

<https://helda.helsinki.fi>

---

## Swedish and Finnish Pre-Service Teachers Perceptions of Summative Assessment Practices

Hilden, Raili

Multidisciplinary Digital Publishing Institute  
2022-01-05

---

Hilden, R.; Oscarson, A.D.; Yildirim, A.; Fröjdendahl, B. Swedish and Finnish Pre-Service Teachers Perceptions of Summative Assessment Practices. Languages

---

<http://hdl.handle.net/10138/349188>

---

*Downloaded from Helda, University of Helsinki institutional repository.*




*This is an electronic reprint of the original article.*

*This reprint may differ from the original in pagination and typographic detail.*

*Please cite the original version.*

## Article

# Swedish and Finnish Pre-Service Teachers' Perceptions of Summative Assessment Practices

Raili Hilden <sup>1,\*</sup>, Anne Dragemark Oscarson <sup>2</sup>, Ali Yildirim <sup>2</sup> and Birgitta Fröjdendahl <sup>3</sup><sup>1</sup> Department of Education, University of Helsinki, 00014 Helsinki, Finland<sup>2</sup> Department of Pedagogical, Curricular and Professional Studies, University of Gothenburg, 405 30 Gothenburg, Sweden; anne.dragemark@ped.gu.se (A.D.O.); ali.yildirim@gu.se (A.Y.)<sup>3</sup> Department of Language Education, Stockholm University, 106 91 Stockholm, Sweden; birgitta.frojdendahl@isd.su.se

\* Correspondence: raili.hilden@helsinki.fi

**Abstract:** Summative assessments are an exercise of authority and something that pupils cannot easily appeal. The importance of teachers being able to assess their pupils correctly is consequently both a question of national equivalence and individual fairness. Therefore, summative assessment is a paramount theme in teacher education, and we aimed to investigate the perceptions and competence of student teachers regarding common summative assessment practices. The study was conducted at three universities, two in Sweden and one in Finland involving prospective language teachers responding to an online survey (N = 131). In addition, interviews were carried out with 20 Swedish and 6 Finnish student teachers. The analysis of the data indicates that student teachers value practices that enhance communication and collaboration as well as the curricular alignment of summative assessments. With respect to perceived competence, the respondents in general felt most confident with deploying traditional forms of summative assessment, while they were more uncertain about process evaluation and oral skills. Regarding significant differences in the participants' perceptions of competence among the three universities, Finnish university students reported higher levels in all variables. However, room for improvement was found at all universities involved.

**Keywords:** language teaching; assessment literacy; summative assessment; teacher education; pre-service teacher education



**Citation:** Hilden, Raili, Anne Dragemark Oscarson, Ali Yildirim, and Birgitta Fröjdendahl. 2022. Swedish and Finnish Pre-Service Teachers' Perceptions of Summative Assessment Practices. *Languages* 7: 10. <https://doi.org/10.3390/languages7010010>

Academic Editors: Dina Tsagari and Henrik Bøhn

Received: 26 September 2021

Accepted: 16 December 2021

Published: 5 January 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

### 1.1. Background

The main function of summative assessments (SA) is, in essence, to sum up development and achievement and thus determine the level of students' knowledge. These assessments are often the basis for decisions teachers make both regarding further teaching strategies and individual help to learners. They are also often high stakes in the form of intermittent or final grades and those responsible for making these judgments need to be both competent and knowledgeable. Furthermore, teachers need to be able to make ethical decisions as these assessments most often have a crucial impact on the future lives of learners, being the basis for further education and future career choices (Yildirim et al. 2021). This exercise of authority in the form of, for example, setting final grades, is something which pupils cannot appeal and thus teacher assessment literacy (TAL) is of vital importance both for national equivalence and individual fairness.

A number of studies have been carried out regarding the grading of national tests in Sweden and several problem areas have been identified, for example, that "the assessment of a student's performance on the tests can be clearly linked to the teacher who makes the assessment" (Riksrevisionen 2011, p. 55), and that assessment and grading may be influenced by factors not set by the national criteria (Klapp Lekholm 2011; Thorsen and Cliffordson 2012). In another study, Oscarson and Apelgren (2011) found that a nationally

representative sample of Swedish teachers used free written production tasks and classroom observations as their main determinants for assessments and grading while at the same time they found setting grades “difficult” or “very difficult.” Inconsistency in teacher grading has also been identified in Finland. For example, teachers tend to include attitude in school grades (Hildén and Rautopuro 2014). Furthermore, the correspondence between grades and learning outcomes seems to be insufficient in some school subjects (Hilden and Fröjdendahl 2018). In both Sweden and in Finland, teachers’ assessment literacy appears to be an area of concern due to inconsistencies in grading. In both countries, these indications are a challenge to equitability and point to the need for further investigation.

To address this gap, a Swedish Research Council funded project aimed to provide further knowledge about how pre-service teachers and, subsequently, novice teachers develop summative assessment literacy (SAL) during the initial phase of their profession, and in this way increase teachers’ ability to make fair and reliable assessments of their pupils’ language proficiency and results.

The present study is part of this larger project carried out at three universities in Sweden and Finland. All three are large universities when it comes to teacher education, and the project specifically focuses on their programs for language teachers. They are here denoted as University A, B (Sweden), and C (Finland). In Sweden and Finland, teacher education (TE) is part of university and not a separate institution as in some other countries. All basic undergraduate programs are 3 years while some professional programs may be somewhat longer. When it comes to teacher training in language education for lower- and upper secondary schools, the TE programs in Sweden are either a 4.5–5.5-year program course or a 1.5-year certificate course if the student already has an undergraduate degree. In Finland, all subject TE programs include 1 year of studies in education at the Masters level on top of their undergraduate degree. In both countries one year of studies equals 60 credits and both national and local criteria define the level of assessment literacy required.

Both formative and summative assessment are focused on course documents (syllabus) and a learning culture rather than a testing culture is advocated. At the two Swedish universities both embedded and “stand-alone” courses, i.e., courses exclusively dedicated to assessment, are given. In Finland, on the other hand, assessment is embedded and addressed in all courses dealing with curriculum development and teaching. In both countries, school practice courses are also expected to provide development of the pre-service teachers’ assessment skills (Yildirim et al. 2021) (see Appendix A for descriptions of courses dealing with assessment literacy).

## 1.2. Purpose

In this article we aim to investigate pre-service teachers’ perceived values and competencies of language assessment literacy (LAL) in three teacher education programs in Sweden and Finland.

The specific research questions set out to explore:

1. How do pre-service teachers perceive the importance of SAL practices, and how do these values differ by university?
2. How do pre-service teachers evaluate their own competence in applying SAL practices, and how do these evaluations differ by university?

The larger project is seen as imperative, since despite the growing number of studies on assessment literacy, follow-up studies are relatively scarce. This paper depicts the first phase of a longitudinal trajectory. Secondly, summative assessment focus is scantily scrutinized in pre-service teacher research. Moreover, there are indications of unnecessary tensions between formative and summative assessment resulting in summative assessment being perceived as outdated and thereby, it has been neglected in research endeavors (Lau 2016).

In the current paper, the composition of pre-service language teachers’ assessment literacy (LTAL) and the efficacy of their study programs were analyzed through *perceptions*. In most of the research literature, they all share an affective and cognitive component that is

modified by action or practice (Barcelos and Kalaja 2011; Borg 2006) and teacher education courses are found to have an impact on teacher beliefs and the way they recognize and verbalize them (Borg 2011).

In educational literature, there is a mass of definitions of perceptions, cognitions, conceptions, and beliefs, some of which rely more on the respondents' experience, others on their emotions and attitudes. In previous research on teachers' dispositions, these terms and constructs have been used partly interchangeably and partly without exact definitions. In this article we use the term perception to refer to a relatively long-lasting belief or opinion, an intuitive insight and understanding, or an interpretation based upon such understanding (Oxford English Dictionary Online 2021). Perception is the way someone thinks or feels about an object based on and affected by sensorial stimuli, experience, or an organization, and an individually and culturally determined interpretation (Qiong 2017).

Since the study regards university students in teacher education, we henceforth use the term "student teacher" synonymously with "pre-service teacher." "Pupils" is used to specifically refer to school children.

## 2. Literature Review

The following review covers conceptions of assessment literacy as well as previous studies on pre-service teachers' assessment literacy.

### 2.1. Perceptions of Assessment Literacy

The traditional conceptions of assessment refer exclusively to summative assessment, whereas classroom-based measures to give feedback and guide learning have been recognized and termed as assessment only in the last few decades (Scriven 1967). Yet the crucial determinant of whether we are talking about summative or formative assessment is not about the place or technique of assessment, but the purpose it is intended to serve and the consequences that assessment decisions have for the pupils and other stakeholders. In the case of formative assessment, the decisions are low-stakes, which means that they do not affect the pupils' future in any significant way.

The evidence base for formative decisions is also typically flexible and easily cancelled or modified; the student can do better next time and obtain more favorable feedback for the effort. In general, the major purpose of formative assessment is to promote learning rather than issue ultimate statements about the existing level of knowledge and proficiency (Brookhart 2011; Cizek et al. 2019).

