JYVÄSKYLÄ STUDIES IN HUMANITIES
329

Pirjo Pollari

(Dis)empowering Assessment?

Assessment as Experienced by Students in Their Upper Secondary School EFL Studies

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN HUMANITIES 329

Pirjo Pollari

(Dis)empowering Assessment?

Assessment as Experienced by Students in Their Upper Secondary School EFL Studies

Esitetään Jyväskylän yliopiston humanistis-yhteiskuntatieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi Normaalikoulun auditoriossa 3005
lokakuun 7. päivänä 2017 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities and Social Sciences of the University of Jyväskylä,
in the Teacher Training School, auditorium 3005, on October 7, 2017 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

(Dis)empowering Assessment?

Assessment as Experienced by Students in Their Upper Secondary School EFL Studies

JYVÄSKYLÄ STUDIES IN HUMANITIES 329

Pirjo Pollari

(Dis)empowering Assessment?

Assessment as Experienced by Students in Their Upper Secondary School EFL Studies

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

Editors
Paula Kalaja
Department of Language and Communication Studies, University of Jyväskylä
Pekka Olsbo, Sini Tuikka
Publishing Unit, University Library of Jyväskylä

Jyväskylä Studies in Humanities
Editorial Board

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-7178-6

URN:ISBN:978-951-39-7178-6
ISBN 978-951-39-7178-6 (PDF)
ISSN 1459-4331

ISBN 978-951-39-7177-9 (nid.)
ISSN 1459-4323

Jyväskylä University Printing House, Jyväskylä 2017

ABSTRACT

Pollari, Pirjo
(Dis)empowering assessment? Assessment as experienced by students in their upper secondary school EFL studies
Jyväskylä: University of Jyväskylä, 2017, 144 p.
(Jyväskylä Studies in Humanities
ISSN 1459-4323; 329 (print) ISSN 1459-4331; 329 (PDF))
ISBN 978-951-39-7177-9 (print)
ISBN 978-951-39-7178-6 (PDF)

Assessment has a great deal of power over students. However, there is little research on how students experience assessment and its power in the school context. The purpose of this mixed-methods study is therefore to examine how students in one Finnish upper secondary school experienced assessment and (dis)empowerment in their EFL studies. The present study, which situates itself within the realm of teacher research, also aims to experiment with alternative assessment methods in order to investigate whether they could foster empowerment in upper secondary EFL studies. The study comprises five articles and a monograph, and is divided into two parts, each with its own research aims. Part 1 and its three articles focus on students' experiences of assessment and (dis)empowerment and explore what factors might predict disempowerment in assessment. In addition, Part 1 focuses on feedback as well as stress and test anxiety in connection with high-stakes testing as possible predictors of disempowerment. The data for these articles was gathered in March 2014 by means of a web-based questionnaire. The aim of Part 2 is to explore whether less traditional assessment methods could promote students' empowerment in assessment. The first article in Part 2 focuses on cheat-sheet tests as a way of engaging and empowering students. The second article explores individual choice in corrective feedback. These teaching experiments took place in six upper secondary groups in 2013-2016. The third study in Part 2 is a monograph describing an earlier portfolio programme in EFL teaching.

The present study shows that although most students were quite satisfied with the assessment in their EFL studies, a significant minority of students found the assessment disempowering. Several factors, such as inadequate or unhelpful feedback or stress and anxiety caused by assessment, predicted assessment disempowerment. However, students seemed to react to assessment as well as to these factors in a highly individual way. Furthermore, although the alternative assessment methods investigated in the teaching experiments proved useful and also empowering additions to the EFL assessment repertoire, students experienced them in different ways. There should therefore be a range of assessment methods to cater for different assessment purposes as well as for students' different learning strategies, needs and personalities.

Keywords: student empowerment, assessment, upper secondary school, EFL, feedback, portfolio, cheat-sheet test, test anxiety, corrective feedback

Author's address Pirjo Pollari
Teacher Training School
University of Jyväskylä
pirjo.pollari@norssi.jyu.fi

Supervisors Prof. Paula Kalaja
Department of Language and Communication Studies
University of Jyväskylä

Prof. Ari Huhta
Centre for Applied Language Studies
University of Jyväskylä

Reviewers Dr. Raili Hildén, Adjunct Professor
University of Helsinki

Prof. Jouni Välijärvi
Institute for Educational Research
University of Jyväskylä

Opponent Dr. Raili Hildén, Adjunct Professor
University of Helsinki

ACKNOWLEDGEMENTS

Writing this dissertation has been an amazing education. In brief, trying to juggle research with full-time teaching has been rewarding but also extremely taxing. At times it has been an emotional roller-coaster ride, with moments of true joy, learning and empowerment interspersed with doubt, frustration and disempowerment. In that sense the project has given me first-hand experience of the power of assessment. However, it has also made me take things at a much slower pace for a couple of years: at the hectic pace of school life, one certainly does not have time to think about one idea or one sentence for long. I really enjoyed that chance. All in all, I do not regret embarking on this journey for a moment although I am glad it is now coming to an end. And, I am sure, so are several people who have helped and supported me on this endeavour. They deserve my most sincere thanks.

