Chapter 3
Changes in Language Assessment Through the Lens of New Materialism

Ari Huhta and Nettie Boivin

Abstract In this chapter, we analyze English tests that are part of two computerised assessment systems, the Finnish Matriculation Examination and the Danish National Tests. Language assessment is a fruitful field to explore from the perspective of materiality, to better understand what materialities exist in modern language tests and how students interact with such systems. Within the assessment and test-taking space, material objects exist that are imbued with political values and force test-takers to perform in specific ways. We explore what new materialism has to offer for interpreting current trends in language assessment and to what extent these perspectives allow for new insights to emerge. We describe the changes in language assessment concerning material developments and focus on the aspects of computerization that pertain to formal tests and examinations. Computerization has increased human-computer interaction during the assessment process, as well as automated analysis and scoring of test-takers' responses. This implies that the computerized system assumes some degree of agency.

Keywords Agency · Computerised assessment · Material relationship · Finnish Matriculation Examination · Danish National Tests

Introduction

Assessment is an interesting and under-explored aspect of language education to investigate from a materialist angle because it may involve a wide range of material objects including pens, papers, test booklets, recordings, and computers. While

A. Huhta (✉)
Centre for Applied Language Studies, University of Jyväskylä, Jyväskylä, Finland
e-mail: ari.huhta@jyu.fi

N. Boivin
Department of Language and Communication Studies, University of Jyväskylä, Jyväskylä, Finland
e-mail: nettie.l.boivin@jyu.fi

J. Ennser-Kananen, T. Saarinen (eds.), *New Materialist Explorations into Language Education*, https://doi.org/10.1007/978-3-031-13847-8_3

these objects do not necessarily differ from those in a language class, their purpose and relationship during testing is worth investigating. Importantly, the intertwined nature of these objects, particularly the digital ones, with the human participants in the assessment process is highly interesting to study (for a discussion of such digital-human assemblages, see Thorne, 2016). Whereas the purpose (assessment, teaching, learning) is largely conceptual and immaterial, the spaces and conditions that separate most assessments from teaching and learning activities are at least partly material. Computer-based assessments also introduce the interesting question of whether digital content is material. After all, computer programmes and digital tasks correspond to test booklets in the paper-and-pencil world (see also, e.g., Bezemer & Kress, 2016). In our chapter, we therefore view digital assessment content as a material equivalent to traditional physical writing implements (see also Burnett et al., 2014 on the complexity of distinguishing between material and immaterial in the digital world). Furthermore, computers also blur the line between subjects (learners, teachers) and material objects (computerised tests) and suggest that the agential cut (Toohey, 2018) between the two may be even more difficult to draw than the traditional one between the learner and textbook (see Saarinen & Huhta, Chap. 9, this volume). This blurring is reinforced by the fact that technology provides access to socio-cultural embedded language context via videos and audios. Moreover, the computer may have different value and affect for the young, 'digital' generation than the material tools of the paper-and-pencil world (Prensky, 2001; Heydon, 2012). Thus, the computer may provide the test-taker with a more lifelike socio-cultural context that paper test-takers are not afforded. This concept will be unpacked later in the chapter.

The clearest examples of how assessment differs from teaching and learning materially are large-scale examinations. Examinations take place in special settings such as large halls whereas teaching often happens in smaller spaces such as classrooms – and learning can happen anywhere. However, many tests, particularly teachers' own tests, are administered in the same classrooms where teaching takes place. Therefore, space is not only about the size and familiarity of the setting but also about the objects that are present and how they are used that distinguish assessment from teaching.

The placement of such material objects as desks and chairs is important in many assessments. In written examinations, desks are placed well apart to prevent examinees from seeing each other's responses but in teaching/learning contexts, learners' desks are often close to each other to enable collaboration. The different spatial arrangements may reflect different learning paradigms: modern teaching/learning is often based on learner collaboration, and therefore, the traditional arrangement of examination desks may appear a relic of the teacher-centered era. In oral tests, the placement of chairs and recording equipment can be based on a careful consideration of their effect on the atmosphere of the interaction (Huhta & Suontausta, 1993). Oral tests differ from written tests also in that they are usually administered in small, quiet rooms with only 2–3 persons present rather than in bigger spaces (e.g. Fulcher, 2003; Luoma, 2004).

The relationship between human participants and non-human objects is regulated in many assessment contexts, particularly in formal examinations and tests. More specifically, the range of material objects test-takers may utilise is restricted; therefore, the material aspects of assessment not only concern which objects are present but also which objects cannot be present. For teaching and learning, anything considered helpful for learning can be used. In contrast, test-takers in paper-based tests are only allowed to bring their writing tools. Everything else is given to them, and any other material found in their possession could be considered cheating. Furthermore, test-takers often have to hand back all the materials given to them after the test.

