LESLLA Symposium Proceedings

Recommended citation of this article

Hamnes Carlsen, C. (2017). Giving LESLLA Learners a Fair Chance in Testing. LESLLA Symposium Proceedings, 12(1), 135–148. https://doi.org/10.5281/zenodo.8058941

Citation for LESLLA Symposium Proceedings

This article is part of a collection of articles based on presentations from the 2016 Symposium held at Universidad de Granada in Grenada, Spain. Please note that the year of publication is often different than the year the symposium was held. We recommend the following citation when referencing the edited collection.

Sosiński, M. (Ed.) (2017). Alfabetización y aprendizaje de idiomas por adultos: Investigación, política educativa y práctica docente/Literacy education and second language learning by adults (LESLLA): Research, policy and practice. Universidad de Granada. https://lesllasp.journals.publicknowledgeproject.org/index.php/lesllasp/issue/view/476

About the Organization

LESLLA aims to support adults who are learning to read and write for the first time in their lives in a new language. We promote, on a worldwide, multidisciplinary basis, the sharing of research findings, effective pedagogical practices, and information on policy.

LESLLA Symposium Proceedings

https://lesllasp.journals.publicknowledgeproject.org

Website

https://www.leslla.org/

GIVING LESLLA LEARNERS A FAIR CHANCE IN TESTING

CECILIE HAMNES CARLSEN
Skills Norway/Kompetanse Norge

ABSTRACT: LESLLA learners have two specific challenges, which both affect their results on language tests: their lack of general literacy on the one hand, and their lack of test literacy on the other. Both challenges need to be taken into account when large-scale language test developers design their tests in order to give LESLLA learners a fair chance to show their language abilities. This paper shows how *Skills Norway (Kompetanse Norge)* has worked to construct a standardized language test that gives this group of learners a fair chance. The results presented in this paper show that despite some room for improvement, we are on the right track towards a fair test for this group of learners. A main point of the paper is that in order to construct a fair test for LESLLA learners, collaboration between test developers and LESLLA teachers and researchers is necessary.

KEYWORDS: standardized tests, LESLLA learners, fairness, justice, test literacy.

1. INTRODUCTION

LESLLA learners have long been part of the immigrant population, but until recently, they have not formed a significant part of the population who take large scale, standardized tests. The past five to ten years, however, this has begun to change and it has become increasingly common for policy makers to set formal language requirements for citizenship and permanent residency, as well as for entrance to the labour market (Extramina et al., 2014). This is the case in Norway as well as the rest of Europe and beyond. Such requirements apply to all immigrants, LESLLA learners included. Adult L2-learners with little or no prior schooling and limited, literacy skills, have some specific challenges when it comes to learning a second language (e.g. Tarone et al., 2009) as well as when it comes to performing well on language tests (Allemano, 2013; Carlsen et al., 2013). The focus of this paper, is to show how test developers at *Skills Norway*

(Kompetanse Norge), are working in order to give LESLLA learners a fair chance in testing by taking this group and into account when planning and developing the test of Norwegian for adult immigrants.

2. JUSTICE AND FAIRNESS IN LANGUAGE TESTING

As language testers, we develop tests that have a great impact on the lives and opportunities of certain members of society. However, language testers do *not* usually make the political decision that a test be introduced or decide who has to take it, what function the test will have in society, and how the results will be used and by whom. Integration policymakers, education policymakers, or even employers, make the choices that decide the impact of the tests we make.

Samuel Messick's definition of validity has been highly influential in language testing and assessment since it was first introduced in 1989. Its innovation was its focus on the social consequences of test scores, and its emphasis that validation studies should not limit themselves to investigations of whether or not a test measures what it is supposed to measure, but encompass the interpretation and use of test results. Despite the obvious advantages of including test use and consequences in the definition of validity, it places an enormous responsibility on the shoulders of language test developer. No matter how much we may want our tests to be door openers for those who take them, no matter what we may think about the use of language tests for citizenship or for permanent residence, these are decisions that are out of our hands.

In light of this, I find McNamara and Ryan's distinction between justice and fairness in language testing extremely useful. In their terms, *justice* is a matter of social and political values of test constructs, and it has to do with the way others choose to use the tests or the scores of tests:

> Questions of the *justice* of tests include considerations of the consequential basis of test score interpretation and use but also, and particularly, the social and political values implicit in test construct (McNamara & Ryan, 2011: 167).

