The practice of assessing Norwegian and English language proficiency in multilingual elementary school classrooms in Norway

Abstract

The increasing representation of young language-minority students in school settings around the world and recent insights into multilingualism as a potential resource for language learning and development call for a critical study of theoretical and practical implications for the field of language teaching and assessment (Jessner, 2008; Ortega, 2019; Schissel et al., 2019; Shohamy, 2011). Relatively little attention has been devoted to exploring the assessment of very young students' language proficiency in the context of multilingualism. The current study explores the role of multilingualism in language assessment in the Norwegian school context. The study is based on teachers' perceptions and practices as regards the way centrally and locally mandated language assessment is carried out in EFL and Norwegian (language arts) in a multilingual elementary school in Norway. The results of the study identify factors that impact on the enactment of language assessment at the beginner level and the assumptions underlying these practices. The paper contributes to our understanding of issues of validity and social consequences in connection with assessment in a multilingual education environment (Bailey, 2017; Kane, 2016; McNamara & Ryan, 2011).

Keywords: multilingualism, very young language-minority learners, classroom-based formative assessment, Norway, ELT, majority language instruction

Vurdering av språkkompetanse i norsk og engelsk i flerspråklige klasserom på barnetrinnet i Norge

Abstrakt

Andelen unge elever med minoritetsspråklig bakgrunn øker på skoler rundt om i verden. Ny kunnskap om flerspråklighet som potensiell ressurs i språklæring og utvikling krever kritisk analyse av de teoretiske og praktiske følgene dette får for språkundervisning og vurdering

This is an Accepted Manuscript version of the following article, accepted for publication in Language Assessment Quarterly: Mona Evelyn Flognfeldt, Dina Tsagari, Dragana Šurkalović & Therese Tishakov (2020) The practice of assessing Norwegian and English language proficiency in multilingual elementary school classrooms in Norway, Language Assessment Quarterly, 17:5, 519-540, DOI: https://doi.org/10.1080/15434303.2020.1827409. It is deposited under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

(Jessner, 2008; Ortega, 2019; Schissel et al., 2019; Shohamy, 2011). Relativt lite oppmerksomhet har vært viet vurdering av de yngste elevenes språkkompetanse i flerspråklige kontekster. Denne studien bygger på lærernes forestillinger og praktiske realisering av sentrale og lokale føringer for vurdering av språkkompetanse i engelsk- og norskfaget på en flerspråklig barneskole i Norge. Resultatene i studien identifiserer faktorer som har betydning for språkvurderingen på begynnerstadiet og forestillinger lærerne gir uttrykk for omkring den praktiske gjennomføringen i klasserommet. Artikkelen bidrar til vår forståelse av validitetsproblematikk og sosiale konsekvenser i forbindelse med elevvurdering i en flerspråklig utdanningssammenheng (Bailey, 2017; Kane, 2016; McNamara & Ryan, 2011).

Introduction

A *multilingual turn* has taken root in education in the past decade, asserting "multilingualism[1], rather than monolingualism, as the new norm of linguistic and sociolinguistic analysis" (May, 2013, p. 1), and altering the key point of reference for discussing language development. It is generally recognized today that being multilingual brings various advantages, both cognitively and socially (Bialystok, 2016; Cenoz, 2003; Jessner, 1999, 2008). Researchers (Cenoz, 2003; Jessner, 2008) have called attention to the fact that the process of learning features of a *third* or later language is qualitatively different from learning what is generally referred to as a second language (L2). However, many teachers seem unaware of the potential for language learning represented by the individual's complex and dynamic multicompetence (Cook, 2016; Hofer, 2017).

In the Norwegian context, particularly in the Oslo school district, 38.6% of all students attending elementary schools have a language-minority background, ranging from close to 100% at some schools to 2% at others, with over 150 different named languages represented (Oslo Kommune, 2019; Šurkalović, 2014). This state of affairs makes every classroom in the current context linguistically and culturally diverse. Other than developing their proficiency in

[1] We are adopting the term *multilingualism* here to encompass both its general reference to societies where more than two languages are spoken, and what is often referred to as *plurilingualism*, i.e. the linguistic repertoire of individuals who know more than two languages.

the majority language (Norwegian), students start learning English as their first foreign language from the very start in elementary school. Overall, English and Norwegian are taught as separate school subjects, consistent with the traditional conceptualization of the two named languages as discrete entities. However, current research within the field of applied linguistics and multilingualism promotes a more holistic and integrated view of the linguistic system underlying individuals' language practices (e.g. Cenoz & Gorter, 2011; Otheguy et al., 2015; 2019). Ideally, if language users' linguistic systems are unitary with emerging and dynamic reservoirs of meaning-making resources, teachers should become familiar with these new insights in order to help their students make the most of their language identity development and potential for learning, their "full (linguistic) humanity" as multilingual beings (Schissel et al., 2019, p. 1).

A growing amount of research and practice in the field of multilingualism has shown that incorporating multilingual or translanguaging practices can improve minority students' linguistic achievements on a long-term basis (Canagarajah, 2013; Creese & Blackledge, 2010; Flores & García, 2013; Flores & Schissel, 2014; García, 2008; Turner, 2017). Furthermore, if *deep learning* is an educational ambition for all students, linking new content and performance to previous learning should be prompted by teachers (Kunnskapsdepartementet, 2017). This would imply continuous fostering and development of students' multilingual identities. This is an acute concern among educational linguists who want to ensure social justice (Bailey, 2017; Flognfeldt, 2018; Ortega, 2019; Shohamy, 2011, 2017).

However, the practice of effectively allowing and even encouraging students to mobilize all their linguistic resources is likely to be a challenge for many educators. Many teachers in today's classrooms have been professionally socialized into a monolingual approach and a deficit orientation when it comes to multilingualism (Simensen, 2007). Keeping languages separate is still seen as the best option by many, and having to manage more than one language is believed to be a liability or at least a potential problem. For these teachers, a necessary first step would be to fully understand what the multilingual turn implies for their professional practice.

New insights into multilingualism have inspired assessment scholars to challenge the traditional monolingual basis of assessment. This applies to assessment of academic content and language achievement at school. Some argue in favour of conceptualizing multilingualism itself as a construct (Schissel et al., 2019). Other assessment theorists raise questions about validation, most frequently in connection with educational measurement (e.g. Kane, 2016; Saville, 2019; Shohamy, 2011, 2016). Important topics are the quality of test designs,

interpretations and uses of test scores, and the consequences test results may have, be they positive or negative. In the context of linguistically and culturally diverse classrooms, unintended bias in any aspect of assessment can have negative consequences for minority students.That said, careful construal of assessment practices through validation processes might call attention to potentially negative systemic effects in formal tests (Kane, 2016, p. 203). In this study we call on, *inter alia*, concepts deriving from Cronbach (1988) and Messick (1989), and developed further by McNamara and Ryan (2011), to shed light on our context of primarily classroom-based formative assessment.

Norwegian students in elementary schools are entitled to what is referred to as "ongoing assessment."[2] In addition, various formal screening procedures are mandated by regional authorities, and teachers make use of self-made tests and other assessment techniques to check progress and diagnose appropriate instructional support for students who struggle. Our focus is mainly on *classroom-based assessment* (Turner, 2012) as it applies to *emergent multilinguals*. Interviews with our participants have given us a useful overview of their assessment practice. It is our hope that this study will provide valuable insights into contextual factors pertaining to early elementary language learning and assessment for language-minority students in the Norwegian setting, some of which are likely to be of interest to education and assessment professionals working in other world locations.

Literature Review

This study revolves around three major concerns: language assessment, multilingualism, and very young language learners. We will focus on aspects of each in turn as they are discussed in the literature, ending with a clear purpose statement and the research questions that have emerged as a result of our examinations of relevant theoretical contributions and our experience as teacher educators.

