The dawn of assessment literacy – exploring the conceptions of Finnish student teachers in foreign languages

Hilden, Raili

University of Jyväskylä, Centre for Applied Language Studies

2018-02-20

Hilden, R & Fröjdendal, B 2018, 'The dawn of assessment literacy – exploring the conceptions of Finnish student teachers in foreign languages', Apples : Journal of Applied Language Studies, vol. 12, no. 1, pp. 1-24. https://doi.org/10.17011/apples/urn.201802201542

http://hdl.handle.net/10138/238582
10.17011/apples/urn.201802201542

Apples – Journal of Applied Language Studies
Vol. 12, 1, 2018, 1–24

The dawn of assessment literacy – exploring the conceptions of Finnish student teachers in foreign languages

Raili Hildén, University of Helsinki
Birgitta Fröjdendahl, University of Stockholm

The paper addresses Finnish student teachers' conceptions of assessment literacy in foreign languages. Student teachers' assessment literacy (STAL) is a focal constituent of teacher cognition and can, according to prior research, be enhanced by principled instruction (DeLuca & Klinger, 2010; Volante & Fazio, 2007). STAL is suggested to imply knowledge, practice and ethical considerations. The nature and priorities of STAL are guided by local needs. Hence, topical issues in the Finnish language education were taken into account alongside general assessment theory. The research questions targeted firstly the emergent factorial structure of STAL, and secondly, the validity of a predetermined theory-driven model in alignment with official national priorities. The data were gathered on a web-based survey to 77 students prior to the lectures, and to 65 students after the lectures. The survey consisted of 75 statements about attitudes and practices related to various domains of assessment. Mainstream inferential statistics was used to compare the pre- and post-dataset. The componential structure of STAL attitudes remained more stable than the construct of practices across the study unit. The major dimension of both measurements was Acquired confidence in assessment of multiple aspects of language ability in the classroom. The envisaged or real practices underwent a substantial transformation towards a more learner-centered architecture highlighting flexibility and communication. Of the predetermined domains, working skills and professional self-esteem seemed to be most sensitive to a short-term pedagogical intervention. The tentative results pave the way for progressive development in raising the impact of teacher education for improved assessment literacy skills.

Keywords: beliefs, language teacher education, foreign language teaching, assessment

1 Background

The role of assessment literacy in the field of teachers' pedagogical knowledge is focal and in various ways acknowledged in prominent models of teacher cognition (Borg, 2006; Darling-Hammond, 2008; Schulman, 1987). Assessment is also considered

Corresponding author's email: raili.hilden@helsinki.fi
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies
University of Jyväskylä

http://apples.jyu.fi
http://dx.doi.org/10.17011/apples/urn.201802201542

in most update models of teacher educator qualifications (Koster, Brekelmans, Korthagen, & Wubbels, 2005).

For a long time, however, assessment was primarily conceptualized as testing, and managing psychometrics was an expert duty, far separate from a language teacher's daily work. With the rise of constructivist and sociocultural approaches to learning and teaching, educational assessment and learning-related modes of assessment were brought to the forefront. Today, harmony between formative and summative assessment strategies is needed to ensure sufficient quality of various kinds of assessment across educational systems.

According to recent findings, Finnish language teachers are prone to apply rather traditional assessment practices (Hildén, Härmälä, Rautopuro, Huhtanen, Puukko, & Silverström, 2014; Härmälä, Huhtanen, & Puukko, 2014). The focus seems to be on summative assessment and written language, which is at odds with the communicative goals of language teaching. However, the importance of formative assessment was underlined already in the core curricula of 1970 and 1994. Although the guidelines provided in the latest National Core Curricula (FNBE, 2014) provide certain guidelines to teachers' role as assessors, the fundamental knowledge base for carrying out valid assessment in any context, is laid during academic teacher education. However, again, courses in assessment compete with a broad array of other pedagogic and didactic study units, and the amount of credits devoted to assessment has notably diminished during the last ten years. Our study explores assessment-related conceptions of student teachers and maps the impact of a short-term academic study unit on their development with the ultimate aim to improve the quality of language teacher education.

2 A theoretical framework for student teachers' assessment literacy (STAL)

2.1 A short history of assessment literacy

Below, conceptual cornerstones of teacher assessment literacy (henceforth referred to as TAL) will be briefly summarized. The overview, far from being exhaustive, offers a definition of the knowledge base addressed as the main focus of this study, student teachers´ assessment literacy (henceforth STAL). The definition that follows has been modified parallel with operational definitions for various attempts to measure it over time (Gotch & French, 2014). TAL obviously needs a specific definition despite of and in addition to general models of subject teacher competence. Moreover, the complex process of the development of teacher cognition is further challenged by two prominent developments in language pedagogy. On one hand, the construct of literacy has extended towards interpreting and producing a variety of texts referred to as multiliteracy that language teacher should employ when instructing their students. Moreover, the concept of teacher professionalism involves reflection and dynamic interpretation of multiple situational requirements. (Harding & Kremmel, 2016; Taylor, 2013).

Apart from the general notions embedded in teacher competence models, the seminal list of Standards for Teacher Competence in Educational Assessment of Students published by American Federation of Teachers, National Council of Measurement in Education, and National Educational Association (hereafter AFT, 1990) served as a model for a multitude of national, regional and organizational standards. Teachers should be skilled in selecting assessment methods appropriate

to instructional decisions and in developing assessment methods appropriate for these decisions. They are expected to administer, score and interpret results of external and teacher-produced methods, and use results for making decisions about individual students' performances, about planning teaching and about developing curriculum for school improvement. Teachers should be proficient in designing valid grading procedures and in communicating assessment results to various stakeholders. Finally, they should recognize unethical, illegal, and inappropriate assessment methods and use of assessment information. The major novel contribution of the Standards was to recognize the importance of learning-oriented assessment alongside the traditional conception of assessment as equal to testing. Some researchers, however, argue that the Standards omitted two focal points that grew in importance in the beginning of the millennium: accountability movement imposing high-stakes tests teachers, especially in the USA, and the role of formative assessment (Brookhart, 2011; DeLuca et al., 2013).

The contribution of the Standards was developed particularly with teacher education in mind by Popham (2009) providing a collection of teacher- and student-oriented statements incorporating practical assessment knowledge and skills with principles of fairness and justice. In the field of language testing, Brindley (2001) was among the earliest to recognize the importance of assessment literacy for language testers and educators, and the issue started to appear as a theme at conferences and other fora (Hasselgreen, Carlsen, & Hellnes, 2003; Huhta & Tarnanen, 2007). The studies unanimously discovered an increased need of teacher training in more learner- oriented assessment practices.

Davies (2008) proposed an extensive set of goals for assessment literacy education comprising domains of skills, knowledge and principles. Following Brindley's model, Inbar-Lourie (2008) underscored the essence of the aspects of "why", "what" and "how" and outlined a tripartite model aligning practical and theoretical knowledge with socio-historical understanding of implications of assessment. She also underlined the value of intertwining learning and assessment by adhering to assessment-for-learning practices (Black & Wiliam, 1998) and to dynamic assessment (Poehner & Lantolf, 2005) in the field of assessment literacy (Inbar-Lourie, 2008, pp. 387–390).