Justice questions regard matters such as whether or not it is just to use language tests for university admission, whether or not it is just to use language tests as gate-keepers to certain professions or to the labour market in general and whether or not it is just to set language requirements for citizenship, for permanent residency, for family reunification, or for entrance to the host country. Similarly, justice applies to whether it is just or not to set such requirements for all immigrants, the low-educated and refugees.

Fairness, however, has to do with ensuring that all candidates have an equal opportunity to demonstrate their skills, in this case, language skills. It has to do with the absence of bias, i.e. of systematic discrimination of certain groups for reasons other than differences in the skill being measured (Kunnan, 2007; Shaw & Imam, 2013). If a language test favours people with certain professions, one gender over the other, people from western societies over people from other parts of the world etc., the test is unfair. Important fairness questions are for instance whether the test measures language ability in a stable and reliable way, whether test developers provide sufficient preparatory material for candidates to know what is expected of them, whether test scores are communicated

in an understandable way to all users, i.e. to learners, teachers, employers and policy makers, thereby preventing the misuse of test scores due to a lack of understanding of what the scores mean. It is uncontroversial to claim, as we do in this paper, that it is the responsibility of language test developers to guarantee that their test yields fair results, is not biased and gives everybody an equal chance to show their abilities, and herein lies the focus of this paper.

3. LESLLA LEARNERS' DOUBLE CHALLENGE

Several studies have shown that LESLLA learners perform significantly worse on verbal tests than test takers with more schooling (Kim et al., 2014). Ostrosky-Solis et al. (1998) and Allemano (2013), among others, claim that LESLLA learners' lack of success on verbal tests is a consequence, not only of a deficit in the ability being tested, but also of a lack of experience with the testing situation itself. Ostrosky-Solis et al. argue that "[...] testing itself represents a nonsense situation that illiterate subjects may find surprising and absurd" (Ostrosky-Solis et al. 1998: 657), a claim which is echoed by Allemano (2013: 67) who says that "[a] major barrier to assessment of beginner readers seems to be the examination process itself". These studies indicated that, when developing a language test which is fair for LESLLA learners, we need to take into account their double challenge in testing: their lack of general literacy, i.e. lack of reading and writing skills[1] on the one hand, and their lack of testing literacy, i.e. their lack of test experience and test strategies, often referred to as test- wiseness, on the other.

The degree of test literacy necessary to perform a certain test task, will vary according to the kind of task you are asked to perform: task types range from those which are similar to tasks one would perform in so-called "real-life", to those which require a large degree of prior test taking experience in order to understand what is expected. Research results showing which tasks LESLLA learners have particular problems with, are of great value to language test developers. For example, research has shown that LESLLA learners have limited metalinguistic awareness of phonological and other structural features of language (Homer, 2009; Kurvers et al. 2006, Kurvers & Uri 2006; Olson, 2002; Read et al.1986). Connected to this is their poor recognition of pseudo-words (Kosmidis et al., 2004; Tarone, 2010). Knowing this, it is obvious that more artificial, inauthentic tasks like cloze tests, C-tests or nonsense-word tests, discriminate against LESLLA learners. On the other hand, it should be noted that authenticity or real-life-likeness, does not guarantee that a test task is suitable for LESLLA learners. So-called integrated tests, where candidates listen to an audio clip, read a text, look at pictures or graphs, and reply in writing, are popular because of their resemblance with how we use language outside the test situation. Here too we need to be cautious: while integrated tests may work well for advanced learners, for instance as a university admission test, we could argue that it is highly inappropriate as a task type for LESLLA learners in a high stakes test. This is because their lack of reading and writing experience would in

1. In this paper, I build on Tarone et al.'s (2009) definition of literacy as alphabetic print literacy.

such a test, make it impossible for them to show their competence in oral skills. We will return to this point later in this paper.

4. TEACHING AND TESTING OF NORWEGIAN TO ADULT IMMIGRANTS

In Norway, refugees, asylum seekers, and those in the family-reunification program have the right, as well as the obligation, to follow courses of Norwegian and knowledge of society (KOS). The courses consist of 550 hours of language, and 50 hours of KOS, and KOS is given in a language the learners understand. Courses are free of charge. After the courses, participants take a compulsory language test as well as a KOS-test; the KOS-test is developed in 28 minority languages (Vox, 2012). Adult immigrants are divided into three different teaching tracks depending on their degree of prior schooling or education. These tracks have different speeds and different learning goals. Track 3 is for those with a medium to long educational background. The courses are intensive and the aims are relatively high (level B1 in both oral and written skills). Track 2 has medium progression and somewhat lower learning aims (levels A2 or B1 in both oral and written skills). Track 1 is for low-educated learners, and as formulated in the curriculum, this is a heterogeneous group:

> Track 1 is tailored to participants with little or no prior schooling, some of whom will have no literacy skills, while others will be able to read, but have little experience in using the written language as a tool for learning [...]. (Vox, 2012: 8, own translation)

Track 1-learners, or LESLLA learners, can get up to 3000 hours of tuition free of charge, progression is slower than in the other tracks, and learning goals for the written skills are lower (A1 or A2 in written skills), while they are the same as for Track 2 in the oral skills (A2 or B1).