First of all, I would like to thank my supervisors, Prof. Paula Kalaja and Prof. Ari Huhta. Thank you both for guiding and supporting me in this project, and most of all, thank you for letting me do most things my way. I know I cannot have been the easiest of graduate students.

Secondly, I would like to express my thanks to the pre-examiners of this work, Dr. Raili Hildén and Prof. Jouni Välijärvi. Thank you for your insightful and encouraging comments, which gave me and this work invaluable feed forward. I am also most grateful that Dr. Hildén agreed to act as my opponent.

Thirdly, I am indebted to Dr. Anne Pitkänen-Huhta for believing in my work and employing me as a doctoral student at the Department of Languages during the most critical months of this work. Doing post-graduate research as a mature student, and as a teacher with a full-time job, is no easy task, and without financial support both from the Department of Languages and from the *Ellen and Artturi Nyyssönen Foundation*, which allowed me to focus on this work full-time for altogether a year, this dissertation would have remained unfinished.

Next, my most heartfelt thanks go to Prof. (emerita) Pirjo Linnakylä, who has given me not only her time, expertise and much needed advice but also unfaltering support and encouragement throughout the project. I am grateful for all that beyond words. Also, I would like to thank Prof. (emeritus) Antero Malin for his help and patience when guiding me through some rough patches in the depths of research methodology and statistics. My sincere thanks are also due to Eleanor Underwood, Wendy Nelson and Louisa Daffue-Karsten for proofreading the summary and the articles of this study, respectively.

There are also several other people, family, friends, colleagues, who have contributed to this work in many different ways. To name just a few, I would like to thank my closest colleagues, all the Ladies in the English department at *Norssi* and, in particular, Maarit Ilola, for sharing the joys and sorrows of this endeavour with me. I also wish to thank both the past and present head teachers of the Teacher Training School of the University of Jyväskylä, particularly Kirsti Koski, for supporting me in myriad ways during this project.

Outside work, I want to thank Tuula and Minna as well as Wellamo for taking care of both my mental and physical well-being. I would also like to thank my parents: my late mother and my father for understanding my nature and letting me pursue my 'academic ambitions' from an early age. To my sister Nina and especially her children Jonatan and Olivia, thank you for playing with me: you reminded me that all work and no play would make Pipa a dull auntie. Je voudrais aussi remercier très cordialement mes voisines et voisins à Sarpe, pour tous les moments entre amis: sans ces moments de bonheur, de soleil et aussi de bon vin, ce projet aurait été beaucoup plus dur et pénible.

Finally, I want to express my deepest gratitude to the students who participated in these studies. Without you all, this dissertation would not have been possible. And thank you, my dear 'Mökkönen' – without you nothing in my life would be possible.

Jyväskylä, 9th September 2017,

Pirjo Pollari

FIGURES

TABLES

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

Article 1 Pollari, P. (2017a). The power of assessment: What (dis)empowers students in their EFL assessment in a Finnish upper secondary school? *Apples – Journal of Applied Language Studies, 11*(2), 147–175.

Article 2 Pollari, P. (2017b). To feed back or to feed forward? Students' experiences of and responses to feedback in a Finnish EFL classroom. *Apples – Journal of Applied Language Studies, 11*(4), 11–33.

Article 3 Pollari, P. (2016). Daunting, reliable, important or "trivial nitpicking"? Upper secondary students' expectations and experiences of the English test in the Matriculation Examination. In A. Huhta & R. Hildén (eds.) *Kielitaidon arviointitutkimus 2000-luvun Suomessa. AFinLA-e. Soveltavan kielitieteen tutkimuksia 2016/n:o 9*, 184-211.

Article 4 Pollari, P. (2015). Can a cheat sheet in an EFL test engage and empower students? *AFinLAn vuosikirja–AFinLA Yearbook, (73),* 208-225.

Article 5 Pollari, P. (submitted for review to *Innovation in Language Learning and Teaching*) How to make corrective feedback more learner-centred? A feedback experiment in upper secondary EFL studies in Finland

Monograph Pollari, P. (2000). *"This is my portfolio": Portfolios in upper secondary school English studies.* Jyväskylä: Institute for Educational Research. http://urn.fi/URN:ISBN:978-951-39-6898-4

1 INTRODUCTION

> But why didn't I get any points for this item? The actual verb form is correct, isn't it?

Over 20 years ago, my sister Nina, who was just starting her upper secondary school, asked me the question above. I had graduated a few months earlier as an English teacher and had used the same exercise in one of my tests. The item in the gap-fill exercise, testing the use of the right tenses and aspects, ran as follows:

> We had been swiming (swim) all day, and so we were (be) tired.

My sister had filled in the right tense, and the right aspect, but had not added the second letter *m*. As a result, her answer was deemed completely wrong. My sister felt it was unfair and discouraging. Her teacher had explained the decision on the grounds of her marking system, where "it's all or nothing – just as it used to be in the Matriculation exam in the 1980s".

I was puzzled. I had marked similar answers, which I saw as almost completely correct, quite differently. I tried to find clear and practical guidelines on assessment, and also an answer to the question how the gap-fill above should have been marked, and why. I reread my teacher education notes, checked the core curriculum, glanced through a few highly acclaimed books on language education, and found very little. I asked my colleagues – and got nearly as many answers as there were colleagues. I was even more puzzled.

