Students' perceptions of teacher assessment practices in foreign language emergency remote teaching during Covid-19 in Finland

Toni Mäkipää
University of Helsinki

The purpose of this study was to explore the perceptions of Finnish general upper secondary students of teacher assessment practices in foreign language emergency remote teaching (ERT). A total of 251 students answered an online survey, and both quantitative and qualitative methods were used in the analysis of the data. The results show that students found essays, listening comprehension tests, and self-assessments to be the most suitable practices for ERT, while learning diaries, peer assessments, and portfolios were deemed the least suitable. Differences were detected in students' perceptions in terms of their previous course grades. Most students expressed positive views about their language teachers' assessment practices in ERT. The results pave the way for expanding and developing current assessment practices in foreign language ERT.

Key words: foreign language teaching, emergency remote teaching, assessment practices

Introduction

The overriding purpose of the present study was to examine general upper secondary students' perceptions of teacher assessment practices in foreign language emergency remote teaching in Finland. As of March 2020, schools at all levels in Finland operated in emergency remote teaching (henceforth, ERT) due to Covid-19. However, teaching has not been conducted online constantly, and variation between regions has been great. Following the orders of the Finnish government, all teaching was ERT during spring 2020. The only exceptions were students in the lowest grades and students with special needs who were still entitled to contact teaching. As of August 2020, teachers have taught both in the classroom and online. Rapid switches to ERT have occurred when the pandemic has worsened in a

Email address for correspondence: toni.makipaa@helsinki.fi

specific region. In other words, teaching has been contact teaching, ERT, or hybrid teaching in which either the teacher is at home and the students are at school, or some students are at home while others are at school.

Assessment in education is powerful (National Academy of Education, 2005; Stobart, 2008); with assessment, teachers form a comprehensive picture of what knowledge students already possess, how they work in relation to their goals, and what support and guidance they need from the teacher (Dorfman & Dougherty, 2017). This guidance can be provided as feedback; teacher feedback on student achievement is a powerful way of advancing learning (Carless & Boud, 2018; Hattie & Clarke, 2019). Nevertheless, most teachers find assessment unpleasant due to grading (Linnakylä & Välijärvi, 2005; Thorndike & Thorndike-Christ, 2010; Tsagari & Vogt, 2017). Because of institutional and policy contexts, assessment manifests itself in a range of ways across educational systems (Scarino, 2013). Therefore, context-specific research on assessment is required.

In the interest of enhancing teaching practices, it is important to focus on ERT. As Anderson (2021) states, Covid-19 will not be the last pandemic affecting education and the world, so future teachers need to be better equipped with the pedagogy of ERT. Teachers also need more support for switching between contact teaching and remote teaching (Alderson, 2021). As the pandemic has placed constraints on education and assessment, research on both teachers' and students' perspectives is crucial for increasing our knowledge of effective teaching practices in ERT. Therefore, the purpose of this study was to shed light on the suitability of various assessment practices to foreign language ERT, and to investigate students' experiences with their teachers' assessment practices. Throughout this paper, ERT refers to teaching conducted temporarily online out of necessity, and this teaching was originally planned to be contact teaching (Hodges et al., 2020). In other words, ERT is involuntary (Hodges et al., 2020).

Literature review

Teacher assessment practices

In the literature, assessment is often discussed from the perspectives of summative and formative assessment. Even though a generally accepted definition of formative assessment is lacking (Dorfman & Dougherty, 2017), most scholars agree that summative assessment

measures what students have learnt at the end of the learning period, while formative assessment supports and solidifies learning during the learning period (Van den Branden, 2020). Fundamentally, summative and formative assessment are processes; the first summarizes the evidence of the learning period (Taras, 2005). With formative assessment and feedback, the latter indicates the gap between what the student has accomplished and what was required (Taras, 2005).

This bifurcation of assessment has been criticized; Sadler (1989) argues that the difference between these two types of assessment relates to the impact and use of assessment data, not when the assessment takes place. Put differently, what is of concern to teachers is the purpose of assessment. For example, summative assessment encompasses several purposes, such as selection decisions, plans for students' educational future, and political decisions, whilst formative assessment includes gauging whether students have understood the material and supporting them going forward (Thorndike & Thorndike-Christ, 2010). It is also crucial to note that summative and formative assessment are referred to as assessment of learning and assessment for learning, respectively. When teachers decide the pertinent assessment practices, they must consider the learning goals: what was supposed to be learnt, and how will students be able to demonstrate how successfully they have attained the goals (National Academy of Education, 2005). Furthermore, the assessment practices are shaped by teaching practices, and they should be in harmony to form a continuity for the student (Atkinson & Lazarus, 2013).

Central to formative assessment is feedback, which refers to comments regarding the quality or other distinct aspects of a performance (Yoon & Burton, 2020). Feedback has traditionally been viewed as either explicit or implicit; the former refers to feedback with corrections, while the latter is concerned with negotiation strategies, such as clarification requests (Gass & Mackey, 2015). When used accordingly, feedback can have a considerable impact on learning (Hattie & Clarke, 2019; Hattie & Timperley, 2007; Carless & Boud, 2018). For example, teacher feedback enhances self-regulated learning and reflection (Hattie & Clarke, 2019; Hattie & Timperley, 2007). When teachers provide feedback, students receive additional information on how to adjust their learning (Gass & Mackey, 2015). Students should also be active in receiving and interpreting feedback, as feedback cannot be reduced to a monologue from the teacher (Carless & Boud, 2018; Hattie & Clarke, 2019). However,

it seems that in practice, teacher feedback most often pertains to short comments and remarks (Vogt et al., 2020).