The major function of summative assessment, on the other hand, is to determine the level of student knowledge or/and the effectiveness of a study program. Typical instances of summative assessments are end-of-unit or end-of-term (course, school years, educational level) grades, national evaluations of learning outcomes, school-leaving exams, and university admission tests (Yildirim et al. 2021). The evaluations, for example, are assumed to sum up the development or achievement preceding the implementation. The results can be used for accountability purposes, or they can merely serve informative and guiding ends (Pizorn and Huhta 2016, p. 243). In addition to tests and exams, summative evidence can incorporate diverse data evaluated both quantitatively and qualitatively (Taras 2005). Educational policy, curricular work, and teaching and learning are modified based on summative assessment (Pellegrino 2018, p. 411). Due to the decisive role that summative assessments play for individual students, it is of utmost importance that teachers are competent, professional assessors with high-level ethical principles from the early stages of their education as a major component of their assessment literacy.

Assessment literacy can be conceptualized as a sub-component of teacher cognition, which is a dynamic and socially determined construct. Borg (2015, p. 1) defines "teacher cognition as what teachers "think, know and believe—and its relationship to teachers' classroom practices." A key factor behind the increased research in teacher cognition in general, not just in language education, has been recognition of the fact that teachers are active and reflective decision-makers who play a central role in shaping classroom events

(Borg 2015, p. 1). This foundational statement perfectly fits the ethos of LAL and qualifies it as one of the cornerstones of teachers' professional competence (Caena 2011, p. 24).

Assessment of the literacy of teachers in general and language teachers in particular has gained increased importance alongside the rapidly changing cultural and educational environment and the challenges these circumstances pose for teaching professionals. Originally, TAL was defined by Stiggins (1991) and Popham (2011) in an appeal for proper training in assessment strategies in teacher education programmes. Popham's argument rested on the conviction that "assessment literacy consists of an individual's understanding of the fundamental assessment concepts and procedures deemed likely to influence educational decisions" (Popham 2011, p. 267).

Since these contributions, a mass of theoretical models of language teacher assessment literacy have been proposed, some of them more empirically based than others (for a more comprehensive account, see Pastore and Andrade 2019). The basic structure reiterated in subsequent LTAL literature expanded from knowledge to capture practical skills (Davies 2008), and further to comprise principles as well. The principles section dates back to Messick's (1989) conception of consequential validity and has inspired ever expanding ethical issues incorporated in the frameworks of test usefulness (Bachman and Palmer 1996), argumentation (Bachman and Palmer 2010; Chapelle et al. 2008; Kane 2013), or justice and fairness (Kunnan 2018) in validity conceptualizations. The role of ethical principles, fairness, and consequences has become something of a watershed in the field between empiricist and post-modernist philosophies around TAL. As Fulcher (2012, p. 117) points out, sound reasoning is preferred to avoid extreme pendulum swings in one direction or the other. There is also a delay in the appearance of valid theoretical innovations in the space of assessment literacy constructs, and an even longer delay until they are adopted and amalgamated in teachers' everyday work.

As the early models of LTAL were primarily theoretically founded, the recent developments have aimed to shape the theory through empirical inquiry (Table 1). In his conceptualization of LTAL, Fulcher (2012), incorporated practical knowledge, theoretical knowledge, and socio-historical understandings of assessment-related activity, while Taylor (2013) proposed different profiles of LTAL for different stakeholder groups, such as teachers, researchers and policymakers. The major components of her conceptualization were knowledge of theory, technical skills, principles and concepts, language pedagogy, sociocultural values, local practices, personal beliefs, scores, and decision making.

Kremmel and Harding (2020) based their study on Taylor's conceptualization when conducting a large-scale empirical survey covering 77 different countries to illuminate the various needs perceived and the dimensionality of language assessment literacy across multiple groups. This study provided an interesting comparison between the hypothesized profiles suggested by Taylor (2013), and the factorial solution yielded through empirical study. The results revealed both matches and deviations from the Taylor scheme. The factorial composition comprised nine dimensions: (1) developing and administering language assessments, (2) assessment in language pedagogy, (3) assessment policy and local practices, (4) personal beliefs and attitudes, (5) statistical and research methods, (6) assessment principles and interpretation, (7) language, structure, use and development (8) washback and preparation, and 9) scoring and rating.

In the North American context, DeLuca and his colleagues (2016) based their instrument to support LTAL on eight major themes in assessment literacy (DeLuca et al. 2016): assessment purposes, processes, communication of results, fairness, ethics, measurement theory, assessment for learning, and support and education for teachers. Other models were revisited by Xu and Brown (2016), whose knowledge constituents incorporated disciplinary knowledge and pedagogical content knowledge, knowledge of assessment purposes, content and methods, knowledge of grading, knowledge of feedback, knowledge of peer and self-assessment, knowledge of assessment interpretation and communication, and knowledge of assessment ethics. This knowledge base and teacher conceptions layers provide a useful frame to guide our study.

As [Pastore and Andrade \(2019\)](#) state “the field is now cluttered with competing definitions and models, some of which overlap with related concepts, whose distinctions have not been clearly articulated” (p. 130). Therefore, they sought to merge and re-conceptualize the most relevant contents into a model of three related dimensions: conceptual, praxeological, and socio-emotional that aimed to cover and cut across the necessary knowledge, skills, and dispositions. The empirical strand of their study entailed a survey carried out with a Delphi-approach, resulting in a model that bears a lot of resemblance with earlier models that fundamentally build on the same triad.

Despite the miscellaneous definitions, the basic structure of LTAL incorporates assessment-related knowledge and skills, accompanied with dispositions (occasionally labelled in other terms). Knowledge concerns the underlying principles of various types of assessment, skills deploy the knowledge in the one’s context of work, while dispositions refer to hosting beneficial perceptions and a readiness to reflect ethical issues to deepen one’s knowledge base and to develop and modify one’s competences in accordance with perceived needs.

Since there are no pre-determined standards for teacher assessment literacy in the Nordic countries, we rely on the well-established notions of knowledge, skills, and dispositions as major categories. With regard to each of them, student teachers need to develop a knowledge base and the skills to execute and to hold favorable dispositions. The cross-cutting themes or key components addressed in the current study are criterion and norm referenced testing, test construction, grading, validity, reliability, alignment, transparency, and cooperative assessment. These dimensions conceptualize the summative strand of assessment for pre-service education in the research project this article belongs to.

**Table 1.** Overview of components of language teacher assessment literacy (along dimensions based on Pastore and Andrade 2019).

	<b>Conceptual Knowledge What Conceptions a Teacher Has of Assessment</b>	<b>Praxeological Dimension How to Monitor, Judge, and Manage the Teaching–Learning Process</b>	<b>Socio-Emotional Dimension How Teachers Manage the Social and Emotional Aspects</b>
<a href="#">Fulcher (2012)</a>	* theoretical knowledge	* practical knowledge	* socio-historical understanding of assessment-related activity
<a href="#">Taylor (2013)</a>	* knowledge of theory principles and concepts	* technical skills, language pedagogy, local practices * scores and decision making.	* sociocultural values, * personal beliefs
<a href="#">Xu and Brown (2016)</a>	* Disciplinary knowledge and pedagogical content knowledge, * Knowledge of assessment purposes, content and methods	* Knowledge of grading, * Knowledge of feedback * Knowledge of peer and self-assessment * Knowledge of assessment interpretation and communication	* Knowledge of assessment ethics
<a href="#">DeLuca et al. (2016)</a>	* Assessment purposes, * Measurement theory * Assessment for learning	* Processes * Communication of assessment results * Education and support for teachers	* Fairness * Ethics
<a href="#">Pastore and Andrade (2019)</a>	* theories * models * purpose, object * methods, * data analysis * reporting and communication	* alignment of learning and assessment, * data-gathering * interpreting evidence * curriculum adaptation * communication, stakeholders engagement, * scaffolding pupils	* collaboration across stakeholder groups * consciousness of one’s role as assessor, * ethical aspects, power and impact issues
<a href="#">Kremmel and Harding (2020)</a>	* assessment policy and local practices * statistical and research methods * language, structure, use and development	* developing and administering language assessments * assessment in language pedagogy washback and preparation, scoring and rating	* personal beliefs and attitudes * assessment principles and interpretation



## 2.2. Previous Studies on Pre-Service Teachers' Assessment Literacy

Student teachers' cognition in general, and TAL in particular, seem to be promoted by their academic studies (Borg 2015; DeLuca and Klinger 2010; Volante and Fazio 2007), with behavioral and cognitive changes detected pre- and post-course designs. Since training enhances teachers' self-esteem and identity as expert assessors (DeLuca et al. 2013; Volante and Fazio 2007), it essentially should belong to teachers' qualifications and requirements, and increased scholarly endeavors to chart the quality and usefulness of pre- and in-service education have been elusive (Coombe et al. 2020).

Studies on student teachers' assessment literacy (STAL) have been carried out in several countries and contexts during the two last decades. The majority of them have targeted in-service language teachers' knowledge-base and needs, but a steadily increasing number of scholars have directed their interest towards prospective language teachers. Some studies also engage instructors, and more rarely, researchers and administrators.