My sister's test item was the very beginning of this study as it sparked an interest in assessment in me. The next incentive came when I was invited to join a *portfolio* project two or three years later. The project changed my views on assessment quite profoundly. I also encountered the concept of *student empowerment* for the first time. Through the portfolio project I saw that assessment could empower students, and give students freedom and the power to express themselves. I also learnt that the purpose of assessment could be to give students real feedback in order to encourage and guide them and their learning, and not just to give them marks or grades in order to rank them.

After over 20 years in the field of teaching English, I am still puzzled by assessment, which I find extremely complex. That is the personal raison d'être for this study.

There is a more general reason for the present study as well. Nowadays it is agreed, at least in professional literature and the national core curricula, that the main purpose of assessment is to guide and encourage students' learning and studying. There is also ample research evidence indicating that assessment affects students in many ways: for instance, it significantly affects students' studying and learning as well as their motivation and self-efficacy (e.g. Crooks, 1988; see also Brown & Hirschfeld, 2008). Hence, Crooks (1988, p. 467) concludes that assessment (or evaluation) "deserves very careful planning and considerable investment of time from educators". We Finns, however, seem to take assessment for granted and pay little attention to it. Every school year hundreds of thousands of pupils and students are assessed in our classrooms and millions of tests are drawn up, taken and marked in schools across Finland (see e.g. Atjonen, 2007, p. 10). These tests and other assessment assignments as well as the comments on them, their marks and grades have varying degrees of effect on students. Sometimes, when they are used for selecting students for further education, the grades may have a great influence on the individual's future. Yet we know very little about these tests and how students experience them. Does assessment really guide and encourage students' learning and studying as it should, or does it discourage them? Does assessment empower or disempower students?

Since the Matriculation Examination is the only external examination in the Finnish school system, our educational assessment system relies almost entirely on the assessment teachers carry out in their classrooms (e.g. Huhta & Hildén, 2016; Sahlberg, 2007). Despite this, our past and current national core curricula say rather little on assessment and give few practical guidelines or instructions for classroom assessment. Moreover, there has been little research on student assessment in Finland in general, and on upper secondary school assessment it is particularly rare. Furthermore, research on how Finnish foreign or second language education is implemented in classrooms is "surprisingly scarce" (Harjanne & Tella, 2009, p. 136), and so is research on student assessment in foreign language education in Finland. Thus, the present study attempts to make its contribution in the field of Finnish student assessment, and in particular, in the study of English as a foreign language (EFL) in upper secondary school. Moreover, this study aims to indicate that more research is needed.

1.1 The setting and research questions of the present study

Assessment is a fundamental part of education (e.g. Taras, 2005). It is also a vast and complex topic, on which, "there is a lack of commonality in the definition of the terminology relating to it" (Taras, 2005, p. 466). Yet, basically, assessment

and evaluation mean judging the worth, value and importance of something (see e.g. Atjonen, 2007, pp. 19-20; Linnakylä & Välijärvi, 2005, p. 16), judging its "goodness" (Stake, 2004, p. 8). For instance, according to the Cambridge Online Dictionary (http://dictionary.cambridge.org/dictionary/english/assessment), assessment[1] is "the act of judging or deciding the amount, value, quality or importance of something, or the judgment or decision that is made".

The concept has two different aspects or levels, *a judgement* and *a decision*, and at least the former – the judgement – is made. The judgement means the determined value of the given thing whereas the decision concerns the use of the assessment judgement, in other words the decision, for instance an action or a process, which it enables (Newton, 2007). However, Newton (2007) argues that assessment has a third level as well, the impact level, which concerns the intended impacts of the assessment.

Indeed, the judgement may be the first aspect of assigning value in informal, everyday situations, for example when we say whether we like something or not, as we do not necessarily intend to assess and analyse it for a particular purpose, use or reason.

However, in the school context, student assessment should be intentional, and thus it should have a purpose, the reason why we assess our students' work and learning. That purpose should come first and it should guide a great many subsequent decisions, such as the content, methods and timing of the assessment (e.g. Gipps, 1994, p. 3). The purpose also guides the intended impact: what do we wish to accomplish through the assessment? Generally speaking, the intended impact of student assessment is to let students know how well they have reached the learning goals and then guide students to act upon this information to further their learning.

When discussing the impact of feedback on learning, Hattie (2009) and Wiliam (2012, p. 33) emphasise the individual student's reaction to it:

> Feedback given by a teacher to one student might motivate that student to strive harder to reach a goal, whereas exactly the same feedback given by the same teacher to another student might cause the student to give up.