Other common formative assessment practices in the language classroom include self-assessment and peer assessment. Self-assessment can be viewed as a process during which students ponder their performance and compare it with external criteria and their own personal goals. In essence, self-assessment can be encapsulated as "feedback for oneself from oneself" (Andrade & Du, 2007, p. 160). However, confusion remains in the terminology as some scholars use self-assessment, self-reflection, and self-evaluation as synonyms, even though their meanings are not interchangeable (Andrade & Du, 2007). Peer assessment refers to feedback from an equal-status student who considers the level, value and quality of another students' performance or product (Topping, 2009). Compared with teacher feedback, peer assessment is particularly advantageous due to its greater volume and immediacy (Topping, 2009). If teachers introduce and describe the assessment criteria for students, that enables students to self-assess their performance, resulting in monitoring and regulating one's own performance; in turn, this will help students to provide feedback to peers (Van den Branden, 2020). In other words, presenting the criteria supports students' self-assessment and peer assessment skills (Van den Branden, 2020). According to Kearns, (2012), peer-assessment also enhances community development.

Opting for diverse assessment methods, both summative and formative, is integral to effective teaching (FNBE, 2016), and this can be achieved through several assessment practices common in language teaching. For example, teachers possess a range of practices for assessing oral skills, such as recording students' speaking in class or group activities (Atkinson & Lazarus, 2013). Another typical assessment practice is to maintain portfolios, which refers to a collection of students' skills and accomplishments (Farrell & Jacobs, 2010). With portfolios, students take ownership of their learning and become more independent (Farrell & Jacobs, 2010). In their teaching, teachers should use various assessment practices alongside exams to amplify learning and to accurately measure how students have achieved the learning outcomes (Atkinson & Lazarus, 2013). It is crucial to use alternative assessment practices in conjunction with the so-called traditional assessment practices (exams) to mirror real-life settings (Farrell & Jacobs, 2010). As articulated by Vogt and colleagues (2020), writing and speaking are the core skills most assessed in language teaching.

Assessment in emergency remote teaching

Even though online instruction might be problematic compared to contact teaching, several assessment practices common in the classroom are also applicable in ERT, although some minor adjustments might be needed (Hodges & Barbour, 2021). However, it is more feasible to focus on formative assessment in ERT to foster the learning process instead of implementing closed-book exams at the end of the teaching unit (Hodges et al., 2020). In support of this claim, Cahyadi and colleagues (2021) argue that teachers should prevent fraudulent practices in assessment-related tasks by minimizing the use of written exams and opt for other assessment practices.

The perceptions of various stakeholders have been propelled to the forefront in investigations of ERT and assessment. A study by Karatas and colleagues (2021) focused on Turkish faculty members' experiences of ERT. The results show that most members did not have any training in online teaching, nor had they ever taught remotely before the pandemic. Moreover, the members underscored the relevance of feedback in online teaching, but they believed that assessment could not be conducted similarly in remote teaching as in contact teaching. The authors conclude that effective assessment and feedback practices in remote teaching should be addressed in professional development programs. Another study by Mäkipää and colleagues (2021) investigated Finnish language teachers' perceptions of assessment and feedback practices in ERT. According to the teachers, the practices they used were suitable for ERT. However, compared to face-to-face teaching, assessing students was more challenging, and some teachers gave less feedback. Differences between schools in terms of the amount of feedback were also detected: general upper secondary teachers gave less feedback in ERT than lower secondary teachers.

The studies above focused on teachers, but students' perceptions have also been addressed in research. For example, Mäkipää (2023) studied the perceptions of general upper secondary students to teachers' feedback practices in ERT. The students generally found teacher feedback to be encouraging and instructive, but students with lower course grades perceived teacher feedback to be more demotivating and more discouraging compared to their peers with higher course grades. It also seems that teachers have failed to personalize the feedback. Al Shlowiy and colleagues (2021) compared students' and teachers' perceptions of ERT. Although they agreed on several issues, one significant difference regarding assessment was detected: the teachers were inclined to believe that students

would cheat in online exams, while the students held the opposite view. The writers speculate that the participants might conceptualize cheating differently; basically, this means that some forms of cheating might be so subtle for students that they would not consider themselves to be cheating. Moreover, Sofianidis and colleagues (2021) explored Cypriote secondary students' perceptions of ERT. Their results show that the students perceived the assessment practices to be unsuitable for ERT as they did not reflect their learning. According to the students, the practices were ineffective and unreliable. The authors concluded that teachers need more support for providing feedback in online instruction and for using continuous assessment practices.

Assessment literacy

As assessment is crucial at all levels of education, it is imperative that both students and teachers understand the meaning and use of assessment practices. In other words, students and teachers are expected to be assessment-literate. Being an assessment-literate student is a fundamental property from the point of view of becoming self-regulated (Smith et al., 2013). Student assessment literacy means that students are required to comprehend the purpose of assessment, be cognizant of the processes of assessment, and be provided opportunities for practicing how to judge their own work (Smith et al., 2013). In terms of teacher assessment literacy, Fulcher (2012) defines assessment literacy as the abilities, knowledge, and skills pertinent to designing and evaluating standardized and/or classroom tests. In addition, teachers should be aware of the ethics and principles that underpin practice. They are also required to position their skills and knowledge within several frameworks; hence, they comprehend the evolution of the practices, and they are able to assess the wider roles and consequences of testing (Fulcher, 2012). Previous definitions of assessment literacy emphasized individual characteristics of what teachers should be able to know and do (DeLuca et al., 2019). By contrast, contemporary definitions also consider the context and its interplay with integrating pedagogy and teachers' knowledge of assessment (DeLuca et al., 2019; Xu & Brown, 2016). In other words, the school's curriculum, teachers' background and in-service training as well as other professional activities shape teachers' knowledge of assessment (Xu & Brown, 2016).