What distinguishes assessment from teaching and learning even more clearly than the materials is the rules that govern assessments. In addition to dictating which objects examinees can have, rules regulate participants' behaviour, rights and obligations (also relating to the space and time of assessment) and, thus, determine their agency. In written examinations, test-takers must work alone in silence, they may not move around freely in the space, and they may not ask for help from others, although this may vary depending on the test-takers and purpose of assessment, as our two examples will illustrate. In contrast, many learning activities are based on collaboration between learners with assistance from their teacher.

The material and agential basis of many assessments is, thus, quite different from teaching and learning. However, assessment purposes differ, which affects their material characteristics, too. Assessments that most radically differ from teaching and learning are large-scale, standardised examinations used for certifying examinees' skills and knowledge or achievement of the goals of education. Such examinations are used for gaining entry, for example, into a higher level of education. They are, therefore, used for gatekeeping, to ensure that only persons with specific competences can enter the desired education, profession or position (Nguyen, 2021). However, smaller scale assessments aiming to improve learning at the classroom level are by far the most common purpose of assessment. These formative assessments can be done with test-like tasks but more commonly through homework and continuous teacher observation of the learners in the classroom. Therefore, formative assessment is often embedded in teaching/learning and does not involve obvious material changes associated with examinations. Like the other aspects of language education, also assessment has changed over time. The most relevant changes for our chapter concern the emergence of centralised, national, and large-scale examinations and their recent digitalisation.

Formal written examinations to control education and select civil servants began in the Western countries in the 1800s (Spolsky, 1995). The 1800s also saw the start of the measurement of mental abilities, first to diagnose disabilities but later to select individuals based on their intelligence and other psychological constructs (Spolsky, 1995). Large-scale psychological testing commenced in the USA during WWI to quickly allocate appropriate roles to a large number of recruits. The solution was the multiple-choice and other objectively scorable test formats. The tools of mental measurement, such as the multiple-choice, spread to language assessment, and are now an established part of all testing. Therefore, the current language

examinations are the product of nearly two centuries of centralised examinations and psychological testing. These two traditions largely underlie the material aspects of current examinations as well as participants' agency.

The most important recent material change in assessment is the computerisation of paper-based assessments since the 2000s. This has happened both in large-scale gatekeeping examinations and diagnostic/formative assessment (e.g. DIALANG; Alderson, 2005; for overviews, see e.g. Suvorov & Hegelheimer, 2014). Below, we discuss two English tests from the Nordic countries to illustrate computerisation from the material and agential perspectives. The first test is part of the Finnish Matriculation Examination (ME) and the second is one of the Danish National Tests (NT). While the two test-taking contexts are different, the chapter investigates from a new materialist perspective the similarities between the students' relationship to the material and immaterial computer objects. We refer both to published studies and an interview by the second author of a Danish/American seventh grader who grew up in Denmark. We also make use of the first author's personal experience based on working for the Finnish Matriculation Examination Board.

The Finnish Matriculation Examination

The Finnish Matriculation Examination (ME) is the final (summative) achievement test at the end of general upper secondary education (see https://www.ylioppilastutkinto.fi/en/). It provides students and admission officials in higher education institutions (HEI) with information about individual student achievement. HEIs give ME results considerable weight in their selections and, therefore, the examination is high-stakes for the students (see Table 3.1).

The ME is administered twice a year. Students must pass at least four subjects, but they can choose several additional subjects (ten is a practical maximum). The subject 'mother tongue and literature' is the only compulsory subject, all others can be chosen from among several natural and social science subjects and foreign/second languages. English is not compulsory, but most students select it. Students can spread their ME across a maximum of three consecutive test dates (i.e. they have to complete all components within 1½ years); thus, they can retake any subject once or twice.

The ME was digitized in 2016–2019. The examination is a traditional fixed test, i.e. all students are given the same tasks. Students take the examination in their school using their own laptops. The computerised ME scores the students' multiple-choice responses automatically whereas open-ended tasks are marked afterwards by assessors using a separate online system. The first assessor is the student's teacher and the second is a rater appointed by the ME Board; the ME raters are typically experienced language teachers from different types of educational institutions.

Table 3.1 Main characteristics of the English tests in the Finnish ME and Danish NT

	Finnish Matriculation Examination (English)	Danish National Test (English)
Purpose / use	Achievement (final summative test of general upper secondary education) Gatekeeping (selection to higher education)	Formative (feedback to students, parents and teachers; lower secondary level) National monitoring of achievement
Structure / skills tested and task formats	One test with four sections: Listening, reading, writing, and vocabulary & structures Multiple-choice, constructed response (gap-fill, short-answer), 1–2 extended writing tasks	One test with three sub-domains: Reading, vocabulary, and language & language usage Multiple-choice
Time	6 hours	45 minutes
Space	School's sports hall or equivalent	Computer classroom
Modality	Computerised Fixed test Student's own laptop	Computerised Adaptive test School's computer (desktop)
Agents	External agent (ME board): Test content; system development Teachers: Invigilation, first rating (compositions, short-answer items) ME raters: Double rating (compositions, short-answer items), additional ratings Student: Decide in which order to complete tasks, how long to spend on tasks, whether to revise responses, when to start listening to audio recordings, how many times to watch the video recordings Computer: Automatic scoring of multiple-choice items and some short-answer items	External agent (NT authority): Test content; system development Teachers: Invigilation / guidance; feedback to students & parents; individualising instruction based on NT results Students: Can decide to skip items & how long to spend on items Computer: Automatic scoring of multiple-choice items; calculation of learner ability; selection of items to administer; providing a score / level

