Fairer for all: Teachers identifying and solving problems with the assessment of English language learners in the mainstream classroom

Rosemary Erlam[1]
Waipapa Taumata Rau |University of Auckland

The emphasis in classroom-based assessment on assessment for learning means that the classroom teacher now bears greater responsibility for assessment (Malone, 2008). The success of any assessment process depends on the skill the teacher brings to classroom assessment and their ability to make decisions which are appropriate for their specific learners in their particular contexts (Hill, 2017). This paper focuses on three teachers who conducted an inquiry into an aspect of assessment practice in their teaching context. These inquiries are examined through two theoretical frameworks. The first uses Hill's (2017) Teacher assessment literacy framework to illustrate how teachers' reflections on the three dimensions of oning of assessment practices. The second uses the Assessment Use Argument which outlines the steps which need to be taken to justify or challenge the use of a given assessment. The paper documents the problems that these three teachers identified at different stages of this argument sequence. Through conducting these inquiries the teachers realized that specific assessments used in the mainstream classroom were not valid when used with English language learners (ELLs). They drew conclusions and proposed adaptations, thus demonstrating that teachers can make appropriate assessment decisions to ensure fairer outcomes for these learners in their local contexts.

Key words: teacher assessment literacy, Assessment Use Argument, validity, inquiry, fairness

Introduction

With increased recognition of the importance of formative assessment in promoting student learning and raising achievement, there is a focus on the teacher, on their role

[1] Email address for correspondence: r.erlam@auckland.ac.nz

in implementing effective assessment practice and on how well equipped they may be to do this. Teachers need to make valid and reliable assessment decisions, based on evidence of learning, which are appropriate to purpose and context (Hill, 2017). With respect to the assessment of linguistically and culturally diverse learners, Leung et al. (2018) stress the importance of giving teachers a high degree of trust and autonomy. Gardner et al. (2014) also underline the role of teacher agency in effective assessment practice.

Teacher assessment literacy

The extent to which teachers may be able to make appropriate decisions about assessment practice in their specific contexts will be dependent on their level of assessment literacy. Teacher assessment literacy (TAL) has been defined as: the ability to "design, develop and critically evaluate tests and other assessment procedures" (Vogt & Tsagari, 2014, p.377). Low levels of assessment literacy are problematic because they lead to poor educational decisions with consequent problems for learning and teaching (Green, 2014). One consequence of poor assessment literacy is that teachers may adopt existing assessment practices without questioning their applicability to their own instructional contexts (Vogt & Tsagari, 2014). The risk is that there will be serious repercussions or consequences for the student in terms of learning and achievement.

There is conflicting evidence as to the standard of assessment literacy that the teacher brings to the classroom. Alderson (2005, p. 4) presents a pessimistic view of TAL, claiming that the insight teachers "could offer into achievement, progress, strengths and weaknesses is usually very limited indeed". In an ambitious study investigating the TAL of foreign language teachers across seven European countries, Vogt and Tsagari (2014) document a lack of ability to critically evaluate tests and highlight a need for training in assessment practice. The teachers in their study typically had had no training in formative assessment. More positively they found that teachers did demonstrate ability to learn on the job, although they tended to default to the practice of testing as they themselves had been tested, and there seemed to be a lack of innovative assessment practice. In another study, conducted in Singapore and involving case studies allowing for deep understanding of teachers' classroom assessment practices, Sellan (2017) found that teachers were able to carry out effective assessments. He concludes that teachers are able, under the right conditions, to take greater responsibility for assessment decisions and carry out assessments that are appropriate for their students in their specific teaching contexts. He argues that we need to give teachers autonomy when it comes to assessment, and the space to learn, encouraging them to learn from their experiences. The Tools to Enhance Assessment Literacy for Teachers of English as an Additional Language (TEAL) project, which provided assessment guidance and resources for teachers in Victorian schools, demonstrates that with support teachers can build contextualised and tailor-made assessment practices that put the learner at the centre of the process (Leung et al., 2018).

The type of professional learning support that teachers are provided with in order to develop their understanding of appropriate assessment practice needs to include opportunities for teachers to reflect on and try out, assessment activities, and then receive feedback about these (Gardner et al., 2014). The present study investigates how well teachers, after a course on assessment, taught by the author of this paper, are able to critically evaluate assessment practices in their local contexts and determine to what extent these are appropriate when used with English language learners (ELLs). It documents whether these teachers are able to design and/or implement contextualized and learner-appropriate assessment practices that are likely to lead to improved consequences for their linguistically diverse students in terms of learning and achievement (Leung et al., 2018; Sellan, 2017). A particular focus of these teachers' inquiries, and, indeed, usually the starting point, was the extent to which these assessment practices were fair and free of bias when used with their ELLs.