From the students' point of view, assessment literacy is a major topic, as assessment-literate teachers can choose suitable assessment practices that in turn advance students' learning (Popham, 2011). Assessment-literate teachers also provide opportunities for self-

assessment and peer assessment, which advances student assessment literacy (Smith et al., 2013). Simply put, assessment can even be detrimental for students if the teacher chooses unsuitable assessment practices (Atkinson & Lazarus, 2013; Popham, 2011).

Previous research on foreign language teachers' assessment literacy indicates that teachers do not always employ diverse assessment methods (e.g., Sultana, 2019; Tsagari & Vogt, 2017; Vogt et al., 2020). For instance, teachers might devote a great deal of time to high-stakes assessments, thereby neglecting feedback and oral assessments (Levi & Inbar-Lourie, 2020). Some teachers reduce feedback to brief comments or remarks, which do not necessarily assist in learning (Vogt et al., 2020). A possible explanation is that the amount of training on using diverse assessment practices in teacher education is scant (Kvasova & Kavytska, 2014). With this deficient knowledge of assessment, most teachers rely on the assessments of the publisher of the materials or create their own questions haphazardly (Guskey, 2016). As these results regarding teacher assessment literacy conducted before the pandemic are unfavorable, it is essential to study assessment practices in foreign language ERT to map how well teachers have succeeded in using assessment practices in a novel context (ERT). As research on student assessment literacy is limited (Chan & Luo, 2021; Smith et al., 2013), focusing on how students perceive the suitability of assessment practices is a primary concern.

Methodology

Context of the study

Finnish general upper secondary education lasts three years, at the end of which students take the matriculation examination. As of Spring 2022, students are required to take at least five tests for this exam, which is the only high-stakes assessment in Finland. The core curriculum for general upper secondary education lists several salient learning goals for students (FNBE, 2016). Students are to develop their learning-to-learn skills, critical thinking skills, and independent working skills. The goal is to become a self-regulated, life-long learner.

In terms of assessment, teachers are expected to use an array of assessment practices to assess their students. Using various assessment practices, teachers are required to support learning, encourage students to set goals, and choose pertinent teaching methods. In

addition, teachers are required to provide students with multifaceted feedback in every course, and to employ self- and peer assessment. When a new course starts, teachers are expected to explain the assessment practices used in the course and the criteria for good performance (FNBE, 2016).

Participants

The participants of this study were 282 Finnish general upper secondary students. However, only the responses from 251 students were analyzed as some students answered only the background questions or the questions of the first section. Using purposive sampling, the students came from seven schools. In these schools, ERT was used extensively during the pandemic because of the regulations of the government and municipalities. Due to the great amount of ERT in these schools, it is worthwhile to investigate students from these schools to gain a deeper understanding of ERT and pedagogy. The schools are sizeable with 200–500 students in each. Two of the schools are located in Western Finland, while the others are in Southern Finland. Two of the schools in Southern Finland emphasize foreign language courses in their program. Specific information about the gender balance is not available from every school, but it is safe to assume that most of the students are girls as they outnumber boys in general upper secondary education in Finland (Kupiainen et al., 2018).

In the background questions of the survey, students were asked about their gender, grade, number of language courses completed in ERT, and languages studied in ERT. Table 1 illustrates this background information about the students.

Table 1. Background information about the students

Gender	**Number of students**	**%**
girl	173	69
boy	72	29
non-binary	4	1
did not disclose	2	1
Grade		
10	130	52
11	95	38
12	25	10
did not disclose	1	0
Number of courses in ERT		
1–2	48	19
3–4	100	40
5–6	71	28
7 or more	32	13

Most common languages studied in ERT		
English	246	98
Swedish	228	91
German	44	18
French	42	17
Spanish	37	15
Russian	20	8

As depicted in Table 1, most of the respondents were girls, and they were in the 10th grade. Close to 80% of the students had relatively substantial experience of foreign language ERT as they had completed at least three language courses. Nearly every student had studied English and Swedish, which was not surprising. Finnish students must study at least two languages, which are commonly these two. In other words, the ecological validity of the study is relatively high as the sample is in many aspects akin to the target population (Dörnyei, 2007), although the number of the oldest students was low.

Prior to undertaking the investigation, written consent was obtained from the students. They participated in this study voluntarily, and they were also issued a data protection notice. All the necessary research permits were obtained from the municipalities and the schools. In terms of anonymity, no-one can be recognized from the responses as no personal data were collected.

Study design

The study design of this paper is a survey of a purposive sample of schools. As discussed in the introduction, the use of ERT has varied significantly between Finnish regions during the pandemic. A purposive sample enables the investigation of students in schools that have made considerable use of ERT. Thus, the study design allows an intensive focus on specific students to gain a deeper understanding of larger populations (Dörnyei, 2007). Studying all schools in Finland would not be required as ERT was not used extensively in all schools. Thus, by examining students who have undertaken a great number of language courses in ERT, this paper contributes to the existing literature by providing suggestions on shaping and redesigning future assessment practices in foreign language ERT.

The key research questions of the study were: (1) Which assessment practices do students find to be suitable for foreign language emergency remote teaching? (2) What has students' experience been of the assessment practices in foreign language emergency remote teaching? The questions were addressed using quantitative and qualitative research methods respectively. The mixed methods approach was selected to provide

complementary data about assessment and also to serve as data triangulation (Creamer, 2018; Dörnyei, 2007). For example, the quantitative results showed that students appreciated feedback, and the qualitative data indicated that some students considered the lack of feedback to be a negative issue. To paraphrase, both datasets showed students' appreciation of feedback in foreign language ERT.