The ME in languages has two versions (more difficult and easier, roughly corresponding to high B2 and low B1 levels of the Common European Framework of Reference, respectively) and it covers listening, reading, writing, and grammar and vocabulary, with a range of item formats. Listening tasks are based on audio or video recordings, and pictures are regularly used in reading, listening and writing tasks. Writing involves a 200–250-word composition on one of the four given topics (more difficult test) or two short writing tasks each with two options (easier test).

The Finnish ME is spread over about two weeks, and the students are allowed six hours to complete each subject test. The tests are administered in the students' school at the same time across the country. The venue is a large room such as a sports hall with teachers as supervisors. The students are familiar with the ME exam but are now building a relationship with the digital aspects of the large-scale exam. The Danish exam, while not high-stakes, still shares features with the Finnish ME as it is used for national monitoring purposes. We will next discuss some of these.

The Danish National Test

The Danish NT programme started in 2006 and has been implemented in its present form since 2010. The NTs are part of a more general educational reform recommended by OECD (2004) and a reaction by the Danish educational authorities to disappointing PISA results (Beuchert & Nandrup, 2018). OECD (2004) recommended that evaluation in the schools be improved by creating better (standardised) assessment and feedback instruments for the teachers, and the NTs implement this recommendation. Consequently, to ensure improved assessment results, the Government implemented external testing more regularly, particularly for such subjects as Danish as L1 for which a national test is taken four times between grades 2 and 8. The NT in English is taken by the students only once, however, typically in grade 7. In total, the Danish national testing system covers ten subjects (Høvsgaard, 2019).

The Danish National Tests (NT) have a dual aim (Beuchert and Nandrup 2018; see Table 3.1). First, they help the teacher provide feedback to learners and to design individual teaching plans (see Høvsgaard, 2019, p. 84); thus, this use of the test results can be called formative. The student's parents are also informed about their child's results by comparing the child's performance with the national average on the particular subject and possibly accompanied by more detailed feedback from the teacher (Kousholt, 2016). Thus, the NT provides students, parents, and teachers with information that aims to improve student learning. Second, educational authorities use the results to monitor school and national level achievement in primary and lower secondary education, which suggests that the test may also be used for accountability purposes.

The NT is computer adaptive (CAT); i.e. it adapts to a student's performance and attempts to find the right level of item difficulty for each student, thus providing everyone with an individualised test scenario. The philosophy behind this is based on a key principle in the Danish School Act, namely that "for students to be equal, we need to treat them differently" (Høvsgaard, 2019, p. 85).

The adaptive system scores responses automatically. In addition to marking, the adaptive algorithm calculates a new ability estimate after each response to decide whether to administer an easier or more difficult item next. The algorithm seeks to estimate the learner's level of proficiency by minimising measurement error and by finding a state where the learner's probability of responding correctly to the items is 50%.

Each NT covers three subdomains presented as one test. For English, these are reading, vocabulary, and language usage. The English test uses only multiple-choice questions, which makes automated scoring possible. The number of items in the test and in each subdomain varies between students depending on how fast the algorithm can estimate their proficiency.

The Danish NTs take 45 minutes, but students can be allowed more time to finish. The NT in English is administered only once during students' studies, at the time decided by the school. The students take the NTs in their school's computer studio.

Material Relationships and Agency in the Finnish and Danish Testing Systems

We next compare the two tests by first providing a general account of the agency of the different actors in the assessment process before moving to a more detailed analysis that focuses on the relationships between the two computerised systems (i.e. objects) and the human participants, particularly the students (i.e. subjects).

We use agency as defined by Barad (2007, p. 235) as "an enactment, not something that someone or something has", in other words, "(a)gency is doing/being in its intra-activity." Barad contends that agency emerges from an interaction between material object and human and one does not contain independent agency over the other. Intra-action thus understands agency as not "an inherent property of an individual or human to be exercised, but as a dynamism of forces" (Barad, 2007, p. 141). Our study examines the intra-action of the assessment process with the task at hand, and the material objects involved in the activity. It highlights the idea that agency is the fluidity of intra-action occurring between digital multimodal object and the learners' choices of when and how to utilize it.

As far as the Finnish and Danish assessment contexts are concerned, agency in both is divided between various human participants – test designers, teachers, and students – but also the computer has an agentive role. The roles of the agents vary, however, as does their significance, freedom of action, and influence on the assessment process.

Test designers: In both countries, the assessment system is designed by a centralised national authority that decides on the content and rules of assessment. They also maintain the computer system that delivers the test content.