Most of the psychometrically oriented research articles on *assessment* are concerned with educational measurement, notably associated with formal and standardized tests with reference to validity issues at the level of content, criterion, construct, or a combination of these. A good deal of work is focussed on construct and content validity in relation to new or emerging needs, and the use of test accommodations in cases where validity would otherwise be jeopardized (e.g. Kane, 1990, 2016; McNamara, 2014; Schissel et al., 2018, 2019). We have chosen to approach our empirical study by drawing on McNamara and Ryan (2011) and Kane

[2] The Norwegian term *underveisvurdering* is officially translated into 'formative assessment'. This is not a perfect translation; *underveis* implies a processual approach; *formative* has a more functional meaning.

(2016), since their accounts offer validation frameworks which we think can also be applied to formative assessment, particularly their treatments of *consequential validity* that have explanatory potential for the analysis of language assessment practices in our multilingual context.

According to Kane, language tests in themselves should not be evaluated as valid or invalid. Rather, it is the interpretations stakeholders make based on test results and how these interpretations are used that can impact on validity. Interpretations give rise to decisions, and these decisions are followed by actions, which in turn have different consequences. Kane singles out three possible effects in connection with test assessments: (a) intended effects, (b) unintended effects that may have social consequences, in the form of bias or lack of fairness, also referred to as "differential impact", and (c) systemic effects (2016, pp. 202-203). One example of a negative systemic effect is when a testing program designed to support accountability leads to a practice of "teaching to the test", or when the focus on one favored attainment detracts from other curricular aims. One of the characteristics of Kane's flexible validation framework is the part of his argument-based approach that involves what he calls a "validity argument": interpretations and uses of test results must be formulated and justified as specified claims. These claims are in turn validated in terms of their coherence and plausibility with respect to appropriate empirical evidence. In other words, validation is contingent and particular. It must be anchored in the situation about which claims are made.

When assessment of language development is at stake, the construct *language proficiency* needs to be defined and delineated. The question is whether language proficiency should be understood as a unitary phenomenon or split up into various components. Scholars are divided on this issue. Hulstijn (2011), for instance, proposed a model which distinguished between "Basic and Higher Language Cognition" (BLC and HLC). In his explication of BLC, he included unconscious knowledge of various parts of the linguistic system, and explicit knowledge of vocabulary, including high-frequency words and structures that are used in communication. The fact that the ability to *use* language for communication is often deemed essential challenges Hulstijn's concentration on language proficiency in terms of traditional aspects of linguistic knowledge recognized as vocabulary, grammar, etc. The ability to use language in context needs to be included as well. As far as proficiency in English is concerned, a consequence of this wider conceptualization would involve acknowledging the use of English as a lingua franca (ELF) which can deploy lexical and grammatical resources flexibly (McNamara, 2014). The use of English resources for intercultural human interaction call for the inclusion of pragmatic skills such as the ability to negotiate meaning and accommodate

interlocutors' communicative needs. These skills would also count as aspects of language proficiency. Although Norwegian is not an international language like English, it is the main medium through which academic knowledge in most school subjects is accessed and assessed in our context. By extension, our construct of Norwegian language proficiency includes literacy skills and lexical-conceptual knowledge at progressive levels of abstraction. In school education, the way the different subject curricula define knowledge and skills helps teachers operationalize their understanding of what constitutes language proficiency at relevant stages and at different grade levels. One central difference between the development of proficiency in the majority language and an additional language in a school context hinges upon the fact that a basic vocabulary is generally already in place in someone's L1. The language proficiency construct will consequently have to be delineated differently for assessment of and for the two languages (Utdanningsdirektoratet, n.d.).

In Norway, one line of research in the field of assessment is based on evaluations of a national strategy aiming to ensure effective assessment practices and building an informed assessment culture among teachers (Utdanningsdirektoratet, 2018). A part of this national initiative focused on "Assessment for Learning" (AfL)[3]. Evaluative research in the wake of this initiative concluded with two concerns that have particular relevance for our study: (a) student participation in assessment practices was found to be minimal, and (b) a clear validity chain was found to be missing in many school contexts (Sandvik & Bruland, 2014). Sandvik (2019) explicates the concept 'validity chain' as "a chain of interpretations of curricular aims, criteria, coursebooks, tasks, student achievements, and the consequences of these" (p. 30, our translation). This point seems to echo Shohamy's (2011) conclusion that "there seems to be a lack of coordination between the two disciplines of teaching and testing" (p. 419). If assessment of and for students' learning is of the essence, this lack of coherence and alignment of teaching, learning, and assessment is a real challenge.

A lot of work in Norwegian school settings includes assessment as part of pedagogical practice (Sandvik, 2019). This takes the form of classroom-based formative assessment. It includes teacher's everyday instruction and their different ways of collecting information about their students' learning. There are many methods and techniques of assessment, ranging from brief on-the-spot checks to dynamic process-oriented feedback within different projects. Other

[3] *Assessment for Learning* is a way of enacting policy-supported formative assessment which should be integrated in classroom instruction. The intention is that learners be actively involved in all the phases of assessment (identifying criteria, setting goals, self-assessment, and peer assessment). By collecting evidence of student performance during the process of learning rather than at the end, teachers can adjust their teaching and decide what is the next step towards further learning for the students (Leung & Mohan, 2004).

forms of formative assessment include classroom observation where students are engaged in self- and peer-assessment based on negotiated criteria (Gorter & Cenoz, 2017; Kouvdou & Tsagari, 2018). In her article about assessing the language of young learners, Bailey (2017) states: «While traditional notions of validity and reliability cannot be easily applied to establishing the technical quality of formative assessment approaches, criteria for establishing the effectiveness of formative assessment in the classroom can be created, discussed, tried out, and refined» (p. 337). At first glance, it would seem that the established notions of validity and reliability do not apply in situations involving formative assessment. Rather, for classroom-based formative assessment to be effective, *criteria* need to be created and used in context. Formative assessment is now a common type of language assessment in classrooms across the world, not least because it is eminently suitable for monitoring learning processes over time. Following McNamara and Ryan (2011), we would argue that validity considerations in terms of fairness and justice, in addition to construct and content, are relevant in all forms of language assessment procedures and warrant particular attention. This is especially important when it comes to assessing language-minority students. McNamara and Ryan (2011) make a case for distinguishing between fairness and justice, connecting *fairness* to test-internal and more traditionally evidence-based validity factors, and *justice* to test-external factors like embodied social values and consequences.

Multilingualism is the most prominent "new" factor on the scene. Having to conceptualize language in a more holistic and coherent multilingual way, teachers may gradually be driven towards major changes in their cognitions and ways of teaching and assessing. Linking multilingualism and language assessment, McNamara (2012) and Shohamy (2001, 2006, 2011), among others, have expressed concerns about *educational equity*. Shohamy, in particular, has stressed the importance of updating and aligning language constructs with current views on multilingualism, acknowledging learners' multicompetence and their translanguaging potential (also in Schissel et al., 2019). In the case of young language-minority learners, their L1 proficiency may be called on as a mediating tool, thereby scaffolding their learning process. Even if they are not proficient in their home language(s), simply valuing their diverse linguacultural backgrounds may have the effect of underscoring and celebrating their multilingual identities. Shohamy offers alternative formative assessment methods, some of which are clearly relevant to young learners. One example is a bilingual writing task allowing students to mobilize more than one language in the same text and potentially including other semiotic resources. She is followed in this by Turnbull (2017), who welcomes a multimodal approach to assessment, including performance-based practical tasks "involving contextualized

language use" (p. 8). In formal standardized testing of learning, adjustments can be introduced that might take the form of translation, use of dictionaries, translanguaging, etc. (Ascenzi-Moreno, 2018; Schissel, 2014, 2015). These usually function as *test accommodations* (Abedi, 2009).