Data collection and analysis

The data were collected through an online survey that focused on how the students perceived foreign language ERT in general, the amount and quality of teacher feedback, as well as the suitability and quality of teacher assessment. The data collection took place in May 2021. Students filled in the survey at school, but they were also provided with an option to complete it at home. The data reported in this paper focus on the suitability and quality of teacher assessment in ERT. Ten researchers provided feedback on the initial draft of the questions. To increase the validity of the survey, pretesting was carried out with a group of students (N=25) to ensure the comprehensibility of the items and to study whether important perspectives had been selected for the survey. Based on the feedback from the researchers and the students, minor revisions to the items were made.

For the first research question, the dataset included a list of 14 assessment practices. These were based on the literature on typical assessment practices in online teaching and ERT (e.g., Hodges & Barbour, 2021; Kearns, 2012; Palloff & Pratt, 2008) and on the guidelines for assessment of the national core curriculum (FNBE, 2016). The students were asked how suitable they found the practices for ERT, and they answered on a scale from one (*not at all*) to five (*very suitable*). The students could also choose a response category *Not applicable* if they did not have experience of particular practices in ERT. A five-point Likert scale was chosen because according to Revilla and colleagues (2014), the quality of the data decreases if the scale includes more than five points. However, to date, there has been little agreement on how many points a Likert-scale should include (Revilla et al., 2014).

The 14 assessment practices of the survey were: exam without the use of course material, exam with the use of course material, word quiz, listening comprehension test, oral exam, teachers' written feedback, portfolio, essay, self-assessment, peer-assessment, presentation, learning diary, teachers' oral feedback, and video/recording made by a student. However, it should be noted that the list includes both products (e.g., exams and tests) and processes

(e.g., feedback and self-assessment). For example, feedback cannot exist without formal assessment of an exam or an essay. Therefore, formative assessment cannot precede summative assessment (Taras, 2005). The practices are also often combined in various ways (e.g., portfolio + self-assessment + peer-assessment + teacher written feedback).

Gender and previous course grade were used as independent variables in the quantitative analyses. Previous Finnish research has found that students perceive teachers' teaching practices incongruently in foreign language teaching. High-achieving students pay more attention to classroom activities and report a range of teaching and assessment practices used by their teachers, while low-achieving students who study in the same classroom report fewer practices (Hildén & Rautopuro, 2014). Moreover, how students have perceived teacher feedback in foreign language ERT differs according to students' previous course grades (Mäkipää, 2023). Gender-wise, prior research has found that learning outcomes differ between girls and boys in foreign languages (Hellgren & Marjanen, 2020; Kupiainen et al., 2018). Hence, it was deemed suitable to employ gender and previous course grade as independent variables. One-way ANOVA was applied to compare students' perceptions by gender and grade. However, the non-parametric Mann-Whitney test was also run as the data were not normally distributed. The results of the Mann-Whitney test are reported when a discrepancy was found between the two tests. These methods of analysis were chosen as the aim was to provide information about the spread of the scores and overall tendencies as well as to compare students' perceptions (Dörnyei, 2007). Concerning alternative methods, a cluster analysis would also have yielded salient information about students' perceptions of assessment practices, but the aim of this paper was not to create separate learner groups or profiles.

In Finnish schools, students are assessed on a scale from 4 (failed) to 10 (excellent), and a grade of 8 indicates good knowledge. Based on the students' previous course grades in English and Swedish courses, two groups were formed. These languages were chosen because they are the most common languages in Finnish schools. Finnish students must study Swedish as a second language in school, and students also have to study the advanced syllabus at least in one foreign language. This language is generally English. Table 2 describes the distribution of the students in the two groups.

Table 2. Distribution of students in groups of grades.

	below 8 (group 1)	**8 and above (group 2)**
both grades below 8	30	-
one grade below 8	70	-
one grade 8 and one grade above 8	-	39
both grades 9 or 10	-	106
all	100	145

As Table 2 shows, the first group consisted of students whose previous course grades in the English and Swedish courses were both below 8, or who had at least one grade below 8. The second group included students whose previous course grades were both at least 8. Six students were excluded from the analysis as they did not remember their grades. As a grade of 8 refers to good knowledge, it was decided to use that grade as the cut-off value when the groups were formed (Mäkipää, 2023). In their current form, the groups consist of students who receive low grades (group 1) and good as well as excellent grades (group 2).

The data for the second research question includes students' responses to the open-ended question: What has your experience been of the teachers' assessment practices (exams, essays, presentations, etc.) in foreign language emergency remote teaching? Inductive content analysis (Drisko & Maschi, 2015) was used to analyze the responses, but only 196 responses were analyzed. Out of the 55 excluded responses, two did not pertain to the question, and 53 were blank. As research and theories on assessment in ERT is sparse, inductive content analysis was deemed suitable as selecting well-established codes prior to coding the data would have been impractical (Drisko & Maschi, 2015). However, the disadvantage of the inductive approach is that the lack of a theoretical framework limits the interpretative power of the analysis (Braun & Clarke, 2006).

The guidelines of Braun and Clarke (2006) were followed in the content analysis. First, the author familiarized himself with the data by reading all the responses several times and simultaneously making notes. Second, he coded the data, and recurring topics were grouped under three potential themes: negative, neutral, and positive perceptions. Third, the themes were reviewed to ensure that they worked, and they were subsequently defined and named (Braun & Clarke, 2006).

In terms of reliability, investigator triangulation (Tsagari & Vogt, 2017) in the form of an outside rater was used to calculate the interrater reliability. The rater assessed 10% of the responses. These responses were selected randomly from the 196 responses. The rater was given the three themes used to analyze the data, and she separately analyzed the responses, which allowed for cross-verification of data (Tsagari & Vogt, 2017). The agreement rate was 90%. In terms of validity, both the closed-ended and the open-ended questions were based on robust research, they were pilot tested prior to data collection, and feedback was collected on the items from researchers. All these practices enhanced content validity (Cohen et al., 2007).