Fairness and equity in assessment practice

Fairness in assessment is defined primarily as a lack of bias and a concern for equitable treatment (Fulcher, 2015). Fulcher goes on to specify that an assessment must not discriminate against any subgroup of the population and that all taking part must have a similar experience. Kunnan (2014, p. 8) outlines a series of principles of fairness for assessment practice. These include the lack of "bias against all test-takers, in particular by avoiding the assessment of construct irrelevant matters". Consideration of the fairness of an assessment entails a concern for the consequences of its use (Fulcher, 2015). However, considerations for fairness and the elimination of bias must be "embedded into the earliest stages of test design" (Fulcher, 2015, p. 185) so that these consequences can be justified (Bachman & Palmer, 2010). The Assessment Use Argument (Bachman & Palmer, 2010) is a framework which holds those who use specific assessments accountable for their use by considering, along with its emphasis on assessment validity, principles of fairness and equity. The principle of fairness is specifically mentioned in both the evaluation and generalization inferences (see Tables 2 and 3). The Assessment Use Argument will be explained in greater detail below.

Theoretical frameworks

The teachers' inquiries which are presented in this paper are examined through two theoretical frameworks. These serve as lenses through which the teachers' understanding of assessment and conclusions about assessment practices are presented; they are both outlined below.

Teacher Assessment Literacy (TAL) framework

Aware of the difficulty that teachers can have in recognizing and prioritizing their needs around assessment practice, Hill designed a teacher assessment literacy framework (Hill, 2017, p. 3-4). She drew on what Fulcher (2012) proposes are the three main components of TAL, these are, practice, concepts and frameworks. Hill reinterpreted frameworks as context (Hill, 2017). An earlier version of Hill's framework was used by researchers conducting an ethnographic study of classroom-based assessment (Hill, 2012). The later version had the aim of helping teachers identify and analyse their classroom assessment practices, as a "precursor to thinking about how these might be improved" (p. 3). A second goal that Hill (2017) identified in designing the framework, and one that is more relevant to the present paper, was to validate the skills and experience that teachers bring to the assessment process (Hill, 2017). The framework was initially organised around four main questions, with a fifth question added to help teachers recognise the situated nature of classroom-based assessment. Hill (2017) points out the importance of context in shaping assessment practice and also acknowledges that assessment practices can impact on individuals, institutions and society. This understanding is important as a basis for teachers to reflect on how they might effect change. The five questions are presented below in Table 1 (Hill, 2017, p. 4). In this table Hill (2017, p.4) also shows their relationship to Fulcher's (2012) three main components of TAL listed above: practice, concepts and context.

Table 1. Main questions around which Hill's framework is structured and their relationship to Fulcher's three main dimensions of TAL (Hill, 2017, p. 4)

Fulcher's three main dimensions	**Hill's teacher assessment literacy framework**
practice	1. What do teachers do? 2. What do teachers look for?
concepts	3. What theories and standards do they use?
context	4. What are learners' understandings of assessment? 5. How does the context for teaching shape assessment practices?

While the three teachers whose assessment practices are discussed in this paper were not familiar with Hill's framework, their problematizing of assessment practice in their contexts demonstrates reflection at these three different levels of the framework. This will be demonstrated and discussed later in this paper.

Assessment Use Argument (AUA)

An Assessment Use Argument is a set of claims which justifies "linking assessment performance to interpretations and to intended uses" of the assessment results (Bachman & Palmer, 2010, p. 31). It originates from the idea of weighing evidence (Kane et al., 1999), later illustrated by McNamara's claim that the test/assessment validation process is similar to bringing evidence to trial in a law court (Green, 2014; McNamara, 2000). In building an AUA there are a number of steps one must take, each of which will justify or challenge the use of the assessment as a means of gathering evidence to support any decision that might be taken on the basis of assessment results (Bachman & Palmer, 2010; Kane et al., 1999). Each of the four steps establishes a link in the chain that builds the argument. Green (2014, p. 83) draws on Kane et al. (1999) to depict these steps as "inferential bridges". See Figure 1 below:

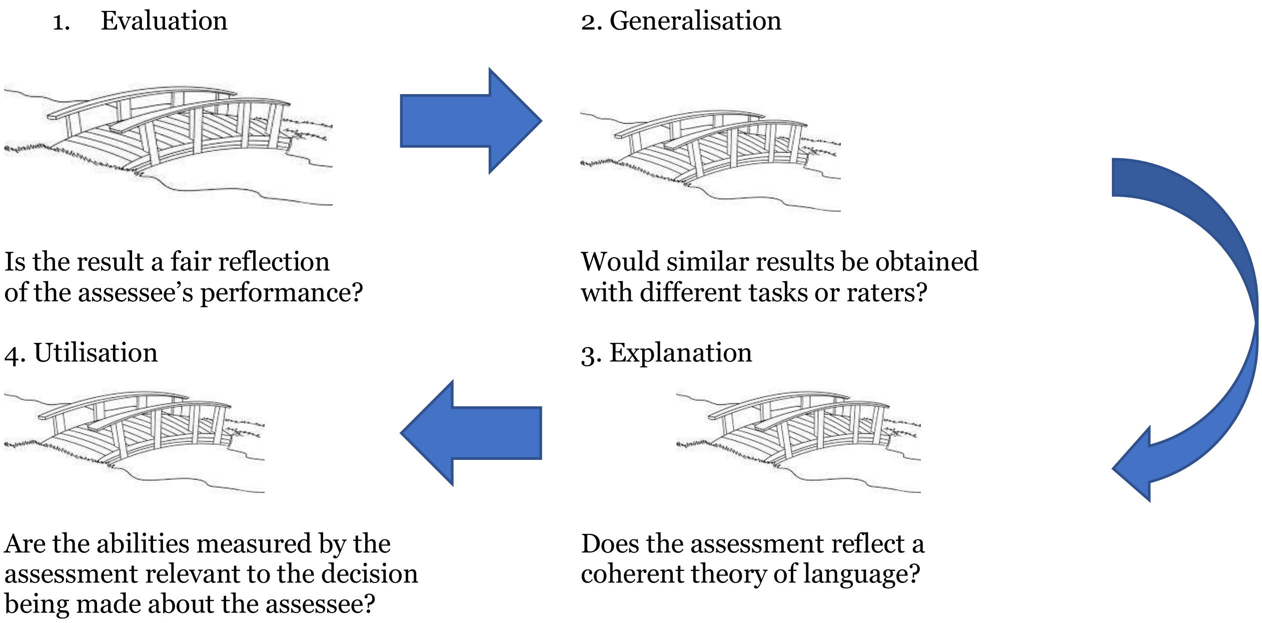
1. Evaluation
2. Generalisation
Is the result a fair reflection of the assessee's performance?
Would similar results be obtained with different tasks or raters?
4. Utilisation
3. Explanation
Are the abilities measured by the assessment relevant to the decision being made about the assessee?
Does the assessment reflect a coherent theory of language?

Figure 1. Inferential bridges in the interpretation of assessment results

More information about the claims that support the inferences made at stages 1 to 3, along with the rebuttals that would challenge or reject these claims, will be outlined below in Tables 2 to 4.

The teachers who are described in this paper were not familiar with the AUA, although as described below, the course they had completed introduced them to important concepts associated with assessment/testing. The author of this paper chose to interpret their inquiries within this framework of an Assessment Use Argument for several reasons. The first was that she wanted to see to what extent the teachers were able to demonstrate principles and choices consistent with this framework, which is influential in test validation, after a short course on language assessment. It was also a useful framework for a group of teachers who were inquiring into the use of assessments with ELLs, when these assessments had not been designed for this population, because of its incorporation of the principles of fairness and equity and its concern for the consequences of assessment use.

Each of the teachers' inquiries is described in detail below and each is used to demonstrate how claims associated with one of the first three inferential bridges, set out in Figure 1 as integral to an Assessment Use Argument, cannot be justified. The paper proposes that, in each case, the final "utilization" claim could not be made because a problem with an earlier inference meant that the argument chain was broken.

The three teacher participants and procedure

The teachers were all students enrolled in the in-service Graduate Diploma in Teaching English in Schools to Speakers of Other Languages (GradDipTESSOL) at the University of Auckland. The GradDipTESSOL (now replaced by the Postgraduate Certificate/Diploma in Teaching Linguistically Diverse Learners) was a specialized programme focusing on the theory and practice of teaching students from language backgrounds other than English. It was suitable for currently practicing early childhood, primary and secondary teachers whose students were not achieving their academic potential because of English language learning needs. It was designed both for teachers teaching mainstream curriculum content, and for those working specifically with ELLs, for example, in withdrawal contexts. The three teachers discussed in this section were in an intensive assessment course (TESSOL: Assessment EDPROFST 375), that they were taking after completion of the core course components of the GradDipTESSOL progrmme. The assessment course, taught by the author of this paper, was an elective and involved around 25 contact hours. It aimed to introduce participants to principles of effective assessment practice with particular focus on the assessment of bilingual students, raising awareness of contextual factors which might impact on the linguistic performance of these learners. Lectures had focused on key assessment concepts, including the topics listed below:

- Introduction to fundamental assessment concepts and principles/ including looking at sample assessments by reference to validity, reliability, fairness, washback
- Purposes of assessment
- Fairness in assessment
- Assessment for bi/multilingual learners
- Cultural bias in assessment
- Assessment of different language skills (reading, writing, listening, speaking)

The teachers had all been introduced to the concept of validity during the course and had had the chance, during lectures, to examine specific assessments and assessment procedures to establish whether there was sufficient evidence to claim that these were valid for use with specific populations.

One of the assignments that the teachers completed as part of their coursework, partway through the course, required them to write a critique of an assessment practice or policy that they were familiar with in their educational context. The brief included:

> Critique of assessment procedure or policy
>
> Choose a test/ assessment that you use at school that you have some concerns about . . . Analyse and critique the test or assessment using the principles that we have covered in the course (including validity). Start by explaining the purpose for which the test/assessment was developed, the intended candidates and the context in which it is being used. As part of your analysis and critique, the following topics can be covered . . . test administration and test environment; candidate instructions (rubric); clarity and simplicity of language; linguistic and cultural bias; L1/L2 consideration; validity, reliability and practicality; uses made of the test/assessment, high stakes/low stakes etc; marking, moderation, feedback. . . . You will conclude by making recommendations as appropriate.

This paper draws on the work that teachers completed for the next and final assignment, called an inquiry, and completed after the lecture component of the course was completed. This assignment, which could be informed/motivated by the work that they had done for the previous assignment, asked them to investigate an issue in depth.