The Test of Norwegian for adult immigrants (*Norskprøven for voksne innvandrere*, hereafter *Norskprøven*), is based on the *Curriculum of Norwegian for adult immigrants* (Vox, 2012) and on the *Common European Framework of Reference for Languages* (hereafter CEFR, CoE, 2001), and it measures at levels Below A1, A1, A2, B1 and B2. *Norskprøven* is a standardized test, administered twice a year, and has around 20 000 test candidates per year. It measures the four language skills: listening, reading, writing and speaking in four separate tests. The tests of listening, reading, and writing are on computer, and the tests of listening and reading are partly adaptive (van der Linden & Glas 2000) in that the concept that all learners, regardless of their level of proficiency, start out with items at the same level, but depending on how well they perform on the first items, they will get a test tailored to their level of proficiency. This way, LESLLA learners avoid having to face tasks that are beyond their reach, and advanced level test takers will not have to perform too many low-level tasks, which they may find dull and irrelevant. The oral test is a paired format test where two candidates talk to each other in certain tasks, and alone in others, to avoid an asymmetrical conversation between a candidate and an examiner, and at the same time ensure that all candidates get a chance to show their abilities. Their oral performance is scored locally by trained raters according to a common rating grid, while the written performances are scored centrally.

Since its introduction, *Norskprøven* has gradually become more high-stakes: Norwegian tests have been compulsory following Norwegian courses since autumn 2013,

but for a while, there were no sanctions if one did not manage a certain level on the test. This changed in 2015 when the government, consisting of the Conservative party and the right-wing Progress Party, introduced a series of restrictions with the purpose of "[...] making it less attractive to apply for asylum in Norway" (Regjeringen 2015). From January 2017, immigrants who want to apply for Norwegian citizenship, LESLLA learners and others, have to prove a certain level of oral Norwegian as well as passing the KOS-test, in Norwegian. Similar requirements have been agreed for permanent residency and family reunification, but with a lower level requirement in Norwegian and a KOS-test in one of the 28 minority language versions of the test.

5. TAKING LESLLA LEARNERS INTO ACCOUNT IN TEST DEVELOPMENT

Making a standardized test for adult immigrants, which takes LESLLA learners into account, was a new experience for the test developers in *Skills Norway*. Before 2013, we only had a test measuring language from level A2 and above, and LESLLA learners only took a test if they wanted to or if their teachers considered it likely that they were at an A2-level in all four skills, which was a prerequisite in order to pass the test. When the test was made compulsory in September 2013, we knew we needed to make some changes for the test to be fair for all test takers, including the new candidate group of low-educated learners. Given our limited experience with this group, we needed help. Therefore, we invited LESLLA teachers to meet and discuss test formats, the structure of the tests and concrete tasks with us. We established a reference group of LESLLA teachers, and we also invited LESLLA learners to give their comments on tasks and task response formats. In September 2015 we carried out a survey among LESLLA teachers to get their opinions about how the test and the tasks worked for LESLLA learners, the results of which will be presented later in this paper.

Let's return to LESLLA learners' two challenges as presented in the introduction of this paper; the lack of general literacy on the one hand, and the lack of test literacy on the other.

For test developers to meet the first challenge, it's paramount to ensure that candidates' limited reading and writing skills do not affect scores on listening and oral production tests (oral skills). To avoid reading skills in the oral tests, we introduced the use pictures both as task prompts and as task responses. We use pictures of a situation, for example a father cooking, a mother setting the table, a girl playing with a cat, a boy watching TV, a brother and sister quarrelling, etc., to allow candidates at lower levels to name objects in the picture, but at the same time providing the opportunity to candidates at A2 and B1-levels to describe what the people are doing, relations between the people in the picture, and, for example, the emotions they are showing. In the listening task, we use the same kind of picture but ask candidates to listen and to follow the instructions: "Click on the cat", or "Click on the person who is cooking". We also use pictures as task responses and distractors, for example in a listening task where candidates listen to a text and then choose one of four pictures that matches what they hear. Skills Norway has hired an illustrator who works full time and in close collaboration with the item writers. This is a great advantage for us. It is hard to find pictures or drawings that are suitable, and having an illustrator working with the item-writers makes it much easier. It also means that we can order pictures containing just the right vocabulary at different

levels of proficiency, and we can make sure the pictures contain no content that may be provocative or sensitive.