The prominence of assessment in in-service and pre-service teacher education continues to manifest itself in assessment literature. Volante and Fazio (2007) advocated a systematic analysis of potential discrepancies between student-teachers´ assessment curriculum and their actual achievements. In Volante and Fazio's view assessment promoting life-long learning, involves three fundamental parts: practice, theory (e.g. reliability and validity issues) and philosophy (links to broader philosophic and pedagogic ideologies). Voss, Kunter, and Baumert (2011) tested an overall framework for student teachers' general pedagogical/psychological knowledge and suggested satisfyingly robust empirical structures entailing the following kinds of knowledge relating to: classroom management, teaching methods, classroom assessment, students' learning processes, and student characteristics. At the same time, Brookhart (2011) published an operationalization of TAL abilities and skills applicable to both classroom assessment and test administration. The sections of our study dealing with feedback, communication and scaffolding student autonomy are informed by Brookhart's work.

The theoretical model of TAL proposed by Fulcher (2012), incorporated practical knowledge, theoretical knowledge, and socio-historical understanding of assessment-related activity. He also appreciated the angle of student teachers

and their experiences of courses and academic study. In reverse to earlier proponents of a broad concept of assessment literacy, Fulcher acknowledged that competence at all the tree levels should not be required form all stakeholders. In addition to the conceptual level of language assessment knowledge (Pill & Harding, 2013), another persistent dilemma with definitions of language assessment literacy seems to be the profiling of various stakeholder constituencies (Harding & Kremmel, 2016). An attempt to solve this dilemma was made by Taylor (2013), when she suggested different profiles for various groups of stakeholders visualized as a spider-net like octagon comprising the following constituents: knowledge of theory, technical skills, principles and concepts, language pedagogy, local practices, personal beliefs/attitudes, and scores and decision-making. The profiles would incorporate varying amounts of each domain, predicating for example more language pedagogy for teachers than for test designers.

The knowledge aspects of previous models are revisited by Xu and Brown (2016) in a large-scale study. The knowledge constituents incorporated are Disciplinary knowledge and pedagogical content knowledge, Knowledge of assessment purposes, content and methods, Knowledge of grading, Knowledge of feedback, Knowledge of peer & self-assessment, Knowledge of assessment interpretation & communication, and Knowledge of assessment ethics. This structure has informed the present article to some extent (see Appendix 1).

The various pieces of research have modified the construct of TAL by giving prominence to different components and combinations of them. One of the most concise definitions frequently referred to is offered by Fulcher (2012), also chosen to guide the study at hand. According to Fulcher, TAL comprises the knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom-based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice.

As most components of TAL can be fostered through assessment instruction (DeLuca et al., 2013; Gotch & French, 2013), the role of STAL in teacher education and in scholarly publications has increased in the current decade, in North America (DeLuca, Lapointe-Mcevan, & Luhanga, 2016a; Kleinsasser, 2013), Europe (Vogt & Tsagari, 2014) and most recently, in Asia (Alkharusi, 2011; Xu & Brown, 2016). The conclusions drawn at different parts of the world tend surprisingly unanimous: teachers call for more targeted and context-sensitive training at all levels and throughout their career especially concerning formative assessment and adequate use of summative large-scale assessments. In one of the most comprehensive accounts of TAL standards to date, DeLuca, Lapointe-Mcevan, and Luhanga (2016b) mentioned a number of recommendations for designing contextually valid instrument to chart the needs. Besides validity and reliability concerns, there should be a consonance between assessment standards and theory and teachers´ actual practices.

2.2 TAL and STAL in the Finnish context

In Finnish educational discourse, standards at large are not very prominent, as we are a culture of wide pedagogical freedom. FNBE (The National Core Curricula) since 2003 introduce the CEFR scales for illustrating student achievement in language ability as an objective for instruction, but for teachers there are no standards except the course requirements of various universities educating teachers. These

are uniform merely in the sense of conforming to the overall Bologna framework, which does not specify subject-related teaching and assessment competences.

Assessment tends to be a rather neglected chapter in subject teachers´ pedagogical syllabi in many corners of the world. This is more or less true for Finland, where research-based teacher education, subject teachers included, is a well-established ideal. The number of credits devoted to issues of evaluation and assessment has diminished slowly but inevitably. Unofficially, students and teachers are familiarized with EALTA and ILTA standards, but there is no obligation to follow these standards.

The operative national core curriculum for basic education (FNBE, 2014) states that the major aim of pupil assessment is to guide and encourage learning and equip pupils with the skills of self-assessment. Creating a supportive atmosphere to promote self- and peer-assessment alongside reflection and cooperation between home and school are strongly advisable. These directives are more precise and consistent with research evidence underpinning effective teaching than in the previous normative documents (FNBE, 2004). Building on pupils existing knowledge base and adapting the instruction to meet their recognized needs are factors that are found to be effective in achieving planned outcomes (Hattie, 2012). Making pupils´ learning process visible is stated explicitly as a condition for fair and ethical assessment. In addition, the cyclic nature of assessment is manifested by the use of obtained assessment information in planning future stages of instruction, and communicating the objectives and achievement criteria to pupils at the onset of a study unit. (FNBE, 2014, pp. 49–50.) Assessment conducted during studies is expected to be a part of learning in the ethos of assessment as learning (Earl, 2003). Assessment also enhances life-long study skills and willingness to maintain and enrich one´s competences, echoing assessment for learning (Gardner, 2006). The primary focus of assessment is progress in proportion to earlier achievements and the agreed objectives. Of vital importance are also working skills including "planning, regulating and assessing one´s own work, acting responsively, and engaging in constructive interaction" (FNBE, 2014, p. 52). Both the mastery of subject-related and working skills are included in summative assessment provided numerically or verbally at the end of the school year. Pupils' behavior, by contrast, is assessed separately and should not be included in verbal assessments or grades in subject mastery. (NCC, 2014, p. 53.)

The final assessment at the end of basic education, as well as of upper secondary education, shall conform to a description of subject mastery required for grade 8 on a 4–10 scale. It is worth noticing that illustrative descriptions only exist for this single grade, while no transparent set of criteria is presented to distinguish between other grades below or above grade 8. For example, the strategic cut-point between pass and fail (grades 4 and 5) is left to the purview of teachers´ pedagogic freedom. Alongside the subject-related achievement, evidenced by versatile performances, the final grade also incorporates an effort component, the share of which remains undefined in the composition of the grade. Teachers are free to weigh these components in different ways.