The brief included:

> Special interest inquiry topic

> This is your opportunity to investigate in some depth an issue connected with the assessment of English language learners. Choose a question that is of interest to you in your classroom/school. You will present your findings to your peers on the 18th June. You will have 10 minutes to speak and 5 minutes to answer questions. You will need to prepare one hand-out for the teachers and lecturer (hand in any supporting material, for example, power point, to the lecturer). The handout should be succinct (no more than two A4 sides).

The following sections will present the issues that three teachers (referred to using pseudonyms) explored to fulfil the requirements for their inquiry assignment. This paper focuses on the inquiries of these three teachers for two reasons, firstly because of the high standard of their work (see Limitations section) and secondly because each inquiry exemplified reflection at different levels of the frameworks described above. The paper discusses each of the issues that the teachers chose to investigate through these two theoretical frameworks, as ways of explaining how teachers first identified, and then took agency, to address assessment problems in their particular contexts.

Profiling Amy and formative assessment of reading comprehension

Amy was teaching in a mainstream classroom in a South Auckland intermediate school (students in Years 7-8, aged approximately 11-12 years) where there was a high proportion of Māori and Pacific students. While she did not provide specific ethnic or linguistic information about the background of the students in her class, Māori and Pacific students in schools like Amy's would typically be overseas or New Zealand born. They could have extensive exposure to Māori or a Pacific language at home or little. For some students therefore, English could be a second language (L2) while, for others, a home or first language (L1). Familiar with research documenting the underachievement of these learners in New Zealand schools, Amy had noticed that her own informal observations and formative assessment of these students' learning often did not accord with results from English-medium reading tests, commonly used in classroom-based assessment, for example, Probe 2 Reading Comprehension Assessment (Parkin & Parkin, 2011) and e-asTTle (Ministry of Education, n.d., b). Amy also documented how, with the end, in 2018, of the requirement for schools to report on National Standards, her school had reconsidered assessment practice:

> as a school management team, we have been able to look into the possibilities of alternative assessment procedures, which increase fairness and reduce bias, encourage self-belief and create a sense of ownership for students.

At the same time, course reading, that she had completed as part of her University study, highlighted how assessment practices valued in the New Zealand educational context were based on Western theories and understanding and how these had led to the marginalisation of some students (Macfarlane, 2009). For both Māori and Pacific students, identity and relationships are valued; these learners need to positively engage in their own learning (Houghton, 2015). The cultural constructs of rangatiratanga and whanaungatanga are both from within a Māori world view (Grace, 2005); the first endorses the notion of the learner taking responsibility for and control over their own learning, while the second recognises that it is important that the learner feels part of a group and it is this sense of belonging that gives him or her agency as an individual (Macfarlane, 2009). The difference between a Western focus on the individual and the importance of the group for her Māori and Pacific students was highlighted for Amy as she became aware of a contradiction between instruction and assessment practice in her teaching context. The school had an emphasis on collaborative group work; this did not seem to accord with the emphasis on individual performance in assessment. Amy decided that this assessment practice was not culturally responsive. She gave an explanation of how she thought that cultural responsiveness would be enacted in her context.

> Teachers who are culturally responsive nurture co-operative and supportive learning environments which lead to increased student empowerment and motivation. This collective learning involves collaboration in groups and pairs and mutual accountability for the success of other members of the class learning community.

In addition, her reading on assessment practice supported the idea that assessment should be a collaborative endeavour between students and the teacher (Ministry of

Education, n.d., a). Amy believed that involving students in this process might lead to mutual accountability and encourage a sense of success and self-belief. The inquiry question that Amy asked was:

> What impact would a culturally responsive collaborative approach have on reading assessment results for Pasifika and Māori learners?

Noting that formative assessment was particularly appropriate for the assessment of Māori students (Houghton, 2015), Amy developed a reading comprehension assessment for her Year 8 students where they could work collaboratively with a buddy to read a text and then answer literal, inferential or evaluative questions. Two texts were chosen at each of two levels (NZ Curriculum level 3 or 4), and allocated to students according to their reading level. Amy divided the class into two groups, with half working collaboratively and the other half completing the assessment individually with one of the two texts. Students then completed the assessment again with the second text under a different condition. This meant that each student had the chance to complete an assessment working individually and another working collaboratively. While her research question considered results only, Amy documented in her written assignment that she also wanted to collect students' views on how they thought that they should be assessed.

While she did not report results quantitatively, Amy documented that the 'paired' reading assessment resulted in increased accuracy, elaborated responses, critical thinking and increased confidence. In terms of student preference, 62% said that they preferred to work with a partner and 66% said that they wanted to be consulted about how they were assessed, and on what.

Amy concluded that paired assessment was a more valid assessment approach given that it corresponded more closely with the emphasis in the classroom on group work. She also noted that the subjective nature of the rating, for both individual and paired assessments, could lead to some compromise in terms of reliability of scores. Interestingly, she did not discuss whether paired assessment scores were allocated to individuals or to pairs.