Teachers: Both Finnish and Danish teachers have different roles that derive from the very different purposes of the two assessments. During test administration teachers' agency is limited to invigilation in both countries; this is particularly important in the high-stakes Finnish ME but also the Danish teachers are expected to ensure that students adhere to the regulations. However, as we will describe later, the Danish teachers may sometimes also guide and encourage their students, particularly the younger students. Where the two contexts differ the most concerns what the teachers are expected to do after the test. In Finland, the teachers also do the first rating of the writing and short-answer tasks for their own students. Although the raters appointed by the ME Board have the final say, they are obliged to forward student performances to another rater, if their rating differs from the teacher's marks by a certain amount. Thus, the teachers' ratings carry some weight in the assessment process. In Denmark, the teacher's role is to interpret the results for the students and also for themselves, and to create study plans for each student (Høvsgaard, 2019). Thus, the teachers are given considerable freedom to turn test scores into feedback and action plans. The Danish teachers' role in the testing process is, thus, directed towards future learning whereas the Finnish teachers judge what students' ability was at the time of the examination,

even if they can try to learn from the current students' performances lessons for future instruction.

Students: Individual students have limited agency in both contexts. Even if the Finnish ME is not mandatory, unlike the Danish NT, students have to pass the exam if they want to enter higher education. However, after completing a specified number of courses, the Finnish students can choose when in the window of three consecutive ME administrations they sit particular subject tests. In Denmark, students must take the NTs when their school decides to administer them. The actual test-taking clearly differs. In the adaptive Danish tests, students must take the items in the order the system administers them, and they cannot return to previous items to change their responses. In the fixed Finnish ME, students can see the outline of the entire test before they start, and they can take the tasks in any order. They can also change their responses. We will discuss student agency in more detail below.

Computer: Finally, the computer can be considered to have some agency, even if the system cannot make free choices since its actions are based on a scoring key or a mathematical formula. However, the system acts independently of the student (and the programmer) when it scores and is not just a platform for delivering content and collecting responses as the paper-and-pencil tests are. In Denmark, the computer both scores and estimates a student's ability after each response in order to decide which item to administer next. In contrast, the Finnish system only scores the multiple-choice items and leaves the rest to humans. Overall, then, the border between the computer and the other agents is somewhat blurred in these assessment systems, particularly in Denmark (see also the discussion about different agential cuts elsewhere in this volume).

Type of Material Relationship – Space, Equipment and Time

We now turn to the material characteristics of the two computerised assessments, such as the place and equipment, because familiarity with these likely affects some test-takers' anxiety. This, in turn, can affect how well they can demonstrate their skills and knowledge.

One of the affordances in both contexts is the venue which is the students' own school rather than an external testing centre. Even the high-stakes Finnish ME is administered in the students' own school with their teachers as invigilators. Admittedly, the largest hall of the school where the ME is administered is not the students' own classroom but, nevertheless, the students have a familiar relationship with the space. What obviously diminishes the familiarity of the venue is the special layout and rules that govern its use for examination.

In the Danish context, too, there is a familiarity and similar relationship with the space. Since the Danish NT is administered in a computer room with a homeroom teacher, the venue is likely to be familiar to the students because of previous teaching. Thus, the physical setting of the NT is somewhat similar to the students' regular

experience with teaching. The students' test-taking behaviour is regulated but this appears to vary depending on the students' age; at primary level (for the NT in L1 Danish and mathematics) the teacher often provides help to students (see Kousholt, 2016).

Both tests are computerized; therefore, computers and related accessories are the key material objects. In Finland, students use their own laptop, but the school lends them the equipment if they need one. Thus, the functionality of the equipment is familiar to the students, including the feel of the keyboard that is important for typing longer responses fast enough. Studies on the ME suggest that both the teachers (Leontjev, in print) and students (Savolainen, 2017) consider typing to be faster than handwriting and that it is easier for the teachers to read and evaluate learners' typed texts. Both students and teachers were, however, worried that typing might increase spelling errors.

The Finnish students take many computerised tests in the years preceding the ME through the digital course examination system Abitti (see https://www.abitti.fi/), created to help students prepare for the examination. This ensures familiarity with the digital testing system. Moreover, through multimodal context in situated context viewed in the videos, audios and visuals that the digitalized test has provides some form of agency over prior group test taking. For example, the student is afforded the time to replay these multimodal (video) affordances which in most tests can only be played once or twice.

Interestingly, decisions by the ME Board to allow students to use their own laptops and to watch video input in listening tests as many times as they like deviate from the principle of standardisation that is so typical of high-stakes examination. The reason for the latter is purely technical: the technology applied in the system allows only one or unlimited number of playbacks of videos, and the once-only option was considered to make video-based task unfairly difficult. Why students were allowed to use their own computers may relate to financial considerations, since it would have been expensive for the schools to provide laptops for all their students. Whatever the ultimate reasons, while decreasing the standardisation of test-taking conditions, these decisions seem to have been beneficial for students' subjective test-taking experience (see Burnett et al., 2014) and possibly given them a fairer chance to demonstrate their language skills. Seen from the New Materialist point of view, this relationship with a familiar object such as one's own laptop provides affordances for the student.

