents' information.

Table 2
Respondents

Total respondents	Gender	Academic background	Language preference
244	Male (9) Female (235)	SPM (2) STPM (1) Diploma (2) Bachelor degree (218) Master's degree (21)	English language (68) Bahasa Melayu (30) Dual language (146)

Separate principal components factor analyses were run on the data to examine some of the psychometric properties of the four scales in the AfLAI. Results from the Kaiser-Meyer-Olkin (KMO) test (none statistically significant) and Bartlett's Test of Sphericity (all statistically significant) supported the application of factor analysis to all four scales. More specific, the KMO Measure of Sampling Adequacy values of LOSC=.93, QCD=.93, FB=.91, and PSA=.94 indicated that the sampling was adequate and factor analysis was useful. Bartlett's Test of Sphericity showed .00 of the significance levels, indicating that the factor analysis was useful. The factor analysis of each scale and inspection revealed one large factor with Eigenvalues ranging in size from 5.8 to 8.1. It was clear that the point at which the curve began to straighten (scree test criterion) occurred at the point of the second factor. It indicated an ideal curve in the scree plots. The analysis also revealed that the proportion of variance explained by the first factor in each scale was large and acceptable in the principal components factor analysis. The Cronbach's alpha reliabilities on LOSC=.91, QCD=.93, FB=.90 and PSA=.94 showed that the AfLAI instrument in this study was

acceptable and satisfactory. There was no instance of any case where removing an item from any scale would improve the overall reliability measure of that scale. The following Table 3 shows the analysis.

In this study, the respondents gave the highest rating on QCD (mean=3.7, SD=0.74) followed by FB (mean=3.7, SD=0.77), sharing LOSC (mean=3.7, SD=0.79), and PSA (mean=3.3, SD=0.87). All the average mean ratings in each scale indicated that the practices in implementing the four key strategies of AfL were 'emerging', moving towards 'established'. Based on this descriptive analysis, the cut-off points from each item were taken from the average mean ratings, which separated the higher and the lower scoring respondents in the AfLAI.

The results showed that the count of responses above the mean rating was: LOSC(n=126), QCD (n=131), FB (n=115) and PSA (n=131). On the other hand, the number of responses below the average mean rating were LOSC (n=118), QCD (n=129), FB (n=129) and PSA (n=113). It was observed that the cut-off point separated the higher and the lower scorers in almost equal respondents. The following Table 4 displays the summary of the descriptive analysis of the AfLAI in this study.

Table 3
AfLAI analysis

Key strategies of AfL in the AfLAI	LOSC	QCD	FB	PSA
Number of items	16	16	12	14
Cronbach's alpha reliability	.91	.93	.90	.94
Factor 1				
Eigenvalue	7.2	8.1	5.8	8.0
Percent of variance explained	45.2	50.7	48.8	57.8
Range of factor loadings	0.41-0.74	0.5-0.73	0.4-0.61	0.45-0.75
Average factor loadings	0.58	0.58	0.48	0.65

Table 4
Descriptive analysis and cut-off points

4 scales	Mean/cut-off point	SD	Value label	Higher scorers (n)	Lower scorers (n)
QCD	3.7	0.74	Emerging	131	113
FB	3.7	0.77	Emerging	116	129
LOSC	3.7	0.79	Emerging	126	118
PSA	3.3	0.87	Emerging	131	113

Based on the cut-off points, four subset respondents (ID=219, 237, 239 and 241) consistently reported higher than average mean ratings in all the four scales of the AfLAI, between 'emerging' and 'embedded'. On the other hand, only one subset respondent (ID=242) scored below the average mean rating between 'sporadic' and 'emerging'. The other three lower scorer subset respondents (ID=205,220,232) had higher scores in at least 1 to 3 scales of the AfLAI, despite having at least one or lower scores. However, they were still categorised as the lower scorers because

of the unavailability and unwillingness of the other lower scorers to consent in participating. The subset average mean rating scores are presented in the following Table 5.

The divergent items with the highest scores and the lowest scores among the higher and the lower scorers were selected as the divergent categories to be explained in this study. These 15 categories were selected because they were divergent from the identified patterns in the results. The following subsections discuss the 15 divergent items according to the

framework of this study: the four scales in the AfLAI (LOSC, QCD, FB, PSA) with the incorporation of the five stages of assessment

in pedagogy (Clear aim, Planning activities, Implementation, Review, Summing up).

Table 5

The subset average rating scores

ID	Overall score moving from...	LOSC	QCD	FB	PSA
219	Established to Embedded	4.9	4.7	4.2	4.4
237	Emerging to Established	4.2	4.1	3.7	3.9
239	Emerging to Established	3.9	4.1	3.7	3.9
241	Emerging to Established	3.7	4.1	3.8	3.5
220	Emerging to Established	3.9	3.9	4.5	3.1
205	Emerging to Established	3.4	3.5	3.6	3.5
232	Sporadic to Established	4.4	4.3	3.8	2.2
242	Sporadic to Emerging	2.9	3.1	2.8	2.0

LOSC. The overall score saw Item 5: ‘Child-friendly language is used to share LO with pupils’ and the subset score on Item 3: ‘Pupils are reminded about the links between what they are learning and the big learning picture’ were rated as Established. The least rated were Item 16: ‘Pupils are given responsibility for checking their own LOSC of lessons’ and Item 13: ‘LOs are available throughout lessons in a manner that is accessible and meaningful for all pupils’, ranked between Never to Sporadic. The divergent categories were labelled as follows: Item 3: Assessment and real-life (also relevant to the themes: Clear aim and Planning activities); Item 5: Simplified language; Item 13: Availability of LOs; and Item 16: Pupils’ responsibility (also relevant to the theme: Implementation and of course LOSC).