Measuring the four language skills in separate tests yielding independent test scores is of paramount importance in order to give LESLLA learners a fair chance in testing. Probably the most important message the reader should take away from this paper is that even though integrated tasks measuring reading, listening, writing, and maybe speaking, in the same task, may well be authentic and well suited for educated learners, it may be disastrous for LESLLA learners, hindering them from showing their real abilities in listening and speaking. In addition, a test measuring the four skills separately allows candidates to re-sit only parts of the test. If for instance candidates get the score they need in speaking and listening, they would not need to take those parts again because they didn't get the score they needed in writing or reading. The tests of Norwegian prior to the current test, only had pass/fail-scoring. This was very demotivating for the LESLLA learners, some of whom after up to 3000 hours did not get a certificate because they failed the written production part and therefore failed it all. A test that takes LESLLA learners seriously should measure also at the lower levels, A1 or below A1. It is particularly important that learners with slow learning progress get a chance to take a test that shows their incremental improvement.

The second challenge we had to take into consideration, was LESLLA learners' lack of test literacy, i.e. their lack of experience with the test situation. It is a central principle in all assessment that the test measure the skill in question, for example language, and be influenced as little as possible by irrelevant skills or abilities. Test-wiseness is a construct-irrelevant factor in a language test (Bachman, 1990:114). This principle is even more important to bear in mind when LESLLA learners form part of the test population. Firstly, we have consciously avoided using artificial task types like C-tests, nonsense-word-tests or cloze-tests. As far as possible, we try to use test types that are authentic in Spolsky's sense, i.e. meaningful and relevant (Spolsky, 1985). We also try, as far as possible, to avoid hypothetical tasks that require candidates to imagine a situation: When given a written production task, for instance, it is easier for LESLLA learners to respond to a prompt like: "Write a text about what you like to eat for dinner", than to a prompt like: "Imagine that you are inviting some friends over. What would you make for dinner?". Our prior experience in test development, as well as LESLLA teachers, have stressed the importance of avoiding hypothetical tasks and making it as simple and concrete as possible to limit the effect of test literacy.

For candidates with limited schooling and little prior test experience, it is more important than for other groups to know what is expected on the day of the test. It is always important to have ample practice material, for test takers, but for LESLLA learners it is indispensable. Before *Norskprøven* was administered for the first time, practice materials illustrating every task format that candidates would meet, were made available online. We also made available benchmark texts written by learners at the different levels of the test, and we video-recorded the oral exam so that candidates could see how this part of the test worked in practice and could get familiar with the tasks types they would encounter in the real test. The purpose of this was two-fold: to make candidates familiar with the test tasks, and thereby reduce the effect of test literacy on test scores, and, to reduce stress and anxiety, which might introduce another source of construct irrelevant variance to the test-score.

Skills Norway has chosen not to allow the use of electronic spell check, grammar check and/or a dictionary in the written production test, as we fear it would introduce another source of construct irrelevant variance, that is, another non-linguistic skill candidates would need to master. We fear that this would be an advantage to the more educated candidates, who might already have these skills, but a disadvantage to the low-educated ones, who would not. To our knowledge, little research has been carried out on the use of electronic aids during computerized tests by LESLLA learners, but at present a pilot study with the aim of gaining more insight into this area being conducted at Skills Norway (Lauvik, 2016).

5.1. Results

So far, this paper has presented how we have worked in order to give LESLLA learners a fair chance when tested. In the next part of the paper, we will look at some results in an attempt to answer the question of whether or not we succeeded. Firstly, we will look at some analyses of how LESLLA-candidates performed as compared to candidates with a higher level of schooling. Secondly, we will look at the results of a survey amongst nearly 60 LESLLA teachers asking them about their opinions related to both the test system, the test tasks and the consequences of *Norskprøven* on LESLLA learners.

5.2. Test scores

Figure 1 displays mean scores across skills and tracks: The CEFR-levels have been transferred into numeric scores to allow for calculation of the mean.