There are studies to attest that grades given by teachers do not adequately reflect the knowledge measured in large-scale evaluations of learning outcomes (Hildén, Ouakrim-Soivio, & Rautopuro, 2016; Ouakrim-Soivio & Kuusela, 2012). In the present curricula a great value is put on versatile methods of assessment for and as learning (FNBE, 2014, 2015). The responsibility of transferring the fairly abstract ideas into the everyday life of schools, is attributed to teachers. In that

sense, Finnish teachers are left to face challenges of assessment competence on a daily basis without due preparation or support.

Although the ethos of assessment advocated by the operative core curricula resonates with the canons of formative assessment, prospective teachers need to be confident with more traditional requirements of test design, and above all, with issues of equity and fairness as exercised in various modes and functions of assessment. Assessment literacy with its multiple layers is to be promoted more strongly at all levels of teacher education, most naturally starting from the university-based pedagogical studies for subject teachers (Vogt & Tsagari, 2014). To make this happen, we need to know more about the process how student teachers shape their concepts of assessment and how their pedagogical thinking with regard to assessment issues develops over time.

As mentioned previously, our study is inspired and advised by examples of colleagues abroad, as research into STAL is not yet grounded in Finland. Even abroad, the validity of TA measures is found to be deficient (Gotch & French, 2014). Having dozens of TAL standards throughout the world DeLuca and his colleagues proposed a new instrument to support teacher assessment literacy based on four major themes in assessment literacy (DeLuca et al., 2016b). These are: assessment purposes, processes, fairness and measurement theory. Connected to the four themes, three subsequent approaches were distinguished, and a set of response options were derived from them (DeLuca et al., 2016b, pp. 250–257). The inventory was initially aimed at in-service teacher training, but it is easily applicable for prospective teachers of languages. Therefore, this blueprint for Assessment Professional Learning Priorities is one of the major inspirations to this paper about the context of teacher education in Finland. Use of the instrument is encouraged across educational systems by the team (DeLuca et al., 2016a, p. 248). Furthermore, Xu and Brown (2016) invite researchers to scrutinize local policies and seek ways to support teachers´ assessment knowledge and skills. The invitation recurs in further publications cited above (Brookhart, 2011; Vogt & Tsagari, 2014; Volante & Fazio, 2007).

3 Research questions

The prior aim of this study is to improve the quality of language teacher education to conform to the current needs recognized in the pedagogical literature and expressed in the National Core Curriculum. The research task is to explore assessment-related conceptions of student teachers and to map the impact of a short-term academic study unit on their development.

Based on the developmental goal of the study, the major research question is:

> What is the impact of an assessment-oriented study unit on student teachers´ assessment literacy (STAL) with regard to beliefs and planned practices?

To illuminate the main research question, two sub-questions are addressed:

> RQ1. What changes are detected in the exploratory factorial structure of STAL at the onset and at the end of the study unit?
>
> RQ2. What changes are detected in the pre-determined curricular domains of STAL during the course?

4 Conducting the study

4.1 Setting, data and method

The authors of this study are university lecturers of foreign language didactics and are in charge of didactic courses and seminars including lectures and group work in assessment and development. In addition to the intention to map the competence development of student teachers, this study aims to scrutinize the impact of an academic instructor´s regular work, and subsequently contribute to the professional development of a teacher educator (Koster et al., 2005). In addition, Cochran-Smith and Fries (2008) voice the benefit of asking important questions, constructing problems and selecting appropriate methods to solve them in teacher education.

This article reports a case study that belongs to an ongoing research project carried out jointly by the universities of Stockholm and Helsinki. The Finnish data derive from a series of 8 lectures (90 minutes each) in a study unit labelled Evaluation and development of teaching. The unit belongs to intermediate studies in pedagogy and is a mandatory component of subject didactic studies for prospective teachers of foreign languages. The scope of the entire unit was 7 ETCS, and the lectures granted 2 credits. The course requirements (syllabus) stated several objectives for the study unit, but this piece of research addresses the following:

The course syllabus dictates that after completing the study unit, students understand the different tasks, objectives, and forms of assessment and evaluation and the importance of feedback and evaluation in different phases of teaching, studying, and learning. They also appropriately use different assessment and evaluation methods for learning and apply opportunities of information and communication technologies in assessment, and critically evaluate their own solutions. (Faculty of Behavioral Sciences, University of Helsinki, 2016)

The titles of the lectures were:
Purposes, core concepts and principles of assessment
Common European Framework of Reference for Languages (CEFR) as the international guideline of assessment
National evaluation of learning outcomes in foreign languages 2013
Assessment in the new national core curricula with formative assessment in focus
Assessment of linguistic skills in various contexts: Listening and reading
Assessment of linguistic skills in various contexts: Speaking
Digitalization of the Matriculation examination
Current and future issues in language assessment

The lectures were delivered twice a week, and the students attended group sessions dedicated to assessment and development at the same time. The groups consisted of about 25 students each and an author of this article instructed two of them, while her colleagues taught the three other groups. At lectures, interaction and cooperative work was encouraged, for example discussions in pairs or in small-groups. At the end of the lecture series, students created a digital mock test (Kahoot) on assessment theory to each other. Focusing on principled application of assessment knowledge, the final exam was taken in pairs.

The students were attending their second semester of pedagogical studies for subject teachers. Their major was commonly a foreign language or the second domestic language Swedish or Finnish, and they had gained admission to the pedagogical studies after their first or second year at the Faculty of Arts. Some of them had already completed their master degree and even worked for a few years. These students were taking their pedagogical studies for their formal teaching qualification. Language programs tend to be dominated by female students and due to the low number of males, gender was omitted as a background variable.

The data were collected on a web-based questionnaire (in Finnish) delivered through the IT-services of the University of Helsinki. The first set (n=77) was collected at the very start of the first lecture (January 2017) to capture the students´ initial conceptions and understanding related to assessment. Nevertheless, these entries were not completely intact because the students already had completed their basic level studies including courses in didactics and the basic practice. Assessment is not explicitly set as a formal objective of those courses, but may well have been sporadically touched upon. The second data (n=65) set was gathered at the end of the last lecture (March 2017). The students responded using their own mobile devices.

The main sections of the questionnaire entailed:

- Background variables (3): duration of studies to date, language of expertise, and length of teaching experience
- Initial definitions (5) of certain core constructs (assessment, grading principles, validity, good assessment, functions of assessment
- Assessment practices (29, current or foreseen) on a scale of 1–5 (almost always, often, sometimes, seldom, never)
- Conceptions of assessment (38): 17 self-efficacy statements "I can…"/ "I know how to…" (8 value statements on grading, 6 case descriptions to elicit ethical reflections)

4.2 Measurement instrument and analysis

The composition of STAL is analyzed through the *conceptions* of teacher students. The scope of this study does not allow for an in-depth analysis of the various definitions of perception, conception and belief. In most research literature, they all share an affective and cognitive component that is modified by action or practice. (Borg, 2005; Barcelos & Kalaja, 2011.) Teacher education courses are found to have an impact on teacher beliefs and the way of recognizing and verbalizing them (Borg, 2011). Here the term "*attitude*" is used to refer to opinions, while "*real or planned actions*" refer to practices.