QCD. Item 19: ‘Questions are used to elicit pupils’ prior knowledge on a topic’ was reported as Established moving towards Embedded. On the contrary, Item 31: ‘Pupils can explain to others what they are learning’ was rated as Emerging and Item 24: ‘Assessment techniques are used to encourage teacher questioning by pupils’ was moving from Emerging to Established. The divergent categories were labelled as Item 19: Questioning strategies; Item 24: Assessment techniques; and Item 31: Pupils’ abilities (also relevant to the theme: Implementation).

FB. The highest-rated were Item 37: ‘Teacher-made tests are used diagnostically to identify strengths and needs in teaching and learning’ (between Emerging and Established) and Item 36: ‘Teachers’

praise of pupils work...is deliberately and consistently supplemented with FB that specifies the nature of the progress made' (Established). On the other hand, the lowest-rated were Item 40: 'FB focuses on one or two specified areas for improvement at any time' (between Never and Established) and Item 44: 'Pupils are provided with information on their learning on a minute-by-minute, day-by-day basis rather than the end of week/month/term' (between Emerging and Established). The divergent categories relevant to the FB, Implementation and Review themes, were: Item 36: Compliment with FB; Item 37: Teacher-made tests; Item 40: Specific FB; and Item 44: Timely FB.

PSA. Item 47: 'Lessons on new topics begin with pupils being invited to reflect on their prior learning' (between Embedded to Established), and Item 55: 'Pupils use each other as resources for learning' (between Sporadic to Embedded) were the highest-rated. Meanwhile, Item 46: 'Pupils are encouraged to record their progress using... e.g., learning logs' (Emerging) and Item 58: 'Pupils have ready access to exemplar materials showing work at different levels of achievement across a range of subject areas (between Never and Embedded) were the lowest rankings. These items were labelled in the divergent categories as Item 46: Pupils' progress report; Item 47: Pupils' reflections; Item 55: Pupils' resources; and Item 58: Differentiated exemplars, which were associated with the themes on PSA, Implementation, Review and Summing up.

The ranking of the items provided

the most divergent categories due to the large gap in different opinions among the respondents. These 15 divergent categories based on the 15 items were sometimes probed if the respondents did not mention them in the semi-structured interviews. More importantly, the 15 divergent categories represented the practices on the nine themes in the framework of this study: the four key strategies of AfL (LOSC, QCD, FB, PSA) and the five stages of assessment in pedagogy (Clear aim, Planning activities, Implementation, Review, Summing up) as displayed in column 1, Table 10. The results from Phase I were further investigated in the quest on RQ2 as follows.

Phase II

Eight English language teachers from Phase I participated in Phase II- Qualitative Strand. The purposeful sampling of the four higher scorers (ID=A1-A4) and four lower scorers (ID=B1-B4) in the AfLAI was based on their results, gender, age-range, teaching experience, academic background and mutual consent. They were located at three primary schools in the Hulu Langat district. Five were females, and the other three were males. The three teachers taught the upper primary pupils aged 10 to 11, and the rest taught lower primary pupils aged 7 to 9. The age range of the participants was between 24 to 45, with the range of teaching experience between 0 to 20 years. Five of the participants were bachelor degree holders, and the other three were two Master's graduates and a Master's candidate. The investigation was wrapped

after six interviews and 12 observations upon the saturation of data. The following Table 6 shows the backgrounds of the participants in Phase II: Qualitative Strand.

Table 6
Background of the participants

ID	Gender	Age range	Teaching experience	Academic background	Contribution
A1-A4	M (1), F (3)	25-45	2-16	Bachelor (2), Master's (2)	4 interviews, 5 observations
B1-B4	M (2), F (2)	24-45	0-20	Bachelor (3), Master's candidate (1)	2 interviews, 7 observations

After analysing the data via within-case and cross-case analyses, the 115 recurring findings from Phase II were divided into three profiles: higher scorers, lower scorers and general consensus. The general consensus comprised 60 categories practiced by both the higher and the lower scorers in the AfLAI. In addition, there were 29 categories unique to the higher scorers and 26 categories to the lower scorers. The following subtopic discusses the three profiles. The findings were framed using the four key strategies of AfL within the five stages of assessment in pedagogy, labelled as nine stages: Clear aims, Planning activities, Implementation, sharing LOSC, QCD, FB, PSA, Review and Summing up.

Profile of the Higher Scorers. The 29 practises in the i higher scorers, whether individually or collectively. For instance, one higher scorer mentioned using SMART strategies in determining the LOs, by saying: ‘*need to have this... SMART (Specific,*

Measurable, Attainable, Realistic...Time)’ as one of the strategies.