One of the earliest documented investigations on language assessment courses was conducted by Bailey and Brown (1995). It was targeted to instructors and entirely dedicated to testing and summative assessment and quite naturally in the time prior to the teacher standard movement. The most frequent activity on the examined courses was test critique, followed by item writing and interpreting test scores. In a follow-up study on the trajectory of ten years, they contended that the field of assessment is dynamic, as mirrored by the new items related to classroom practice and ethics, but at the same time their results indicate "the presence of a stable knowledge base that is evolving and expanding rather than shifting radically" (Bailey and Brown 1995, p. 371). They call for further scrutiny on the heavily growing amount of assessment education globally and locally. The students' attitudes were touched on only indirectly via their instructors' perceptions, but the overall impression transmitted was that students' knowledge, skills, and attitudes towards assessment grew to be more positive during the course implementation.

Pursuits to unpack the factorial structure of STAL have intensified during the last decade. The study by DeLuca and Klinger (2010) focused on Canadian pre-service teachers (288 in number), exploring the factorial structure of their assessment literacy in the customary dimensions of knowledge, skills, and philosophy. Within the practice domain, the five-factor solution revealed design and marking, provincially mandated assessment practices, technical knowledge of summative assessment practices, assessment for learning, and types of assessment as the major components emerging from the inquiry. The knowledge domain consisted of four dimensions, namely, assessment-of-learning theory, theoretical principles of assessment-of-learning and assessment-for-learning, assessment item formats, and statistical techniques for assessment. With respect to the philosophy domain, they found six factors comprising the philosophy of large-scale assessment and classroom assessment, respectively, issues of reliability and validity, and rationale for assessment decisions and practices, and the articulation of a personal philosophy of assessment.

Hilden and Fröjdendahl (2018) conducted a study on 77 Finnish teacher students in modern languages. These took an introductory course in language assessment during their second semester of pedagogical studies, where after they completed their advanced teaching practice at a university training school. The students responded to a survey in the beginning and at the end of the semester. The results indicated that the componential structure of assessment-related conceptions remained relatively stable over the timespan, whereas the real or envisaged practices underwent a more substantial transformation towards learner-centeredness, flexibility, and communication. The short-term pedagogical intervention was most influential with regard to working skills and professional self-esteem.

The focal nature of in-service teacher education is univocally underscored as a mediating factor that shapes pre-service teachers AL through learning experiences, context and personal dispositions (e.g., DeLuca et al. 2020), although the quality and impact do not often meet the everyday needs of teaching professionals. As pointed out by DeLuca et al. (2013), the alignment of in-service teacher education policies, standards, and course curriculum is insufficient and numerous improvements should be introduced into teacher preparation.

Similar findings tend to recur all around the world. Assessment courses do not correspond to the needs of the teaching field (Giraldo and Murcia 2018; Vogt and Tsagari 2014). Topics, such as social dimensions and fairness of assessment, are not sufficiently covered in teacher education programs (Lam 2015). Further gaps have been detected with respect to summative assessment, purposes of testing, and grading (Vogt and Tsagari 2014). Misalignment of theory and practice constitute a widely acknowledged challenge (Inbar-Lourie 2008; Lam 2015). Furthermore, the efficacy of assessment education is assumed to vary according to student orientations due to their background factors and dispositional approaches (DeLuca and Klinger 2010), suggesting a more discerned implementation of delivery (Coombs et al. 2020). All the above mentioned studies are unanimous in voicing the demand for more high-quality training courses and equal opportunities for pre- and in-service language teachers.

An interesting issue worth mentioning here is the concern for the need of a more curriculum-based assessment literacy, originally expressed by Brindley (2001), which has been voiced by numerous scholars since. The situation in Sweden and Finland today seems to be quite the opposite. In contrast with accountability driven educational contexts, formative assessment is accentuated in the language curricula of both countries, and high-stakes tests apart from the national tests are scarce. Many of the national tests in Sweden are only mandatory at the upper secondary school level in the most advanced level and final courses. The others are given only if the school decides to use them. Between 2017 and 2019 for example, there was actually a 22% decrease in the number of tests in English 5 ordered from the National Agency of Education (Skolverket 2021a; Tenfeldt 2021).

Furthermore, even though the national test results are to be given special consideration when the teachers set their pupil's final grades, the test results should not, according to the Swedish National Agency for Education, necessarily be the same grade as that given on the test nor should it be the only basis for the final grade (Skolverket 2021b, 2021c).

In Inbar-Lourie's (2008) terms, the educational cultures in Sweden and Finland today stand out as "learning cultures" rather than "testing cultures". Therefore, we find it paramount to map the state-of-the art of the somewhat neglected summative type of assessment in the teacher education territory.

### 3. Methods

The design of the study involved mixed methods, and both interviews and surveys were used to collect data from student teachers at the three universities in Sweden and Finland.

#### 3.1. Participants

The universities in the study house large language TE programs and are representative as far as they, because of their number of students, have a major impact on TE in each country. The language TE cohorts were small in Sweden during 2019–2021, and involving two Swedish universities helped the researchers reach more student teachers.

A total of 131 student teachers responded to the survey (Table 2). Close to half (47%) studied in Sweden (23% at University A and 24% at University B), and a little more than half (53%) studied in Finland (University C). The majority of the participants were female and the mean age was 35 with a median of 30. Around a third of the participants (34%) reported Swedish as their first language, and close to half (47%) Finnish. Two-fifths (60%) were studying to teach English, and a great majority (89%) assessed their own knowledge of their teaching language as being at the C1–C2 level according to the CEFR scale (Council of Europe 2001). In addition, more than two-thirds (70%) had teaching experience apart from their teacher education ranging from a couple of weeks to 15 years.

Twenty-six survey respondents also agreed to an in-depth interview. Twenty were from Sweden, and 6 were from Finland. Only three of the interviewed students were male.



**Table 2.** Distribution of participants by background variables (N = 131).

Country	Sweden 61 (47%)	Finland 70 (53%)	
University	A (Sweden) 30 (23%)	B (Sweden) 31 (24%)	C (Finland) 70 (53%)
Gender	Male 24 (18%)	Female 104 (79%)	Other 3 (3%)
Language 1	Swedish 45 (34%)	Finnish 61 (47%)	Other 25 (19%)
Self-assessed CEFR level of teaching language	C2–C1 117 (89%)	B2–B1 11 (8%)	A2–A1 3 (3%)
First teaching Language	English 79 (60%)	Swedish as a second language 37 (28%)	Other 15 (12%)

### 3.2. Data Collection

The data for the present paper were collected during the pre-service teachers' last semester of teacher education through an online questionnaire and in-depth interviews.

#### Questionnaire

The questionnaire was developed using an analytical framework involving “knowledge and understanding” (e.g., concepts, purposes, approaches), “skills and practices” (e.g., item writing, grading, statistical analysis), and “dispositions” (e.g., beliefs, values, attitudes) (Yildirim et al. 2021). The selection of different items was based on the theoretical literature and the language curricula in both countries and included criterion and norm referenced testing, test construction, grading, validity, reliability, alignment, transparency, and cooperative assessment. It contained (apart from background information) 14 5-point Likert scale and 4 open-ended questions. The questionnaire was first constructed in English and then translated into Finnish and Swedish and piloted with 10 students from each university. The students could thus choose to answer in the language of their choice. The completed questionnaires were later merged for comparative analysis.

The questions focused on in this paper answer how competent the pre-service teachers perceived themselves to be able to perform certain assessment tasks and what importance they gave to these skills (See Appendix B). They specifically address various summative assessment practices, such as “constructing different types of tests” and “writing different forms of test items,” and values attached to different assessment skills, such as “communicating summative assessment to students and parents.” On both questions, the Likert scale ranged from 1 = “not at all” to 5 = “fully.” An option of “no answer” was also included.

#### Interviews

A semi-structured interview guide consisting of 8 questions was constructed using the same conceptual framework to be consistent with the questionnaire. The researchers conducted the interviews themselves and the respondents could thus choose the language which they felt most comfortable to be interviewed in. The quotes presented in the paper were thus either originally in English or have been translated by the authors who are also teachers of English. The interviews lasted between 30 and 60 min each.

In the interview, the order of the questions was not always strictly followed to provide flexibility in the flow of conversation in line with the study purpose. Two questions specifically covered crucial aspects of summative assessment and their importance as well as how comfortable student teachers perceived themselves to be able to use different methods to assess students' performance in their subject area (including item and test construction as well as alternative assessment practices). The responses to these questions

provided a deeper understanding of the student teachers' questionnaire responses. Their voices are illustrative of what the statistical data represent.

### 3.3. Data Analysis

Descriptive and inferential statistics were employed to analyze the data from the questionnaire to identify both general trends in student teachers' responses and significant group differences based on university and country. The focus was mainly on university programs, and the three universities were recoded into Sweden and Finland to analyze the differences countrywise. For comparisons between groups, ANOVAs were used as well as the Eta-squared test (to measure the proportion of the total variance in each dependent variable, thus measuring the effect size of each). The Eta-squared test is commonly used in ANOVA and t-test designs as an index of the proportion of variance attributed to one or more effects. The statistic is useful in describing how variables are behaving within the researcher's sample (Salkind 2010).