Appendix 1 illustrates the content sections of the survey questionnaire with a reference to the type of knowledge addressed by previous research in general, and by Xu & Brown (2016) in particular. The specific contents of the items were informed by the Finnish National Core Curricula for basic education (2014) in the first place, supplemented with indications of current research findings in regard to effective teaching (Hattie, 2012) and models of investigating STAL in other countries (DeLuca & Klinger, 2010; DeLuca et al., 2013). Items t1–t29 were statements concerning real or envisaged action, such as *I (plan to) inform students about targets and grading criteria at the beginning of courses*. The conventional Likert scale was used for eliciting the frequency (Almost always-often-sometimes-seldom-almost never). Items m1–m38 and v1–v8 implied statements regarding

self-perception (*I know how to design a written test*). The grade of agreement was indicated by a four-point scale ranging from *Strongly agree* to *Strongly disagree*.

In addition to the Likert-scaled statements, the students were asked to provide a verbal state-of-art definition of five key-constructs (e.g. reliability) and to suggest a solution to six scenarios. These sections are beyond the scope of this paper.

The envisaged model of the structure of student teachers' conceptions was subjected to analyses of range, variance and reliability at the scale level, which confirmed the relevance of the scales for Interpretation and communication (6 items), Versatile assessment techniques (12 items), professional self-esteem (17 items), Professional development (7 items), Professional support (3 items), as well as Grading perceptions (11 items). The scales for Feedback (4 items), Working skills (3 items) and Theory of summative assessment (6 items) were weaker but modestly consistent.

The impact of the lecture course was analyzed by means of descriptive and inferential statistics. Variables were explored for their normality, accuracy and suitability for further analyses. The overall reliability (Cronbach alpha) of the pre-questionnaire was 0.92 and for the post-questionnaire 0.93. The scale alphas for standardized items are given in Table 1. The philosophy dimension resorts to extensive qualitative evidence that cannot be tested by statistical means. The mutual correspondence between the single numeric variables and the sum variables can still be calculated. The sum variables aggregated for the sub-scales will be correlated with the background variables included into the questionnaire: years of study in general, the teaching subject and years of teaching experiment in both subsets (pre- and post-).

Although derived from previous research, predetermined categories may not match the actual outcome of a specific study. Hence, an exploratory factor analysis to uncover the componential structure of the two domains, attitudes and real or planned practices was run separately for each section. The potential changes between the two points of measurement (pre- and post-course) are reported separately for the factorial structure of both domains.

The composition of student teachers' assessment literacy (STAL) at the beginning of the assessment course and at the end of it was approached from a dual perspective. Firstly, by running an exploratory factor analysis separately for the entire blocks of variables pertaining to attitudes and practices without forcing any single variable to stand for a predetermined dimension.

To discover the factorial composition of conceptions attached to STAL required for RQ1, the pre-course and post-course datasets were subjected to a main component analysis and a factor analysis. To see whether an explorative factor analysis was in place, a Kaiser-Meyer-Olkin (KMO) test was run for the attitudes and actions data sets in the pre- and post- data respectively. Additionally, an anti-image correlation matrix was computed for all the four data sets.

For further exploration, the Promax rotation was selected, since it allows the components to correlate. In the context of this study, there was scarcely any reason to assume total absence of mutual interdependence between the components, rather the opposite.

Secondly, the anticipated model derived from literature and the particulars of the Finnish educational context, was tested for its immutability overtime and to expose changes that had occurred in single variables during the course.

5 Results

5.1 The factorial composition of STAL

5.1.1 Attitude

The main component analysis carried out for the pre-course survey on attitudes (46 variables), resulted in 14 main components with eigenvalues above 1 that explained 76% of the total variance. Actually, the first six components were most powerful exceeding 5% of the variance each, where after the explanation rate tangibly fell. The six first components counted for 52% of the total variance.

The obtained KMO values were not convincing for attitudes: Pre-course attitude = 0.567, Post-course attitude=0.579, while the recommended KMO value should exceed 0.6.

The number of responses in both datasets (pre n=75 and post n=65) was probably too low for a factor analysis. Moreover, the number of items attempting to cover the diverse field of assessment phenomena was probably too ambitious.

Based on the rate of meaningful explanation, a factor analysis proper was run with a predetermined number of six dimensions with eigenvalues higher than 2.0 for attitudes. The dimensions obtained by the rotated analysis are displayed in Table 1.

In the post-course data on attitudes, the main components were 14 in number and explained 80% of the total variance. In the post-course data, the most powerful six components with eigenvalues higher than 2.0 explained 56%.

The yielded dimensions of the rotated exploratory factor analysis with a preset number of six are reported in Table 1. The scale reliabilities based on standardized item values refer to the respective dataset. The item with the highest loading on the factor is given as an example in each column.

Table 1. Explorative factorial structure of attitudes in pre- and post-course data.

	1	2	3	4	5	6
pre *n*=77	Acquired confidence in assessment of multiple aspects of language ability in classroom *I can assess a student´s multiliteracy skills (to interprete texts, images, and multimedia content).*	Oral communication, individuality and professional support *Assessment methods should consider individual differences even when a single grade is given.*	Curriculum-based assessment, fairness at multiple levels *National school leaving tests are useful at the end of compulsory education.*	Communicative language use and timely feedback *I find that that it is easy to provide prompt feedback for students' achievements.*	Planning and assessment in alignment *It is important to inform students about targets and knowledge requirements early when teaching.*	Effort and process of study *Effort should be included in a course grade.*
	α=0.919	α=0.839	α=0.846	α=0.851	α=0.724	α=0.647

post *n*=65	Acquired confidence in assessment of multiple aspects of language ability in classroom *I can assess a student's ability to produce a written text according to criteria.*	Acquired confidence in classroom assessment and feedback *I know how to apply fair assessment procedures.*	Scaffolding and respect for individuality *Commitment to extramural activities in the target language is important in grading.*	Professional goal-oriented communicative assessment for learning *It is essential to map literacy in the target group.*	Assessing the level of linguistic skills *I can use the CEFR scale when assessing my students.*	Summative testing and grading *Oral tests are important in grading*
	α=0.920	α=0.923	α=0.631	α=0.766	α=0.782	α=0.754

As Table 1 portrays at a first glance, the scale reliabilities are satisfactory, even high. They also diminish in a logical order, the first being the most coherent. The dimensions emerging from both data, show somewhat differing patterns. In both data sets, the primary dimension related to confidence in assessing most of the essential components of communicative language teaching and ability including classroom assessment and cultural issues. That confidence is built through pedagogical studies. In addition to the example in Table 1, the statements addressed evaluation of the traditional four skills (*I can assess a student's multiliteracy skills; ability to produce a written text according to criteria; ability to understand and interpret a spoken text*) completed with perceived capacity of discussing mistakes and reasons for them during the instruction. Statements such as *I know how to cater for learning diversity in my assessment* and *I know how to apply fair assessment procedures* also appeared in this dimension.