The activities suggested in the CEFR-aligned curriculum and the Scheme of Work such as ‘*...quizzes, question papers, prompts, language games...*’ were also included in the Planning Activities and Implementation stages.

It was believed that every pupil must be ‘*...in the same line...*’ in achieving the shared LO. On the other hand, one of the higher scorers prepared a sum chart to record the group achievements in a lesson.

QCD was practiced to encourage the pupils to socialize among themselves. ‘*Two-way communication between the pupils and the teachers...*’ was necessary. According to a higher scorer, ‘*...you can ask questions from minimal level to maximum level like asking Wh questions...*’, in other words, begin with the more straightforward questions, then move to the more difficult questions.

In primary school ELE, ‘*...some lower proficient pupils need extra guidance in the*

practice of giving FB'. The higher scorers provided guided thinking, a scenario, language structures and simplified sign language or gestures to encourage the pupils in giving FB. In addition, one of the higher scorers used mini-surveys so that the pupils could practice exchanging FBs among themselves.

One of the PSA strategies practised by the higher scorers was '*...Parking Lots for the pupils to see their own understanding.*' Self-assessment was indicated as an important skill. One of the suggestions on the SA strategy was checking one's work

with exemplars from the fast finishers. In peer-assessment, the whole class was instructed to check their peer's writing on the whiteboard.

The lesson Review stage was sometimes conducted via QCD or language games.

During the Summing up stage, sometimes the higher scorers recapped the new words learnt on that day or reminded the pupils of their homework or ongoing projects. In addition, the exemplars from the early finishers were sometimes displayed during the Summing up stage. The following Table 7 shows the higher scorers' profiles.

Table 7
Profile of the higher scorers

Stages/themes/categories	Frequency by the higher scorers
Clear aims	
SMART objectives	1
Implementation	
Quiz	1
Question papers	1
Prompts	1
Language games	1
Sharing LOSC	
All the pupils in the same line	1
Sum chart for SC	1
QCD	
Use questions from easy to difficult	1
Use classroom discussions to socialise	1
Ask to clarify any confusions	1
Encourage pupils to ask the teacher	1

Table 7 (Continued)

Stages/themes/categories	Frequency by the higher scorers
FB	
Teacher provides guided thinking or giving scenarios as FB	2
Focus on constructing learning culture via FB	1
FBs from pupils are used to review the lesson	1
Types of FB depending on the lesson–sign or verbal	1
Tangible FB (e.g., stickers, stamps)	2
The culture of pupils being afraid to give FB	1
Pupils are trained to give FB using language structures	1
Use survey activities to train pupils in exchanging FB	1
PSA	
An example of SA–Parking Lots to see own understanding	1
SA is important as a reflection in a pupil	1
Fast finisher asked to read the answers aloud	2
Pupils listen to the fast finisher and check their own work	2
Checking peers writing on the whiteboard	2
Review	
Review via QCD	1
Review with a game	1
Summing up	
Identify new words	2
Homework and/or on-going project instructions	2
The teacher displays an exemplar on the on-going project	1

Profile of the Lower Scorers. The 26 practises unique to the lower scorers were also beneficial to the classroom, as shown in the following Table 8.

Table 8

Profile of the lower scorers

Stages/themes/categories	Frequency by the lower scorers
Implementation	
Colour codes	1
Assessment of learning	1
AfL	1
Sharing LOSC	
Sometimes forget to share LOSC	1
Use LOSC targets for the standardised assessment	1
QCD	
Pupils need clear instructions for discussions	1
Teacher asks questions to do a perception tag on the pupils	1
Frequent questionings so pupils are engaged in the lesson	1
Questions during the teaching point in 15 minutes	1
Lesson time is affected with QCD	1
Pupils are more confident during whole class QCD	1
FB	
FB is given if pupils complete the task	1
FB is given once a week	1
FB is given based on exam papers	1
Sometimes FB is too much for weak pupils	1
PSA	
Hard for lower proficient pupils to do SA	1
Some pupils are confused with PSA	1
Time consuming to make sure each pupil practice PSA	1
Gallery Walk: a PSA strategy	2
Teacher displays a written exemplar from a fast finisher	1
Review	
Check pupils' progress in the lesson using a list	1
Review lessons near examinations	1

Table 8 (Continued)

Stages/themes/categories	Frequency by the lower scorers
Summing up	
Rewards to the pupils	1
Sharing session on the answers of the activities	2
Immediate corrections	1
Taking pictures of pupils and their art work	1

Some differences were found in the Implementation stage. For example, in a lower primary class, a lower scorer used colour codes on the whiteboard to differentiate between the notes that the pupils should read and the notes that the pupils should copy. On the other hand, another lower scorer admitted to the practices of AfL and assessment of learning as being equally important.

In the sharing LOSC stage, sometimes a lower scorer seemed to forget to practice the strategy, but he operationalised the LOSC targets for the standardised assessment.

At least six practices were unique to the lower scorers in the QCD stage. According to a lower scorer, the pupils can be more confident in whole class QCD, and they need clear instructions: *'...in a weaker class, I will use classroom discussions because by doing that, students will be more confident to answer the questions, instead of asking them one by one. When they answer in discussions or giving opinions, they would likely express themselves more freely than if they're asked one by one...'* Sometimes the teacher asked questions to do a perception tag on the pupils. During the first 15 minutes of the lesson, there should be frequent

questioning to engage the pupils. However, it was sometimes perceived that lesson time is affected when the QCD strategy was practised.