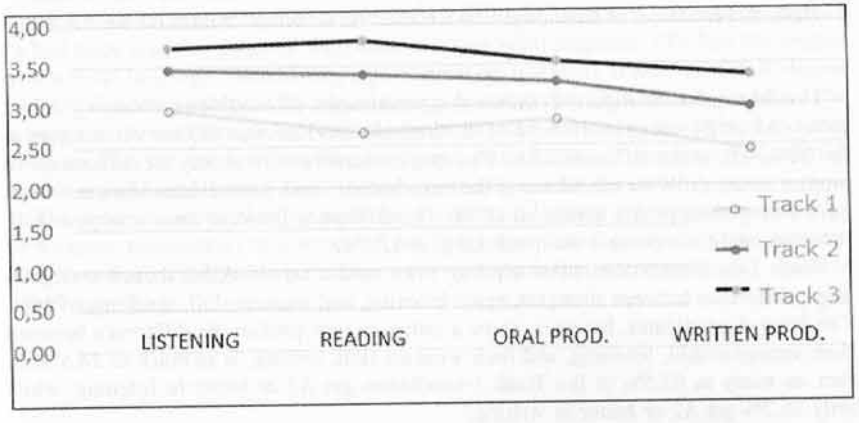

Figure 1: Mean Scores across Skills and Tracks.

Track 1= LESLLA-learners, Track 2=medium school background, Track 3=lengthy school background. CEFR-scores were converted to numerical variables to allow calculation of means: 4=B1, 3=A2, 2=A1, 1= Below A1.

As is obvious from the graph, there are differences in scores between the three tracks for all four skills, and the effect of track on test scores is significant at the p<.001 level for all skills: Track 1 (LESLLA learners) perform the lowest, Track 3 the highest on all skills. In addition, we can see that the profile of the Track 1-candidates differ somewhat from the profiles of the more educated learners: Both Track 2- and Track 3-candidates perform better in the receptive skills than in the productive skills, while the Track 1-candidates perform better in oral production than they do in reading. Track 1-candidates perform better in listening than in reading, while the opposite is true for the Track 3-candidates.

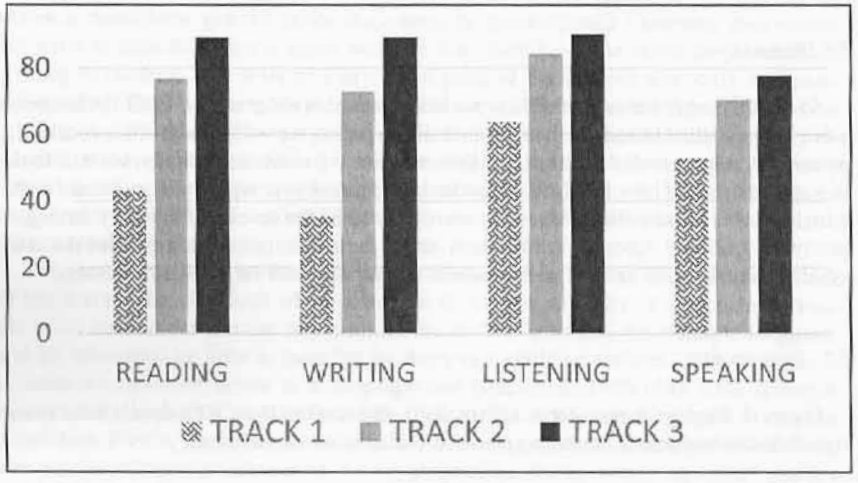

Figure 2: Percentages of Candidates who Obtained A2 or Better across Skills and Tracks.

This histogram in Figure 2 shows the percentages of candidates obtaining A2 or better (A2 or B1, since at the time of the analysis, *Norskprøven* did not yet measure at the B2-level), in the different skills. This graph visualize very clearly the differences in profiles across skills for candidates of the three tracks: Track 3-candidates (dark columns) have a very even profile across all skills. The difference between their strongest skill, listening, and their poorest skill, speaking, is 12.6%.

Track 2-candidates also show a pretty even profile across skills, though a slightly larger difference between strongest, again listening, and weakest skill, speaking, of 14%. The Track 1-candidates, however, show a rather uneven profile: the difference between their strongest skill, listening, and their weakest skill, writing, is as much as 28.5%. In fact, as many as 63.5% of the Track 1-candidates get A2 or better in listening, while only 35.2% get A2 or better in writing.