Overall, the digitalisation of the Finnish ME seems to have been successful, according to the English teachers, even if they have concerns about students' variable computer skills (Leontjev, in print). Similar, rather positive findings were obtained in a study of the ME in geography covering school rectors, teachers, and students (Kari, 2019). However, Hava's (2019) survey of over 700 students across all ME subjects revealed a mixed picture with a number of students who would have preferred a traditional paper-and-pencil exam; unfortunately, Hava's survey did not investigate students' reasons for their preferences.

In Denmark, the computers are not personal but provided by the school, even though potentially familiar to the students as the tests are given in the school's

computer room. The Danish students may seem disadvantaged compared with their Finnish peers as they must work with less familiar equipment. However, they are likely to have taken several NTs (e.g. in L1 Danish) by the time of taking English, even if, overall, schools may vary considerably in how frequently computers are used in teaching. However, comparisons of the effect of familiar vs unfamiliar devices on students' feelings and performance are difficult because of the differences between the tests. The Danish NT for English uses multiple-choice and, thus, requires very simple interaction with the tasks. Therefore, the lack of familiarity with the equipment may not, as such, have a serious impact on Danish students' ability to demonstrate their language skills.

A separate issue is that individual students' familiarity with using computers varies in both countries. The schools and teachers, too, differ in how much homework is on computers, so some students are unavoidably better prepared for the tests than others. Given the high stakes of the Finnish ME and due to the widespread use of the Abitti system, the Finnish students, who are also older since they study at upper secondary level, are probably more experienced in using computers, even if some English teachers have concerns about their students' computer skills (Leontjev, in print).

Other material objects can also be present. In the Finnish ME, students can use paper and pens to take notes, for example, when listening to recordings and planning their written compositions. In contrast, the use of pen and paper is apparently not possible in the Danish NT – on the other hand, such tools would be of limited value since the English test only uses multiple choice items and does not include listening. However, it appears that some Danish students may regard this as a problem because it deprives them of the tactile multisensory mediation (Boivin, 2021) that they are used to in their regular classroom learning. The student interviewed for this chapter mentioned that "some students like the feel of paper" and that she herself likes to "write notes to organize their thinking" (interview 1/3/2020; see also Hava's study of Finnish students' preferences). The NT removes this affordance.

Type of Interaction with Modality

The computerized modality of assessment affects the way test-takers interact with the assessment system that comprises both hardware and software that administers test materials, and in the case of the Danish NT also scores student responses. However, there are significant and interesting differences between student interaction in the two systems. These differences relate to what the students know (or assume) about the test in general, how they monitor their progress through the test in terms of time, what choices they can make, and how they understand success vs failure during the test.

Transparency of and Familiarity with the Test-Taking Process

Fixed tests, paper-based or computerized, are quite easy to understand since everybody takes the same items. In the Finnish ME, the students get an overview of the examination on the first screen of the entire test and can, thus, easily see how many sections and items there are, which helps them to monitor their progress through the test and be aware of how many items are left. They know the time allowed for the whole test and can monitor how much time they have for the remaining tasks. However, it should be noted that the ease with which students "understand" fixed tests is partly due to their socialization to them by participating in an educational system that uses such tests.

Research on the Danish NT, which is a computer adaptive test (CAT), indicates that the adaptivity of the system results in very different interaction between the students and the test compared to fixed tests. Overall, adaptivity, as an entirely new feature of a test, appears to be very difficult for the students to understand, which leads to uncertainty and erroneous assumptions of what happens during the test and what the test result means. Teachers, too, appear to struggle to understand how adaptive tests function (Høvsgaard, 2019, p. 88). The students will not know in advance how long the test is going to be, particularly in terms of the number of items. The NT is planned to take about 45 min, but the students can obviously complete it faster if the algorithm can estimate their skill level more quickly. In this respect, the adaptive test does not differ from fixed tests because in the latter, too, fast and more able test-takers can complete the test well before the maximum time allowed. What makes the difference is that in a CAT, students do not know in advance how many items their version will contain, which makes it difficult to predict the length of their test session.

Since the Danish students cannot know, by counting the number of items, how far they are in the test at a given time, the system indicates progress with color codes. All students start the test with the visual modality of a red light, move into yellow as the algorithm begins to find the right level, and then into green when the algorithm has found a level of proficiency within a specific degree of certainty (Høvsgaard, 2019, p. 87). This creates an impression for the students that a "green screen" signifies that they have managed to complete the test. While the color system was created to help the students to know how far they are in the test, the unpredictable variation in the actual number of items each student has to answer has been found to result in unforeseen and even unfortunate consequences. However, the color system as a familiar indicator used in video games also becomes a multisensory discourse resource for the students navigating the test. This navigation creates a relationship with the material color the computer creates with the student and with time. Therefore, the computer algorithm creates a relationship with time and gaming that the students are familiar with (Allerup & Kjeldsen, 2017).