On the FB stage, the lower scorers gave FB if pupils have completed the task once a week or based on exam papers. Sometimes, FB was regarded as too much for the weak pupils.

It was considered hard for lower proficient pupils to do SA because some became confused with the PSA strategy. Moreover, it would be time-consuming to make sure each pupil practice PSA. Nevertheless, a lower scorer displayed a written exemplar from a fast finisher in the PSA strategy practice.

The Review stage was an opportunity to check the pupils' progress in the lesson using a list. Sometimes, the teacher also reviewed the lessons when the examination was around the corner.

During the Summing up stage, the lower scorers rewarded the pupils after sharing sessions on the answers to the activities. There were also immediate corrections. One of the lower scorers took pictures of the pupils and their artwork as a gesture of appreciation.

The differences between the higher and lower scorers were limited in the stages: Clear aims, Implementation, sharing LOSC, QCD, FB, PSA, Review and Summing up. Therefore, the following General consensus profile covered all nine stages, including the stage on Planning activities.

General Consensus Profile. There were 60 practises in the nine stages, which were applied by both the higher scorers and the lower scorers, as presented in the following Table 9.

Table 9

General consensus

Stages/themes/categories	Frequency by consensus
Clear aims	
One clear aim	8
Specific language skills	6
Refer to the SOW	5
Specific topic	8
Planning activities	
Consider the pupils' levels in real life	5
Link with the previous lessons	3
Lesson plan	4
Learning aids	7
Refer to the SOW	8
Implementation	
Observation	2
Explanation on the activities in the lesson	4
Homework	2
Sharing LOSC	
Tell pupils what they are going to learn	8
SC based on the levels of the pupils	3
At the end of the lesson, check LOSC achievement	3
Keywords in the lesson	3
LOSC in words for proficient pupils	4
LOSC in pictures for less proficient pupils	3
Write LO on the whiteboard	3
Use child-friendly language in LOSC for the young learners	2

Table 9 (Continued)

Stages/themes/categories	Frequency by consensus
QCD	
Use questions to check pupils' understanding	5
Use different kinds of questions	6
Have discussions because pupils like to talk	2
Can't force the pupils to use the target language 100%	2
Need to plan and prepare items for classroom discussions	2
Pupils can critic, can agree and disagree in discussions	2
Pupils need guidance in classroom discussions	3
Need to use Malay language to make sure pupils understand	2
Waiting time for the pupils to answer the questions	6
FB	
Teacher asks questions/probes to get FB from the class	4
Teacher gives specific and simple FB to individual pupils	4
Teacher gives FB in English even when pupils speak Malay	3
Differentiated FB—words, then pictures	5
Scaffolding FB using pictures, then questions, then answers	3
Peers are allowed to give FB	6
PSA	
SA—know your level and what to be done	2
PA—get multiple FB from peers	2
PSA among children need to be guided and trained	4
Pair work—pupils check on their partners	3
PSA mostly on written work (e.g., spelling)	3
Pupils do corrections based on PSA	3
SA based on worksheets or checklists	2
PSA is applicable during reading and speaking lesson	2
Review	
FB from the teacher and pupils near the end of the lesson	4
Repeat the lesson to achieve the intended objectives	4
Check individual pupils near the end of the lesson	2
Enrich, give more examples and explanations as a review	3
Complete the tasks within the review time	2
Summing up	
Discuss the moral values	2
Pupils listen, sing and chant again	3
The teacher compliments the achieved tasks	3
Repeat the language structures in the lesson	3
Reminder on the submission of tasks	2

In short, there was mutual consensus in the practices of having one Clear aim, a specific topic, refer to the SOW and tell the pupils what they will learn. In the other stages, the shared practises are as follows.

All the participants had clear aims and followed the specific topics in the textbooks and the curriculum. They also prepared daily lesson plans based on the SOW. Each aim in the lesson was targeted according to specific language skills.

In the Planning activities stage, the teachers considered the pupils' levels and used suitable learning aids according to the levels. The activities in a lesson were usually linked to the previous lessons.

During the Implementation stage, observations on the pupils were practised. The teachers also provided explanations on the upcoming activities in the lesson, including homework tasks.

To share LOSC, the teachers told the pupils about what they were going to learn verbally, in writing, or using pictures, according to their level of understanding. The Success criteria were also determined based on the pupils' levels. Near the end of the lesson, the LOSC was checked against the pupils' achievements to scan their progress and understanding.

In the QCD stage, the participants used different types of questions to check the pupils' understanding. Moreover, during the QCD stage, the pupils can socialise, criticise, agree and disagree. Since the pupils usually like to talk to each other, QCD was the preferred AfL strategy. The only issue was the use of Bahasa Melayu

instead of the target English language during QCD. Prior to a QCD activity, extra preparations were needed, such as items to start conversations and suitable learning aids. Some of the pupils needed continuous guidance throughout QCD sessions. Some of the participants resorted to Bahasa Melayu, while some maintained using simplified English to accommodate the pupils who were reluctant to communicate in the target language. Using Bahasa Melayu was argued as acceptable systematically in certain situations (Romli & Abd Aziz, 2015; Romli et al., 2021), in this case, for the young learners to learn the target language. Nevertheless, all the participants always gave optimum waiting time in getting responses and interactions with the pupils during QCD.