This graph shows two things: Firstly, it underlines the importance of measuring the four skills separately in order to give LESLLA learners a fair chance to show their abilities, and secondly, it shows that Skills Norway has succeeded in giving LESLLA learners a chance to do exactly this. If we had measured integrated skills, or if we had, like we used to, required candidates to pass all subtests in order to get a test score,

only 35,2% would have succeeded. Instead, LESLLA-candidates and others can obtain good scores in, listening and speaking, whilst working to improve their writing and reading skills, if needed.

N=2082	% of score variance explained by Track	Sig.
Reading	23,9%	p<.000
Writing	18,4%	p<.000
Listening	13,4%	p<.000
Speaking	13%	p<.000

Figure 3: Nominal Regression Analysis of Track Effect on Skills.

Nominal regression analysis shows how much of the score variance is explained by track. The analysis shows that there are significant effects of tracks four all for skills at a p<.000-level. In addition, it shows the largest effects for the two skills that require literacy, reading and writing. Again, this analysis shows that we have succeeded in isolating the oral skills and preventing LESLLA learners' scores to be negatively affected by their lack of general literacy.

5.3. Results – LESLLA teacher survey

In September 2015, a year after *Norskprøven* was introduced, a survey among LESLLA teachers was carried out it solicited their opinions about the test regarding: 1) the test system, 2) the test tasks, and 3) the consequences of the test for LESLLA learners. 53 Track 1-teachers replied to the survey. They were highly qualified and experienced; 45 % had more than five years of experience teaching adult migrants, 75% had Norwegian as a second language (second language acquisition research) in their academic degree and 51% had experience with administering *Norskprøven* for LESLLA learners (The reason why this percentage isn't higher may be because the survey was carried out only one year after the test was first introduced, and many LESLLA learners would still not have reached a level where they would be prepared to take the test).

The teachers were asked to reply on a Likert scale from 1 (disagree completely) to 5 (agree completely) to a series of positively formulated questions about the test. The results are displayed in percentages of respondents that *Agree mostly* (4) or *Agree completely* (5).

	Agree mostly/completely
1. It's good that the test measures the four skills separately	100 %
2. It's good that you can re-sit only parts of the test	100 %
3. It's good that the test is computer-adaptive	98 %
4. It's good that the test is not pass/fail	89 %
5. It's sufficient information about the test	85 %
6. It's good that no electronic aid is allowed	72 %
7. Sufficient example material is provided	47 %
8. It's good that candidates take all four skills the first time	40 %
9. It's good that the test is digital	40 %

Table 1: Questions[2] about the Test System.

As the table above shows, teachers are very pleased that the test measures the four skills separately and that candidates need only re-sit parts of the test. 98% think it is good that the test is adaptive and 89% think it is good that everybody obtains a score instead of pass/fail. Most of the teachers think candidates receive enough information about the test, and 72% agree with the test developers' opinion that it is better not to allow electronic aids for LESLLA-candidates. On a less positive side, only 47% think there is enough practice material, and only 40% are in favour of the (political) decision that everybody has to sit for all four parts of the test the first time they take it. In addition, only 40% think it's good for LESLLA learners that the test is digital. This may however be due to the fact that they were asked only a year after the test was administered the first time, and LESLLA learners may need more time to familiarize themselves with using a computer. Note that this contradicts the responses to the fourth question shown above, that to be able to use a computer is a prerequisite for computer-adaptive testing, which 98% of the LESLLA teachers applauded.

2. *The questions were translated from Norwegian to English for the purpose of this paper.*

	Agree mostly/completely
1. It's good that *Norskprøven* uses pictures as prompts	96 %
2. It's important that the tasks are not hypothetical	96 %
3. It's important that the task aren't provocative or sensitive	96 %
4. It's good that *Norskprøven* uses pictures as task responses	94 %
5. The oral interaction task functions well	82 %
6. It's good that the oral exam uses paired format (candidate-candidate)	79 %
7. The test works well on a whole	77 %
6. Describe picture task works well in the oral production task	77 %
9. It's good that the examiner can be the candidates' own teacher	73 %
10. Describe picture task works well in the written production task	67 %
11. It's easy for LESLLA-candidates to understand what to do on the tasks	62 %
12. LESLLA-candidates have enough time for the written production test	41 %

Table 2: Questions about the Test Tasks.

Almost all teachers are pleased with the use of pictures as prompts (96%) and task responses (94%), and they are relatively pleased with the oral test and the measures we have taken to reduce stress and anxiety, such as including a paired format and allowing the candidates' teacher to be the examiner. The only question regarding test tasks where teachers were more negative, was the time allocated to the written production tasks. Only 41 % of the LESLLA teachers thought their learners had enough time to write. As a consequence of this feedback, we decided to augment the time with 1/3 from 60 to 90 minutes from November 2015.