Amy's concern with the assessment **practice** that was being implemented in her context can be situated at Question 1 (What do you do?) of Hill's framework (p. 5). She had become aware of a mismatch between assessment and instruction, and of a disregard for learner factors such as "social, emotional and psychological attributes" (Hill, 2017, p. 5). At the same time she had realised the importance of learner involvement in assessment decisions, consulting her learners to ascertain their preferences about "how . . they will be assessed" (p. 5). The fifth question in Hill's framework ("How does the context for teaching shape assessment practices?") was also relevant to Amy's inquiry into assessment practice (Hill, 2017, p. 7). Thinking about the characteristics of her learners and their preference for collaborative learning motivated her to modify assessment practice in a way that was responsive to context.

The problem that Amy had identified was at the level of the **evaluation inference** in the Assessment Use Argument. This inference justifies a connection between performance on an assessment and the score or rating that is given as a result of this performance (see Table 2). Amy had noticed that mainstream reading test scores did not, according to her professional teacher judgement, accurately represent students' abilities. There was therefore a risk that the validity of the reading assessment results of her students may be compromised. Amy's reading had made her aware of how the beliefs and values of a dominant culture can be normalized within the education system (Houghton, 2015; Macfarlane & Macfarlane, 2012), and of how assessment needs to be "culturally responsive and . . active in acknowledging and respecting" what students value (Houghton, 2015, p. 11). One of the rebuttals to the evaluation inference is that candidates might be unwilling to perform. An additional rebuttal (underlined) that more specifically speaks to Amy's context is added to Table 2. Amy did not have direct evidence that her students were unprepared to engage in the individual reading assessment, but she did have evidence to suggest that they preferred a collaborative approach. Amy demonstrated a willingness to challenge accepted ideas of, and measures of progress, realising that they were inadequate for the students in her context.

Table 2. The evaluation inference (Green, 2014, p.84)

Inference 1	**claim**	**potential rebuttals**
Evaluation Links assessment tasks and candidate's ability to the score	The score is a fair reflection of performance on the assessment	The candidate cheated The candidate was not willing to perform or didn't understand the instructions The assessment was not scored correctly An individualised assessment mode does not reflect collaborative approaches valued by the home culture and in classroom teaching practices

Profiling Alice and the PM reading benchmarks

Alice was teaching a mainstream class of students in Years 3-4 (aged approximately 8-9 years) and had noticed that her ELLs were decoding and comprehending texts at 7-7.5 years, which was below curriculum expectations. However, she reflected on how often the texts used to assess these learners assumed cultural background knowledge that they did not have. She therefore formulated the following inquiry question:

> What is the difference in achievement on a reading comprehension assessment when using texts with known as opposed to unknown context?

Alice first chose a text about setting up an aquarium for a pet goldfish. She had previously ascertained that none of the three ELLs she was conducting this inquiry with were familiar with what a home aquarium was and that this lack of knowledge was independent of their language knowledge; a home aquarium was not something they had come across in their country of origin. She had the students first read this text independently and then she conducted the assessment using PM assessment (PM Benchmarks, n.d.) principles; students had to retell the story and then answer six literal and inferential questions.

In the next step, Alice chose three texts, one for each student, and each one containing content that she knew was familiar for this student. For example, she chose a text about a bike for a student she knew did a lot of biking. Again the learners first read the book and then completed the same assessment procedure with her when they were ready.

Alice found that with the texts that had unknown context, none of the ELLs were able to find all the information necessary to answer the literal questions fully. In answering the inferential questions, they were not able to justify their inferences, lacking understanding of unfamiliar words and phrases due to the lack of background knowledge. On the other hand, in working with texts with familiar content, the ELLs answered literal and inferential questions in detail and more confidently. They also demonstrated greater confidence and coherence in retelling their stories.

Alice's concern was also related to **practice** and relevant to the first two questions of Hill's framework. The first question requires consideration of the background knowledge of the learner and how that might need to be taken into consideration in assessment practice; Alice had realized that, because of cultural differences, her ELLs might not have background knowledge assumed by the assessment. In thinking about question two, Hill explains that teachers would need to ask themselves whether "the skills, knowledge and behaviours comprising the focus of assessment are consistent with the intended learnings (construct relevance)" (Hill, 2017, p. 9).

Alice was quite aware of the impact of task, or here, text choice, on learner performance. While she did not use the metalinguistic term construct irrelevance she had nonetheless implicitly recognized that performance could potentially "involve factors other than the intended construct" (Green, 2014, p. 230) and that these could impact significantly on the assessment process. She had been introduced to the concept of construct irrelevant variance in her prescribed readings and had been given an example in class; that of the requirement that participants completing the PM Benchmark Reading Assessment Resource orally retell what they had read and orally answer comprehension questions (Clark & Erlam, 2019). Interestingly, while she identified for this inquiry that background knowledge could constitute an example of construct irrelevant variance she did not, in her assignment, discuss the fact that the requirement for students to retell content and answer questions orally in a reading assessment was another example of construct irrelevant variance.