Regarding the FB stage, the participants mostly agreed that the pupils had to be trained not to be afraid to exchange FBs. Most of the time, the teachers would ask questions and probe the pupils to get their FB. Nevertheless, the participants gave simple and specific FB throughout the lessons to individual pupils when necessary. Written FBs were given when the participants marked their pupils' books or on the whiteboard during a lesson. One time, when the pupils could not read the words on the whiteboard and copied the wrong spelling, the teacher would use bigger fonts for the pupils to do the corrections and recall the keywords. Verbal compliments and gestures as FB from peers were also encouraged when necessary, such as whole class applause. Some of the younger pupils

appreciated tangible FBs such as stickers or stamps. Similar to the Questioning strategy, the FB strategy was also differentiated and scaffolded to consider the pupils' reactions and comprehension. There was a need to translate the FB into Bahasa Melayu for the lower proficient pupils from time to time. In cases when the pupils could not give FB even after some waiting time, their peers were allowed to chip in.

The PSA stage was defined as the pupils knowing their level and what to be done to improve their learning. Peer-assessment served a purpose to get multiple FBs from the peers in the classroom. PSA among the children needed much guidance from the teachers. One of the most frequent strategies was checking a partner's spelling as PSA, followed by self-corrections. SA was also done by giving the pupils a worksheet or a checklist of tasks with the teacher's guidance.

In the Review stage, some participants repeated the gist of the lesson, some checked the individual pupils finishing their tasks, and some exchanged FBs with the pupils. Sometimes, the teacher included enrichments with extra examples and explanations. However, most importantly, the pupils were usually required to complete the class tasks during the Review stage.

Lastly, during the Summing up stage, some participants held mini QCDs on moral values. Others instructed the pupils to repeat the language structure, the song or the chant taught earlier. The pupils were also reminded to complete their ongoing projects during the Summing up stage.

The general consensus provided some evidence that regardless of whether the teachers were higher or lower scorers in the AfLAI, they had practised the four key strategies of AfL and the five stages of assessment in pedagogy in their respective classrooms. It should also be noted that this data was based on self-reports and observations. Therefore, the data may not be perfectly accurate (Black, 2015; Lysaght, 2009), but at least some of the interactions recorded during the observations proved that as humans, we tend to be positive and choose to show our best foot forward (Mohamed et al., 2019; Ryan et al., 2012). Further discussions on the quantitative and qualitative data follow.

The Interpretation of How the Phase II Findings Explain the Phase I Results

Phase I had answered RQ1, and Phase II resonated with RQ2. Consequently, RQ3 was the mixed method research question in this study. As discussed earlier, 15 divergent categories from the AfLAI results in Phase I would be explained with qualitative data from Phase II quotes. The following joint display: Table 10 serves a purpose to answer RQ3.

From the joint display, there were 11 similarities and four differences between the quotes from the lower scorers and the higher scorers of the AfLAI. The discussion on the similarities and the differences follows.

DISCUSSION

Based on the joint display (Table 10), there were more similarities than differences

Table 10

Quotes related to the divergent categories among the lower and the higher scorers in the AfLAI

Divergent categories	Quotes
1: Assessment and real life (Similar)	
Lower AfLAI score	“...I only sometimes have one aim in my lesson, for example: ‘to be able to write a paragraph’... if they could write a paragraph by the end of the lesson, with my guidance, I’ll be happy enough.”
Higher AfLAI score	“...need to have this SMART (Specific, Measureable, Attainable, Realistic, Time) [in the objectives].”
2: Simplified language (Similar)	
Lower AfLAI score	“...it must be comprehensible by the pupils. If they do not understand it, there is no point for you to tell them.”
Higher AfLAI score	“...among lower proficient pupils, I usually don’t share the LOSC in words, but pictures...”
3: Availability of LOs (Different)	
Lower AfLAI score	“...sometimes I will share [LOSC] but sometimes I will forget that.”
Higher AfLAI score	“I will write [the LO] on the board.”
4: Pupils’ responsibilities (Similar)	
Lower AfLAI score	“I think it’s quite good to share the LO with them because they will know what they are learning and they will not feel clueless about what they will learn in the classroom. SC, I’m not sure. I will not make the students feel expected to do the SC, which means they will learn on their own and will not be pressured by the SC: what they need to do...the end of the lesson.”
Higher AfLAI score	“...before I start my lesson, I will tell my students...the objectives, then the success criteria...at the end, I will repeat, I’ll ask the pupils: they achieved the objectives? ... introduce a sum chart [for group achievements] for SC.”
5: Questioning strategies (Different)	
Lower AfLAI score	“I need to ask questions very frequently so that pupils are engaged in the lesson, and I can do a perception tag each and every time I’m giving them a new information in the classroom.”
Higher AfLAI score	“I use different kinds of questions to check back whether they could understand, depending on the pupils at that time.”