Thus, student assessment seems to have a fourth level, i.e. *the actual impact* that student assessment has on the learner since, ultimately, the impact of student assessment depends on how the learner reacts to the assessment and its

---

1 In this study, I will talk about *assessment*. Some readers as well as authors may make a distinction between the concepts of *assessment* and *evaluation*. The definitions of these two can vary quite considerably, depending on the background, language variety and discipline of the language user. Some American English authors consider assessment to be more formative, giving information on learning in order to improve teaching, but when grades are given, it is considered evaluation. However, most British English authors of today do not see a similar distinction. For many British authors, evaluation means mainly evaluating schools or systems, not students (see e.g. Harlen, 2007; Newton, 2007; Wilcox, 1992). Thus, as there are no clear-cut definitions for these concepts, I will use the word *assessment* whenever discussing any form of student assessment taking place in a school environment, whether it involves grades and marks or not.

outcome (Hattie, 2009; Wiliam, 2012). This assessment feedback on their own learning may, hopefully, empower students. However, it may also disempower students. The actual impact therefore depends on various factors, but always also on the individual student.

In order to encapsulate the four levels of assessment and their relationship, Figure 1 depicts the concept of student assessment as I see it: assessment starts with its purpose (the decision), includes the judgement as well as the intended impact of assessment, and finally leads to the actual impact of assessment. Figure 1 also summarises the process of student assessment in a nutshell.

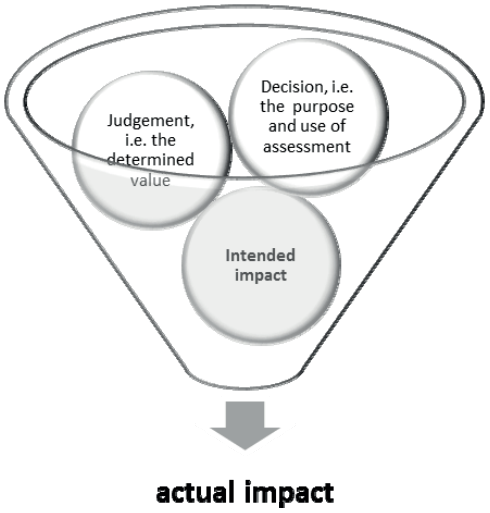

FIGURE 1 Student assessment as a concept and also as a process.

Figure 1 also helps to explain the outline of the present study (see Figure 2). The overall research aim of this study is to find out what *students' experiences of assessment and (dis)empowerment are in their upper secondary EFL studies.*

To do this, Part 1 of the present study and its three articles will focus on the *actual impact* of student assessment. They will concentrate on students' experiences of and reactions to assessment in their EFL studies in one Finnish upper secondary school. Thus, the research question for Part 1 is: *Do students experience assessment in their upper secondary EFL studies as (dis)empowering? What explains potential (dis)empowerment in assessment?*

The second part of the present study, Part 2, will report teaching experiments with some assessment methodology. Consequently, Part 2 will shift the focus to *the intended impact* of student assessment. The intended impact is not only to let students know how well they have reached the learning goals but also to actively attempt to empower students through the chosen assessment methods. Hence, the research question for Part 2 is: *Could some assessment methods foster student empowerment in EFL studies? If yes, how?*

Finally, on the basis of both Part 1 and Part 2, the present study will explore how assessment empowerment and disempowerment manifest themselves.