Interview responses were coded for analysis using the common set of categories set out in our conceptual framework and also used in the questionnaire. Statements were coded under the following headings: SA conception and understanding; SA competence; Significance attached to SA; Weight of different types of input for SA; Summative assessment in TE; External Testing; and SA understanding and skills developed in school. Subheadings could be, for example, purposes of SA, aspects of validity, test construction, assessment culture in school, and further learning needs. This enabled us to identify similar and different aspects that pre-service teachers across the three universities brought into focus. To ensure reliability, an internal coding reliability was not calculated, but all codings were cross-checked by the partners as well as an outside researcher for consistency. By attaining consensus through mutual discussions and an approach to highlight emergent concepts and themes recurring in the responses and to refine the coding accordingly, greater trustworthiness can be achieved (McDonald et al. 2019, pp. 13–14).

## 4. Results

The results are presented according to the research questions and cover both questionnaire and interview responses.

### 4.1. Pre-Service Teachers' Perceptions of the Importance of SAL Practices

The first research question addressed the importance that student teachers assign to various summative assessment skills and practices. The students were asked to express their agreement on a statement referring to the importance of different practices on a Likert scale ranging from 1 (not at all) to 5 (fully). It is worth stating that the questionnaire mapped intended practices, not the real experience or frequency of implementation.

#### Questionnaire results

Table 3 depicts the practices perceived as most important among the student teachers when they were envisaging SA in their future work. They found it most essential on average to inform their pupils about learning targets and grading criteria ( $M = 4.58$ ). The second place was awarded to the use of national curriculum guidelines in designing SA tasks ( $M = 4.27$ ). Furthermore, the respondents found it paramount to communicate summative assessment to pupils and parents ( $M = 4.15$ ). Moreover, cooperation with other teachers in planning/conducting assessment ( $M = 4.14$ ), considering individual differences (e.g., special needs, interests) ( $M = 4.14$ ), and participating in summative assessment in-service training ( $M = 4.07$ ) were assigned a "largely" or "fully" agreement regarding their importance.

**Table 3.** Pre-service teachers' perceived importance of summative assessment practices.

	N	Mean *	SD
informing students about targets and grading criteria	129	4.58	0.693
using national curriculum/guidelines for summative assessment tasks	127	4.27	0.821
communicating summative assessment to students and parents	127	4.15	0.892
cooperating with other teachers in planning/conducting assessment	129	4.14	0.808
considering individual differences (e.g., special needs, interests)	129	4.14	0.817
participating in summative assessment in-service training	126	4.07	0.869
adjusting classroom teaching based on results of summative assessments	127	3.99	0.812
profiling language proficiency	126	3.69	0.784
using national test results for grading	117	3.37	0.988
using online testing	127	3.2	1.202
using ready-made/published tests when assessing students' performance	119	3.1	1.069
using self/peer-assessment when grading	126	3.1	1.101

\* In this table and the following ones, the mean is based on a scale ranging from 5 = "fully" to 1 = "not at all".

The methods regarded as most important share certain resemblances to formative assessment practices. This is true for the use of summative assessments to adjust teaching, inspire collegial collaboration, and respect individual differences. In fact, the formative use of summative assessments has been considered in the literature (Broadbent et al. 2018; Taras 2009) and can be seen as a reasonable way to bridge unnecessary gaps between the two purposes of assessment. Yet, in the questionnaire all items referred to the implementation of SA in forms such as course assessments or interim checkpoints.

On the other hand, some practices were judged as only moderately important, such as the use of ready-made/published tests when assessing students' performance ( $M = 3.10$ ), and peer and self-assessment ( $M = 3.10$ ). These may not be conceived to be as reliable as teacher assessment, and of course, in most educational systems the teacher is the only person that can be held legally responsible for making summative decisions regarding student achievement.

When it comes to differences between the three universities, altogether six variables displayed significant differences (Table 4). The practical significance was ensured by the effect size estimate eta squared values, which indicate a small effect at a range 0.01–0.06, a medium effect at a range 0.07–0.14, and a large effect at a range 0.15 or above (Salkind 2010). The importance of communicating summative assessment results to pupils and parents was judged as largely more important by students at the Finnish university C than the Swedish university A. Modest differences were revealed with regard to informing pupils about targets and grading criteria and the use of peer- and self-assessments when grading between universities A and C in favor of C. Small differences appeared in profiling language skills and in reporting assessment outcomes to pupils and parents between the Swedish universities A and B in favor of B, and lastly in considering individual differences between universities A and B in favor of A.

As well as comparing the three universities as single entities, we considered it interesting to look into the differences between Swedish and Finnish universities (Table 5), even if they cannot be said to be wholly representative for either country. Both of the Swedish universities (A and B) believed the practice of communicating SA to students to be of moderate importance and gave even more modest support to the use of online testing. The Finnish university's (C) students gave significantly more value to both practices. Significant differences also appeared between the Finnish university C and the Swedish university A in relation to the use of peer- and self-assessment in summative assessment.

**Table 4.** Pre-service teachers’ perceived importance of summative assessment practices by university.

	Univ A (Swe)			Univ B (Swe)			Univ C (Fin)			Significance Test
	N	Mean	SD	N	Mean	SD	N	Mean	SD	
Informing (pupils) about targets and grading criteria	30	4.13	0.9	31	4.61	0.667	68	4.76	0.492	F = 9.9 df = 2 p = 0.0001 etasq = 0.12
Communicating SA to (pupils) and parents	28	3.61	0.916	31	3.77	1.023	68	4.54	0.584	F = 11.6 df = 2 p = 0.0001 etasq = 0.23
Using self/peer assessment when grading	29	2.55	1.152	29	3.24	1.123	68	3.28	1.005	F = 5.0 df = 2 p = 0.008 etasq = 0.08
Profiling language proficiency	28	3.43	0.92	29	3.97	0.626	69	3.68	0.757	F = 3.5 df = 2 p = 0.034 etasq = 0.05
Considering individual differences	30	4.27	0.691	31	3.77	0.956	68	4.25	0.760	F = 4,3 df = 2 p = 0.016 etasq = 0.054
Using online testing	29	3.00	1.254	30	2.73	1.172	68	3.49	1.126	F = 4.6 df = 2 p = 0.009 etasq = 0.028

**Table 5.** Differences between the Swedish and the Finnish universities in how important the pre-service teachers considered summative assessment practices.

	Inform Students about Targets and Grading Criteria		Communicate Summative Assessment to Students and Parents		Use Online Testing	
	Swedish Unis	Finnish Uni	Swedish Unis	Finnish Uni	Swedish Unis	Finnish Uni
N	61	68	59	68	59	68
Mean	4.38	4.76	3.69	4.54	2.86	3.49
Std. Deviation	0.820	0.492	0.969	0.584	1.210	1.126
F	19.028		18.914		0.075	
t	−3.294		−6.067		−2.978	
df	127		125		119.512	
Sig. (2-tailed)	0.00128		0.00000001		0.004	
Cohen’s d	0.667		0.787		1.166	
effect size	medium		medium		large	

Three noticeable differences were detected in the student evaluations of importance between the Swedish universities and the Finnish one. The largest difference was found for the perceived importance of using online testing. Furthermore, the Finnish students acknowledged the importance of informing pupils about course targets and grading more

strongly than their Swedish counterparts. The result for communicating SA results to pupils and their parents was of the same size.

#### Interview results

There were few differences that could be discerned between the interviewed students with regard to universities, and in any case the sample of students interviewed was relatively small in relation to the larger cohort. Still, regarding the importance that student teachers assign to different summative skills and practices, it is apparent that teacher education has had an impact with regard to summative assessment knowledge and intent. This was especially noteworthy when it came to communicating and explaining summative assessment decisions to the pupils, cooperative assessment, and individual differences.

With regard to communication, students often referred to their experience of practice teaching at schools, which had changed their understanding of summative assessment practices or made theory clearer. This communication was often linked to the curricular goals and or grading criteria. One student expressed it in this way:

During my first teaching practice [VFU], where there was some kind of a problem, where pupils misunderstood, where I think that they did misunderstand the summative grade they got, the grade, at the end of the term, that's where I even talked about this in class, about my dilemma, you know, how can the grade be communicated to the children, and what is lost in translation . . . and through summative, formative process which lead to the summative. So I think, a way I thought to deal with that was to really make the learning intentions and the goal of lessons and even the term, as a whole, make it clear enough to the students, which would in turn lead to them understanding what it ended up in, what that summative thing. (B1:1)

It was also viewed as important by the student teachers to be transparent in their communication, that is "that you would inform them beforehand about what you are going to assess and that they would know what kind of things they need to learn before and that you inform them about all these things". (B3:2) Here we see how student teachers expressed their understanding of the importance of alignment and being able to communicate their summative assessments to their pupils in a clear and distinct manner.