Alice's assessment problem meant that there was not adequate support for **the generalization inference** (see Table 3)**.** This inference rests on the assumption that the score is a true reflection of the assessees' abilities and that scores would be similar with different tasks (here texts), something that Alice was able to demonstrate was not the case at all. For the ELLs in her class, performance on the reading assessment was adversely impacted by lack of background knowledge.

Table 3. The generalization inference (Green, 2014).

Inference 1	**claim**	**potential rebuttals**
Generalization The score is a fair reflection of assessees' true abilities	Scores would be similar with different tasks, different people scoring them	The candidate got a very different score on another test of the same abilities

Profiling Abby and Structured Literacy assessment

Abby was the Reading Recovery and ESOL teacher in a primary school where approximately 45% of the students in Years 2-3 (aged 7-8 years old) were ELLs. In the year she conducted her inquiry, a Structured Literacy reading approach had been introduced to students in Years 0-3. From her discussions with teachers, Abby had established that opinions of Structured Literacy (SL) and of the usual method of teaching reading (i.e., the Three Cueing System, Goodman, 1982) tended to be very polarised.

In thinking about the assessment practices associated with each approach to reading, Abby wondered whether one was more suited to ELLs, asking the following question for her learning inquiry:

> Which is a fairer reading assessment for young English language learners – a Structured Literacy test or a Running Record?

Abby first established the main differences between the two approaches to teaching reading. She identified that Structured Literacy focuses on the development of phonological and phonemic awareness, and knowledge of the alphabetic system (often considered as a "basic literacy skills" approach). On the other hand, the three cueing system (Goodman, 1982) teaches students to attend to meaning (semantics), language structure (syntax) and visual (graphic-phonic) information as they read a text (an integrated approach to literacy). Abby discussed the relative merits of each approach, stating that decoding and word recognition were essential for reading and were recognised as important by each. She then claimed that reading was, however, considered to also involve interaction between bottom-up and top-down skills (Alderson, 2000). At this stage, Abby had concluded that there were some limitations to the theory of reading on which the Structured Literacy assessment (Stone, 2020) was based, given that reading involved a range of strategies and skills, not just decoding ability.

Abby explained that in a Structured Literacy assessment a student is asked to read aloud to the teacher a short, unsighted, one page story with minimal illustrations. A record is taken of the number of words read correctly, including the number of 'heart words' or high frequency words (HFW). Reading behaviours are observed but do not influence the results. The assessment is discontinued if the student makes three errors in a line of text. The student's responses to two comprehension questions, asked orally, provide a measure of fluency and comprehension. The assessment which corresponds to the three cuing system approach is known as a Running Record (Clay, 2000). In a Running Record assessment a student reads aloud to the teacher a text that is levelled and believed to be developmentally appropriate. Ideally, they have previously read this text. The number of words read correctly is recorded. Errors and self-corrections are noted and analysed to establish whether they are related to meaning (semantic), structure (syntax) or visual (grapho-phonic) information. Notes are made on the pace, fluency and phrasing of the reading. The student is asked to retell the story after they have finished reading it. Drawing on her knowledge of each of these different assessment approaches to reading, Abby identified the construct (using this term) of both. These are presented in Table 4.

Table 4. Comparison of the constructs of each reading assessment

	Construct	Assessment tasks/practice
Structured literacy assessment	Decoding discrete words, instant word recognition	Recording of recognition of decodable and 'heart' (non-decodable, e.g., 'my') words. Reading behaviours observed but don't influence results Two comprehension questions asked
Running record (3 cueing system)	Integrated strategies, making meaning	No of words read correctly, errors, self-corrections & other reading behaviours are recorded and analysed Notes made of pace, fluency & phrasing Discussion of the text assesses meaning-making

Abby also focused on differences in the type of text chosen for each assessment, analysing each against the descriptive categories (see Table 5) in the English Language Learning Progressions (ELLP), (Ministry of Education, 2008). She used an online tool (Compleat Lexical Tutor, n.d.) to establish differences in the level of vocabulary difficulty for each text. She looked at the percentage of words that were identified, according to this tool, as high frequency, along with the percentage of words that were not on any of the established vocabulary lists. Key differences Abby identified between the two texts are summarised in Table 5.

Table 5. An analysis of texts against descriptive categories of the English Language Learning Progressions

	Structured Literacy Text: Stage 4	Running Record Levelled Text: Level 7
Text characteristics	‘heart’ & content words not well supported by illustrations	High frequency words (HFW) & content words well supported by illustrations
Topic development	Short text – 59 words Problem is explained NOT shown	Longer text – 162 words Problem is shown & simply explained
Language structures	Mostly simple sentences	Simple sentences
Vocabulary	2-4 letter words 71% of words are HFW 29% are off list	1-8 letter words 83% of words are HFW 17% are off list
ELLP stage	1B but without illustrations and high % of off list words	1B with a low % of off list words

Abby concluded that the use of a Structured Literacy (SL) assessment and text was problematic for ELLs for a number of reasons:

- lack of prior familiarity with the text used for assessment could compromise validity.
- illustrations, which would help ELLs establish meaning, were sparse in the Structured Literacy assessment texts.
- SL texts often presented unusual language and contexts – in the SL text she chose, students had to read about ‘wigs, bugs and jugs of pop’ inside a ‘mud hut’. On the other hand, she felt that the language in the Running Record text was more natural.
- there was a lower percentage and range of HFW and a greater percentage of ‘off-list’ words in the SL text.
- self-corrections are not considered in SL assessment, yet they provide evidence to teachers showing what strategies learners are using when reading.