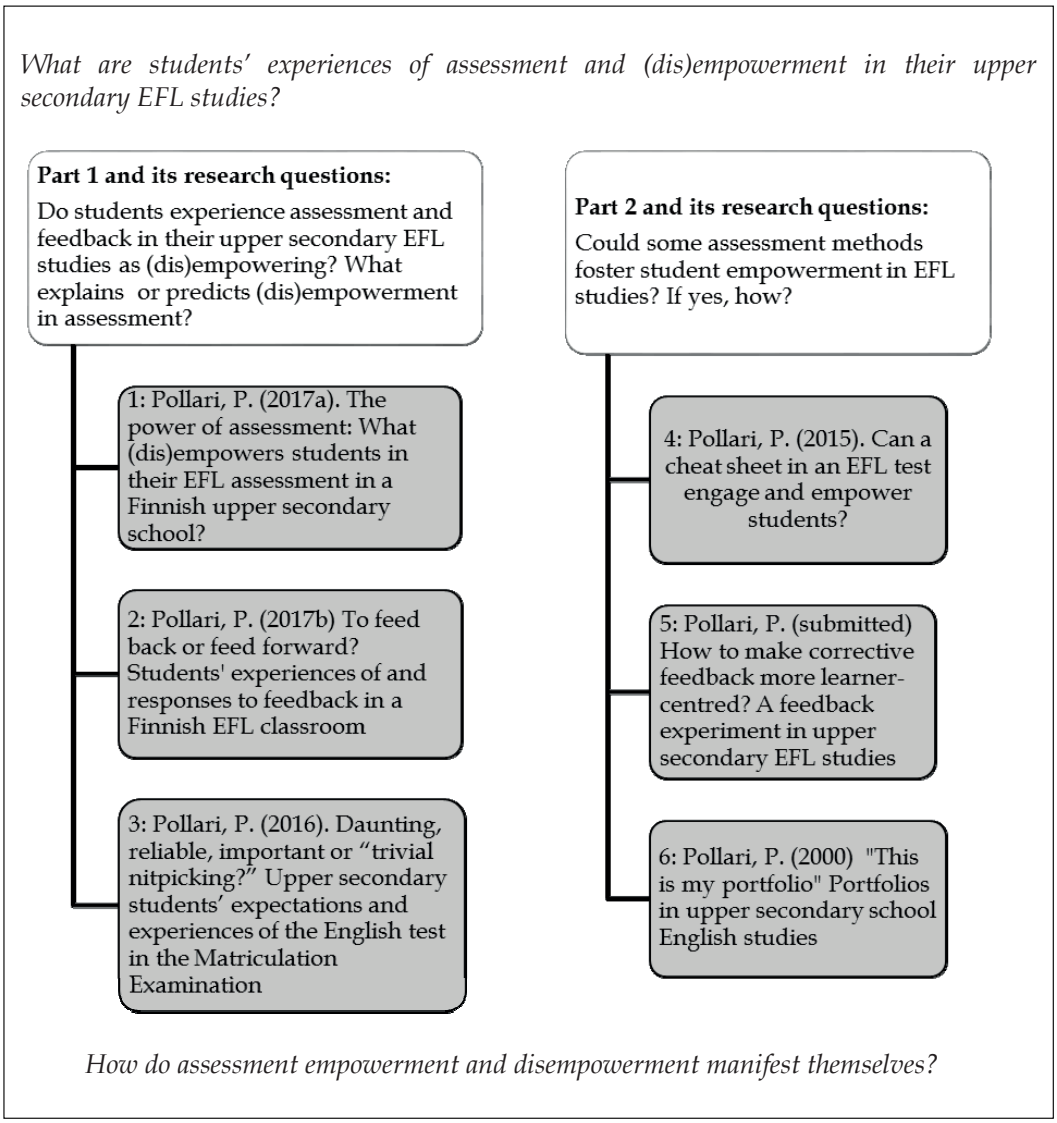

FIGURE 2 The outline of the present study and its sub-studies.

1.2 The structure of the present study

The starting point as well as the raison d'être of this study is pedagogical. Language education, and therefore also assessment in language education, play a significant role in Finnish upper secondary education. Thus, the theoretical framework lies heavily within pedagogy and education. Although I dislike categorising research into different fields or paradigms, I will situate my research at a crossroads. It lies where language, education and assessment meet one another. It is also at the crossroads of theory and practice. It is at school.

I will first discuss the concept of student assessment. The purpose of this more general discussion is to conceptually explain and situate Finnish assessment culture, as it differs quite significantly from the assessment culture of many other countries (e.g. Ouakrim-Soivio, 2013, p. 216; Sahlberg, 2007). However, since assessment is such a huge and complex topic that it is impossible to even touch on its every aspect within the limits of this study, I will try to limit this discussion so that it bears direct relevance to this study, i.e. to student assessment in EFL studies in Finnish upper secondary education.

Next, in Chapter 3, I will define the concepts of empowerment and disempowerment as understood in this study. I will also discuss assessment and (dis)empowerment as well as prior research related to it.

In Chapter 4, I will return to student assessment, but this time look at it in the context of Finnish upper secondary education only. To do so, I will first concentrate on what the Finnish core curricula for upper secondary education say, and have said, about assessment, and then focus on what prior research has found out about assessment in Finnish upper secondary education. Finally, I will present my own evaluative summary of assessment in Finnish upper secondary school studies, and in EFL studies in particular. I will also indicate the gaps in research that my study will attempt to address.

The present study and its sub-studies will be reviewed in Chapter 5. As explained above, Part 1 of the present study will focus on the actual impact of assessment, as experienced and explained by students. The first article seeks to find out whether students find assessment empowering or disempowering. Also, it aims to discover what predicts *assessment (dis)empowerment* and how assessment disempowerment and empowerment manifest themselves. Article 2 builds on the findings of Article 1, and focuses on the actual impact *feedback* has on students by looking at student responses to feedback from the perspective of empowerment. Article 3, also building on the predictors of assessment (dis)empowerment, considers *pressurised and high-stakes tests*, namely the Matriculation Examination and its English test, and its possible link to assessment (dis)empowerment.

Part 2 of the present study will explore whether some less traditional assessment methods could foster student empowerment. Thus, Part 2 and its teaching experiments have a clear pedagogical goal as they aim to both study and develop assessment methodology that could promote students'

empowerment. First, Article 4 will concentrate on experimenting with *cheat-sheet tests* as a vehicle for student engagement and empowerment. Article 5 will explore students' preferred methods of *corrective feedback* on their EFL writing. Sub-study 6 will focus on *portfolios* as a vehicle for comprehensive student empowerment. Even though this teaching experiment took place a long time ago, I have decided to include it in this study for several reasons. First of all, it was a brave and even radical experiment: it was very student-centred and self-directed and something completely new at that time. Thus, it gives a good point of reference for the other two experiments in search of student empowerment that are included in the present study. Secondly, the portfolio project had a strong effect on my own views on assessment. It also introduced teacher-research to me. Without the portfolio project, I do not think the present study would ever have taken place. Therefore, I revisit the portfolio project with a sense of nostalgia but also pride.

In Chapter 6 I will summarise and discuss the findings as well as their practical and theoretical implications. The limitations of the present study as well as suggestions for future research will also be discussed.