Kousholt (2016), Allerup and Kjeldsen (2017) and Høvsgaard (2019) report on research on the NT test-taking process in the primary, and the second author interviewed a lower secondary school student for this chapter. These studies show that

many students want to complete the NT as quickly as possible and that the test can turn into a competition of who can finish first. Furthermore, the students regularly compare how many items they have answered. What further affects the test-taking process is that the testing conditions seem to vary across grade levels and probably across schools and teachers. Kousholt (2016) observed a primary school teacher actively helping struggling students but added that such help was probably not given in the secondary schools. However, she maintained that students are very interested in comparing their NT color codes, number of items taken, and finishing times with their peers at all grade levels. How openly they do this varies.

Transparency of the test is also a matter of the relationship between the number of correctly answered items and the overall test result, and in this, too, fixed and adaptive tests differ. The relationship is straightforward in fixed tests: the more items you get right the better your overall score will be. Item weighing may slightly affect this (see Alderson, 2005).

However, computer adaptive tests work in a way that makes learners' prior experience based on fixed tests invalid. CATs give test-takers items that are likely to match their level. The Danish NT aims to give students items where their chance of answering them correctly is about 50% (Allerup & Kjeldsen, 2017, p. 112) because such items yield the most information about test-takers' ability. At the start of the test, this is not possible since nothing is known about the student's ability but with more items the estimation becomes more accurate. In the Danish NT, the CAT stops when the algorithm estimates that the student's probability of answering the next item correctly is exactly 50% (with a certain amount of error). In fixed tests, students who answer more items correctly get better results, whereas in a CAT, both low and high ability students may answer an equal number of items correctly even if their overall result is very different. Allerup and Kjeldsen (2017, p. 115) agree that this is conceptually very different from what the students, and teachers, are used to and can, thus, confuse them, since the assumption that a larger number of correct answers leads to a better result does not hold.

In addition to a certain lack of transparency, CATs seem to result in a different approach to time and speed than fixed tests. The NTs were not designed to measure speed but students' skills and knowledge. However, as described earlier, they often appear to turn into speeded tests probably because of several reasons. One reason is likely the uncertain number of items that a student encounters. Another is the color coded indication of progress, which is apparently easy to spot by other students sitting nearby and which may lead to competition about who is fastest. Kousholt (2016) observed that in primary schools, at least, this was more typical of boys than girls. She argued that students' attention to speed and the number of items they can answer probably comes from computer games where speed is a key factor for success. The student interviewed for this chapter also said that many students "look at numbers, if you are the last person still yellow you don't feel good, you feel slow and stupid" (interview 1/3/2020). Høvsgaard (2019, p. 87) reported that teachers often remind the students "to keep a good pace", which can also contribute to the speeded nature of the test (see Helsper & Eynon, 2010, on learners' age, gender, experience and education as predictors of computer skills).

An interesting finding by Kousholt (2016) and (Høvsgaard, 2019) that some students attempted to reach the green light as fast as possible by skipping all the items they considered too difficult suggests another failure to understand CATs. Høvsgaard (2019, p. 87) reports that skipped items are counted as wrong answers which can result in a too low overall result and at the very least means that the test takes longer simply because the system struggles to estimate the student's level due to his/her inconsistent replies and, thus, needs to administer more items.

Awareness of Success and Failure During the Test

It is probably easier for test-takers to be aware of how successful they are in completing the test tasks in fixed tests than in CATs. This is because fixed tests contain a number of items that are either quite easy or quite difficult to most test-takers since such tests target average students. Thus, less advanced students encounter a lot of very difficult items, whereas advanced students come across many items that are easy for them. Whatever the students' ability, they are aware, to some extent, which items they certainly got right and which they simply had to guess or leave unanswered. More generally, in Finland, the students practise by taking retired ME tests and can therefore develop quite accurate expectations about their typical performance on such tests. Since the English NT in Denmark is taken only once, students do not have similar points of comparison to base their expectations on.

Besides the relative difficulty of the tasks, the task type may matter when it comes to test-taker awareness about success. In multiple-choice items, it is always possible to guess so that even in the most difficult items there is a reasonable chance of answering correctly and, therefore, apart from very easy items, test-takers cannot be entirely sure whether they have managed to make the right choice. In tasks requiring free production, test-takers have to create their own responses and it may be easier to be aware of how successfully one has addressed the task. There appears to be no systematic research on this matter but the first author's own experience in rating student performances in the Finnish ME suggests that weak students often leave short-answer questions unanswered but very seldom do the same in multiple-choice questions.

Test-takers' awareness of their success in a CAT is bound to be different from a fixed test for the basic characteristic of CATs, namely that they aim at administering such items to the students that are neither too easy nor too difficult. Thus, students constantly encounter items where they cannot be quite sure if they got them right or not. The multiple-choice nature of the English NT in Denmark may further add to students' uncertainty about how well they are doing on the test.