- the potential complexity of the reading comprehension questions used in Structured Literacy assessment could compromise any conclusions about comprehension (Clay, 2000).

Abby's assessment issue was focused on the **conceptual** dimension of teacher assessment literacy (Fulcher, 2012, p. 125) and at Question 3 (What theories and standards do they use?) of Hill's framework (p. 6). It was concerned with the nature of the subject of assessment, in this case reading. Abby was very able to evaluate and discuss her beliefs about reading and to analyse which of two reading assessment practices most closely accorded with these beliefs and with what she considered better practice for her ELL students. Abby believed that reading in a L2 is different to reading in a L1, and that ELLs need a repertoire of skills and not just decoding ability (Ministry of Education, 2008).

Abby challenged the use of Structured Literacy assessment with her ELLs at the level of **the explanation inference**, in that she did not consider that the score obtained adequately corresponded to a coherent theory of reading which could explain this performance (see Table 6). She was concerned that lower reading scores could be attributed to less developed English, rather than to poor performance in reading. Interestingly, Abby concluded in her inquiry:

> There is no single perfect test to assess reading, but, overall, it is considered that a Running Record is a fairer reading assessment for young ELLs than a Structured Literacy test.

Table 6. The explanation inference (Green, 2014, p. 84)

Inference 1	**claim**	**potential rebuttals**
Explanation The score reflects the theoretical construct the assessment is designed to measure	The results of the assessment reflect a theory of language knowledge, skills or abilities	The theory the test is based on has been discredited

Discussion

The assessment problems that the teachers, profiled in this paper, identified, correspond to four of the five main questions around which Hill's teacher assessment literacy framework is structured. Furthermore, they demonstrate reflection at all dimensions of Fulcher's three main components of TAL, that is, practice, concepts and context. Given that the question that motivated their inquiries required them to consider an assessment issue "that is of interest to you in your classroom/school", it is to be expected that the teachers actually started their reflection by engaging with Fulcher's third component of context. Furthermore, the prompt that they were 'to investigate in some depth an issue connected with the assessment of English language learners' meant that it is very likely that they also started their reflection by engaging with Hill's fifth question, that of how assessment practice is shaped by their specific context of teaching. The particular focus relevant to their teaching context, consistent with Hill's recommendation that one of the factors that teachers need to consider is student attributes, was of course the issue of the English language proficiency of their learners from other linguistic backgrounds.

Each teacher identified a problem with an assessment practice in relation to one of the first three inferential bridges which together provide the necessary links to build an argument for the use of the assessment. Their recognition that this problem meant that the assessment practice had the potential to impact negatively on ELLs, resulted of course, in the fact that the last inference, **the utilisation inference** could not be supported.

Limitations

Some caveats are in order. While this exploratory paper suggests that teachers are able to make valid assessment choices that serve English language learners in their local contexts (Sellan, 2017), it is important to remember that these teachers had all participated in an in-service course focused on assessment and also, that they were studying in a programme which emphasized the importance of linguistic and cultural responsiveness and inclusion. It is not possible to conclude that other teachers might be able to make similar assessment decisions without professional development. Furthermore, the assignments which document the inquiries these three teachers conducted into the assessment practices they chose to focus on were highly graded. In this respect, they represent best practice, and it is not possible to claim that all teachers would demonstrate these levels of assessment literacy following a course of professional development on assessment. In addition, this paper draws exclusively on these assignments as evidence of teachers' reflection upon assessment issues in their instructional contexts. More convincing and robust evidence of the teachers' conceptions and understanding of assessment practice might have been obtained from, for example, interviews. Lastly, while the teachers' proposed solutions demonstrate that innovative classroom and school-based assessment is possible, we are not able to be sure of the extent to which these solutions were implemented and/or sustained over time.

Conclusion

While none of these teachers were familiar with either Hill's framework (2017) or with the Assessment Use Argument, they were, nonetheless, able to identify problems with assessment practice in their teaching contexts. Furthermore, they were able to realise that these problems did not justify the ongoing utilisation of these practices; they therefore proposed solutions with the aim that these would make the assessments fairer for their English language learners and lead to decisions and consequences that could be justified. The teachers' conclusions support literature that claims that it is important to acknowledge teachers' expertise (Black et al., 2004) and to take account of their professional experience, perspectives and knowledge (Leung, 2005; 2014). They also corroborate one of the aims that Hill (2017, p. 12) identified in drawing up her framework, that of "validating the skills and experience" that teachers bring to assessment practice.

References

Alderson, J. C. (2000). Assessing reading. Cambridge University Press.

Alderson, J. C. (2005). Diagnosing foreign language proficiency: The interface between learning and assessment. Continuum. https://doi.org/10.5040/9781474212151

Bachman, L. F. & Palmer, A. S. (2010). Language assessment in practice. Oxford University Press.