Results

The results are presented in the order of the research questions. First, students' perceptions of suitable assessment practices for ERT are presented (RQ 1). They are displayed by gender and grade. Second, students' responses concerning their experiences with language teachers' assessment practices are shown (RQ 2). To illustrate the responses, direct quotes from the dataset will be presented. The quotes were translated from Finnish to English.

Suitable assessment practices by gender

The aim of the first research question was to ascertain which assessment practices are suitable for foreign language ERT, according to general upper secondary students. The responses are shown by gender in Table 3.

Table 3. Suitability of assessment practices for ERT by gender.

	boys		girls		all					
item	**M**	**S.D.**	**M**	**S.D.**	**M**	**S.D.**	**df**	**F**	**p**	**η^2**
1.exam without the use of course material	3.24	1.17	2.94	1.14	3.04	1.17	1	2.986	.085	.01
2.exam with the use of course material	3.46	0.86	3.63	1.06	3.58	1.00	1	1.175	.280	.01
3.word quiz	3.21	1.17	3.21	1.19	3.21	1.18	1	.000	.991	.00
4.listening comprehension test	3.59	1.10	3.88	1.03	3.79	1.06	1	3.276	.072	.02
5.oral exam	2.92	1.22	3.13	1.23	3.06	1.24	1	1.217	.271	.01

6.teachers' written feedback	3.59	1.01	3.90	0.93	3.81	0.96	1	4.581	.034	.02
7.portfolio	2.94	1.06	2.85	1.23	2.89	1.18	1	.252	.616	.00
8.essay	3.73	1.08	4.18	0.82	4.06	0.93	1	10.739	.001	.05
9.self-assessment	3.46	1.00	3.67	1.06	3.61	1.04	1	1.710	.192	.01
10.peer-assessment	2.94	1.05	2.80	1.21	2.85	1.16	1	.617	.433	.00
11.presentation	3.05	1.16	2.97	1.15	3.00	1.16	1	.187	.666	.00
12.learning diary	2.38	1.08	2.68	1.20	2.61	1.18	1	2.899	.090	.01
13.teachers' oral feedback	3.27	0.95	3.44	0.98	3.40	0.97	1	1.412	.236	.01
14.video/recording made by a student	2.97	1.12	2.98	1.24	2.98	1.21	1	.004	.952	.00

Note: M = mean, S.D. = standard deviation, η^2 = partial eta squared

From the data in Table 3, it is apparent that from the students' perspective, the most suitable assessment practices for foreign language ERT were essays, teachers' written feedback, and listening comprehension tests. In contrast, learning diaries, peer-assessments, and portfolios were deemed the least suitable. Two statistically significant differences were found: girls found teachers' written feedback and essays more suitable than boys. However, the effect sizes were small.

Suitable assessment practices by grade

In line with the first research question, students' perceptions of the suitability of assessment practices were also examined by grade. Table 4 provides the results obtained from the one-way ANOVA.

Table 4. Suitability of assessment practices for ERT by grade.

	below 8		**8 and above**					
item	**M**	**S.D.**	**M**	**S.D.**	**df**	**F**	**p**	**η^2**
1.exam without the use of course material	3.14	1.15	2.94	1.19	1	1.401	.238	.01
2.exam with the use of course material	3.53	1.01	3.61	1.00	1	.279	.598	.00
3.word quiz	3.42	1.08	3.05	1.20	1	5.255	.023	.03

4.listening comprehension test	3.55	1.04	3.94	1.05	1	7.316	.007	.03
5.oral exam	2.81	1.08	3.23	1.28	1	6.445	.012	.03
6.teachers' written feedback	3.47	0.97	4.06	0.85	1	21.785	<.001	.10
7.portfolio	2.48	1.11	3.19	1.13	1	20.486	<.001	.09
8.essay	3.70	0.95	4.32	0.81	1	25.074	<.001	.11
9.self-assessment	3.39	1.06	3.74	1.02	1	5.951	.016	.03
10.peer-assessment	2.66	1.03	3.00	1.20	1	4.608	.033	.02
11.presentation	2.70	1.04	3.24	1.17	1	11.681	.001	.06
12.learning diary	2.49	1.12	2.68	1.18	1	1.437	.232	.01
13.teachers' oral feedback	3.19	0.91	3.56	0.96	1	7.691	.006	.04
14.video/recording made by a student	2.57	1.06	3.30	1.21	1	20.684	<.001	.09

Table 4 is revealing in several ways. First, students with lower course grades found essays, listening comprehension tests, and course exams with the use of course material to be the most suitable assessment practices for ERT. Similarly, students with higher course grades considered essays and listening comprehension tests to be suitable. However, instead of course exams with the use of course material, they opted for the teacher's written feedback. Second, in terms of the unsuitable assessment practices, students with lower course grades chose portfolios, learning diaries, and video/recordings made by a student. In contrast, students with higher course grades found learning diaries, course exams without the use of course material, and peer assessment to be unsuitable for ERT. Third, statistically significant differences were found in 11 of the 14 items. The items without a statistically significant difference were course exams without the use of course material, course exams with the use of course material, and learning diaries. Fourth, further analyses showed that the largest effect sizes were in oral presentations, video/recordings made by a student, portfolios, teacher's written feedback, and essays. These effect sizes were medium, while the others were small (Cohen, 1988). The item with the largest effect size was essays.