Table 10 (Continued)

Divergent categories	Quotes
6: Assessment techniques (Similar)	
Lower AfLAI score	“...assessment, it depends, if it is assessment of learning it will be examinations, tests, and if it is AfL, for a lesson to take place, it will be based on the homework that I gave them, what I’ve taught them.”
Higher AfLAI score	“...you can do many things, for example from your own observation or mini quiz, question papers, and homework. I give them certain things that may show whether they could understand or not. Playing games can be used as assessment like matching, puzzles...”
7: Pupils’ abilities (Similar)	
Lower AfLAI score	“...in a weaker class, I will use classroom discussions because by doing that, students will be more confident to answer the questions, instead of asking them one by one. When they answer in discussions, or giving opinions, they would likely to express themselves more freely than if they’re asked one by one.”
Higher AfLAI score	“...from the discussion, we could see how the pupils could understand or not, the lesson, the level, whether they have that kind of thing to know their level.”
8: Compliment with FB (Different)	
Lower AfLAI score	“I will give FB by giving them the compliments and telling their friends are doing the correct job and you need to follow their example.”
Higher AfLAI score	“I think, it’s good in terms of motivating them, but I think for constructing thinking, I think that’s way better... I think if we are focusing on constructing learning culture in them, I think, giving them a constructive FB is very important.”
9: Specific FB (Similar)	
Lower AfLAI score	The teacher then gave a clear instruction on the importance of listening carefully the second time, and gave specific warning to those who did not follow the instruction.
Higher AfLAI score	The teacher asked the pupils to refer to a specific page.
10: Timely FB (Similar)	
Lower AfLAI score	“I will give FB only if they finish the work given... because, the lack of time, I focus on them to finish the work first, so, FB is likely to be done once a week.”
Higher AfLAI score	“...we have to look and see what are the FB on certain things so that we know whether we could proceed further with the lesson or we could review back at whatever we need to review. Depends on the lesson or the activity.”

Table 10 (Continued)

Divergent categories	Quotes
11: Teacher-made test (Similar)	
Lower AfLAI score	“...assessment, it depends, if it is assessment of learning it will be examinations, tests, if it is AfL, for a lesson to take place, it will be based on the homework that I gave them, what I’ve taught them...”
Higher AfLAI score	“...some aspects we can assess using exams, some can’t be assessed via exams...”
12: Pupils’ progress report/self-assess (Different)	
Lower AfLAI score	“I will check through a checklist, give the pupils a checklist and ask them to tick the things that they have achieved.”
Higher AfLAI score	“...you know you, you know your level and what to be done... self-assessment among students, for example the Parking lot system, from where they will see how much the lessons have helped them in understanding certain area. I think SA among students should be guided... I think SA is very important as a reflection for themselves.”
13: Pupils’ reflections (Similar)	
Lower AfLAI score	“...if it is AfL, for a lesson to take place, it will be based on the homework that I gave them.”
Higher AfLAI score	The teacher made sure the pupils were able to recall the previous lesson before beginning a new lesson.
14: Pupils’ resources (Similar)	
Lower AfLAI score	During the QCD, the pupils corrected each other when necessary.
Higher AfLAI score	“...let them talk peer-to-peer”. The pupils corrected their peer who wrote on the whiteboard, when necessary.
15: Differentiated exemplars (Similar)	
Lower AfLAI score	During the writing session, the teacher showed an exemplar from the pupil who did the written task neatly.
Higher AfLAI score	In a speaking lesson, the teacher wrote language structures on the whiteboard. Some pupils used the language structures to present what they have learnt and their peers listened. Some referred to the structure, some say their own sentences.

among the practices of the higher scorers and the lower scorers in the AfLAI. This finding suggested that the inclination of self-bias in giving self-reports were possible (Ryan et al., 2012).

The eleven similarities were on: Assessment and real-life; Simplified language; Pupils' responsibilities; Assessment techniques; Pupils' abilities; Specific FB; Timely FB; Teacher-made test; Pupils' reflections; Pupils' resources; and Differentiated exemplars.

The first difference was in the Availability of LOs, whereby the lower scorer claimed that he '*...sometimes will share [the LOSC] but sometimes... forget that*'. On the contrary, the higher scorer made sure the LO was shared.

The second difference was on the Questioning strategies that involved more frequent questionings by the lower scorer. Meanwhile, the higher scorer alleged that she '*...use[s] different kinds of questions...*' to assess the pupils' understanding. This finding challenged Sardareh and Mohd Rashid (2013) data, whereby the lower primary teachers were claimed to have used questions '*...below the pupils' Zone of Proximal Development (ZPD) and did not help their thinking skills.*'

Next, the difference was on Compliment with FB. It was interesting to note that the lower scorer complimented his pupil and made the work an exemplar. On the other hand, the higher scorer decided that '*...giving them a constructive FB is very important...*' compared to complimenting them, although '*...it's good in terms of motivating them...*'.

The last difference was about the Pupils' progress report/self-assess, on which the lower scorer provided a clear-cut example of using checklists for the pupils to '*...ask them to tick the things that they have achieved....*' The higher scorer suggested more guidance for the pupils to produce their progress reports. The different strategies could be suggested for future studies.