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. Phi Beta Kappan, 86(1), 8-21. https://doi.org/10.1177/003172170408600105

Clark, J., & Erlam, R. (2019). A fairer assessment of the reading level of English language learners. SET 2019 (2), 29-41. https://doi-org/10.18296/set.0127

Clay, M. (2000). Running records for classroom teachers. Heinemann.

Compleat Lexical Tutor (n.d.) https://www.lextutor.ca/cgi-bin/vp/eng/output.pl

Fulcher, G. (2012). Assessment literacy for the language classroom. Language Assessment Quarterly, 9(2), 113–132. https://doi.org/10.1080/15434303.2011.642041

Fulcher, G. (2015). Re-examining language testing: A philosophical and social inquiry. Routledge. https://doi.org/10.4324/9781315695518

Gardner, J., Harlen, W., Hayward, L., & Stobart, L. (2014). Engaging and empowering teachers in innovative assessment practice. In J. Gardner (Ed.) Assessment by teachers. Assessment in education. Vol. 1. SAGE.

Goodman, K. S. (1982). Process, theory, research. (Vol. 1). Routledge & Kegan Paul.

Grace, W. (2005, November). He mapuna te tamaiti: Maori ecologies to support the child. Paper presented to the Commentary Group on the New Zealand Curriculum Framework Key Competencies, commissioned by the Ministry of Education, Wellington.

Green, A. (2014). Exploring language assessment and testing: Language in action. Routledge. https://doi.org/10.4324/9781003105794

Hill, K. (2012). Classroom-based assessment in the school foreign language classroom. (Language Testing & Evaluation series, Vol. 27). Peter Lang. https://doi.org/10.3726/978-3-653-01984-1

Hill, K. (2017). Understanding classroom-based assessment practices: A precondition for teacher assessment literacy. Papers in Language Testing and Assessment, 6(1), 1-17. https://doi.org/10.58379/yiwz4710

Houghton, C. (2015). Underachievement of Māori and Pasifika learners and culturally responsive assessment. Journal of Initial Teaching Inquiry, 1, 10-12.

https://ndhadeliver.natlib.govt.nz/delivery/DeliveryManagerServlet?dps_pid=IE25612749

Kane, M. T., Crooks, T. & Cohen, A. (1999). Validating measures of performance. Educational Measurement: Issues and Practice, 18(2), 5-17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kunnan, A. J. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), The companion to language assessment. Wiley. https://doi-org/10.1002/9781118411360.wbcla144

Leung, C. (2005). Classroom teacher assessment of second language development. In E. Hinkel (Ed.), Handbook of research in second language teaching and learning (pp. 869-888). Lawrence Erlbaum. https://doi.org/10.4324/9781410612700

Leung, C. (2014, October). Learning from feedback: Conception, reception and consequences. Paper presented at the TCCRISLS Conference, New York.

Leung, C., Davison, C., East, M., Evans, M., Liu, Y., Hamp-Lyons, L., & Purpura, J. (2018). Using assessment to promote learning: Clarifying constructs, theories, and practices. In J. Davis, J. Norris, M. Malone, T. McKay & Y. Sun. (Eds.). Useful assessment and evaluation in language education (pp. 75-91). Georgetown University Press. https://doi.org/10.2307/j.ctvvngrq.8

Macfarlane, S. (2009). Te Pikinga ki Runga: Raising Possibilities, Set (2), 42-50.

Macfarlane, S., & Macfarlane, A. (2012). Diversity and inclusion in early childhood education: A bicultural approach to engaging Māori potential. In D. Gordon-Burns, A. Gunn, K. Purdue & N. Surtees, (Eds.). Te Aotūroa Tātaki: Inclusive early childhood education. Perspectives on inclusion, social justice and equity from Aotearoa New Zealand, (21-38). New Zealand Council for Educational Research.

Malone, M. E. (2008). Training in language assessment. In E. Shohamy & N. Hornberger (Eds.), Language testing and assessment (pp. 225-239). Encyclopedia of language and education. Springer. https://doi.org/10.1007/springerreference_60087

McNamara, T. (2000). Language testing. Oxford University Press.

Ministry of Education (n.d., a). Assessment online. https://assessment.tki.org.nz/

Ministry of Education (n.d., b). e-asTTle. https://e-asttle.tki.org.nz/

Ministry of Education (2008). The English Language Learning Progressions. Learning Media.

Parkin, C., & Parkin, C. (2011). Probe 2: reading comprehension assessment kit. Triune Initiatives.

PM Benchmarks (n.d.) Available from https://assessment.tki.orgnz/Assessment-tools-resources/

Sellan, R. (2017). Developing assessment literacy in Singapore: How teachers broaden English language learning by expanding assessment constructs. Papers in Language Testing and Assessment, 6(1), 64-87. https://doi.org/10.58379/xgnu8346

Stone, V. (2020). Little Learners Love Literacy: Test of phonological awareness for little learners. Little Learners assessments. https://littlelearnersloveliteracy.com.au

Vogt, K. & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. Language Assessment Quarterly, 11(4), 374-402. https://doi.org/10.1080/15434303.2014.960046