	Agree mostly/completely
1. *Norskprøven* has a positive washback effect on teaching and learning for this group	67 %
2. It's motivating for LESLLA-learners to take *Norskprøven*	64 %
3. *Norskprøven* contributes to raising LESLLA-learners status	59 %
4. *Norskprøven* contributes to giving LESLLA-learners priority (access to computer room)	51 %
5. Taking *Norskprøven* is not a scaring experience for LESLLA-learners	44 %

Table 3: Questions about the Test Consequences.

As stated in the introduction, it is not up to the test developers to decide how the test is used and its impact on peoples' lives. However, it is interesting to know whether LESLLA teachers are in favour of a test for this group and how they think the test influences them: We were pleased to learn that almost 70% think the test affects LESLLA learners' learning process positively and that the effect on classroom activities is positive. 64% thought it

was motivating for this group to take *Norskprøven*, and almost 60% find that it contributes to raising LESLLA learners' status in the school. Unfortunately, still only 44 % agree that it is not a daunting experience for LESLLA learners to take the test. Nevertheless, several teachers did indeed comment on the opposite effect, ie. that LESLLA learners felt they were taken seriously, as this quote from one of the teachers shows:

> Being met with certain expectations by the teacher, by the school or by society is experienced by most LEL2-learners as positive. That way, they feel they are given the same opportunities, even though their point of departure is different (Respondent 56, own translation).

6. DISCUSSION & CONCLUSION

This paper has described how Skills Norway has worked to make a standardized, high-stakes test of Norwegian for adult immigrants a fair test for LESLLA learners. A comparison of test scores of LESLLA-candidates and candidates with more schooling showed that it is indeed possible to give LESLLA learners a chance to show their skills in a standardized test if certain measures are taken from the start. A good test for these learners needs to measure the four language skills in separate parts which yield independent scores. The results of the study presented in this paper show that *Norskprøven* does give LESLLA learners a chance to perform well at the listening and speaking tests, which do not rely on their limited reading and writing skills. Furthermore, it is important to avoid hypothetical and artificial tasks in order to prevent test scores from depending too heavily on test literacy. LESLLA teachers in the survey presented in this paper underline the importance of this, and the majority agree that *Norskprøven* is a good test for LESLLA learners on the whole.

The evidence presented in this paper shows that we are on the right track, but there is still some room for improvement: For example the LESLLA teachers who responded to the questionnaire are particularly unhappy with the fact that learners need to sit for the four parts of the test the first time they take it. This is a political decision, but something the test developers may try to change. In addition, they feel there is not enough practice material to prepare LESLLA learners for the test. This is something we will have to take into account and work to improve. In addition, around 50% of the LESLLA teachers surveyed fear that the test is a daunting experience for their learners. This too needs to be addressed, and we can see how more practice material may help making candidates feel more familiar, less stressed about taking the test, and change teachers' views of how daunting the test is.

In her presentation at the 2016 LESLLA-symposium, Gonzalves touched upon an important dilemma when assessing LESLLA learners: standardized tests are often not suited to LESLLA learners and, if they have a choice, LESLLA teachers therefore often choose to develop their own tests for this group. LESLLA teachers, however, may know the learner group well, but do not necessarily know how to develop a good test and often refer to their assessment as gut-feeling based. This is reminiscent of Charles Alderson's important argument that language testing is too important to be left to language teachers, but also too important to be left to language testers (Alderson, 2001). One of the main purposes of this paper is therefore to argue in favour of closer collaboration between LESLLA teachers and researchers on the one hand, and large scale test developers on the other. We need to draw upon each others' competence in order to take LESLLA learners into account when designing large scale tests – tests that are increasingly used by policymakers as part of integration policy, and which may have serious consequences on immigrants' lives.

WORKS CITED

Alderson, Charles (2001). "Testing is too important to be left to testers". In Coombe, Christine (Ed.) *Alternative Assessment*. TESOL Arabia, pp. 1-14.

Allemano, Jane (2013). "Testing the Reading Ability of Low-educated ESOL Learners". *Apples – Journal of Applied Language Studies*, 7(1): pp. 67-81.

Bachman, Lyle (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Carlsen, Cecilie Hamnes, Edit Bugge & Ann-Kristin Helland Gujord (2013). "The Effect of Internal and External Variables on Language Learning & Test Scores». Paper presentation at *Language Testing Research Colloquium (LTRC)* June 4-6, 2014, Amsterdam.

Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Extramina, Claire, Reinhilde Pulinx & Piet Van Avermaet (2014). *Linguistic Integration of Adult Migrants: Policy and Practice. Final Report on the 3rd Council of Europe Survey*. Council of Europe: Language Policy Unit.

Gonzalves, Lisa (2016). "When Standardized Tests fail: Informal Assessment of LESLLA Learners in California Adult School". Paper presentation at *LESLLA-Symposium*, September 8-10, 2016, Granada, Spain.

Homer, Bruce (2009). "Literacy and Metalinguistic Development". In David Olson & Nancy Torrance (Eds.), *The Cambridge Handbook of Literacy* (pp. 487–500). Cambridge: Cambridge University Press.

Kim, Jung Wan, Ji Hye Yoon, Soo Ryon Kim & Hyang Hee Kim (2014). "Effect of literacy level on cognitive and language tests in Korean illiterate older adults". *Geriatr Gerontol Int.* 2014, 14 (4), pp. 911-7.

Kosmidis, Mary, Kyrana Tsapkini, Vasiliki Folia, Christina Vlahou, and Grigoris Kiosseoglou (2004). "Semantic and phonological processing in illiteracy". *Journal of the Neuropschological Society*, 10, pp. 818-27.

Kunnan, Anthony (2007). "Introduction: Test fairness, test bias and DIF". *Language Assessment Quarterly*, 4(2), pp. 109–112.

Kurvers, Jeanne & Helene Uri (2006). "Metalexical Awareness. Development, Methodology or Written Language? ". *Journal of Psycholinguistic Research* 35, pp. 353-367.

Kurvers, Jeanne, Roelan van Hout, & Ton Vallen. (2006). "Discovering Features of Language: Metalinguistic Awareness of Adult Illiterates". In Ineke van de Craats, Jeanne Kurvers, & Martha Young-Scholten (Eds.), *Low-educated second language and literacy acquisition: Proceedings of the inaugural symposium*. Tilburg 2005 (pp. 69–88). Utrecht: LOT.

Lauvik, Hanne (2016). "Kartlegging av bruk av hjelpemidler ved gjennomføring av Norskprøven – delprøven i skriftlig framstilling" ("Mapping of the use of digital aids at Norskprøven – written production"). Project in progress.

Messick, Samuel (1989). "Validity". In Robert Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan.

McNamara, Tim & Kerry Ryan (2011). "Fairness versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test". *Language Assessment Quarterly*, 8(2), pp. 161–178.

Olson, David (2002). *"What writing does to the mind"*. In Amsel, Eric and James Byrnes (Eds.). *Language, Literacy, and Cognitive Development: The Development and Consequences of Symbolic Communication.* Mahwah, NJ: Lawrence Erlbaum Associates.

Ostrosky-Solis, Feggy, Alfredo Ardila, Mónica Rosselli, Gabriela Lopez-Arango & Victor Uriel-Mendoza. (1998). "Neuropsychological Test Performance in Illiterate Subjects". *Archives of Clinical Neuropsychology, 13(7), pp- 645-660.*

Read, Charles, Yun-Fei Zhang, Hong-Yin Nie, & Bao-Qing Ding (1986). "The Ability to Manipulate Speech Sounds Depends on Knowing Alphabetic Spelling". *Cognition*, 24, pp. 31-44.

Regjeringen (2015). Innstramminger.)

Shaw, Stuart & Helen Imam (2013). "Assessment of International Students Through the Medium of English: Ensuring Validity and Fairness in Content-Based Examinations". *Language Assessment Quarterly*, 10 (4), pp. 452–475.

Spolsky, Bernard (1985). "The Limits of Authenticity in Language Testing". *Language Testing*, 2(1),pp. 31-41.

Tarone, Elaine, Martha Bigelow & Kit Hansen. (2009). *Literacy and Second Language Oracy.* Oxford: Oxford University Press.

Tarone, Elaine (2010). "Second Language Acquisition by Low-Literate Learners: An Under-Studied Population". *Language Teaching*, 43(1), pp. 75-83.

van der Linden, Wim & Cees Glas (Eds.) (2000). *Computerized Adaptive Testing: Theory and practice.* Dordrecht, The Nederlands: Kluwer Academic Publishers

Vox – Norwegian Agency for Lifelong Learning (2012). *Curriculum of Norwegian for adult immigrants*, 2012: 8.