2 STUDENT ASSESSMENT

Student assessment is a prominent feature of school life in all educational contexts (e.g. Rust, O'Donovan, & Price, 2005; Taras, 2005). However, it is not easy to find a general but comprehensive definition for student assessment in the literature. Is the concept taken for granted, or divided into several definitions depending on the purpose, scale or method of assessment, for example, or does every educational system give their own definition for student assessment (see e.g. Wiliam, 2011)?

One general definition, however, is provided by the Glossary of Education Reform (http://edglossary.org/assessment/ read 5.12.2015):

> In education, the term assessment refers to the wide variety of methods or tools that educators use to evaluate, measure, and document the academic readiness, learning progress, skill acquisition, or educational needs of students.

In this chapter I will attempt to frame the student assessment landscape. As this landscape is vast, and also varied, I will have to limit my discussion rather severely. Many interesting and important areas will therefore be left undiscussed. For instance, I will not discuss language (proficiency) testing, nor will I discuss the concepts of validity and reliability as such, however important they are in the field of testing and evaluation in general.

Instead, I will approach assessment through concepts or categories that play a central role in the process of assessment. The discussion is structured around Figure 3, which aims to give a visual presentation of the whole process of assessment from its starting point, the purpose, to its actual impact.

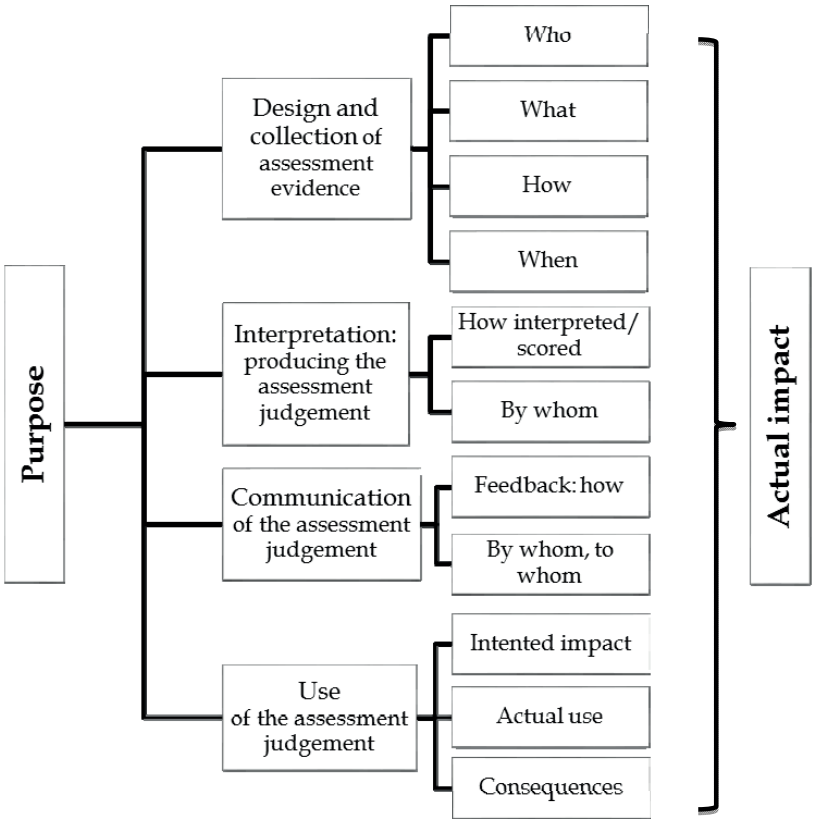

FIGURE 3 Assessment process from its purpose to actual impact.

As the purpose of assessment is the most vital consideration in the assessment process and it should determine all the following steps or decisions (e.g. Bachman & Palmer, 1996; Gipps 1994; Linnakylä & Välijärvi 2005, p. 5, 22; McMillan, 2000, pp. 4-5), the purpose was placed in the most prominent position in this figure.

Another figure, by Pickford and Brown (2006, p. 4), which illustrates five interlocking questions *(What? Why? How? Who? When?)* underpinning assessment design, served as an incentive for the design and collection of assessment evidence in Figure 3. However, designing assessment is only one part of the assessment process. The following definition of assessment and evaluation by Harlen (2007, p. 12) helped me to visualise the assessment process in its entirety:

> Assessment and evaluation both describe a process of collecting and interpreting evidence for some purpose. They both involve decisions about what evidence to use, the collection of that evidence in a systematic and planned way, the interpretation of the evidence to produce a judgment, and the communication and use of that

> judgment. The evidence, of whatever kind, is only ever an indication or sample of a wider range that could be used.

Acknowledging this, Figure 3 attempts to combine the various aspects of assessment from its purpose to the actual impact.

The figure also works as a lens through which I will discuss assessment in this study. First, it serves as a framework for the brief descriptions of some of the key concepts of student assessment. The purpose of the figure as well as of the descriptions is to help the reader to situate Finnish student assessment on the 'map' of student assessment in general. However, the figure is by no means exhaustive, nor is the descriptive list of assessment concepts. Many important concepts had to be omitted as I attempt to concentrate on concepts relevant to Finnish student assessment in upper secondary education.