CONCLUSION

This sequential explanatory research had framed a part of the English language teachers' practises and voices regarding their practices on the four key strategies of AfL and the five stages of assessment in pedagogy (the nine stages/themes/categories). In Phase I: Quantitative Strand, AfLAI respondents self-reported that their practises on the key strategies of AfL were *Emerging: Happens 50% of the time* in their lessons. Furthermore, 15 divergent categories were also derived from the highest and the lowest mean average rating on the 58 items. Next, the participants in Phase II: Qualitative Strand had provided in-depth data via interviews and lesson observations on the nine stages/themes/categories by providing a total of 115 recurring practises in the assessment strategies. In sequence, the 15 divergent categories from Phase I were explained with the in-depth data from Phase II to observe the similarities and the differences between the AfL and assessment in pedagogy practises among the selected different scorers in the AfLAI. These similarities determined that the higher or lower scores in the AfLAI did not affect

the participants' understanding and practises on the 11 divergent categories. On the other hand, the differences in the divergent categories became evidence for the multiple realities in the practices of the concept.

Regardless in the education or assessment system changes, these practices could always be improvised, especially when the teachers participate in Professional Learning Communities at their site as researched (Lysaght, 2009). Moreover, these practises are not limited to the English language subject per se but are applicable across disciplines and in online lessons. Nevertheless, it could be observed that all the assessment practises in SBA are not limited to examination-orientations because the higher target would be able to inculcate self-regulated learners on a lifelong basis.

ACKNOWLEDGEMENT

This article was written with the support of the supervisors, the feedback from the reviewers and the editors, and two semesters of scholarship from MyBrain15. The data were collected with the approval from the Ministry of Education Malaysia, the Selangor State Education Department, the Hulu Langat District Education Department, the headteachers and the participants. Much appreciation and thanks to all.

REFERENCES

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115-129. <https://doi.org/10.1093/applin/14.2.115>

Alla Baksh, M. A., Mohd Sallehudin, A. A., Tayeb, Y. A., & Norhaslinda, H. (2016). Washback effect

of school-based English language assessment: A case-study on students' perceptions. *Pertanika Journal of Social Sciences & Humanities*, 24(3), 1069-1086.

- Azman, H. (2016). Implementation and challenges of English language education reform in Malaysian primary schools. *3L: The Southeast Asian Journal of English Language Studies*, 22(3), 65-78. <http://doi.org/10.17576/3L-2016-2203-05>
- Babbie, E. (2007). *The practice of social research* (11th ed.). Wadsworth/Thomson.
- Black, P. (2013). Formative and summative aspects of assessment: Theoretical and research foundations in the context of pedagogy. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 167-178). SAGE Publications. <http://doi.org/10.4135/9781452218649.n10>
- Black, P. (2015). Formative assessment - an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice*, 22(1), 161-177. <https://doi.org/10.1080/0969594X.2014.999643>
- Brown, G. T. L. (2019). Is assessment for learning really assessment? *Frontiers in Education*, 4(64), 1-7. <https://doi.org/10.3389/feduc.2019.00064>
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). Routledge/Taylor & Francis Group. <https://psycnet.apa.org/record/2007-05446-000>
- Creswell, J. W. (2014a). *A concise introduction to mixed methods research*. SAGE Publications.
- Creswell, J. W. (2014b). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publications, Inc.
- Creswell, J. W., & Creswell, J. D. (2017). *Research Design: qualitative, quantitative and mixed methods approaches* (5th ed.). SAGE Publications.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

- Hallam, S., Kirton, A., Peffers, J., Robertson, P., & Stobart, G. (2004). *Evaluation of Project 1 of the Assessment is for Learning Development Programme: Support for Professional Practice in Formative Assessment – Final Report*. The Scottish Government. <https://discovery.ucl.ac.uk/id/eprint/1507418/>
- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Taylor & Francis.
- Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge/Taylor & Francis Group.
- Hattie, J., & Clarke, S. (2019). *Visible learning: Feedback*. Routledge.
- Hattie, J., & Zierer, K. (2018). *10 Mindframes for visible learning: Teaching for success*. Routledge. <https://doi.org/10.4324/9781315206387>
- Hattie, J., & Zierer, K. (2019). *Visible learning insights*. Routledge.
- Khan, A. B. M. A., Abd Aziz, M. S., & Stapa, S. H. (2019). Examining the factors mediating the intended washback of the English language school-based assessment: Pre-service ESL teachers' accounts. *Pertanika Journal of Social Sciences & Humanities*, 27(1), 51-68.
- Lysaght, Z. (2009). *From Balkanisation to boundary crossing: Using a Teacher Learning Community to explore the impact of assessment on teaching and learning in a disadvantaged school* [Master's thesis, Dublin City University]. <http://doras.dcu.ie/22527/>
- Lysaght, Z., & O'Leary, M. (2013). An instrument to audit teachers' use of assessment for learning. *Irish Educational Studies*, 32(2), 217-232. <https://doi.org/10.1080/03323315.2013.784636>
- Lysaght, Z., & O'Leary, M. (2017). Scaling up, writ small: Using an assessment for learning audit instrument to stimulate site-based professional development, one school at a time. *Assessment in Education: Principles, Policy & Practice*, 24(2), 271-289. <https://doi.org/10.1080/0969594X.2017.1296813>
- Lysaght, Z., O'Leary, M., & Ludlow, L. (2013). Measuring teachers' Assessment for Learning (AfL) classroom practices in elementary schools. *International Journal of Educational Methodology*, 3(2), 103-115. <https://doi.org/10.12973/ijem.3.2.103>
- Mansor, A. N., Leng, O. H., Rasul, M. S., Raof, R. A., & Yusoff, N. (2013). The benefits of school-based assessment. *Asian Social Science*, 9(8), 101-106. <https://doi.org/10.5539/ass.v9n8p101>
- Marshall, B., & Drummond, M. J. (2006). How teachers engage with Assessment for Learning: lessons from the classroom. *Research Papers in Education*, 21(2), 133-149. <https://doi.org/10.1080/02671520600615638>
- Ministry of Education Malaysia. (2012). *Malaysia Education Blueprint 2013-2025 (Preschool to Post-Secondary Education)*. http://www.moe.gov.my/cms/upload_files/articlefile/2013/articlefile_file_003108.pdf
- Ministry of Education Malaysia. (2016). *Quick facts 2016 Malaysia Educational Statistics*. Education Data Sector, Educational Planning and Research Division. <https://www.moe.gov.my/muat-turun/penerbitan-dan-jurnal/terbitan/buku-informasi/1586-quick-facts-2016/file>
- Mohamad Uri, N. F., & Abd Aziz, M. S. (2018). Implementation of CEFR in Malaysia: Teachers' awareness and the Challenges. *3L: The Southeast Asian Journal of English Language Studies*, 24(3), 168-183. <http://doi.org/10.17576/3L-2018-2403-13>
- Mohamed, M., & Abd Aziz, M. S. (2018). Juxtaposing the primary school assessment concepts and practices in Singapore and Malaysia. *International Journal of Engineering & Technology*, 7(3.21), 552-556. <http://doi.org/10.14419/ijet.v7i3.21.17232>