Cooperative assessment, that is, working together with a colleague was viewed as important, but also as a great help, especially as a soon to be novice teacher. "Yes, I think it's good, especially for us now, new, when we are new teachers. So, we'll need that quite more than it's done now in school". (B3:1). It was also seen as a guarantee for a fair assessment:

[cooperative assessment] is an important part because you can easily be blinded by what you feel towards a pupil [...] you need to help each other in difficult cases". [...] I find it absolutely imperative when you are new and inexperienced because . . . it takes real life experience and you need to practice, practice, practice all the time to become good at assessment. That's why I need help . . . (B4: 3–4)

This and other similar comments on cooperative assessment emphasized the student teachers' awareness of being new in their role and their trust in forthcoming help from future colleagues.

Individual differences were also something that many students had reflected on when it came to making assessments and was something which teacher education had brought to light. "In the psychology of learning class, much was done about the learner as an individual, individual needs and their considerations" (F3:1), one of them said while two other student teachers expressed:

Previously, I didn't really think about them [pupils] as weak . . . but now I am more, what should I say . . . I am more . . . Now I understand that they also need attention, more than the others. Adaptations . . . (A3:1)

Well, of course, you should also give them a grade because if you think they have ADHD, then it's a factor that affects your grading in such a way that they



may have difficulty concentrating in the test. It won't tell you anything about intelligence, but it will tell you [...] that it is difficult to focus. (F4:1)

On the other hand, several felt that individual differences and their role in making summative assessments was not sufficiently emphasized in TE courses. For example, one said, "When we had the special education course we brought up quite a lot concerning different handicaps but how it then related to assessment, that link wasn't very clear in my opinion." (A7:3).

In summary, the questionnaire and interview data aligned as the latter provided depth and detail to the perceived importance student teachers attached to summative assessment practices. The interviewed student teachers had a somewhat more complex and nuanced understanding than the perceptions explored in the questionnaire data. It was only when it came to in-service-training that there seemed to be no greater need to comment. As the student teachers are still within the educational context it is quite understandable that further theoretical education is not what they may find most important at this point in time. The focus on praxeological skills regarding summative assessment on the other hand is apparent.

#### 4.2. Pre-Service Teachers' Perceptions of Their Competence in Applying SA Practices

With our second research question we sought to gain knowledge about the student teachers' experienced competence in exploiting a variety of SA-related practices. They may have had an opportunity to try out some of these themselves, while others they may only have heard about at lectures, observed during their teaching practice, or become familiar with through the course literature. The question aimed to capture their perceived competence as a result of their experiences as well as their confidence to put these techniques into practice.

##### Questionnaire results

At a glance (Table 6), it is notable that the total averages of all items in the competence section remained below the "largely" level (point 4 out of 5 Likert steps).

**Table 6.** Pre-service teachers' perceived competence in summative assessment practices.

	N	Mean	SD
Writing true–false items	131	3.92	0.865
Writing short-answer questions	130	3.85	0.836
Writing multiple-choice items	130	3.81	0.916
Constructing written tests	131	3.56	0.860
Writing open-ended/essay questions	130	3.55	0.881
Grading standardized items	128	3.52	0.972
Assessing students' written performance	131	3.52	0.844
Assessing students' reading skills	130	3.51	0.865
Grading written tests	131	3.46	0.806
Assessing students' listening skills -	130	3.38	0.893
Constructing oral tests	131	3.20	0.940
Grading homework/project assignments	127	3.19	0.932
Assessing students' oral performance	130	3.13	0.848
Assessing students' interaction	130	3.11	0.856
Assigning reliable grades	131	3.11	0.834
Grading open-ended/essay questions	131	3.08	0.771
Grading oral tests	131	2.96	0.845
Weighting different language skills statistically when grading	124	2.77	1.052
Using basic statistics when analyzing test results	129	2.62	1.126
Grading portfolios	127	2.57	1.005

The respondents felt they were most competent in designing true/false items (M = 3.92), followed by the perceived capability of writing short answer questions (M = 3.85). The

third place was taken by perceived skill in writing multiple choice items ( $M = 3.81$ ), which is far from a simple task (Jones 2020), although laymen and beginners tend to assume so.

Student teachers' confidence on average was at its lowest with regard to grading portfolios ( $M = 2.57$ ) and using basic statistics ( $M = 2.62$ ). They found it most difficult to give statistical weights to the different language skills when grading.

Table 7 depicts the significant differences between the three universities regarding perceived competence in employing SA practices. The perceived competence in writing multiple choice items and true–false items was judged higher by student teachers at the Finnish university C than in the Swedish university A. For the multiple choice items the effect size was small and for the true–false items medium. A small size effect was also detected for the competence in assessing reading skills, reported again as higher by the Finnish students than by students at the Swedish university B. For the perceived competence in using basic statistics, a medium size effect was found between the Finnish university C and both Swedish universities. The Finnish students expressed higher confidence in their competence than the Swedish students, but the overall average of this item remained fairly low in all universities.

**Table 7.** Pre-service teachers' perceived competence in summative assessment practices by university.

	Univ A (Swe)			Univ B (Swe)			Univ C (Fin)			Significance Test
	N	Mean *	SD	N	Mean *	SD	N	Mean *	SD	
Writing multiple choice items	30	3.5	1.0	30	3.73	0.9	70	3.97	0.8	F = 3.0 df = 2 p = 0.053 etasq = 0.045
Writing true-false -items	30	3.53	0.9	31	3.90	0.83	70	4.10	0.78	F = 4.8 df = 2 p = 0.010 etasq = 0.070
Assessing reading skills	29	3.34	0.72	31	3.26	0.99	70	3.69	0.83	F = 3.4 df = 2 p = 0.036 etasq = 0.051
Using basic statistics	30	2.30	1.21	30	2.33	0.88	69	2.88	1.13	F = 4.3 df = 2 p = 0.016 etasq = 0.064

\* Based on a scale ranging from 5 = fully to 1 = not at all.

The differences between the two countries were also analyzed by merging the two Swedish universities (A and B) to represent "Sweden", while the Finnish university C exemplified "Finland". In total, significant differences were found in eight competencies and all of them displayed large effect sizes as indicated by a Cohen's d coefficient higher than 0.8 (Table 8).

**Table 8.** Comparison of pre-service teachers’ perceived competence in summative assessment practices in Sweden and Finland.

		N	Mean	Std. De- viation	Std. Error Mean	F	df	Sig. (2-Tailed)	Cohen’s d	Effect Size
Write multiple-choice items	Swedish unis	60	3.62	0.993	0.128	5.126	128	0.027	0.902	large
	Finnish uni	70	3.97	0.816	0.098					
Write true-false items	Swedish unis	61	3.72	0.915	0.117	2.481	129	0.012	0.847	large
	Finnish uni	70	4.10	0.783	0.094					
Write short-answer questions	Swedish unis	60	3.67	0.914	0.118	7.051	128	0.018	0.821	large
	Finnish uni	70	4.01	0.732	0.088					
Grade portfolios	Swedish unis	59	2.37	1.032	0.134	2.699	119.27	0.043	0.992	large
	Finnish uni	68	2.74	0.956	0.116					
Weight different language skills statistically when grading	Swedish unis	56	2.55	1.159	0.155	8.865	122	0.041	1.039	large
	Finnish uni	68	2.94	0.929	0.113					
Assess students’ listening skills	Swedish unis	60	3.22	0.865	0.112	0.519	128	0.047	0.882	large
	Finnish uni	70	3.53	0.896	0.107					
Assess students’ reading skills	Swedish unis	60	3.30	0.869	0.112	0.083	126.46	0.004	1.094	Large
	Finnish uni	70	3.69	0.826	0.099					
Use basic statistics (e.g., means, correlations) when analyzing test results	Swedish unis	60	2.32	1.049	0.135	0.030	127	0.004	1.094	Large
	Finnish uni	69	2.88	1.132	0.136					

The findings showed a higher confidence for the Finnish respondents when it came to writing multiple choice questions, true–false statements, and designing questions to open-ended answers, as well as grading portfolios. With regard to traditional linguistic skills, Finnish students trusted their competence in grading receptive skills, listening and reading, and also in weighting the multiple skill components in a summative grade more than the Swedish respondents. In Finland, confidence in using basic statistics in SA also proved to be less vague than in Sweden, although the general level of perceived competence in this strand was rather modest in both countries.

Most often the significant differences relating to university programs were detected between Swedish university A and Finnish university C, the latter ranking higher on the likert scale. These comprised the following items:

- competence in writing multiple choice items: M(A) = 3.50; M(C) = 3.97; F = 3.001; p = 0.053; eta squared = 0.045, small effect size;
- competence in writing true –false items: M(A) = 3.53; M(C) = 4.10; F = 4.782; p = 0.010; eta squared = 0.07, modest effect size;
- competence in using basic statistics: M(A) = 2.30; M(C) = 2.88; F = 4.288; p = 0.016; eta squared = 0.064, modest effect size.

Between the two Swedish universities, one significant difference was revealed in the perceived competence in assessing reading skills: M(B) = 3.26; M(C) = 3.69; F = 3.411; p = 0.036; eta squared = 0.051, small effect size.