In the post-course data the most well-established linguistic skills turned even more prominent than in the pre-data. The confidence in assessing multiliteracy skills pertained, but the first factor in the post-data showed increased belief in mastering summative and skill-oriented forms of assessment. Logically, this tendency is underlined by the only negative correlation for the statement *The major function of assessment is to promote learning*.

The second major attitude dimension in both datasets brought forth aspects of spoken communication and language use and the idea of diversity and individually adapted assessment. While the first dimension highlights the summative function of assessment, the next powerful one underscores the ongoing and learning-centered function based on transparent criteria. The negative correlation for the statement *All pupils should be assessed by the same principles when a single grade is given* underlines the overall formative spirit of this dimension. There is an additional focus on communicative language use in and out of school.

In the post-data the classroom orientation prevailed with slightly lower intensity, though. The need of professional support for teachers was present in the pre-data, while at the end of the course students strongly believed that their pedagogical studies had equipped them for conducting better assessment in classroom.

In the third dimension, there was a more obvious structural difference between the two data sets. The responses provided prior to the lectures focused curriculum-based guidelines and large-scale evaluations (e.g. *National evaluations are important*). On the contrary, individual feedback, scaffolding and assessment-as-learning aspects of classroom life predominated the post-data.

Communicative language use in classroom context supported by timely feedback were the major characteristics of the fourth dimension in the pre-data (e.g. *Giving regular feedback is feasible).* The classroom activity also implied assessing the learner's ability to respond and act appropriately in encounters with the target language culture and taking students' attitudes into account. The responses given after the course revisited language use, but underlined the need for support in professional development and purposively planned and conducted learning processes.

The theory-based alignment of planning and assessment was the theme of the fifth factor dimension in the pre-data. Statements such as *It is important to find out where pupils are in their learning process when planning courses or sequences of courses* and *It is important to inform pupils about targets and knowledge requirements early in your teaching* showed high consistency with this dimension.

For the post-data, on the other hand, the fifth dimension was notably influenced by the idea of dimensionality present in the CEFR scales voicing a more test-centered approach. This result aligned with the specific course syllabuses that put more emphasis on the process of teaching and formative assessment in the beginning of the pedagogical studies. The responses given at the onset of the assessment course naturally echoed the contents of the preceding courses. The goals of the assessment course incorporating more systematic forms of assessment were in turn reflected in the post-course survey.

In the last of the six dimensions effort and process of study were emphasized in the pre-course data (*Effort should be included into a course grade; Homework is important in setting a grade).* In the post-data, the sixth dimension accentuated summative forms of assessment and in-class (written and oral tests are important when setting a grade).

5.1.2 Practices

The analysis of real and intended practices (29 items) reported at the beginning and at the end of the lecture series was carried out parallel to the analysis of attitudes. The KMO values for real and intended actions were more satisfactory: 0.73 for pre- and 0.70 post-course data. Furthermore, the Bartlett test value (p<0.001) corroborated the suitability of the data for main component analysis.

In pre-course action data nine main components with eigenvalues above 1, explained 73% of the variance. After the six first components the explanatory power faded.

The primary main components counted for promoting cooperation and feedback. The consistency and appropriateness of the initial main component model was higher with regard to action variables than with attitude, but a more consistent pattern of six factors was computed the make the structure more transparent. Promax-rotation was selected, since it allows the components to correlate. The outcome of the rotated solution is displayed in Table 1. The alphas for scale reliability are adhered to as well as the statement with the highest loading to the particular factor.

In the post-course data for real and intended practices the main component analysis produced nine components with eigenvalue above 1 that explained 77% of the variance. An eigenvalue limit of 1.4 yielding six factors was chosen to enable comparison with the attitude strand and at the same time to maintain a reasonable explanation rate (61%). The rotated exploratory factor analysis with a preset number of six, yielded the dimension displayed in Table 2. The scale reliabilities refer to the relevant dataset.

Table 2. Explorative factorial structure of real and planned practices in pre- and post-course data.

	1	2	3	4	5	6
pre ***n*=77**	Cooperation and systematic action *I (plan to) cooperate in planning or carrying out assessment with other language teachers in my school.*	Feedback and communication *I (plan to) give regular feedback to every student during a course*	Versatile theory-based evidence *I (plan to) base my grading on evidence presented to students and their parents.*	Versatile summative assessment *I (plan to) include free production into my tests*	Flexible planning and goal orientation *I (plan to) inform students about targets and grading criteria at the beginning of courses.*	Summative assessment *I (plan to) inform students about targets and grading criteria at the beginning of courses.*
	α=0.890	α=0.896	α=0.806	α=0.865	α=0.893	α=0.738
post ***n*=65**	Cooperation and flexible planning based on feedback *I (plan to) provide prompt feedback to students for their achievements.*	Flexible systematic planning involving pupils *I (plan to) include free production into my tests.*	Systematic planning, professional development and communication *I (plan to) record assessment scores to track the performance of my students over a timespan.*	Modern and versatile theory-based evidence *I (plan to) use digital devices and web resources in assessment.*	Feedback, cooperation and communication *I (plan to) design a variety of assessment tasks that allow students to choose how they will demonstrate their achievement*	Summative assessment *I (plan to) weight spoken and written language skills equally when grading the students.*
	α=0.893	α=0.854	α=0.877	α=0.844	α=0.900	α=0.803

The scale alphas for the sum variables incorporated in the practice factors were relatively high, higher than for attitudes. The number of items (29) was lower and the entire instrument more concise.

The most powerful factor of practice statements in the pre-data involved co-operation and systematic action: *I (plan to) cooperate in planning or carrying out assessment with other language teachers in my school; I check or plan to check the national core curricula for aims and objectives when planning course assessment.* High value was

also put on professional development as a teacher and an assessor (*I [plan to] record assessment scores to track the performance of my pupils over a timespan)* and cognizant of theoretical principles of assessment.

In the post-data the formative ethos was intensified by plans to be flexible in planning and carrying it out jointly with colleagues. Prompt feedback on cooperative settings of classroom work was considered favourable as well as regular contacts with parents informing them about the progress of their child.

The second dimension in pre-data greatly centered on feedback and alternative, non-traditional modes of assessment. In the post-data, statements including purposive and well-prepared assessment based on the curricula bore highest correlations. The students intended to use tests, but these should include free communicative production.

In the pre-data, the third dimension of real or planned practice incorporated use of digital tools and net-based resources along with principled statistic evidence that is presented to the parents on a regular basis. The systematic manner of action recurred in the post-data enriched by recording assessment scores and evidence to track the progress of students and to develop one´s own assessment skills as a teaching professional.

In the pre-data the fourth factor centered around versatile summative assessment and grading procedures including student self-assessment. Free communicative production was planned as a part of language test given by the prospective teachers. At the end of the course, students additionally intended to compile theory-based accounts of assessment utilizing digital devices and cross-curricular projects.