In Norway, the multilingual turn materialized half a decade ago in ELT, with a growing number of research studies and development projects addressing multilingualism and its impact on English language pedagogy. Work by Šurkalović (2014), Dahl & Krulatz (2016), Iversen (2017), Burner & Carlsen (2017), Flognfeldt (2018) and Krulatz et al. (2018) showed the need for further development of teachers' professional preparedness for work in diverse multilingual classrooms. Multilingualism was recognized as an important factor in the teaching and learning of Norwegian a lot earlier (e.g. Golden, 2006; Hvistendahl et al., 2009).

In this paper, our focus is on *very young* language-minority learners (age range 7-8). Research into the assessment of these *young* students' language learning was rather scarce up until the beginning of this century, when Penny McKay published her book *Assessing Young Language Learners* (2006). McKay attributed the new research focus to increased interest in formative classroom assessment. Butler (2016) stressed that when assessing young learners, it is important to identify age-appropriate topics, tasks, and assessment formats that align with learners' socio-cognitive developmental levels. Concern for the realities of young language learners resonates with Cameron's (2001) notion of *dynamic congruence*. This concept refers to the need to choose learning material which is appropriate to a child's lifeworld, affording language that will grow with the child and serve as a base for other purposes as the child develops.

At the beginner level of elementary school, the main objectives regarding language development in EFL center around the development of vocabulary and basic language skills, notably in relation to listening and speaking from the start, progressing to reading and writing (Utdanningsdirektoratet, n.d.). Accordingly, the language proficiency construct teachers need to assess is young learners' oral skills, their emergent vocabularies, and their command of useful sentence structures that can function as productive patterns. To meet curricular aims in Norwegian (language arts), teachers need to prepare students systematically and effectively for progressively demanding academic tasks. They also need to help them learn how to learn. At the very beginning of formal schooling, a number of general factors, for instance, limited working memory capacity, short attention span, holistic learning needs, and the need for play and physical movement, have to be taken into account.

Another concern for stakeholders is to take sufficient care to consider the vulnerability of young learners. Butler's (2016) admonition echoes Black and Wiliam's (2018) call for sensitivity in assessment work. For instance, teachers need to be responsive to what they perceive as their learners' capabilities and adjust their instruction accordingly. This aspect is vital if we are to understand the influence or washback effect of assessment on teaching and learning (Tsagari & Cheng, 2016), also recognized as a dimension of *consequential validity* (Messick, 1989; Kane, 2016). We need to know more about how assessment affects young learners' motivation, confidence, and other affective features (Butler, 2009).

As we have seen, a considerable number of critics of traditional assessment practices agree that monolingual ideologies have been dominant for a long time. Working with our growing understanding of the benefits of multilingual language use, it is time to allow language-minority students to leverage their multilingualism and use it as a resource at school for assessment as well as in language development (see Heugh et al., 2017; Sachtleben, 2015). To be able to facilitate this, a sufficient level of multilingual and assessment literacy is required in teachers (Dahl & Krulatz, 2016; Fulcher, 2012; Inbar-Lourie, 2008; Taylor, 2009; Vogt & Tsagari, 2014). In other words, teachers need to develop and sustain a repertoire of linguistically and developmentally appropriate assessment strategies.

In our study, based on the insights gained from reviewing the research into multilingualism and assessment, we wished to explore whether students in our linguistically diverse classroom are given fair and just learning conditions, and to what extent classroom-based assessment practices produce valid outcomes (Kane, 2016; McNamara & Ryan, 2011). Our aim was to gain a better understanding of this complex issue based on a case study of teaching and assessment practices in Norwegian (language arts) and English classes in 2nd grade. The following research questions informed our investigation:

RQ 1: How is language assessment enacted in English and Norwegian classes?
RQ 2: To what extent do teachers take the language background of language-minority students into account in their assessment?
RQ 3: What dimensions of validity are evidenced in the teachers' perspectives and practices?

Context and Methods

Since our intent was to conduct an exploratory in-depth study of teaching and assessment practices in a natural setting, our approach was largely qualitative (J. W. Creswell and J. D. Creswell, 2018). The strength of qualitative research is that it gives researchers an opportunity

to access participants' meanings and perspectives. By choosing the design of a case study, we were able to apply a multi-method way of collecting data, including a pre-observation survey of the participants, classroom observations, and post-observation individual and group interviews. The chosen approach aimed at triangulating data from these different sources, enabling us to answer our research questions while also enhancing the validity of our interpretations. (Turner, 2014). Relevant permissions from the Norwegian Centre for Research Data (NSD) were obtained, and all three participants gave written consent to participation. Observations took place over a two-month period at the end of 2018, with interviews conducted a couple of months afterwards. We used a purposeful sampling strategy. We picked the school for our research primarily because we knew it had a multilingual student population. Even though the 1st grade would be even closer to our wish to study language development at the very beginning of schooling, we knew that very little time is generally allocated to English in 1st grade. The principal picked out three teachers in 2nd grade and their two classrooms for our project.

We were four collaborating researchers, who all work as teacher educators in English language and language pedagogy. Our research specializations vary and include educational measurement and assessment literacy, differentiated instruction, 1st-4th elementary education with a focus on multilingualism and English. Our background as language teacher educators shape our approach to data analysis and interpretation; we were mindful of the need to be critical and to identify areas of practice that might benefit from change.

In 2nd grade at the school of our study, English is formally allocated one hour per week. Norwegian, on the other hand, was taught six hours a week, two of which were organized as learning stations[4]. Norwegian literacy holds a place of great importance in the educational system and is seen as a key to students' overall educational success. Language-minority students' proficiency in Norwegian is carefully monitored. According to the Norwegian Education Act §2-8 (Kunnskapsdepartementet, 1998), students with a mother tongue other than Norwegian or Sami[5] are entitled to specially adapted Norwegian language instruction until they are sufficiently proficient in the language to follow grade-level Norwegian-medium instruction.

[4] The organization of the classroom into learning stations means that groups of students rotate between different timed activities in the course of a lesson. For instance, one station may be a writing station, another could be a play station, a reading stations where students pick a book they want to read, and a teacher-led station devoted to guided reading or development of phonological awareness. In some cases, different school subjects are included, not just one.

[5] *Sami* is the collective name for languages spoken by the Sami population. There are three Sami languages in use in Norway. The Sami are recognized as an indigenous or autochthonous people, according to the ILO Convention, with rights safeguarded by the Norwegian constitution.

If necessary, based on screening results, bilingual subject and/or mother tongue instruction may be offered. Before language-minority children start elementary education, the school district is mandated to screen students' level of proficiency in Norwegian in order to assess whether and when students are ready to be part of mainstream Norwegian-medium instruction. Decisions are made by the municipality about individual students for one year at a time.

When it comes to *types* of assessment in elementary school, regulations relating to the Education Act (Kunnskapsdepartementet, 2006) state that students in public schools in Norway have a right to formative and final summative assessment. No grades are given in elementary school (1st-7th grade). Compulsory and standardized national tests are administered in reading, mathematics, and English in 5th grade. In addition to the policy-mandated national screening and basic skills tests, teachers make use of self-designed classroom tests for diverse formative and summative purposes.

Our research school is located in a linguistically and culturally diverse part of Oslo and has a multilingual profile slightly above the average for Oslo in terms of the percentage of students with language-minority backgrounds and students receiving Norwegian language support. We observed English and Norwegian lessons in two 2nd grade classes. One class had 23 students, the other 24, all around seven years of age. Seven students in each class were registered as language-minority students based on initial screening. According to the teachers, many speak Norwegian with their parents, and all except two were born in Norway. Schools receive additional funding to provide adapted language education based on screening results.