Even if the Danish students struggle to understand CATs, they nevertheless try to find ways to figure out how well they are faring. Completing the test as fast as possible appears to be a sign of success for some students. Another clue that students seem to use is the number of items they have taken, but they appear to interpret that information in two contradictory ways. Allerup and Kjeldsen (2017, p. 115) report

of the students' views that "it is considered prestigious to be presented with as few items as possible". However, Kousholt (2016) found that some young, primary level learners confused the number of the items they had taken with the number of items they had answered correctly. Even though the teacher told her students that they could not know how many items they had responded correctly, the erroneous interpretation persisted among some learners.

Freedom of Action and Student Agency During Assessment

The two tests differ in what choices students can make. In the ME, students can complete the tasks in any order, although analyses of the log files indicate that many take the items in the order they are listed. The students can return to previously completed items and change their answers. These are design features since computerised fixed tests can obviously be designed so that these actions are not possible.

In contrast, computer adaptive tests force test-takers to answer items in the order determined by the adaptive algorithm. Students cannot go back and change their answers as that would distort the calculations of student ability. However, in both the ME and NT, students can skip items but with somewhat different consequences. In the fixed ME, a skipped item automatically lowers the student's total score, whereas in a CAT skipping results in the test becoming longer as the system has to administer more items. In the NT, skipping may also lower the final score, as was mentioned earlier.

The Finnish students can also use pen and paper for planning, which adds another dimension to their interaction with the digital materials. However, in the Danish context the students' relationship with modality is much more constraining, since the students can do little else than select options in multiple-choice items.

Computerisation has also increased student agency in the listening tasks in the Finnish ME by allowing students to take as long as they like to read the questions before listening to the related recording; in the pre-digital listening tests, there were fixed length pauses for students to read the task before the recording commenced automatically. As to the listening tasks based on a video, the students can play them as many times as they want to. Because the English test in Denmark does not include listening, direct comparisons cannot be made, but the nature of the CAT, particularly its high degree of automatisation of scoring and standardisation makes it unlikely that test-takers could be given as much freedom of action – and agency – as in fixed tests.

If the Danish students do not have much agency when taking the NT, does this imply that the computer adaptive test has some agency or even more agency than the student? The answer probably depends on how independent the computer is considered and how we define independence. Some might argue the computer algorithm provides the material object (computer) with independence in the relationship between student and computer. After all, a computer programme such as a CAT algorithm certainly interacts with the student very differently from a textbook.

However, Burnett et al. (2014) findings revealed "…that the world of Google is a constructed one, and so on. In this sense, Street View…It is produced elsewhere, it is pre-selected and in order to read it we have to do two important things. We have to operate at the interface, and we have to believe in it by mapping it on to our unfolding experience" (p. 96). Therefore, programming is a language that is pre-structured and created by human coding, and thus, a computer is not independent but a component in the intra-action. Ultimately, computer agency probably depends on the degree to which their programmes can simulate human thinking. CAT algorithms are clearly more advanced than those applied in fixed tests since they do much more than just count correct answers. Systems that can automatically recognise and evaluate language learners' speaking are even more complex than CATs (e.g. Zechner & Evanini, 2020). All such developments increase computer agency and independence, but it is difficult to determine the amount of such agency and compare it with human agency.

Furthermore, one could argue in a new materialist vein that the children taking the Danish NTs have a relationship with the social semiotic representation of color as the computer projects their position in the test. These children have grown up with videogames (in conversation from student participant) and see color and time as being connected (Prensky, 2001). The children respond to the computer's shift in color as communicating where they are in the "race" (test). Therefore, the relationship with time (color), the young test-taker and the computer is established. This highlights, as Barad (2003) argues, that the relationship with materiality '"incorporates important material and discursive, social and scientific, human and nonhuman, and natural and cultural factors" (p. 808). How the children understand and race towards the meaning of color as if it was a videogame raises interesting questions about test familiarity and success.

Conclusion

This chapter explored what new materialism has to offer for interpreting current trends in language assessment by analysing two computerised assessment systems that differ in their design and implementation. Assessment materials have changed from purely concrete objects to a combination of concrete objects (computers, earphones) and digital materials (software, digital content), thus broadening the meaning of "material". Furthermore, the intra-action during the assessment context highlights a new agential cut between the different actors of the assessment process.

The two assessment systems illustrate how the general term "computerised testing" can mask considerable differences in interaction with the test and in agential relationships between stakeholders. The analyses also shed light on how test-takers' assumptions based on their experience with "normal" fixed tests affect their expectations about computer adaptive tests and how these expectations can lead to problems for both the testing system as well as the learners and

their teachers. However, intertwined with test-takers' expectations of what language tests should be like is their often extensive experience with new technologies and new media in general. Such technologically savvy young people are sometimes called digital natives (Prensky, 2001) whose way of communicating and learning differs from that of older generations. More recently, scholars (e.g. Helsper & Eynon, 2010) have argued that age alone does not explain why younger generations interact with computers in particular ways and that learners' prior experience, education, and gender also need to be considered. Our analysis of the two computerised assessment contexts has shed light on the similarities and differences in the participants' agency and interaction with the computer and other material aspects of the assessment. However, to obtain a deeper understanding of how test-takers experience, understand and interpret their interaction with the different digital assessment systems, more comprehensive investigations paying attention to the factors proposed by Helsper and Eyron (2010), among others, are needed.