The fifth factor accentuated objectives and assessment criteria as foundations of joint planning of teaching and learning in the pre-data. After the course feedback, the most important statements voiced a broad variety of working and evaluation methods applied in cooperation and communication *(*e.g. *I provide peer response to collaborative work; I [plan to] create opportunities for peer feedback whenever possible.)*

The sixth factor dealt mostly with summative assessment in both datasets. The planned composition of tests (for example *I [plan to] use multiple choice questions in my tests.)* was addressed in both. In the post-data the perspective was broader including weighing of certain linguistic skills (productive), and allowing students to have a word in setting the course grade.

Revisiting the course syllabus in light of the results the students seem to have moved in the intended direction reasonably well. At the end of the course, they are more aware of "the different tasks, objectives, and forms of assessment and evaluation and the importance of feedback and evaluation in different phases of teaching, studying, and learning." They also intend to apply a variety of assessment and evaluation methods for learning and apply opportunities of information and communication technologies in assessment, as well as critically evaluate their own solutions.

5.2 Predetermined domains

Next we turn to the predetermined sections of the survey. As presumed by RQ2, the theory-driven grouping of items of curricular relevance were scrutinized for their development during the lectures. The scale reliabilities are provided in Appendix 2. The dimensions intended to be measured were Feedback, Interpretation and communication, Versatile assessment techniques, Working skills, Professional self-esteem, Professional development and Professional support. No item was assumed to belong to more than one conceptual subfield.

As documented in Appendix 2, there were only few disparities between pre- and post-course data. The number of responses was lower at the end of the course, but the range and standard deviations remained rather stable. Increase of mean scores during the course occurred most notably in Versatile techniques and Professional self-esteem, whereas the mean of Interpretation and communication of assessment results and Grading diminished. The outcome is expected regarding the content of lectures and the parallel seminar studies emphasizing instruction in working skills supporting learner autonomy and formative feedback on one hand, and systematic training in item-writing and assessing the various strands of language education on the other hand. The drop of Interpretation and communication is slightly surprising, but may be explained by the practical limitations of an academic course without opportunities of meeting pupils or their parents in real life. This finding necessitates further exploration over a longer timespan involving internship. The Grading component covered priorities in setting a course grade and the sum score was primarily influenced by the decreased average score for the statement *Effort should be included in a course grade* from pre 3.16 to post 2.26.

The predetermined subfields were correlated for mutual dependency in pre- and post-data. Spearman correlation was applied because of the small size of the dataset. The item/respondent ration was compromised as the sample size should be three- to five-fold in relation to the number of items, which means that this questionnaire with 75 items in all would require a sample of 300–400 to meet the assumptions of mainstream statistical operations (Cattell, 1978).

The strongest mutual alignments in both datasets depicted by Appendix 3 pertain to the domains of Feedback, Interpretation and communication of assessment results, and Versatile assessment techniques. Working skills and grading joined this group in the post-course data. This may be due to the conceptual overlap of the domains operationalized in the questionnaire despite an attempt to keep them apart. The bonds even strengthened during the lecture series, incorporating assessment of learner working skills more largely than at the onset of the course. The second set of large mutual connections were, quite naturally, those related to aspects of teacher professionalism. The alignments of these domains were further strengthened in the post-course data.

Significance of the change during the course was tested with t-test and Mann-Whitney test for two samples across all the nine domains. T-test produced statistically significant changes in perceptions of confidence in Working skills (p=0.20; *eta*=0.039 a small effect), Professional self-esteem (p=0.018; *eta*=0.042, a small effect), Professional development (p=0.142; *eta*=0.017, a small effect), Summative assessment (p=0,013, *eta*=0.012, a small effect), and Grading (p=0.024, *eta*=0.038). Mann-Whitney test for small samples corroborated the impact of the lecture series on Working skills (0.014), Professional self-esteem (0.005) and Grading (0.014). No statistically significant change was detected in Feedback (p=0.313), Interpretation & communication (p=0.312), Versatile techniques (p=0.809), Professional support (p=0.315).

5.3 Changes in single statements

Furthermore, the changes detected in single statements were scrutinized by the means of Mann-Whitney test and paired samples t-test. Appendix 4 documents results of the two analyses for the items concerned.

Regarding the guidelines given by the National core curriculum and the course requirements responding to that, the aims of the lecture series were met in regard to the forms of assessment in different phases of teaching, studying and learning. We can assume that prerequisites for effective teaching employing formative assessment for learning were met, since student teachers expressed their intentions to engage pupils with planning their work, and equipping them with versatile feedback. The intention to inform pupils about targets and grading criteria, adjusting the teaching to the pupils' prior knowledge, and to offer optional assessment tasks to choose among were surprisingly declined across lectures. Considering that the ideology of versatile forms of assessment was accentuated throughout the lectures, the finding obviously necessitates further investigation. In contrast, the increased use of multiple-choice questions barely confirmed by the Mann-Whitney test, was not particularly encouraged during the lectures.

With respect to attitudes, the students, all of whom had recently taken the matriculation examination, appreciated national tests at the end of upper secondary education. By contrast, they had adopted a more critical stance towards the guidelines of the national curricula both in terms of its supportive value in guiding assessment generally and more specifically in proposing a composed grade including effort and knowledge. Most convincingly, the students stated their confidence in designing written tests, assessing speaking, conducting classroom assessment, and using the CEFR scales – all pivotal course priorities. They were also convinced of their ability to carry out reliable and valid procedures in grading. The findings are by and large consonant with those of DeLuca et al. (2013) and Gotch & French (2014) in attesting impact of assessment instruction to increase self-efficacy and mutual bonds between components of STAL. In this study, however, changes in the factorial structure could not be disentangled, as the structure changed during the lecture series. Students' perceptions about the impact of pedagogical studies with regard to assessment were not corroborated by the t-test at the end of the lectures. Further examination is needed to see the impact of seminar sessions and advanced internship.

6 Summary

The paper at issue documents a small-scale pilot of a localized development in quality work on assessment instruction to embellish the assessment literacy of prospective language teachers.

The main findings testify the claim of previous scholars (Borg, 2011; DeLuca et al., 2013) that student teachers' conceptions can be modified by research-based instruction. The study unit at hand had most impact on the attitudes and planned practices connected to assessment of working skills and professional development. The componential structure of STAL attitudes remained more stable than the construct of practices across the study unit. The major dimension at both points of time was Acquired confidence in assessment of multiple aspects of language ability in classroom. The envisaged or real practices underwent a more substantial transformation towards a more learner-centered architecture highlighting flexibility and communication. By and large, the results suggest that the course objectives resonating the operative national curricula and topical scholar knowledge of STAL were attained fairly well. The course also contributed to a more sophisticated insight in the state of the art of the STAL of language teacher

candidates. Among the predetermined domains, the most tangible changes were detected with respect to working skills and professional self-esteem that seemed to be the most sensitive domains in a short-term pedagogical intervention.