### Interview results

The feelings of competence that the student teachers experienced in using a variety of SA practices were dependent on whether they had had an opportunity to try these out themselves, or whether they had only heard about these in class. The interview questions targeted similar thematic areas as in the questionnaire to capture both real experience as well as their imagined skills of putting these techniques into practice.

In the interviews we could see that one of the most important things that the students felt to be lacking in their education was first of all the ability to construct tests themselves. This seems to be partly due to the fact that there are few, if any, courses directed specifically towards summative assessment and testing at any of the universities. It may also be due to the fact many teachers in our school systems, that is the student teacher mentors, do not seem to use their own tests to any large degree or use them for summative purposes. When this is the case it means that the students do not get a chance to practice test construction when doing their school practice sessions. This concern is evident in the following quotations:

Generally, not competent at all. Or, well I feel that it's an area where I have limited knowledge and especially with regard to my own ability to make test. I am sure we've had something about it in our education but I have difficulties in remembering clearly. I am sure we've gone through what you should think about and so on, something I remember is maybe that, that when you make your own tests you should start with easy items and things like that . . . basic stuff. but generally I feel that I don't have enough with me there and whatever we had during that course, it is, for my part anyway . . . it went too fast, passed by quickly and then we didn't return to it. (A4:4)

And then, the times when I have tried to construct tests and homework quizzes I realize how, how easily it goes wrong. So I think, I mean I . . . it's not that I feel, that I feel that it is impossible to construct different types of tests but, I haven't practiced. Neither during school practice nor during our education here at the university, so in this respect I do not feel competent, not really. (A4:3)

Interestingly, this gives a different picture from previous research ([Oscarson and Apelgren 2011](#)) in Sweden where teachers stated that using their own tests was a regular basis for grading even if other forms such as the national tests were considered more commonplace. The following quote mirrors an everyday experience that shows how more formative practices have taken precedence over testing:

No, not really. I did some kind of . . . if it's like, this online quiz that they can do. I did some kind of test construction to make them test their own . . . how much they know about things. It's like—because at the school I was at they didn't really work with tests like for an assessment-assessment. It's more for a self-assessment that you know how much you have learnt about something. (B3:1)

On the other hand, when it came to item construction, the interviews also gave the general picture that the student teachers' feelings of competence in designing true/false items, short answers, and multiple-choice questions were not as open-and-shut as the questionnaire answers may have led us to believe. The answers reflected that they instead found them challenging due to their complex nature. "Yeah, but for example in a multiple choice question—how do I find three different distractors where it is not super evident which is correct?" (A2:5); "[...]; "especially things like multiple choice questions, that's why it is so important—the choices are so important, because if you write an inappropriate distractor there, you can make the answer really super evident and then, it isn't reliable" (B2:1) and "It was surprisingly complicated to think about how to get one [distractor] that does not somehow lead in any direction, or that all alternatives are, in principle, equally as good looking". (C1:1) The responses illustrate that they have reflected on the apparent simplicity of these types of items. It was obvious from both the questionnaire and interview responses that pre-service teachers' confidence was at its lowest with regard to grading

portfolios and using basic statistics. In general they felt that they had not had enough instruction and training which touched upon these forms of assessment. This was, for example, explicitly expressed in the following manner: “Show and process portfolios—it, I am very interested in portfolios, I want to learn more about them. I don’t feel that I know enough yet.” (A1:3) and when it came to statistics “No, we haven’t covered this at all” (A8:1); “No, really, no clue. [ . . . ] never in relation to assessment and to summative assessments. No”.(A4:6) and “Interpretation of statistical data? No, I don’t think in Sweden we do that, all that much” (B1:2). Even if portfolio work may not be that common in schools, a basic knowledge of statistics with regard to understanding test results seems a serious failing.

Regarding the grading of reading skills brought up in the survey responses, the interviewed student teachers did not delve into the specific problems within this area, merely stating that it was difficult or that they had not received enough practice to feel competent in this area. These praxeological skills seem to need a lot more attention for student teachers to feel fully competent when entering their teaching profession.

## 5. Discussion

In this section we first reflect on the findings in the light of the [Pastore and Andrade \(2019\)](#) triadic model as a prerequisite for our first goal to map the strengths and challenges of the student teachers’ assessment literacy. The results in the present study mapped intended practices, rather than the real experience or frequency of implementation, and it is the student teachers’ perceived knowledge and approaches to issues about SA and to how they evaluate their own knowledge that is in focus.

### 5.1. Perceptions on Conceptual Knowledge, Praxeological Dimension, and Socio-Emotional Issues

In general, the responses reveal that informants in this study believed in the vital role of SA and of being competent in the field. Competence by definition comprises knowledge and skills, i.e., the praxeological know-how ([Ufer and Neumann 2018](#), pp. 433–35). The overall competence perceived by the students tended to be relatively evenly spread across the inquired practices to design tests and grade the different domains and skills. The students seemed to be more self-confident in designing tasks than in grading, but the difference is negligible. However, it is a matter of concern that the pre-service teachers articulated a lack of ability with regard to test construction. Another concern is that the overall findings pertain to the low confidence in grading oral skills and portfolios, as well as the conceived incompetence in using basic statistics. Similar issues have surfaced in earlier studies by [Vogt and Tsagari \(2014\)](#) and [DeLuca and Klinger \(2010\)](#) among others.

Oral skills are crucial in communicative language teaching, a longstanding essence in the national curricula of both countries. All student teachers in all universities should be equipped with the adequate knowledge and skills to include speaking proficiency in summative assessment. Portfolio assessment is a manifestation of a diverse and deliberative stance on assessment, which deserves a more impactful status also as a component in summative assessment ([Abrar-ul-Hassan et al. 2021](#)). None of the three universities seems to have succeeded in embodying this skill in the repertoire of prospective language teachers. The most obvious reason is the time constraints of pedagogical studies that seldom allow teacher students to orchestrate longer sequences than a few teaching hours per group, while capturing the idea portfolio assessment takes more time to get to know the pupils and preferably even working with them themselves. In the same vein, [DeLuca and Lam \(2014\)](#) maintain the primacy of inclusiveness of formative and summative assessment education to enable the students to acquire both analytic and interpretive aspects of assessment in theory and practice.

Despite the informative and interactional ethos suggested by the curricula and the pedagogical studies, the perceived competence in deploying the various practices is heavily biased towards very traditional item types (true/false, short answers, and multiple-choice items) derived from the early psychometric tradition of assessment. This finding is in accordance with studies by Bailey and Brown as early as in the 1990s. A simple reason



for their conceived primacy may be found in the course programs including practical training in writing short items, and secondly, in the lack of the knowledge base assuming “simple items” to be simple to design as well. On the other hand, students in both countries expressed their awareness of the complexity and challenge of designing multiple-choice items in the interviews. LAL is a complex entity that requires multiple competencies including subject-specific language ability alongside general AL, as previously stated by [Levi and Inbar-Lourie \(2020\)](#). Item construction according to theory-driven perspectives on reliability and validity was identified as a matter of specific attention also by [DeLuca and Klinger \(2010\)](#).

Statistics tend to fall short in linguistic disciplines and in the AL of their future teachers. Yet in the modern world, a basic understanding of inferential principles is important when trying to cope with an overwhelming supply of data and information, let alone the obligation of a teacher to guide pupils in this demanding task. Basic statistics are also important for teachers to not only use in their own practice but also to understand the basis for how national and global tests are produced and be understood. Nonetheless, the amount and quality of teaching statistical principles to prospective language teachers deserves more attention at all the investigated universities.

The socio-emotional dimension in the importance section addresses students’ values and preferences regarding curriculum alignment, communication, collaboration, and appreciation of individual differences. The results of our study suggest that student teachers considered it important that summative assessment was aligned with national curriculum guidelines. It is further paramount to cooperate with colleagues, to inform and communicate assessments and to consider individual differences. They also appreciated in-service training in their future work. These findings open up encouraging prospects on a future with responsible teaching professionals who appreciate transparency and collaborative dialogue in their assessment related activity.

### *5.2. Differences among the Three Teacher Education Programs*

The second goal of this paper was to explore comparatively how well the three teacher education programs equip the student teachers with adequate and timely knowledge, skills and beneficial disposition regarding summative assessment literacy.

Differences between universities were mainly detected between the Finnish university C and one of the Swedish universities, that is, either A or B. A rather straight-forward explanation to the higher perceived values of importance and the assumed competence to deploy them, might be found in the temporal structure of pedagogical studies. In Finland, the entire set of 60 credits is taken during a single study year, as an ultimately focused and laborious endeavor without many other studies or activities in the students’ schedule. In the second and last semester of the pedagogical studies, assessment-related themes are among the focal points, and as the questionnaire was administered in spring, assessment issues must have been more on the surface of Finnish students’ memories. At the Swedish universities, the corresponding courses may have been taken longer ago and thus knowledge may have faded somewhat. On the other hand, Swedish students may have several courses, both embedded and specifically dedicated to assessment. For many, these course are spread over 5 years and thus they have a longer time to actually grasp and digest the knowledge.