References

Alderson, J. (2005). *Diagnosing foreign language proficiency*. Continuum.

Allerup, P., & Kjeldsen, C. (2017). Standard setting in Denmark: challenges through computer-based adaptive testing. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard setting in education: The Nordic countries in an international perspective* (pp. 101–121). Springer.

Barad, K. (2003). Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs: Journal of Women in Culture and Society, 28*(3), 801–831.

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Beuchert, L., & A. Nandrup. (2018). The danish national tests: A practical guide. *Economics Working Papers* No. 2014–2025.

Bezemer, J., & Kress, G. (2016). *Multimodality, learning and communication: A social semiotic frame*. Routledge.

Boivin, N. (2021). Homescape: Agentic space for transmigrant families' multisensory discourse of identity. *Linguistic Landscape, 7*(1), 37–59.

Burnett, C., Merchant, G., Pahl, K., & Rowsell, J. (2014). The (im) materiality of literacy: The significance of subjectivity to new literacies research. *Discourse: Studies in the Cultural Politics of Education, 35*(1), 90–103. https://doi.org/10.1080/01596306.2012.739469

Fulcher, G. (2003). *Testing second language speaking*. Pearson.

Hava, K. (2019). *Lukiolaisten näkemyksiä lukiosta, digitaalisuudesta ja hyvinvoinnista* [Students' views on the upper secondary school, digitalisation and wellbeing]. Master's thesis. University of Helsinki. http://urn.fi/URN:NBN:fi:hulib-201906142932

Helsper, E., & Eynon, R. (2010). Digital natives: Where is the evidence? *British Educational Research Journal, 36*(3), 503–520. https://doi.org/10.1080/01411920902989227

Heydon, R. M. (2012). Multimodal communication and identities options in an intergenerational art class. *Journal of Early Childhood Research, 10*(1), 51–69. https://doi.org/10.1177/1476718X11402751

Høvsgaard Maguire, L. (2019). Adapting to the test: Performing algorithmic adaptivity in Danish schools. *Discourse: Studies in the Cultural Politics of Education, 40*(1), 78–92. https://doi.org/10.1080/01596306.2018.1549705

Huhta, A., & Suontausta, T. (1993). Suullisen kielitaidon testausmenetelmiä. [Testing methods in oral skills assessment]. In S. Takala (Ed.), *Suullinen kielitaito ja sen arviointi. [Oral proficiency and its assessment]* (pp. 227–266). Institute for Educational Research, University of Jyväskylä.

Kari, A. (2019). *Kokemuksia sähköisestä ylioppilaskoeympäristöstä. Maantieteen ylioppilaskoe systeeminä* [Experiences on the digital Matriculation Examination. Geography test as a system]. Master's thesis. University of Turku. http://urn.fi/URN:NBN:fi-fe2019062622058

Kousholt, K. (2016). Testing as social practice: Analysing testing in classes of young children from the children's perspective. *Theory & Psychology, 26*(3), 377–392. https://doi.org/10.1177/0959354316641911

Leontjev, D. (in print). Finnish Matriculation Examination, National Curriculum, and teachers' attitudes, perspectives, and practices: When the two assessment cultures meet**.** *AFinLA Yearbook*.

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Nguyen, P. (2021). Uses of language assessments. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal1237.pub2

OECD. (2004). *Reviews of national policies for education: Denmark – Lessons from PISA 2000*. OECD Publishing.

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon, 9*(5), 1–6.

Savolainen, J. (2017). *Digitalize it: Upper secondary school students' views on the digitalized matriculation examination.* Candidate thesis. University of Jyväskylä.

Spolsky, B. (1995). *Measured words*. Oxford University Press.

Suvorov, R., & Hegelheimer, V. (2014). Computer-assisted language testing. In A. Kunnan (Ed.), *The companion to language assessment*. Wiley.

Thorne, S. L. (2016). Cultures-of-use and morphologies of communicative action. *Language Learning & Technology, 20*(2), 185–191.

Toohey, K. (2018). *Learning English at school. Identity, socio-material relations and classroom practices* (2nd ed.). Multilingual Matters.

Zechner, K., & Evanini, K. (2020). *Automated speaking assessment*. Routledge.

Ari Huhta is Professor of Language Assessment at the Centre for Applied Language Studies, University of Jyväskylä in Finland. His research interests include assessments that support language learning such as diagnostic and formative assessment, computer-based assessment, and self-assessment, as well as research on the development of reading, writing and vocabulary knowledge in a foreign or second language.

Nettie Boivin is Associate Professor at the Department of Language and Communication Studies, University of Jyväskylä, Finland. Presently, she is involved in a nine-country Horizon 2020-Migration project utilizing a co-creation approach and multimodal narratives with paperless youth. Her expertise and specialization are in the areas of ethnography, decolonizing ethnographic research practices, her newly defined concept of multisensory discourse resources analysis for inclusivity, and homescape.