- Mohamed, M., Abd Aziz, M. S., & Ismail, K. (2019). The validation of assessment for learning audit instrument: A mixed methods approach. *3L: The Southeast Asian Journal of English Language Studies*, 25(4), 209-226. <http://doi.org/10.17576/3L-2019-2504-13>
- Mohd Don, Z., Abdullah, A. C., Abd Samad, A., Abdullah, M. H., Singh, K. K. K., Lee, B. H., Pillai, J. P. @ L., Abu Bakar, M., Salim, S., & A. Alias, M. K. (2015). *English Language Education Reform in Malaysia: The Roadmap 2015-2025*. Ministry of Education Malaysia.
- Nasri, N., Roslan, S. N., Sekuan, M. I., Abu Bakar, K., & Nor Puteh, S. (2010). Teachers' perception on alternative assessment. *Procedia-Social and Behavioral Sciences*, 7, 37-42. <https://doi.org/10.1016/j.sbspro.2010.10.006>
- Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, 4(1), 56-78. <https://doi.org/10.1177/1558689809355805>
- Ratnam-Lim, C. T. L., & Tan, K. H. K. (2015). Large-scale implementation of formative assessment practices in an examination-oriented culture. *Assessment in Education: Principles, Policy & Practice*, 22(1), 61-78. <https://doi.org/10.1080/0969594X.2014.1001319>
- Romli, N. H., & Abd Aziz, M. S. (2015). The use of Bahasa Melayu in the English language classroom by 'non-optionist' English teachers. *Procedia - Social and Behavioral Sciences*, 172(2015), 770-777. <https://doi.org/10.1016/j.sbspro.2015.01.431>
- Romli, N. H., & Abd Aziz, M. S., & Krish, P. (2021). Investigating the utilisation of the micro-functions of Bahasa Melayu by english teachers. *3L: Language, Linguistics, Literature*, 27(1), 34-46. <https://doi.org/10.17576/3L-2021-2701-03>
- Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small- and medium-scale evaluations. *American Journal of Evaluation*, 33(3), 414-430. <https://doi.org/10.1177/1098214012441499>
- Sardareh, S. A., & Mohd Rashid, M. S. (2013). Malaysian primary school ESL teachers' questions during assessment for learning. *English Language Teaching*, 6(8), 9. <https://doi.org/10.5539/elt.v6n8p1>
- Talib, R., Kamsah, M. Z., Abu Naim, H., & Abdul Latif, A. (2014). From principle to practice: Assessment for Learning in Malaysian school-based assessment classroom. *International Journal of Social Sciences and Education*, 4(4), 8. <http://ijsse.com/sites/default/files/issues/2014/v4-i4-2014-1/Paper-11.pdf>
- Tashakkori, A., & Creswell, J. W. (2007). Editorial: The new era of mixed methods. *Journal of Mixed Methods Research*, 1(1), 3-7. <https://doi.org/10.1177/2345678906293042>
- Tayeb, Y. A., Abd Aziz, M. S., Ismail, K., & Khan, A. B. M. A (2014). The washback effect of the General Secondary English Examination (GSEE) on teaching and learning. *GEMA Online Journal of Language Studies*, 14(3), 83-103. <https://doi.org/http://dx.doi.org/10.17576/GEMA-2014-1403-06>
- Wiliam, D. (2018). *Embedded Formative Assessment* (2nd ed.). Solution Tree Press.