Three teachers participated in our study. Lars was responsible for the planning and teaching of English in both classes, using the same teaching sequences. He did not hold formal qualifications in English language teaching. Ida and Kate collaborated in planning the Norwegian lessons and used the same lesson plans and learning materials. They both have training in how to use assessment tools for reading development. Below is an overview of the teachers' professional experience, functions, formal qualifications, and amount of in-service training in testing/assessment (see also Appendix 1).

Table 1 Overview of the participating teachers

Pseudo-nyms	**Elementary school experience**	**Teaching experience E/N**	**Lessons per week**	**Functions**	**Formal qualification: ECTS**	**Training testing/ assessment**
Lars *(male)*	1 year, 3 months	5 months (E)	1 x 2 (E)	Resource teacher (N)	0 (E) TE and pre-school education	None

Kate *(female)*	19 years	3 years (E) 16 years (N)	6 (N)	Lead classroom teacher of 2B group	TE 1-10 60 ECTS (N)	Course in assessment/screening of reading development
Ida *(female)*	1 year, 3 months	1 year (E)	7-8 (N)	Lead classroom teacher of 2A group Resource teacher (E)	TE 1-7 (N, E) 30 ECTS (N) 30 ECTS (E)	Course in assessment/screening of reading development

(E=English, N=Norwegian, TE=Teacher Education; ECTS=credits according to the European Credit Transfer System standards)

The school allowed for a variety of teaching support in the classroom, including a teaching assistant, a language specialist, and extensive use of team teaching.

Observations give researchers first-hand experience of enacted pedagogical practice. In the classrooms, we were non-participants, but in some cases, students and teachers made occasional reference to our presence. Observation data from altogether 22 lessons were collected and typed out, comprising 16 Norwegian and 6 English lessons. In the majority of cases, two researchers observed together taking fieldnotes, which were later checked and compared for systematicity. We started out using an observation protocol template, intended to facilitate the recording of the lessons as they unfolded, with a column for recognizable assessment events. Due to the assessment practices being heavily integrated in all parts of the lessons, the procedure changed to writing down as many details as possible about what went on, identifying salient moments of assessment later during the typing out of our notes.

Given the exploratory nature of our study, the research tool that was chosen for the collection of teachers' perceptions about their assessment practices and strategies was interviews. Cohen et al. (2007) note that interviews allow for exploration of participants' meanings in great depth in comparison to other methods of data collection. We adopted semi-structured individual interviews (II) using an interview guide. There were two rounds of interviews. The initial interviews with the participants individually took place after the classroom observations, two researchers being present each time. Afterwards, the three teachers were interviewed together as a group (GI), with all four researchers present (see interview guides, Appendices 2 and 3). The second interview enabled us to ask follow-up questions about the teachers' assessment practices and other local policy requirements. The interviews were conducted in Norwegian. This enabled the participants to express their perspectives more freely. All interviews were audio-recorded and transcribed. The reliability of the interview questions was attained by means of a piloting round (Silverman, 1993) with a small sample of

English teacher educators. Samples of classroom assessment tools and teaching materials were supplied by the teachers.

Assigning codes to our observation data involved multiple steps of analysis. From the start, the observation data were coded inductively, with categories emerging from our interpretations of the observed activities. More activity features than those directly associated with assessment were included at first. This process helped us see the complex reality of the classroom. Later steps produced more refined categories as determined partly by the research questions. The interpretations based on the analysis of the classroom data helped us develop the interview questions for the teachers, more subject-specific ones for the individual interviews and more general and policy-oriented ones for the group interview. Some codes in this part of the study were related to the research questions and thus deductive. Critical incidents (Cohen et al., 2007) in the data, that is, instances that typified or illuminated a particular aspect in the teachers' behavior or instructional style, were identified and coded in both the observation and interview materials. Similarly, working with recent research on language assessment and general assessment theory, we approached our data with an enquiring mind, noticing unexpected actions and utterances, and becoming more sensitive to what our own reflections might imply. This is arguably one of the strengths of a qualitative approach involving iterative steps of data analysis.

Findings

This study has given us a picture of the rich variety of assessment events that played out during the eight weeks of classroom observations of Norwegian and English lessons. One striking observation was that classroom management played a prominent part in teacher discourse throughout. The teachers had a repertoire of routine responses to help students find their seats, keep quiet when instructions or messages were given, etc. Behavioral rules were often sanctioned by points given to or taken away from the students, who were seated in groups around color-coded tables. From an observer's point of view, these frequent managerial actions/reactions were interruptions in learning sequences, but for these very young students, acquiring classroom routines is a vital part of their social learning. The affective climate in the two classrooms was characterized by patience, support, respect, warm relations, and routines.

Our study aimed to identify what assessment activities were employed and how far these assessments were supportive of learning English and Norwegian. Our intent was also to find out to what extent the language backgrounds of the 2nd grade language-minority students were

taken into account by the teachers in their assessment practice. A final aim was to see what aspects of assessment validation could be recognized in the teachers' perspectives and practices. In the coding of the observation data, their classroom practices were to some extent unavoidably filtered through our researcher perspectives, since what we studied was non-contrived classroom actions that can be linked to teacher cognition and decisions. In the interviews, the participants were offered an opportunity to express their views more directly.

The presentation of our findings takes two forms: The results from our analysis of observation data generally take the form of vignettes, short selected scenes from the classroom, whereas interview data are represented as quotes. These teacher utterances were translated by us, with false starts, hesitations, affirmative noises, etc. left out.

Language assessment in the Norwegian and English classes (RQ 1)

All through elementary school in Norway, students are expected to experience formative assessment. As mandated by the Oslo school district, our school had begun a development project focusing on *Assessment for Learning* (AfL) as a school-wide strategic initiative. The teachers received in-service training on these topics and were encouraged to use formative assessment with increased student participation. As Table 2 below demonstrates, teachers made use of a variety of assessment activities. Some of these were mentioned in the interviews; others were observable in lessons. When asked how they conceptualized AfL, the teachers mentioned their wish to include students more when defining criteria for different tasks. They wished to learn more about this learning-centred way of assessing proficiency, not least in order to verbalize practices Ida said they might already be enacting. A characteristic that turned up frequently in their discourse about assessment was "systematic". For instance, Lars described his assessment methods as somewhat impressionistic. He kept referring to his practice as involving "getting a feel" for where the students stood. His plan was to engage in more systematic record-keeping.

Table 2 Overview of assessment activities

Standardized tests (timed)	**Dynamic screening tool[6]**	**Teacher-made assessments (tests)**	**Learning-oriented feedback to students**	**Self-assessment tasks**	**Student-response techniques**

[6] We are using the descriptive label "dynamic screening tool " here about a particular tool created for formative monitoring of reading development over time, seen as the interplay between elements like phonological awareness, decoding of words, reading fluency, reading comprehension, and reading interest (Lundberg & Herrlin, 2008; Michaelsen, 2018). We are aware that the term is used by other scholars with reference to a test-intervention-retest design (e.g. Poehner & Infante, 2016).