I will start with the purpose of assessment, i.e. the question *Why*? Then, I will proceed, following the figure, to the design and collection of assessment evidence as well as the interpretation, communication and use of the assessment judgement. However, assessment decisions and judgements are intertwined with one another and do not proceed as neatly in reality as in a theoretical, two-dimensional figure. The following descriptions will therefore merge and overlap at several points.

2.1 Purpose of assessment: Summative and formative

In the assessment and evaluation literature, the distinction between summative and formative assessment/evaluation has been prominent for the past 50 years. While formative assessment refers to assessment whose purpose is to support learning, teaching and studying, summative assessment means assessment the purpose of which is to measure and report learning outcomes.

The origins of the distinction between summative and formative assessment have been attributed to Scriven (1967), to Bloom (1969) and to Bloom, Hastings and Madaus (1971) (see e.g. Bennett 2011, Gardner 2012a; Leahy & Wiliam 2012; Newton 2007). While Scriven was mainly discussing programme evaluation and its different approaches (Bennett 2011, Newton 2007, Scriven 1991), Bloom, Hastings and Madaus (1971), in their *Handbook of formative and summative evaluation of student learning,* identified three characteristics according to which formative and summative student assessment could be distinguished. These characteristics were the purpose, i.e. the expected uses of the assessment outcomes, the timing, and the level of generalisation, which refers to the scope and generalisability or transferability of the skills assessed:

> We have chosen the term 'summative evaluation' to indicate the type of evaluation used at the end of a term, course, or program for purposes of grading, certification, evaluation of progress, or research on the effectiveness of a curriculum, course of study, or educational plan. … Perhaps the essential characteristic of summative

> evaluation is that a judgment is made about the student, teacher, or curriculum with regard to the effectiveness of learning or instruction, after the learning or instruction has taken place. - - Formative evaluation is for us the use of systematic evaluation in the process of curriculum construction, teaching, and learning for the purpose of improving any of these three processes. (Bloom et al., 1971, p. 117)

> The main purpose of formative observation (there are other useful ways of observing behavior besides testing) is to determine the degree of mastery of a given learning task and to pinpoint the part of the task not mastered. Perhaps a negative description will make it even clearer. The purpose is not to grade or certify the learner; it is to help both the learner and the teacher focus upon the particular learning necessary for movement towards mastery. On the other hand, summative evaluation is directed toward a much more general assessment of the degree to which the larger outcomes have been attained over the entire course or some substantial part of it. (Bloom et al., 1971, p. 61)

While earlier scholars highlighted the timing and the generalisation when making the distinction between summative and formative assessment (Gardner, 2012a; Newton, 2007), most scholars now agree that the purpose, use or function of the assessment is the defining factor (see e.g. Bachman & Palmer, 1996, 2010; Lloyd-Jones, 1986, p. 2; Newton, 2007; Sadler, 1989, 1998). In a nutshell, *summative assessment could be characterised as assessment that is used to summarise and report learning outcomes* (that have taken place), *formative to form, help and guide learning as well as teaching* (that is taking place). To illustrate the two assessments, Stake (2004, p. 21) gives an everyday analogy: "When the chef tastes the soup, it's formative evaluation, and when the guest tastes the soup, it's summative evaluation."

Accordingly, formative assessment has one, clear function: its purpose is to improve learning. Over the past two or three decades several studies have concluded that formative assessment is, indeed, a powerful tool for enhancing learning (see e.g. Leahy & Wiliam, 2012). The seminal work in the field, the meta-analysis by Black and Wiliam (1998a, 1998b), indicated that formative assessment improved learning significantly. In particular, the meta-analysis indicated that formative assessment improved the learning of "low achievers more than other students – and so reduces the range of achievement while raising achievement overall" (Black & Wiliam, 1998b, p. 141).

Summative assessment, on the other hand, has no single purpose as such but rather a cluster of various purposes. In the school context, probably the most common purpose, or use, of summative assessment is grading. Often grading is also the only tangible feedback given on learning. Summative assessments and their grades may also serve the purpose of reporting and giving information on students' learning to their parents. Sometimes summative assessment is used to gauge and also compare the student's attainment with a larger student population. Summative assessment can also serve a selective purpose when examination results or final grades are used for entry to further education (e.g. Huhta & Hildén, 2016; Newton, 2007). However, when the examination results are used for some other purpose, such as ranking or labelling schools or teachers, or selecting in which school to enrol one's

children, the use, or impact, is not what was intended when the assessment was designed (e.g. Jones, Jones & Hargrove, 2003; Stobart, 2008).

Sadler (1989), for instance, discusses the different impact of these two forms of assessment. While considering the effect of formative assessment on learning to be potentially powerful, Sadler (1989, p. 120) sees that summative assessment is "essentially passive and does not normally have immediate impact on learning, although it often influences decisions which may have profound educational and personal consequences for the student".