Students' experiences of language teachers' assessment practices

The second research question focused on students' experiences of the assessment practices in foreign language ERT. To answer this question, content analysis was used to group the students' responses thematically. As a result, three groups emerged from the analysis: negative, neutral, and positive perceptions. Most students manifested positive perceptions (49%) of teachers' assessment practices. The students found the practices to be good (38%), fair (5%), diverse (2%), and equal (2%). In essence, the students felt that their teachers had succeeded in assessing them well using multifaceted and flexible practices. According to the students, the core skills in language learning (reading, listening, writing, and speaking) were also considered in the assessment practices. The following excerpts illustrate students' responses:

> "Pretty good considering that it's difficult to assess in emergency remote teaching." (girl, 10th grade)
>
> "[Teachers assessment practices] have measured learning well." (girl, 10th grade)
>
> "Assessment has been good." (boy, 10th grade)

Although nearly half of the students mentioned positive examples, close to one third of the students (32%) pointed out negative issues in teachers' assessment practices. A variety of perspectives were expressed in these responses. According to some of the students, assessment had been one-sided (12%), too strict (6%), unfair (2%), and unclear (2%). In addition, the amount of feedback had been insufficient (9%). According to the students, teachers had primarily used essays and exams, but the grading of the assignments had been extremely strict. Some students raised concerns about the likelihood of cheating as it was effortless to cheat with online translators. Excerpts illustrating students' negative perceptions are shown below.

> "Exams: too long, difficult, unclear and the exams could be much simpler so basically the questions wouldn't be in different places. For instance, the questions in Teams are unclear." (girl, 11th grade)
>
> "They [Teachers' assessment practices] have not worked well because in emergency remote teaching, it's easy to use online translators and cheat." (girl, 11th grade)
>
> "I would have wanted to receive more feedback and comments about why I lost points in an essay." (boy, 10th grade)

In terms of neutral (19%) perceptions, most students pointed out that teachers' assessment practices in ERT did not differ from the practices in face-to-face teaching (11%). However, this contrasts with a small number of students (2%) who claimed that differences existed between the practices; the students emphasized that exams and other tests in ERT were deliberately designed to be more difficult as teachers assumed that students would cheat. Moreover, the use of various practices varied between the teachers (2%). The excerpts below illustrate students' responses.

> "In my opinion, assessment has been the same in emergency remote teaching and contact teaching." (non-binary, 10th grade)
>
> "Of course, it depends on the teacher." (boy, 12th grade)
>
> "Due to emergency remote teaching, teachers have had to execute assessments differently." (girl, 10th grade)

Discussion

The purpose of the current study was to ascertain which assessment practices students find suitable for foreign language ERT, and how they experienced language teachers' assessment practices in ERT. From the student perspective, it can be deduced that the most suitable assessment practice for foreign language ERT was the essay. As foreign language teachers mostly assess writing among other skills (Vogt et al., 2020), this finding is not surprising. It also seems that written feedback, self-assessment, and exams with the use of course material were more suitable assessment practices than oral feedback, peer assessment, and exams without the use of course material. However, it is useful to bear in mind that formative assessment practices, such as teacher feedback, come second after formal assessment (Taras, 2005). Even though students opted for written feedback, the teacher must complete formal assessment of student achievement before providing feedback. The students pointed out neutral perceptions on the suitability of presentations and oral exams. Interestingly, students found the learning diary to be an unsuitable assessment practice for ERT. Using only traditional exams should be avoided in assessment (Farrell & Jacobs, 2010), and the results show that students opt for several other assessment practices in ERT. This implies that students have acknowledged the multifaceted nature of assessment and that closed-book exams are not the only suitable assessment method.

In terms of course grades, students with higher course grades found nearly all assessment practice to be more suitable for ERT than did their peers with lower course grades, and most of the differences were statistically significant. These findings match those in Mäkipää (2023), who found that students' perceptions of teacher feedback in foreign language ERT differ by grade. This raises the critical question of why students with higher course grades express more positive perceptions of the suitability of assessment practices. Hildén and Rautopuro (2014) suggest that high-achieving students are more attentive to teaching and assessment practices in the classroom than low-achieving students. Therefore, it is possible that students with lower course grades have similarly overlooked teaching practices in ERT, and consequently they fail to recognize the suitability of some teacher assessment practices. This is a conundrum from the perspective of student assessment literacy, as assessment-literate students can use teacher assessment to monitor and enhance learning (Smith et al., 2013). If students with lower course grades are unable to recognize suitable assessment practices for foreign language ERT, one wonders how assessment-literate they are.

Surprisingly, teachers' written feedback and essay were the items in which the only statistically significant differences were detected by gender and they had the largest effect sizes in terms of previous course grades. This suggests that these two items involve contrasting perceptions by the respondents. In brief, girls and students with higher course grades found teachers' written feedback and essays to be more pertinent assessment practices for ERT than did boys and students with lower course grades. The reasons for these conflicting results are not clear, but regarding gender, the reason might be attributed to grades. As girls receive better grades than boys in Finland in basic education (e.g., Hellgren & Marjanen, 2020), it is probable that girls express greater interest in teacher feedback, and they continue this practice in general upper secondary education. Therefore, they regard it as a more crucial assessment practice compared to boys. On the other hand, boys receive higher grades in the English test of the matriculation examination than girls (Kupiainen et al., 2018). This might mean that, due to their poorer English skills, girls take on more in English courses and consequently rely more on teacher feedback.

This study has revealed that most students have positive experiences with teachers' assessment practices. This concurs with previous research regarding language teachers' perceptions of assessment in ERT in Finland (Mäkipää et al., 2021). As most of the students perceived teacher assessment practices positively, it suggests that teachers have succeeded

in choosing germane assessment practices for ERT, and they have also implemented them successfully. This finding was surprising as most language teachers elsewhere have been found to lack assessment literacy (Sultana, 2019; Tsagari & Vogt, 2017; Vogt et al., 2020). As contextual aspects lie at the heart of assessment literacy (DeLuca et al., 2019; Xu & Brown, 2016), it is probable that aspects such as curriculum, professional training and reflection, have supported Finnish teachers' assessment literacy and guided them in creating assessment practices suitable for ERT. Conversely, one-third of the students encountered negative experiences with assessment practices, which had mainly been one-sided and too strict. It also seems that teachers had focused on summative assessment, as feedback was lacking. This goes against the guidelines of the national core curriculum (FNBE, 2016). Assessment should be multifaceted and include both summative and formative aspects, but it seems that this was neglected in some language courses.