N: **Initial** screening test of language proficiency: *NSL* [=Norwegian as a language for learning] N: **Annual** national screening tests (Norwegian): Intended to measure reading comprehension, vocabulary, orthography E: Optional, but mandated by Oslo school district: Listening- and reading test **in 3rd grade**	N: *Good Reading Development*: Gradual screening of aspects of phonological awareness and other aspects of reading (word decoding, reading fluency, reading comprehension, Interest in reading)	N: Word/concept graphic organizers N: Dictation E: "Simon Says" (Action-based response to cues: receptive focus) E: "Checks" - often self-assessed E: Greeting every student personally at the door (diagnostic) E: Learning ticket[7] (based on communicated aims and criteria)	N: Monitoring writing tasks by means of active prompts N: Process-oriented differentiated feedback in longer projects based on shared criteria E: Immediate comments from T based on monitoring of individual tasks in the classroom G: Open questions	N: Self-assessment tasks based on criteria (for instance, letter shapes) G: Medals - gold – silver – bronze based on student-made achievement targets	N: 3-2-1 (three things I learned, two things I found exciting, one that I did not understand N: "Tweet game" Signal at unknown words in oral text G: Red-yellow-green (droplets, bricks)[8] G: Thumbs (up, to the side or down) G: Check-out (post-it notes, exit pass)

(E=English, N=Norwegian, T= teacher, Ex=example, G=in general, no distinction made between the two languages)

During our observations in the classrooms, we were at times wondering whether the assessment we saw playing out did in fact assess *language learning* or functioned more as brief on-the-spot responses of a more affective kind. Thumb signals would, for instance, show whether the students *liked* an activity or whether they found it easy/difficult, or something in between. Kate commented herself that some of the feedback techniques were more social than learning-focused. However, we have to juggle with learning outcomes, the age and maturity of the students, their affective needs and age-appropriate behavioral characteristics in our reflections.

[7] Each learner is given a piece of paper in the shape of a ticket with, for instance, five "I know…" or "I can…." statements. The learners take the learning tickets home with them and make a parent or carer check their performance and then sign the ticket.

[8] This refers to a self-assessment scheme mirroring traffic lights, where showing the color red signals that the learner finds a task or content too difficult, yellow means that they struggle to grasp it, and green signals that they find it easy and manageable.

Sustaining motivation and engagement is essential and necessary in the flow of teacher-student interaction.

In the group interview, we asked the teachers about feedback they gave in Norwegian and English (see question 3a in Appendix 3). Their responses were interesting, as they referred to students' comments on the effects of their instruction, rather than to examples of how they themselves gave feedback on and for learning to the students.

Language backgrounds and assessment (RQ 2)

In the interviews, the teachers acknowledged that the language-minority students differ from the others in specific aspects of learning. Kate had witnessed a huge gap between the two groups when it comes to specific elements in Norwegian language proficiency: prepositions, vocabulary range, synonymy, salient speech sounds (notably the vowel /ʏː/), and sentence structure. Differences in lexical comprehension were detected as well. Some of these challenges had been identified during classroom observation; other teacher statements were based on different screening results.

The teachers saw their own lack of knowledge about the structure of the different home languages in their classes as a disadvantage when it comes to helping learners and their parents make the most of multilingual resources: "One could have used the students' strengths a bit more. This is perhaps what is often missed" (Ida, II, 31:23). Ida had observed that the language-minority students were quicker than the other students in her class to learn new concepts. Her own explanation was that it is easier to connect new words to meanings that are familiar in another language than to learn something completely new.

When it comes to learning English, both Kate and Ida commented that most of the learners were presumably in the same position, since English was a new language to them. When teaching single words, according to Ida, "if you were to translate from Russian into English, or from Japanese into English, or from Norwegian into English, this would in a way be equally difficult" (Ida, II, 31:50). She goes on to say that when the teachers give explanations, work with themes in English, or discuss word classes, "we use many words to explain things, then we have to be more aware, because then of course it is easier for the *Norwegian* students to understand" (Ida, II, 32:03). Kate experienced a revelation when she realized that linking English vocabulary to Norwegian equivalents did not necessarily benefit Urdu-speaking students: "it has to go from English to Norwegian; we don't translate into Urdu…They don't understand it, because they don't understand the Norwegian word" (Kate, II, 29:10).

The strategies reported by the teachers vis-á-vis language-minority students who struggled were the following: In the screening situations, they chose not to offer accommodations so as not to compromise test results. However, in their follow-up activities, they adapted and used the tests diagnostically by either supplying explanations or letting students do the test again without the time limit. As a result of this diagnostic assessment, they used flexible grouping of 3-4 students to provide concrete needs-based instruction. This was possible since the general practice in 1st and 2nd grade is team-teaching with an additional teaching assistant in most lessons.

Kate argued in favor of keeping students whose Norwegian vocabulary is limited in the mainstream classroom rather than adhering to a general pull-out practice for these learners. They have a strong need for rich language input, so a *contingent* needs-based pull-out option seemed to be more effective. The teachers used various differentiation and accommodation strategies in class. Kate was sensitive to the fact that students' vocabularies differed. She strategically named most of the objects in thematic pictures when starting a new topic. Ida made use of interactional modification by adjusting her talking speed and simplifying her sentences (Oh, 2001). She also intentionally used elaboration techniques like rephrasing and repetition. When communicating with language-minority students, she would be concerned about the accuracy of her own language production, using correct articulation and complete sentences and recasts when learners made mistakes that matter.

Validity considerations (RQ 3)

From the point of validity in the assessments referred to in this study, several participant reflections and observed incidents spring to mind. These have to do with content, construct, criterion, and consequential validity.

The overall learning situation is a case in point once we acknowledge that multilingualism is an asset, and that students who are multilingual would benefit from being allowed to capitalize on the linguistic resources they possess. We saw that the teachers recognized the potential for learning that comes with knowing more than one language. However, with the overwhelming focus on Norwegian and the taken-for-granted reliance on Norwegian even in English lessons, there were few opportunities and resources to use other students' L1s. We did not observe any supportive use of visual tools. Ida mentioned the potential benefit of visual support when giving explanations in class, as this would help language-minority students more than having to process concepts through Norwegian. The

teachers deplored the fact that they did not know enough about the structure of other languages, since this would help them identify likely problems for the learners.

In the individual interviews, Kate and Ida commented on the two types of mandated language screening tests. One was the timed standardized initial screening of Norwegian language proficiency (NSL)[9] designed for language-minority students only. Ida's criticism of the NSL was that it is static and involves a set of subject-specific vocabulary and logical thinking. If a student does not arrive at the expected conclusion, having reasoned in another way, the answer will count as wrong, affecting the test score. She deplored the fact that test accommodation is not allowed: "It is a bit painful, of course, when you have to sit there a full hour, not understanding what to do (Ida, II, 27:31) .

The other standardized assessment was the annual national screening of Norwegian competence. Kate saw this language screening as "unfair" or "unjust"[10]: "Well, they [these screening tests] are a bit unfair as far as [language-minority students] are concerned. There are words and concepts that Norwegian children don't use so much" (Kate, II, 23:23. The example she gave was *stillas* (= 'scaffold'). Kate did not find this vocabulary item useful in 2nd grade. The two teachers' reactions reflect affective and value-oriented perspectives that give rise to validity concerns. Admittedly, for a very young child to have to sit for an hour feeling incompetent and not knowing what to do, the experience may have unfortunate emotional consequences. Clearly, the assessment format itself is a challenge that affects their learning conditions adversely. Similarly, the vocabulary used in the national screening of Norwegian competence for 2nd grade was perceived by the teacher as not the most appropriate for this age group (Butler, 2016). In contrast, the dynamic reading tool "Good Reading Development" (Lundberg & Herrlin, 2008) was singled out by the teachers as very useful. Unlike the standardised screening tests described above, "Good Reading Development" takes place as a dialogue between individual students and the teacher, who keeps a record of their progression and thereby facilitates AfL practices (see Table 2 for a description of this tool).