On the one hand, a contributing factor is that a new curriculum for upper secondary education was issued in Finland in 2019, and at the same time, the final grading criteria for basic education were updated, and both topics were strongly presented in media and in teacher education. The language curricula in both countries underscore interaction, communication and collaboration, primarily in the formative assessment mode, but the overarching ethos of the operative curricula may reflect itself in the responses by the Finnish student teachers. On the other hand, the deliberative and interactive approaches are also emphasized in the Swedish national curricula, and should have an equal impact on the Swedish students’ responses. Still, during teaching practice, they are more likely to learn

about how the national test is conducted and graded rather than being involved in the final grading of pupils' language achievement. Our findings accord with the claim by [Tavassoli and Farhadi \(2018\)](#) that assessment literacy deserves serious reconsideration and support in teacher education programs, also in countries, in which high-quality education is a commonplace.

The few differences that appeared between the two Swedish universities are most probably not due to national factors as there is not one form of TE in Sweden. The explanation is likely to be found in local traditions or the various teaching cultures. Perhaps the problematic differences found in recent reports and research concerning the validity and reliability of teacher assessments are a product of the different forms teacher education takes within different programs and are due to the fact that every university develops their own particular profile.

It is obvious that all three teacher programs give the pre-service teachers a good sense of competence and of what assessment skills are important for their future careers. In the interviews, we can also see that many of them realized that they lack knowledge and skills with regard to in-depth reflections on various problems that teachers may face when making assessments, especially summative ones. There is also the dilemma of misalignment of theory and practice (see also [Giraldo and Murcia 2018](#)) that cannot easily be resolved during a fairly short-term intervention.

Since student teachers find summative assessment important but formative practices are emphasized in TE, more attention should be paid to in-depth knowledge building and training to employ diverse practices of timely summative assessment.

Teacher education should make an effort to bridge the gap between summative and formative purposes and uses of assessment to incorporate techniques that capture the learning process in summative assessment (e.g., portfolios).

The sense of self-confidence acquired during teacher education in assessment aspects, such as curricular alignment and appreciation of communication and diversity, are to be maintained, while certain areas (test construction, grading e.g., oral skills, statistics) are in dire need of increased attention.

### 5.3. Limitations of the Study

Despite the beneficial contribution that the students' questionnaire and interview responses make to chart the venue of pre-service AL in Sweden and Finland, we are aware of certain limitations that one will need to take into account in future studies. The issue of social desirability is always present in survey studies based on self-reported statements. In this case, the students responded anonymously and after a completed course, which added to the trustworthiness of the entries. Another limitation to the study is that there were fewer interview responses from Finland than from Sweden which may not do Finnish students teachers justice when it comes to expressing their perceptions of importance and competence. On the other hand, due to the set up of the Finnish program, there were also difficulties in reaching these students after the completion of their pedagogical studies as they move on to another faculty. It is also vital to consider that the responses merely represent the perceptions and intentions of the students, rather than their real activity. Therefore, the authors of this paper have incorporated a follow-up phase into the project span, in which the same students will be surveyed and interviewed as novice teachers after their first summative grading experience.

**Author Contributions:** Conceptualization, R.H., A.D.O., A.Y. and B.F.; methodology, R.H., A.Y. and A.D.O.; software, A.Y. and A.D.O.; validation, R.H., A.Y. and A.D.O.; formal analysis, R.H., A.Y. and A.D.O.; investigation, R.H., A.Y. and A.D.O.; resources, University of Gothenburg, University of Helsinki; data curation, University of Gothenburg, University of Helsinki; writing—original draft preparation, R.H. and A.D.O.; writing—review and editing, R.H., A.D.O., A.Y. and B.F.; visualization, R.H. and A.Y.; supervision, A.Y.; project administration, A.D.O. and A.Y.; funding acquisition, A.D.O., A.Y., R.H. and B.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Swedish Research, grant number (2019–2021, 2018-04008).

**Institutional Review Board Statement:** Approval Waived by University of Helsinki Ethical Review Board in Humanities and Social and Behavioral Sciences.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The project is not reported yet. Results to be reported in March 2023.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Appendix A**

Courses	University A Sweden	Courses	University B Sweden	Courses	University C Finland
Common for all	Assessment, analysis and evaluation of learning and development	Common for all	Language education, curriculum theory, grading and assessment (General Level)	1 year program Common	Planning, implementation and assessment of teaching
Long program	English learning, teaching and assessment I–III	Long program	Language education, curriculum theory, grading and assessment (Advanced Level)		Curriculum and development of educational institution
	Assessment and grading for teachers in secondary and upper secondary school		Language assessment (focus on ie. Swedish as a Second Language)		Teacher as a researcher/ Teacher as a researcher Didactics
					Social, cultural and philosophical foundations of education
School practice	School practice 1–4	School practice	School practice 1–3	School practice	School practice basic and advanced
Short program	Steering, organization and assessment for teachers in secondary school and upper secondary school	Short program	Summative assessment module		
School practice	School practice 1–3				

**Appendix B**

How Important Do You Think It Is to . . .	Conceptual Knowledge	Praxeological Dimension	Socio-Emotional Dimension
inform students about targets and grading criteria			x
use national curriculum/ guidelines for summative assessment tasks		x	
communicate summative assessment to students and parents			x
cooperate with other teachers in planning/ conducting assessment			x
consider individual differences (e.g., special needs, interests)			x
participate in summative assessment in-service training		x	
adjust classroom teaching based on results of summative assessments		x	
profile language proficiency (e.g., report separately for the “four skills”)		x	
use national test results for grading		x	
use online testing		x	
use ready-made/ published tests when assessing students’ performance		x	
use self/ peer-assessment when grading		x	
write true–false items	x	x	
write short-answer questions	x	x	

How Important Do You Think It Is to ... How competent do you feel to ...	Conceptual Knowledge	Praxeological Dimension	Socio-Emotional Dimension
write multiple-choice items	x	x	
construct written tests	x	x	
write open-ended/essay questions	x	x	
grade standardized items	x	x	
assess students' written performance	x	x	
assess students' reading skills	x	x	
grade written tests	x	x	
assess students' listening skills	x	x	
construct oral tests	x	x	
grade homework/project assignments	x	x	
assess students' oral performance	x	x	
assess students' interaction	x	x	
assign reliable grades	x	x	x
grade open-ended/essay questions	x	x	
grade oral tests	x	x	
weight different language skills statistically when grading	x	x	
use basic statistics (e.g., means, correlations) when analyzing test results	x	x	
grade portfolios	x	x	

## References

- Abrar-ul-Hassan, Shahid, Dan Douglas, and Jean Turner. 2021. Revisiting Second Language Portfolio Assessment in a New Age. *System* 103: 102652. [\[CrossRef\]](#)
- Bachman, Lyle F., and Adrian Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.
- Bachman, Lyle F., and Adrian Palmer. 2010. *Language Assessment in Practice. Developing Language Assessments and Justifying their Use in the Real World*. Oxford: Oxford University Press.
- Bailey, Kathleen M., and James D. Brown. 1995. Language testing courses: What are they? In *Validation in Language Testing*. Edited by Alister H. Cumming and Richard Berwick. Clevedon: Multilingual Matters, pp. 236–256.
- Barcelos, Ana Maria Ferreira, and Paula Kalaja. 2011. Introduction to beliefs about SLA revisited. *System* 39: 281–89. [\[CrossRef\]](#)
- Borg, Simon. 2006. The distinctive characteristics of foreign language teachers. *Language Teaching Research* 10: 3–31. [\[CrossRef\]](#)
- Borg, Simon. 2011. The Impact of In-Service Teacher Education on Language Teachers' Beliefs. *System: An International Journal of Educational Technology and Applied Linguistics* 39: 370. [\[CrossRef\]](#)
- Borg, Simon. 2015. *Teacher Cognition and Language Education: Research and Practice*. Bloomsbury Classics in Linguistics Edition. London: Bloomsbury Academic.
- Brindley, Geoff. 2001. Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing* 18: 393–407. [\[CrossRef\]](#)
- Broadbent, Jaclyn, Ernesto Panadero, and David Boud. 2018. Implementing summative assessment with a formative flavour: A case study in a large class. *Assessment & Evaluation in Higher Education* 43: 307–22.
- Brookhart, Susan M. 2011. Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practice* 30: 3–12. [\[CrossRef\]](#)
- Caena, Francesca. 2011. *Literature Review. Teachers' Core competences: Requirements and Development*. European Commission Thematic Working Group Professional Development of Teachers. Strassbourg: European Commission.
- Chapelle, Carol A., Mary K. Enright, and Joan M. Jamieson. 2008. *Building a Validity Argument for the Test of English as a Foreign Language*. London: Routledge.
- Cizek, Gregory J., Heidi L. Andrade, and Randy E. Bennett. 2019. Formative Assessment, History, Definitions, and Progress. In *Handbook of Formative Assessment in the Disciplines*. Edited by Heidi L. Andrade, Randy E. Bennett and Gregory J. Cizek. New York and London: Routledge, pp. 3–19.
- Coombe, Christine, Hossein Vafadar, and Hassan Mohebbi. 2020. Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia* 10: 1–16. [\[CrossRef\]](#)
- Coombs, Andrew, Christopher DeLuca, and Stephen MacGregor. 2020. A person-centered analysis of teacher candidates' approaches to assessment. *Teaching and Teacher Education* 87: 102952. [\[CrossRef\]](#)

- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Davies, Alan. 2008. Textbook trends in teaching language testing. *Language Testing* 25: 327–47. [CrossRef]
- DeLuca, Christopher, and Chi Yan Lam. 2014. Preparing Teachers for Assessment within Diverse Classrooms: An Analysis of Teacher Candidates' Conceptualizations. *Teacher Education Quarterly* 41: 3–24.
- DeLuca, Christopher, and Don Klinger. 2010. Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education* 17: 419. [CrossRef]
- DeLuca, Christopher, Teresa Chavez, Aarti Bellara, and Cao Chunhua. 2013. Pedagogies for Preservice Assessment Education: Supporting Teacher Candidates' Assessment Literacy Development. *Teacher Educator* 48: 128. [CrossRef]
- DeLuca, Christopher, Danielle LaPointe-McEwan, and Ulemu Luhanga. 2016. Approaches to Classroom Assessment Inventory: A New Instrument to Support Teacher Assessment Literacy. *Educational Assessment* 21: 248–66. [CrossRef]
- DeLuca, Christopher, Christoph Schneider, Andrew Coombs, Marcela Pozas, and Amirhossein Rasooli. 2020. A cross-cultural comparison of German and Canadian student teachers' assessment competence. *Assessment in Education: Principles, Policy & Practice* 27: 26–45. [CrossRef]
- Fulcher, Glenn. 2012. Assessment Literacy for the Language Classroom. *Language Assessment Quarterly* 9: 113–32. [CrossRef]
- Giraldo, Frank, and Daniel Murcia. 2018. Language Assessment Literacy for Pre-service Teachers: Course Expectations from Different Stakeholders. *GIST Education and Learning Research Journal* 16: 56–77. [CrossRef]
- Hilden, Raili, and Birgitta Fröjndendahl. 2018. The dawn of assessment literacy—Exploring the conceptions of Finnish student teachers in foreign languages. *Apples: Journal of Applied Language Studies* 12: 1–24.
- Hildén, Raili, and Juhani Rautopuro. 2014. *Saksan Kielen A-ja B-Oppimäärän Oppimistulokset Perusopetuksen Päätöväiheessä 2013*. Helsinki: Finnish National Board of Education.
- Inbar-Lourie, Ofra. 2008. Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing* 25: 385–402. [CrossRef]
- Jones, Glyn. 2020. Designing Multiple-Choice Test Items. In *The Routledge Handbook of Second Language Acquisition and Language Testing*. Edited by Winke Paula Marie and Tineke Brunfaut. New York: Routledge, pp. 90–101.
- Kane, Michael T. 2013. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement* 50: 1–73. [CrossRef]
- Klapp Lekholm, Alli. 2011. Effects of School Characteristics on Grades in Compulsory School. *Scandinavian Journal of Educational Research* 55: 587–608. [CrossRef]
- Kremmel, Benjamin, and Luke Harding. 2020. Towards a Comprehensive, Empirical Model of Language Assessment Literacy across Stakeholder Groups: Developing the Language Assessment Literacy Survey. *Language Assessment Quarterly* 17: 100–20. [CrossRef]
- Kunnan, Antony John. 2018. *Evaluating Language Assessments*. New York: Routledge.
- Lam, Ricky. 2015. Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing* 32: 169–97. [CrossRef]
- Lau, Alice Man Sze. 2016. Formative good, summative bad?—A review of the dichotomy in assessment literature. *Journal of Further and Higher Education* 40: 509–25. [CrossRef]
- Levi, Tziona, and Ofra Inbar-Lourie. 2020. Assessment Literacy or Language Assessment Literacy: Learning from the Teachers. *Language Assessment Quarterly* 17: 168–82. [CrossRef]
- McDonald, Nora, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *ACM on Human-Computer Interaction* 3: 1–23.
- Messick, Samuel. 1989. Validity. In *Educational Measurement*, 3rd ed. Edited by Robert Linn. American Council on Education. Washington, DC: Macmillan.
- Oscarson, Mats, and Britt Marie Apelgren. 2011. Mapping language teachers' conceptions of student assessment procedures in relation to grading: A two-stage empirical inquiry. *System* 39: 2–16. [CrossRef]
- Oxford English Dictionary Online*. 2021. Oxford: Oxford University Press.
- Pastore, Serafina, and Heidi L. Andrade. 2019. Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education* 84: 128–38. [CrossRef]
- Pellegrino, James W. 2018. Assessment of and for learning. In *International Handbook of the Learning Sciences*. Edited by Frank Fischer, Susan M. Goldman, Cindy E. Hmelo-Silver and Peter Reiman. New York: Routledge, pp. 410–21.
- Pizorn, Karmen, and Ari Huhta. 2016. Assessment in educational settings. In *Handbook of Second Language Assessment*. Edited by Tsagari Dina and Jay Banerjee. New York: Mouton De Gruyter, pp. 239–254.
- Popham, James W. 2011. Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator* 46: 265–73. [CrossRef]
- Qiong, Ou. 2017. A brief introduction to perception. *Studies in Literature and Language* 15: 18–28.
- Riksrevisionen. 2011. *Lika Betyg, Lika Kunskap? En Uppföljning av Statens Styrning mot en Likvärdig Betygsättning i Grundskolan*. Stockholm: Riksrevisionen.
- Salkind, Neil J. 2010. *Encyclopedia of Research Design*. Thousand Oaks: Sage, vol. 1.
- Scriven, Michael. 1967. The methodology of evaluation. In *Perspectives of Curriculum Evaluation*. Edited by Robert Gagne, Ralph W. Tyler and Michael Scriven. Chicago: Rand McNally, p. 39.
- Skolverket. 2021a. Available online: <https://www.skolverket.se/undervisning/gymnasieskolan/nationella-prov-i-gymnasieskolan/provdatum-i-gymnasieskolan> (accessed on 12 December 2021).



- Skolverket. 2021b. Available online: <https://www.skolverket.se/undervisning/grundskolan/nationella-prov-i-grundskolan/genomfora-och-bedoma-prov-i-grundskolan#h-Provresultatetsbetydelseforbetyget> (accessed on 11 December 2021).
- Skolverket. 2021c. Available online: <https://www.skolverket.se/undervisning/gymnasieskolan/nationella-prov-i-gymnasieskolan/genomfora-och-bedoma-prov-i-gymnasieskolan#Provresultatetsbetydelseforbetyget> (accessed on 12 December 2021).
- Stiggins, Richard J. 1991. Relevant Classroom Assessment Training for Teachers. *Educational Measurement, Issues and Practice* 10: 7–12. [CrossRef]
- Taras, Maddalena. 2005. Assessment—Summative and Formative—Some Theoretical Reflections. *British Journal of Educational Studies* 53: 466–78. [CrossRef]
- Taras, Maddalena. 2009. Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education* 33: 57–69. [CrossRef]
- Tavassoli, Kobra, and Hossein Farhadi. 2018. Assessment Knowledge Needs of EFL Teachers. *Teaching English Language* 12: 45–65.
- Taylor, Lynda. 2013. Communicating the Theory, Practice and Principles of Language Testing to Test Stakeholders: Some Reflections. *Language Testing* 30: 403–12. [CrossRef]
- Tenfeldt, Torbjörn. 2021. Många Gymnasieskolor Ratar Nationella Proven. *Ämnesläraren*, 11 November. Available online: <https://www.lararen.se/amneslararen-matte-no/nationella-prov/manga-gymnasieskolor-ratar-nationella-proven> (accessed on 10 December 2021).
- Thorsen, Cecilia, and Christina Cliffordson. 2012. Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research & Evaluation* 18: 153–72. [CrossRef]
- Ufer, Stefan, and Knut Neumann. 2018. Measuring Competencies. In *International Handbook of the Learning Sciences*. Edited by Frank Fischer, Cindy E. Hmelo-Silver, Susan R. Goldman and Peter Reimann. Abingdon: Routledge, pp. 433–43.
- Vogt, Karin, and Dina Tsagari. 2014. Assessment Literacy of Foreign Language Teachers: Findings of a European Study. *Language Assessment Quarterly* 11: 374. [CrossRef]
- Volante, Louis, and Xavier Fazio. 2007. Exploring Teacher Candidates' Assessment Literacy: Implications for Teacher Education Reform and Professional Development. *Canadian Journal of Education* 30: 749–70. [CrossRef]
- Xu, Yueting, and Gavin L. T. Brown. 2016. Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education* 58: 149–62. [CrossRef]
- Yildirim, Ali, Anne Dragemark Oscarson, Raili Hilden, and Birgitta Fröjdendahl. 2021. Teaching summative assessment literacy: A comparative curriculum analysis of three teacher education programs in Sweden and Finland. Paper presented at American Educational Research Association Conference, Orlando, FL, USA, April 8–12.