Being barely a pilot attempt in its strand in Finland, the validity of this study is subject to criticism. The sample size should be larger and students should receive individual codes when entering their responses to capture the level of progression for each individual student. The items need to be explored with more sophisticated tools based on the item-response theory, and new sets of adjusted statements should be presented to students at future courses. The connections of factors and domains to background variables are also of focal interest, as well as the conceptual analysis of the constructed verbal responses to be published later on. Following the recommendation by DeLuca et al. (2016) the target population and relevant stakeholders will be consulted to enhance and cultivate the questionnaire by additional items. The Finnish and Scandinavian community of assessment experts and teacher educators are also invited to suggest relevant supplements germane to the Finnish and Nordic context of language education.

References

Alkharusi, H. (2011). An analysis of the internal and external structure of the teacher assessment literacy questionnaire. *International Journal of Learning, 18*(1), 515–528.

American Federation of Teachers, National Council of Measurement in Education, & National Educational Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: National Council on Measurement in Education.

Barcelos, A. M. F., & Kalaja, P. (2011). Introduction to beliefs about SLA revisited. *System, 39*(3), 281–289.

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31.

Borg, S. (2005). *Teacher cognition and language education: Research and practice.* London: Continuum.

Borg, S. (2006). The distinctive characteristics of foreign language teachers. *Language Teaching Research, 10*(1), 3–31.

Borg, S. (2011). The impact of in-service teacher education on language teachers' beliefs. *System: An International Journal of Educational Technology and Applied Linguistics, 39*(3), 370–380.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 126–136). Cambridge: Cambridge University Press.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3–12.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum Press.

Darling-Hammond, L. (2008). The case for university-based teacher education. In M. Cochran-Smith, K. E. Demers, S. Feiman-Nemser, & D. J. McIntyre (Eds.), *Handbook of research on teacher education: Enduring questions in changing contexts (3rd edition)* (pp. 333–346). New York: Routledge and Association of Teacher Educators.

Davies, A. 2008. Textbook trends in teaching language testing. *Language Testing*, 25: 327–348.

DeLuca, C., Lapointe-Mcevan, D. & Luhanga, U. (2016a). Approaches to classroom assessment inventory: A new instrument to support teacher assessment literacy. *Educational Assessment, 21*(4), 248–266.

DeLuca, C., Lapointe-Mcevan, D. & Luhanga, U. (2016b). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251–272.

DeLuca, C., Chavez, T., & Cao, C. (2013) Establishing a foundation for valid teacher judgement on student learning: the role of pre-service assessment education. *Assessment in Education: Principles, Policy & Practice, 20*(1), 107–126.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education, 17*(4), 419–438.

Earl, L. (2003). *Assessment as learning*. Thousand Oaks: Corwin.

Faculty of Behavioral Sciences, University of Helsinki. (2016). *Degree requirements for the English medium subject teacher education program 2016–2017*.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

Gardner, J. (2006). Assessment for learning: a compelling conceptualization. In J. Gardner (Ed.), *Assessment and learning* (pp. 197–204). London: Sage.

Cochran-Smith, M., & Fries, K. (2008). Research on teacher education: Changing times, changing paradigms. *Handbook of Research on Teacher Education: Enduring Questions in Changing Contexts, 3*, 1050–1093.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice, 33*(2), 14–18.

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. In J. Banerjee & D. Tsagari (Eds.), *Handbook of second language assessment* (pp. 413–428). Handbooks of Applied Linguistics. Boston: Mouton de Gruyter.

Hasselgreen, A., Carlsen, C., & Hellnes, H. (2003). *European Survey on Language Testing and Assessment Needs. Report: part one – general findings*. Retrieved from http://www.ealta.eu.org/documents/resources/survey-report-pt1.pdf

Hattie, J. (2012). *Visible learning for teachers: A maximizing impact on learning*. London: Routledge.

Hilden, Raili; Härmälä, Marita; Rautopuro, Juhani; Huhtanen, Mari; Puukko, Mika; Silverström, Chris. (2015). *Outcomes of language learning at the end of basic education in 2013.* Helsinki: Finnish Education Evaluation Centre; Finnish National Board of Education, 2015. 28 s. (Information materials; Nro 2015:1).

Hilden, R., Ouakrim-Soivio, N., & Rautopuro, J. (2016). Kaikille ansionsa mukaan? Perusopetuksen päättöarvioinnin yhdenvertaisuus Suomessa [Fair marks for all! Equal and equitable grading in the end of basic education in Finland]. *Kasvatus*, *47*(4), 342–357.

Huhta, A., & Tarnanen, M. (2007). *Assessment and feedback practices in the Finnish comprehensive school*. Paper presented at conference Language, Education and Diversity. Hamilton, New Zealand, 21–24 November 2007.

Härmälä, M., Huhtanen, M., & Puukko, M. (2014). *Englannin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013* [Learning outcomes in English advanced syllabus in the end of basic education]. Publications 2014:2. Helsinki: Finnish Education Evaluation Centre.

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

Koster, B., Brekelmans, M., Korthagen, F., & Wubbels, T. (2005). Quality requirements for teacher educators. *Teaching and Teacher Education: An International Journal of Research and Studies, 21*(2), 157–176.

Poehner, M. E., & Lantolf, J.P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research, 9*(3), 233–265.

FNBE. (2004). *National Core Curriculum for Basic Education 2004*. Helsinki: Finnish National Board of Education.

FNBE. (2003). *National Core Curriculum for Upper Secondary Education 2003*. Helsinki: Finnish National Board of Education.

FNBE. (2014). *National Core Curriculum for Upper secondary Education 2015*. Helsinki: Finnish National Board of Education.

Ouakrim-Soivio, N., & Kuusela, J. (2012). *Historian ja yhteiskuntaopin oppimistulokset perusopetuksen päättövaiheessa 2011* [Learning outcomes in history and social sciences in the end of basic education]. Helsinki: Finnish National Board of Education.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381–402.

Popham, W.J. (2009) Assessment Literacy for Teachers: Faddish or Fundamental? *Theory Into Practice, 48*(1), 4–11.

Schulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. *Harward Educational Review, 57*(1), 1–22.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412.

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education, 30*(3), 749–770.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374.

Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology, 103*(4), 952–969.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 2016*(58), 149–162.

APPENDICES

Appendix 1. The composition of the instrument for measuring Finnish student teachers' conceptions of STAL.

Domains of STAL (DeLuca & Klinger, 2010; DeLuca, 2016a; Xu & Brown, 2015)		
	Attitudes	Real or planned actions/practices
Practice		
Feedback		t5, t6, t7,t28
Interpretation & communication	m32	t22, t25, t10, t11, t12,
Versatile assessment techniques	m1, m2, m3, m4, m34	t1, t8, t9, t13, t23, t24, t27
Working skills		t2, t3, t4,
Professional self-esteem	m5 – m15, m19, m28–29, m30, m33, m38	
Professional development	m35, m36	t19, t20, t21, t26, t29
Professional support	m_18, m_20, m_21	
Theory		
Summative	m16–17, m27	t16, t17, t18,
Grading	d2, m37, v1–v8	t14, t15,
Philosophy		
Ethics and fairness	m22, m23, , c2, c4–6	
Diversity	m31, c1, c3	
Validity	d3, m_24,	
Reliability	m25, 26,	
Construct of assessment	d1, d4, d5	

Appendix 2. Descriptive statistics on the sum variables of pre-determined subfields of STAL.