One of the uses of the NSL has been to determine placement with respect to group membership. If students score lower than a cut-off point, a decision is made triggering an offer of language support of some kind. One option for school leaders is to spend the earmarked allowance on pull-out strategies for these students. As we have seen, this was not the policy adopted at our school. Resources were spent on having two teachers in all lessons with the

[9] *NSL* stands for 'Norwegian as a language for learning' (*Norsk som læringsspråk*). As a matter of fact, the NSLtest no longer has the unique status it used to have at this school; other and more formative assessments are encouraged.
[10] The Norwegian adjective «urettferdig» matches both English *unfair* and *unjust*.

addition of an assistant and the weekly service of a language specialist. The decision to give all learners the benefit of these resources rather than targeting the language-minority students exclusively is controversial. However, the justification for the decision to let language-minority students spend most of their time in the classroom with the rest of the students was to maximize their exposure to Norwegian. In other words, the consequence of the test results in this case is arguably positive. The emotional effects of the test situation itself, however, raises serious concerns when it comes to validation.

Two particular incidents are noteworthy. The first has to do with a comment Lars made about a father asking him not to make any distinctions between his Arabic-speaking son and the other students. Lars assured him that he would not "differentiate", that his son would not be treated differently. What seems to be at stake here is a potential conflict between inclusion in the classroom community and recognition of the student's multilingualism. Differentiation seems to be conceptualized as a problem. This is a critical incident which we will return to below.

The other incident that has relevance from a validity point of view came as a response to our question whether the teachers were under any kind of pressure: "Well, perhaps those screening tests that our leadership is anxious [that we] do as well as possible – which perhaps goes slightly against the purpose they really have…" (Ida, II, 36:53). In terms of Kane's category *systemic effects* (2016), this seems to be a consequence of accountability and its possibly competitive edge.

Multilingualism in classroom-based teacher formative assessment: A project in progress?

In this article we have attempted to understand the scope of the experiences of our participant teachers as regards language assessment through classroom observations and the teachers' reflections about their practice. Our main interest lies in the interface of multilingualism and assessment, particularly with respect to validation issues. Through the teachers' reflections, we could see that they were attentive to the physical, cognitive, and socio-emotional development of their students. They would use various scaffolding techniques and other formative practices to enhance their learners' language proficiency.

There were times when we noted missed opportunities of using multilingualism as a resource. One incident was when the topic in English was polite greetings, and one student volunteered a greeting in Spanish. The teacher instantly said, "But that's not English!", then immediately appeared to acknowledge it as a relevant comment after all from the perspective of multilingualism as a resource. There may have been an element of reactivity here, since the

three teachers were aware of our interest in the multilingual realities of their context. From the point of view of multilingual theory, this was an incident where the teacher's instinctive reaction was to keep languages separate. We recognize this drive as a possible effect of previous language-didactic orthodoxy (Simensen, 2007); not long ago, we would recommend English teachers to stick to one language and preferably English in order to maximize exposure to the target language.

Another point relating to the theme of multilingualism is Kate's matter-of-fact comment that the students' various language backgrounds was a potential problem. An Urdu speaker, for instance, needed an extra step, since translation from English allegedly "had to go via Norwegian". This is a critical incident in our view, in that the role of Norwegian is uncritically taken to be central to learning an additional language. If deep learning of important words is an essential aim, a more conducive strategy would be to allow the Urdu-speaking learner to make a direct link between the English word-to-be-learned and its concept via Urdu, using multilingual resources, or at least visual support.

A third reflection regarding multilingualism is based on a comment made by Ida that her impression was that the language-minority students in her class learned concepts more easily. She is obviously not yet in a position to evaluate this reflection as directly and positively related to current knowledge about multilingualism, that being multilingual brings certain cognitive and other advantages (Bialystok, 2016; Cenoz, 2003; Jessner, 2008). This is a reminder that teachers need to develop their knowledge base and their awareness when it comes to multilingualism as a resource. We will come back to this below as a case in point when we consider our empirical findings through the lens of consequential validity and the issue of justice.

The three teachers willingly produced a list of assessment techniques that they saw as useful tools in connection with assessment for learning (AfL). Apart from the strange fact that many of the examples they gave were *student* responses more that learning-oriented feedback from the teachers to the learners, some of the assessment activities mentioned in Table 2 seemed to be effective. One instance from an English lesson was when the teacher took time to greet the children individually at the door in English, eliciting polite greetings and conversation from each student in turn. This was an instance where a clear validity chain as we have defined it was manifest: the ability to produce polite greetings is a curricular aim, which had been operationalized by the teacher leading to a meaningful communicative activity (Sandvik, 2019). He assessed each learner in the process, also building upon the knowledge already obtained about his group of students. What he admitted and deplored himself was that he had not

systematically recorded his assessments. Similarly, in Norwegian classes, a validity chain was visible in a writing project involving a descriptive text about an animal, which aligned with grade-relevant curricular aims, including the development of shared literacy criteria, followed by timely feedback from the teachers at a teacher-led learning station a day or two later.

Questions were raised by a teacher about *justice* in connection with the annual national screening tests. The way the participant used the term had more to do with what is traditionally referred to as *construct validity* (Kane, 2016). The teacher questioned the relevance of the vocabulary in the test in relation to very young learners' language proficiency. A feeling of irrelevance and distance from one's own lifeworld may have the undesired consequence of leading to lack of motivation. Affective consequences must be taken seriously especially when working with very young learners.

Considerations about fairness and justice (McNamara & Ryan, 2011) become pertinent when we consider the learning environment as a whole for language-minority students. If multilingualism is to be understood as a resource for effective learning, a situation where developing proficiency in the majority language is granted a hegemonic position jeopardizes opportunities for effective language learning for all learners. A critical incident appeared in our context when a language-minority father asked that his son's language background *not* be taken into consideration. This wish indicates that his interpretation of equality and inclusion entails a disregard for difference or diversity as positive values. There can be no doubt that this father wanted the best for his son; what he is not aware of is that his boy will stand a better chance of academic achievement if his emergent multilingualism is taken seriously and treated as an asset.

Consequential validity has to do with social values underlying assessment interpretations, which may be implicit or explicit (Kane, 2016; McNamara & Ryan, 2011). These interpretations have social consequences, and our concern is that teachers in multilingual educational contexts become aware of what implications multilingualism as a phenomenon has for their students so that they can enhance the learning conditions for all students. Based on our experience, recognizing multilingualism itself as a construct in assessments is an aim for the future. As a first step, we need to help teachers challenge inherited monolingual ideologies and develop their awareness of the potential for deeper language learning that lies in acknowledging multilingualism as a facilitative factor. This is ultimately a prerequisite for language education built on and creating social equity.

References

Abedi, J. (2009). Linguistic Factors in the Assessment of English Language Learners. In G. Walford, E. Tucker, & M. Viswanathan (Eds.) *The Sage Handbook of Measurement* (pp. 129–149). Sage.

Ascenzi-Moreno, L. (2018). Translanguaging and responsive assessment adaptations: Emergent bilingual readers through the lens of possibility. *Language Arts, 5*(6): 355-369.

Bailey, A. L. (2017). Assessing the language of young learners. In E. Shohamy (Ed.), *Language testing and assessment* (pp. 323-342). Springer International Publishing AG.

Bialystok, E. (2016). The signal and the noise. *Linguistic Approaches to Bilingualism, 6*(5), 517-534. https://doi.org/10.1075/lab.15040.bia

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in education: Principles, policy & practice*, *25*(6), 551-575. https://doi.org/10.1080/0969594X.2018.1441807

Butler, Y. G. (2009). How do teachers observe and evaluate elementary school students' foreign language performance? A case study from South Korea. *TESOL Quarterly, 43*(3), 417-444. https://doi.org/10.1002/j.1545-7249.2009.tb00243.x

Butler, Y. G. (2016). Assessing young learners. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 359-375). Mouton de Gruyter.

Burner, T., & Carlsen, C. (2017). English instruction in introductory classes in Norway. In K. Kverndokken, N. Askeland, & H. H. Siljan (Eds.), *Kvalitet og kreativitet i undervisningen – ulike perspektiver på undervisning* [Quality and creativity in instruction – different perspectives on teaching] (pp. 193-208). Fagbokforlaget.