Compared with previous research, some distinct differences were found. Some students raised the question of cheating as they were adamant that students would cheat and use online translators. In contrast, the students in Al Shlowiy and colleagues' (2021) study did not agree with this. Another distinct difference was the fact that in Sofianidis and colleagues' (2021) study, the students perceived assessment negatively, as the practices had been ineffective and unreliable. The results of the study at hand were to the contrary.

Conclusions

As the pandemic has caused staggering changes to education worldwide, research on the changes is of primary importance, if we are to advance and develop future teaching. The abrupt change to ERT within days was unprecedented and stressful for both teachers and students worldwide. Given the absence of physical teacher-student interaction in ERT, the importance of valid assessment practices and teacher feedback are magnified (Hodges et al., 2020). The results of this study indicate that these practices in general seem to have worked well, and the study proffers two immediately reliable applications for foreign language ERT.

First, it points to the suitability of specific assessment practices for ERT. From the students' perspective, essays, listening comprehension tests and written feedback are the most suitable assessment practices. These practices are multifaceted as they can be used to assess various skills in language learning (e.g., writing and listening), which is in line with the core curriculum (FNBE, 2016). This corroborates the recommendations of Hodges and

colleagues (2020) and Cahyadi and colleagues (2021), who argue that closed-ended exams should be avoided, and teachers should opt for continuous assessment practices. From the student perspective, choosing multifaceted and germane assessment practices is key, as unsuitable assessment practices can impede learning (Atkinson & Lazarus, 2013). In essence, when teachers design the assessment practices of their courses in foreign language ERT, they should consider implementing essays, listening comprehension tests, and written feedback in particular.

Second, the results suggest that teachers have mostly succeeded in implementing their assessment practices to ERT. Assessment practices can intensify learning (Popham, 2011), which is why it is a primary skill for every teacher to employ suitable practices. Nevertheless, some teachers' assessment practices have generated negative perceptions in students, which is troubling; inappropriate use of assessment practices can have adverse ramifications for students. ERT and the pandemic have been strenuous for teachers and students. Thus, to accelerate learning in ERT, training for in-service teachers to support teachers' growth in pedagogical skills is recommended. This will enhance teachers' assessment literacy (Fulcher, 2012; Tsagari & Vogt, 2017; Vogt et al., 2020). More consideration should also be given to pre-service teachers' assessment courses as the teaching of assessment in teacher education is limited (Kvasova & Kavytska, 2014). Focusing on pre-service teachers is important because hybrid teaching might become a common practice in foreign language teaching in the future, thanks to the new technological skills learned during the pandemic. Therefore, pre-service teachers need training in using assessment practices efficiently in hybrid teaching.

This study extends our knowledge of assessment in foreign language ERT. However, the current study disregarded teachers' perspectives and only examined students' self-reported data. Another drawback of the study is that its participants were not nationally representative of all Finnish students. In addition, proficiency in language learning was measured by previous course grades, which do not reflect all the core skills of language learning fairly. Lastly, the 14 items of the survey include both products and processes, which needs to be considered in the interpretation of the findings.

Further investigation into assessment in ERT is strongly recommended. As this study focused on students' perceptions, future research is needed to ascertain how assessment practices work in practice in ERT. For example, observation studies could focus on the

implementation of various assessment practices in ERT. With such a study, the pedagogy of ERT could be developed and reshaped.

References

Al Shlowiy, A., Al-Hoorie, A. H., & Alharbi, M. (2021). Discrepancy between language learners and teachers concerns about emergency remote teaching. *Journal of Computer Assisted Learning*, *37*(6), 1528–1538. https://doi.org/10.1111/jcal.12543

Anderson, L. W. (2021). Schooling interrupted: Educating children and youth in the covid-19 era. *CEPS Journal*, *11*, 17–38. https://doi.org/10.26529/cepsj.1128

Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education*, *32*(2), 159–181. https://doi.org/10.1080/02602930600801928

Atkinson, T., & Lazarus, E. (2013). Assessment. In A. Swarbrick (Ed.), *Aspects of teaching secondary modern foreign languages: Perspectives on practice* (pp. 200–210). *Routledge/Falmer*. https://doi.org/10.4324/9781315013206

Van den Branden, K. (2020). Measuring task-based performance. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 316–325). Routledge. https://doi.org/10.4324/9781351034784-34

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cahyadi, A., Hendryadi, Widyastuti, S., & Suryani. (2021). COVID-19, emergency remote teaching evaluation: The case of Indonesia. *Education and Information Technologies*, *27*(2), 2165–2179. https://doi.org/10.1007/s10639-021-10680-3

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Chan, C. K. Y., & Luo, J. (2021). A four-dimensional conceptual framework for student assessment literacy in holistic competency development. *Assessment & Evaluation in Higher Education*, *46*(3), 451–466. https://doi.org/10.1080/02602938.2020.1777388

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

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

Creamer, E. G. (2018). *An introduction to fully integrated mixed methods research. SAGE.* https://doi.org/10.4135/9781071802823

DeLuca, C., Coombs, A., MacGrevor, S., & Rasooli, A. (2019). Toward a differential and situated view of assessment literacy: Studying teachers' responses to classroom assessment scenarios. *Frontiers in Education*, *4,* 1–10. https://doi.org/10.3389/feduc.2019.00094

Dorfman, L. R., & Dougherty, D. (2017). *A closer look: Learning more about our writers with formative assessment*. Stenhouse Publishers.

Drisko, J. W., & Maschi, T. (2015). *Content analysis*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190215491.001.0001

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.