Domain (number of items)	N pre/post	Mean pre/post	SD pre/post	Scale alpha pre/post
feedback (4)	74/60	14.97/15.30	1.96/1.73	0.58/0.59
interpretation & comm (6)	71/62	**19.87/19.18**	3.97/3.91	0.84/0.84
versatile techniques (12)	73/59	40.29/40.46	3.75/4.30	0.70/0.74
working skills (3)	75/63	**12.89/13.51**	1.53/1.53	0.57/0.69
prof_ self-esteem (18)	73/59	**45.77/48.71**	6.94/7.10	0.90/0.92
prof_ development (7)	72/59	18.24/18.93	2.70/2.67	0.70/0.73
summative assessment (6)	71/61	20.15/20.69	2.42/2.46	0.52/0.55
grading (12 items)	75/60	**10.28/9.68**	1.60/1.38	0.68/0.78
prof_support (3)	74/61	6.61/6.90	1.61/1.80	0.61/0.78
Valid N (listwise)	62/51			

Appendix 3. Correlations (Spearman) across the sum variables of pre-determined domains of STAL in the pre- and post data.

Pre	feedback	int&com	versatile	work	prof_self	prof_dev	summat	grading	prof_supp
Feedback	1	.491**	.535**	.472**	.420**	.361**	.217	.296*	.17
int&com	.491**	1	.731**	.491**	.175	.217	.335**	.437**	.052
Versatile	.535**	.731**	1	.461**	.327**	.361**	.166	.386**	.02
Work	.472**	.491**	.461**	1	.03	-.002	.08	.313**	-.279*
prof_self	.420**	.175	.327**	0.03	1	.654**	.291*	0.224	.476**
prof_dev	.361**	.217	.361**	-.002	.654**	1	.22	.291*	.658**
Summat	.217	.335**	.166	.08	.291*	.22	1	0.233	.184
Grading	.296*	.437**	.386**	.313**	.224	.291*	.233	1	-.016
prof_supp	.17	.052	.02	-.279*	.476**	.658**	.184	-.016	1
Post	feedback	int&com	versatile	work	prof_self	prof_dev	summat	grading	prof_supp
Feedback	1	.490**	.557**	.666**	.355**	.326*	.326*	.550**	.177
int&com	.490**	1	.659**	.505**	.194	.351**	.168	.332**	.173
Versatile	.557**	.659**	1	.489**	.318*	.388**	.336*	.481**	.079
Work	.666**	.505**	.489**	1	.290*	.383**	.438**	.460**	.262*
prof_self	.355**	.194	.318*	.290*	1	.657**	.344**	.285*	.506**
prof_dev	.326*	.351**	.388**	.383**	.657**	1	.304*	.305*	.713**
Summat	.326*	.168	.336*	.438**	.344**	.304*	1	.483**	.24
Grading	.550**	.332**	.481**	.460**	.285*	.305*	.483**	1	.097
prof_supp	.177	.173	0.079	.262*	.506**	.713**	.24	.097	1

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Appendix 4. Statistically significant changes detected in planned practices and assessment related attitudes prior to the lecture series and after.

Domain	Statement	Phase pre/post	N	Mean	SD	M-W	t-test Cohen´s d
Working skills	I (plan to) inform students about targets and grading criteria at the beginning of courses.	pre	75	4.77	0.509		
		post	63	4.38	0.705	000	000 d=0.06
Working skills	I (plan to) find out where students are in their learning process when planning courses or sequences of courses	pre	75	4.39	0.715		
		post	63	4.71	0.551	003	003 d=0.050
Working skills	I (plan to) invite language learners to participate in the planning of my courses.	pre	75	3.73	0.827		
		post	63	4.41	0.687	000	000 d=0.09
Feedback	I (plan to) adjust my teaching whenever necessary. for example as the result of feedback or course evaluations.	pre	75	4.19	0.608		
		post	61	3.9	0.831	040	027 d=0.048
Feedback	I (plan to) give regular feedback to every student during a course.	pre	75	3.93	0.777		
		post	63	4.24	0.756	015	022 d=0.040
Versatile assessment techniques	I (plan to) design a variety of assessment tasks that allow students to choose how they will demonstrate their achievement.	pre	74	4.22	0.668		
		post	63	3.94	0.716	023	020 d=0.040
Feedback	I (plan to) create opportunities for peer feedback whenever possible.	pre	75	3.44	0.702		

		post	63	3.75	0.74	024	014 d=0.043
Summative assessment	I (plan to) use multiple choice questions in my tests.	pre	74	3.18	1.052		
		post	63	3.49	0.896	040	059 d=0.037
Summative assessment	National school leaving tests are useful at the end of upper secondary education.	pre	74	2.93	0.709		
		post	62	3.21	0.604	021	017 d=0.081
Professional support	The guidelines of the national core curricula are sufficient to support teacher assessment work.	pre	75	2.57	0.64		
		post	62	2.31	0.759	048	030 d=0.031
Professional self-esteem	The teaching to date (on campus and during school placement or practicum) with regard to assessment and grading has met my needs as a student teacher	pre	75	1.96	0.761		
		post	62	2.55	0.694	000	070 d=0.081
Versatile assessment techniques	I can assign grades for tests and examinations using procedures that are reliable and valid.	pre	74	2.32	0.599	0.07	
		post	61	2.52	0.673	0.035	086 d=0.031
Philosophy (reliability)	All students should be assessed by the same principles when a single grade is given.	pre	75	2.52	0.811		
		post	62	2.82	0.713	029	000 d=0.039
Professional self-esteem	I know how to design a written test.	pre	74	2.64	0.821		
		post	62	3.03	0.572	002	023 d=0.055
Professional self-esteem	I know how to assess students' oral proficiency.	pre	74	2.66	0.688		

		post	62	2.9	0.646	038	001 d=0.355
Professional self-esteem	I feel competent to conduct classroom assessment.	pre	75	2.27	0.759		
		post	62	2.53	0.671	017	037 d=0.036
Versatile assessment techniques	Giving regular feedback to every pupil is feasible.	pre	75	2.85	0.608		
		post	62	2.32	0.763	000	034 d=0.077
Grading	Effort should be included in a course grade.	pre	75	3.16	0.736		
		post	61	2.26	0.794	000	000 d=0.118
Professional self-esteem	I can use the CEFR scale when assessing my students.	pre	73	1.92	0.722		
		post	62	2.74	0.788	000	000 d=0.064

Received August 22, 2017
Revision received December 14, 2017
Accepted February 13, 2018