Cameron, L. (2001). *Teaching languages to young learners*. Cambridge University Press.

Canagarajah, S. (2013). *Literacy as translingual practice: Between communities and classrooms*. Routledge.

Cenoz, J. (2003). The additive effect of bilingualism on third language acquisition: A review. *The International Journal of Bilingualism, 7*(1), 71-87. https://doi.org/10.1177/13670069030070010501

Cenoz, J., & Gorter, D. (2011). A holistic approach to multilingual education: Introduction. *The Modern Language Journal, 95*(iii), 339-343. https://doi.org/10.1111/j.1540-4781.2011.01204.x.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6 ed.). Routledge.

Cook, V. (2016). *Working definition of multi-competence*. Retrieved January 12, 2019, from http://www.viviancook.uk/Writings/Papers/MCentry.htm

Creese, A., & Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching. *The Modern Language Journal, 94*(i), 103-115. https://doi.org/10.1111/j.1540-4781.2009.00986.x

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative and mixed methods approaches*. Sage.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity*, (pp. 3-17). Lawrence Erlbaum Associates, Inc.

Dahl, A., & Krulatz, A. M. (2016). Engelsk som tredjespråk: Har lærere kompetanse til å støtte flerspråklighet? [English as a third language: Do teachers have the competence to support multilingualism?]. *Acta Didactica Norge, 10*(1):1-18. https://doi.org/10.5617/adno.2397

Flognfeldt, M. E. (2018). Teaching and learning English in multilingual early primary classrooms. In K. Palm & E. Michaelsen (Eds.), *Den viktige begynneropplæringen: En forskningsbasert tilnærming* [The important beginner education: A research-based approach] (pp. 229-248). Universitetsforlaget.

Flores, N., & García, O. (2013). Linguistic third spaces in education: Teachers' translanguaging across the bilingual continuum. In D. Little, C. Leung, & P. Van Avermaet (Eds.), *Managing diversity in education: Key issues and some responses* (pp. 243–256). Multilingual Matters.

Flores, N., & Schissel, J. L. (2014). Dynamic bilingualism as the norm: Envisioning a heteroglossic approach to standards-based reform. *TESOL Quarterly, 48*(3), 454–479. https://doi.org/10.1002/tesq.182

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

García, O. (2008). Multilingual language awareness and teacher education. In J. Cenoz & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Volume 6: Knowledge about language* (pp. 385-400). Springer.

Golden, A. (2006). Minoritetselever og norsk [Minority learners and Norwegian]. *NOA – Norsk som andrespråk* [NOA – Norwegian as a second language], *2*, 101-103.

Gorter, D., & Cenoz, J. (2017). Language education policy and multilingual assessment, *Language and Education, 31*(3), 231-248. https://doi.org/10.1080/09500782.2016.1261892

Heugh, K., Prinsloo, C., Makgamatha, M., Diedericks, G., & Winnaar, L. (2017). Multilingualism(s) and system-wide assessment: A southern perspective. *Language and Education, 31*(3), 197-216. https://doi.org/10.1080/09500782.2016.1261894

Hofer, B. (2017). Emergent multicompetence at the primary level: A dynamic conception of multicompetence. *Language Awareness, 26*(2), 96-112. https://doi.org/10.1080/09658416.2017.1351981

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*(3), 229-249. https://doi.org/10.1080/15434303.2011.565844

Hvistendahl, R. (Ed.). (2009). *Flerspråklighet i skolen* [Multilingualism in schools]. Universitetsforlaget.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385–402. https://doi.org/10.1177/0265532208090158

Iversen, J. (2017). The role of minority students' L1 when learning English. *Nordic Journal of Modern Language Methodology, 5*(1), 35-47. https://doi.org/10.46364/njmlm.v5i1.376

Jessner, U. (1999). Metalinguistic awareness in multilinguals: Cognitive aspects of third language learning. *Language Awareness, 8*(3-4), 201-209. https://doi.org/10.1080/09658419908667129

Jessner, U. (2008). Teaching third languages: Findings, trends and challenges. *Language Teaching, 41*(1), 15-56. https://doi.org/10.1017/S0261444807004739

Kane, M. T. (1990). *An argument-based approach to validation* (ACT Research Report Series 90-13). The American College Testing Program.

Kane, M. T. (2016). Explicating validity. *Assessment in education: Principles, policy & practice, 23*(2), 198-211. https://doi.org/10.1080/0969594X.2015.1060192

Kouvdou, A., & Tsagari, D. (2018). Towards an ELF-aware assessment paradigm in EFL contexts. In N. Sifakis & N. Tsantila (Eds.), *English as a lingua franca for EFL contexts* (pp. 227-246). Multilingual Matters.

Krulatz, A., Dahl, A., & Flognfeldt, M. E. (2018). *Enacting multilingualism: From research to teaching practice in the English classroom*. Cappelen Damm Akademisk.

Kunnskapsdepartementet (1998). *Education Act.* Retrieved February 7, 2019 from https://www.regjeringen.no/en/dokumenter/education-act/id213315/

Kunnskapsdepartementet. (2006). *Forskrift til opplæringslova.* [Regulations relating to the Education Act] Retrieved February 7, 2019, from https://lovdata.no/dokument/SF/forskrift/2006-06-23-724

Kunnskapsdepartementet. (2017). *Core curriculum - values and principles for primary and secondary education*. Retrieved January 19, 2019 from
Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: assessment as discourse and assessment of discourse. *Language Testing, 21*(3), 335-359. https://doi.org/10.1191/0265532204lt287oa
Lundberg, I., & Herrlin, K. (2008). *God leseutvikling: Kartlegging og øvelser* [Good reading development: Screening and exercises]. Cappelen Akademisk.
May, S. (2013). *The multilingual turn: Implications for SLA, TESOL, and bilingual education*. Routledge.
McNamara, T. (2012). Language assessments as shibboleths: A poststructuralist perspective. *Applied Linguistics, 33*(5), 564–581. https://doi.org/10.1093/applin/ams052
McNamara, T. (2014). 30 years on – evolution or revolution? *Language Assessment Quarterly, 11*(2), 226–232. https://doi.org/10.1080/15434303.2014.895830
McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test, *Language Assessment Quarterly, 8*(2), 161-178. https://doi.org/10.1080/15434303.2011.565438
McKay, P. (2006). *Assessing young language learners*. Cambridge University Press.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). Washington: American Council on Education and National Council on Measurement in Education.
Michaelsen, E. (2018). Dynamisk kartlegging av leseutvikling de første skoleårene [Dynamic screening of reading development during the first schoolyears]. In K. Palm & E. Michaelsen (Eds.), *Den viktige begynneropplæringen: En forskningsbasert tilnærming* [The important beginner education: A research-based approach] (pp. 139-162). Universitetsforlaget.
Oh, S.- Y. (2001). Two types of input modification and EFL reading comprehension: Simplification versus elaboration. *TESOL Quarterly, 35*(1), 69-96. https://doi.org/10.2307/3587860
Ortega, L. (2019). SLA and the study of equitable multilingualism. *The Modern Language Journal, 103 (Supplement 2019),* 23-38. https://doi.org/10.1111/modl.12525
Oslo Kommune (2019). *Andel minoritetsspråklige elever i grunnskolen 2018/2019*. Retrieved December 10, 2019 from http://statistikkbanken.oslo.kommune.no/webview/
Otheguy, R., García, O., & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review, 6*(3), 281-307. https://doi.org/10.1515/applirev-2015-0014
Otheguy, R., García, O., & Reid, W. (2019). A translanguaging view of the linguistic system of bilinguals. *Applied Linguistics Review, 10*(4), 625-651. https://doi.org/10.1515/applirev-2018-0020
Poehner, M. E., & Infante, P. (2016). Dynamic assessment in the language classroom. In D. Tsagari & J. Baneree (Eds.), *Handbook of second language assessment* (pp. 275-290). *De Gruyter Mouton.*
Sachtleben, A. (2015). Pedagogy for the multilingual classroom: interpreting education. *The International Journal for Translation & Interpreting Research, 7*(2), 51-59. https://doi.org/10.12807/ti.107202.2015.a04
Sandvik, L. V. (2019). Vurdering som bidrag til dybdelæring [Assessment as a contribution to deep learning]. *Bedre Skole* [Better Education], *3*, 27-31.
Sandvik, L. V., & Bruland, T. (Eds.). (2014). Vurdering i skolen. Utvikling av kompetanse og fellesskap. Sluttrapport fra prosjektet Forskning på individuell vurdering i skolen (FIVIS) [Assessment in schools: Development of competence and community. Final report from the project Research on individual assessment in schools]. Trondheim: NTNU, program for lærerutdanning og SINTEF. Retrieved January 20, 2019 from