Farrell, T. S. C., & Jacobs, G. M. (2010). *Essentials of successful English language teaching*. Continuum. https://doi.org/10.5040/9781474212205

FNBE 2016 = Finnish National Board of Education. (2016) *National core curriculum for general upper secondary schools 2015*. Publications 2016:8.

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

Gass, S. M., & Mackey, A. (2015). Input, interaction and output in second language acquisition. In B. VanPatten, & J. Williams (Eds.), *Theories in second language acquisition* (pp. 180–206). Routledge.

Guskey, T. R. (2016). How classroom assessments improve learning. In M. Scherer (Ed.), *On formative assessment: Readings from educational leadership* (pp. 3–13). ASCD.

Hattie, J., & Clarke, S. (2019). *Visible learning: Feedback*. Routledge. https://doi.org/10.4324/9780429485480

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hellgren, J., & Marjanen, J. (2020). *Svenska och litteratur i slutet av årskurs 9: Resultat av en utvärdering* [Learning outcomes in Swedish language and literature in the final stage of basic education in 2019]. Finnish Education Evaluation Centre, Publication 18/2020.

Hildén, R., & Rautopuro, J. (2014). *Ruotsin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013* [Outcomes of language learning in A level Swedish at the end of basic education in 2013]. Finnish National Board of Education.

Hodges, C. B., & Barbour, M. K. (2021). Assessing learning during emergency remote education. *Italian Journal of Educational Technology*, *29*(2), 85–98. https://doi.org/10.17471/2499-4324/1208

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). *The difference between emergency remote teaching and online learning*. Available: https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning (Accessed Mar 15, 2022).

Karatas, F. Ö., Akaygun, S., Celik, S., Kokoc, M., & Yilmaz, S. N. (2021). Challenge accepted: Experiences of Turkish faculty members at the time of emergency remote teaching. *CEPS Journal*, *11*, 141–163. https://doi.org/10.26529/cepsj.1136

Kearns, L. R. (2012). Student assessment in online learning: Challenges and effective practices. *Journal of Online Learning and Teaching*, *8*(3), 198–208.

Kupiainen, S., Marjanen, J., & Ouakrim-Soivio, N. (2018). *Ylioppilas valintojen pyörteissä. Lukio-opinnot, ylioppilastutkinto ja korkeakoulujen opiskelijavalinta* [Dizzy with choices: Students deciding on general upper secondary studies, the matriculation examination and higher education]. Suomen ainedidaktisen tutkimusseuran julkaisuja: Ainedidaktisia tutkimuksia 14.

Kvasova, O., & Kavytska, T. (2014). The assessment competence of university foreign language teachers: A Ukrainian perspective. *CercleS*, *4*(1), 159–177. https://doi.org/10.1515/cercles-2014-0010

Levi, T., & Inbar-Lourie, O. (2020). Assessment literacy or language assessment literacy: Learning from the teachers. *Language Assessment Quarterly*, *17*(2), 168–182. https://doi.org/10.1080/15434303.2019.1692347

Linnakylä, P., & Välijärvi, J. (2005). *Arvon mekin ansaitsemme: Kansainvälinen arviointi suomalaisen koulun kehittämiseksi* [Worthy of recognition? International assessment and the development of the Finnish school]. PS-Kustannus.

Mäkipää, T. (2023). Feedback practices in foreign language emergency remote teaching in Finland. *Apples – Journal of Applied Language Studies, 17*(1), 1–18. https://doi.org/10.47862/apples.113732

Mäkipää, T., Hahl, K., & Luodonpää-Manni, M. (2021). Teachers' perceptions of assessment and feedback practices in Finland's foreign language classes during the COVID-19 pandemic. *CEPS Journal, 11*, 219–240. https://doi.org/10.26529/cepsj.1108

National Academy of Education. (2005). *A good teacher in every classroom: Preparing the highly qualified teachers our children deserve*. Jossey-Bass.

Palloff, R. M., & Pratt, K. (2008). *Assessing the online learner: Resources and strategies for faculty*. Jossey-Bass.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265–273. https://doi.org/10.1080/08878730.2011.605048

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods & Research*, *43*(1), 73–97. https://doi.org/10.1177/0049124113509605

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/bf00117714

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing*, *30*(3), 309–327. https://doi.org/10.1177/0265532213480128

Sofianidis, A., Meletiou-Mavrotheris, M., Konstantinou, P., Stylianidou, N., & Katzis, K. (2021). Let students talk about emergency remote teaching experience: Secondary students' perceptions on their experience during the COVID-19 pandemic. *Education Sciences, 11*, 1–23. https://doi.org/10.3390/educsci11060268

Stobart, G. (2008). *Testing times. The uses and abuses of assessment*. Routledge.

Sultana, N. (2019). Language assessment literacy: An uncharted area for the English language teachers in Bangladesh. *Language Testing in Asia*, *9*(1), 1–14. https://doi.org/10.1186/s40468-019-0077-8

Taras, M. (2005). Assessment – summative and formative - some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466–478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Pearson.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice*, *48*(1), 20–27. https://doi.org/10.1080/00405840802577569

Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges, and future prospects. *Papers in Language Testing and Assessment*, *6*(1), 41–63.

Vogt, K., Tsagari, D., Csépes, I., Green, A., & Sifakis, N. (2020). Linking learners' perspectives on language assessment practices to teachers' assessment literacy enhancement (TALE): Insights from four European countries. *Language Assessment Quarterly*, *17*(4), 410–433. https://doi.org/10.1080/15434303.2020.1776714

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149–162. https://doi.org/10.1016/j.tate.2016.05.010

Yoon, H-J., & Burton, J. D. (2020). Measuring L2 writing. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 295–304). Routledge. https://doi.org/10.4324/9781351034784-32