https://www.udir.no/globalassets/upload/forskning/2015/fivis-sluttrapport-desember-2014.pdf
Saville, N. (2019). How can multilingualism be supported through language education in Europe? *Language Assessment Quarterly, 16*(4-5), 464-471. https://doi.org/10.1080/15434303.2019.1676246
Schissel, J. L. (2014). Classroom use of test accommodations: Issues of access, equity, and conflation. *Current Issues in Language Planning*, *15*(3), 282-295. https://doi.org/10.1080/14664208.2014.915458
Schissel, J. L. (2015). What new (summative) assessments are being developed under the CCSS and how are ELLs/emergent bilinguals to be included? What are test accommodations and which, if any, are most effective for emergent bilinguals? In G. Valdés, K. Menken, and M. Castro (Eds.), *Common Core and ELLs/emergent bilinguals: A guide for all educators*. (pp. 249–251). Carlson Publishing.
Schissel, J. L., De Korne, H., & López-Gopar, M. (2018). Grappling with translanguaging for teaching and assessment in culturally and linguistically diverse contexts: Teacher perspectives from Oaxaca, Mexico. *International Journal of Bilingual Education and Bilingualism*. https://doi.org/10.1080/13670050.2018.1463965
Schissel, J. L., Leung, C., & Chalhoub-Deville, M. (2019) The construct of multilingualism in language testing, *Language Assessment Quarterly, 16*(4-5), 373-378. https://doi.org/10.1080/15434303.2019.1680679
Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Routledge.
Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. Routledge.
Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *Modern Language Journal*, *95,* 418–429. https://doi.org/10.1111/j.1540-4781.2011.01210.x
Shohamy, E. (2016). Critical language testing. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 1-15). Springer International.
Shohamy, E. (2017). Critical language testing. In E. Shohamy and N. H. Hornberger (Eds.), *Encyclopedia of language and education (3 ed.), Volume 7: Language testing and assessmen*t (pp. 441-454). Springer.
Silverman, D. (1993). *Interpreting qualitative data: Methods for analyzing talk, text and interaction.* Sage Publications.
Simensen, A. M. (2007). *Teaching a foreign language: Principles and procedures* (2 ed.). Fagbokforlaget.
Šurkalović, D. (2014). Forbereder grunnskoleutdanningen engelsklærere for undervisning i engelsk som tredjespråk i Norge? [Does initial teacher education prepare English teachers for teaching English as a third language in Norway?]. *Acta Didactica Norge, (8)*2. Art 6. https://doi.org/10.5617/adno.1129
Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21–36. https://doi.org/10.1017/S0267190509090035
Tsagari, D., & Cheng, L. (2016). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.). *Encyclopedia of language and education (3 ed), Volume 7: Language testing and assessment* (pp. 1-13). Educational Linguistics: Springer. https://doi.org/10.1007/978-3-319-02326-7_24-1
Turnbull, B. (2017). Towards new standards in foreign language assessment: Learning from bilingual education. *International Journal of Bilingual Education and Bilingualism*, 488-498. https://doi.org/10.1080/13670050.2017.1375891
Turner, C. E. (2012). Classroom assessment. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 65–78). Routledge.

Turner, C. E. (2014). Mixed methods research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1403-1417). John Wiley & Sons, Inc.
Turner, M. (2017). Integrating content and language in institutionally monolingual settings: Teacher positioning and differentiation. *Bilingual Research Journal, 40*(1), 70–80. https://doi.org/10.1080/15235882.2016.1276029
Utdanningsdirektoratet [Norwegian Directorate for Education and Training] (n.d.) *Læreplanverket for Kunnskapsløftet* [Curriculum for the Knowledge Promotion Reform] Retrieved February 10, 2020, from https://www.udir.no/globalassets/upload/larerplaner/fastsatte_lareplaner_for_kunnskapsloeftet/5/prinsipper_lk06_eng.pdf
Utdanningsdirektoratet (2018). Observations on the national assessment for learning programme (2010-2018): Skills development in networks. Retrieved March 5, 2019 from https://www.udir.no/contentassets/977da52955c447bca5fc419d5be5e4bf/the-norwegian-assessment-for-learning-programme_final-report-2018.pdf
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

Appendix 1

Initial Questions

1. How long have you been teaching in public primary schools in Norway?
2. How long have you been teaching English and/or Norwegian in public primary schools in Norway?
3. How long have you been teaching in this school?
4. How many lesson periods to you teach English and /or Norwegian per week?
5. How many study points in English and/or Norwegian do you have? For what level (1-7, 5-10, other)?
6. Have you received any training in language testing and/or assessment? If yes, what kind?
7. Do you feel you have enough training to teach English and/or Norwegian in lower primary school?

Appendix 2

Questions for Individual Interviews

We would like to talk about assessment of language learning (språklæring/språkutvikling).

1) How are you required to assess your students' language learning during your English/Norwegian classes?

a) Do you use other ways of assessing their language learning in your teaching? If yes, what ways?

i) E.g. How do you use questions to monitor your students' learning? What kinds of questions do you think are successful?

b) What are you happy/satisfied with in the way you assess language learning?

c) Which challenges do you face (if any), in the way you assess?

d) Do you share practices and collaborate with other teachers teaching the same group of students? How?

e) What would you like more training on when it comes to assessment of language learning?

2) To what extent does your students' language background affect the way you assess their language learning and development?

a) In what way? Can you give examples?

b) Is this different in Norwegian and English as a subject, and how?

3) What do you do with the language assessment information you collect?

a) How do you act on it / follow up on it?

b) How do you communicate it to the students and parents?

c) Do you share and collaborate with other teachers teaching the same group of students?

Appendix 3

Questions for Group Interview

1. What resources do you use to assess? Do you refer to any national documents or resources to help you plan for assessment?
2. Do the school's thinking tools help with teaching? With assessment of language development / skills? Do they create any challenges in teaching? In assessing?
3. In the interviews, you mentioned something about new AfL practices at the school. Can you tell us more?

   a. Can you give some examples of feedback you give in English and Norwegian? In what forms? About what?

   b. Ida mentioned a template for lesson planning – with aims, questions, expected student responses. Can you tell us more about this template? How is it used? Is this a school-wide focus?

   c. Tell us about the types of questions you use.
4. How do your school work with §2-8? What procedures do you follow?
5. What are the common topics you discuss in regards to language assessment, for instance, at your grade level team meetings?
6. What types of challenges with language assessment do you discuss?

7. To what extent does having an assistant, a language specialist, or an additional teacher in the room aid with language assessment?
8. To what extent is English part of development conferences with students?
9. Can you provide us with samples of assessment tools you use?