Some scholars also mention *diagnostic* assessment in conjunction with formative and summative assessment. Once again, the definitions vary. For some scholars, its main purpose is to assess learners' entry performances in order to support the planning of the instruction (see e.g. Linnakylä & Välijärvi, 2005; Takala, 1997), or learners' placement in different programmes or groups, and thus takes place either before or at the very beginning of instruction; others see its purpose as diagnosing learning difficulties and identifying their causes during instruction (e.g. Miller, Linn, & Gronlund, 2013, pp. 55-63). Lately, the concept has been quite prominent in second or foreign language assessment literature (see e.g. Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja, 2015; Harding, Alderson, & Brunfaut, 2015; Huhta, 2008; Jang & Wagner, 2013; Lee, 2015). In that context, the concept has two primary goals: "to identify language learners' weaknesses and deficiencies, as well as their strengths, in the targeted language domains and provide useful diagnostic feedback and guidance for remedial learning and instruction" (Lee, 2015, p. 295). Thus, the purposes of diagnostic language assessment are very close to the purpose of formative assessment in general.

Table 1, by Linnakylä and Välijärvi (2005, p. 26), summarises the concepts and their core characteristics in education:

TABLE 1 Assessment in the education and learning process (Linnakylä & Välijärvi, 2005, p. 26).

	Diagnostic	*Formative*	*Summative*
Purpose	To canvass and strengthen the prerequisites for learning and education, to support planning	To form, motivate and guide learning, teaching and education	To collect data on learning and educational outcomes; to evaluate and grade, to predict further learning
Timing	At the beginning of education or study period, or when problems emerge	During education or studies, monitoring and supporting progress	At the end of education or study period
Criteria	Criterion- or norm-referenced	Criterion-referenced or the learner's prior performance level	Criterion-referenced or comparison with other learners

(continues)

TABLE 1 (continues)

Methods	Tests, tests created by the teacher, questioning, self-assessment, discussions, observation	Questioning by the teacher, observation, homework, tests, learning logs, portfolios, self-assessment, assessment discussions	Tests, demonstrations, exams, comparative national and international evaluations, final examinations
Feedback	For the education planner, teacher	For the educator and teacher and, above all, for pupils/students themselves	For external decision-makers, for the providers of further education; for the pupil/student, teacher and school

In real life – and in academic discussion as well – the distinction is not necessarily very clear cut, and the "complex relationship between formative and summative assessment has been an ongoing concern" (Baird, Hopfenbeck, Newton, Stobart, & Steen-Utheim, 2014, p. 44). While some scholars have wanted to keep the terms separate and distinct from each other, the consensus now seems to be that they are interrelated (e.g. Baird et al., 2014; Bennett, 2011; Biggs, 1998; Black, Harrison, Lee, Marshall, & Wiliam, 2003; Harlen, 2007; Harlen & James, 1997; Taras, 2005). This possible overlap and interrelation has, however, caused concern, too. Harlen and James (1997), for instance, were concerned that if the terms, and also the purposes, of formative and summative assessment are blurred and conflated, there might be very little genuinely formative assessment, or when trying to meet both these purposes, there would be "assessment overflow" (Harlen & James 1997, p. 365). Hence, they called for a way of linking these together that would still recognise and preserve their different functions and characteristics. Instead of a clear dichotomy, Harlen (2012a) proposes a *dimension* of formative/summative assessment purposes and practices. This continuum might do away with some conceptual problems of the terms. Stake (2004, p. 21), although agreeing, offers a word of warning, and also a piece of advice: "Formative and summative evaluation can happen together, but the roles of formatively looking forward and summatively looking back are worth keeping separate."

Finally, some scholars, perhaps Newton most clearly, challenge the whole distinction and argue that it is not conceptually valid:

> Since the earliest days, most commentators have assumed that there *is* a meaningful distinction to be drawn between summative and formative. At least to my mind, though, no one has yet managed to nail a definition. I believe that there is a simple reason for this: the term 'summative' can only meaningfully characterize a type of assessment judgement (i.e. it operates at the judgement level of discourse), while the term 'formative' can only meaningfully characterize a type of use to which assessment judgements are put (i.e. it operates at the decision level of discourse). The terms belong to qualitatively different categories; to attempt to identify

> characteristics that distinguish them—within a single category—is to make a category error. (Newton, 2007, pp. 155-156, emphasis in the original).

Newton (2007) goes on to present 18 different categories of assessment purposes or uses. Of these purposes, only one is formative (results are used to identify the students' learning needs in order to direct subsequent teaching and learning) and the others range from social evaluation to placement, and from selection to resource allocation and national accounting (Newton, 2007). Most scholars and educators, however, would probably label the majority of these purposes or uses as summative and some of them also as diagnostic.

Perhaps because the terms formative and summative assessment have become conceptually somewhat muddled over time, the terms *assessment for learning (AfL)* and *assessment of learning (AoL)* have recently become common (Leahy & Wiliam, 2012) and are often used instead of, or interchangeably with, the terms formative and summative assessment.

Nevertheless, although many authors do use the terms formative assessment and assessment for learning interchangeably (see e.g. Baird et al., 2014, pp. 39-40), some make a slight difference between the two (see e.g. Atjonen, 2014). For instance, when discussing and explaining assessment for learning, Black et al. (2003, p. 2) write as follows: "The focus here is on any assessment for which the first priority is to serve the purpose of promoting students' learning." "Such assessment becomes *formative assessment* when the evidence is used to adapt the teaching work to meet learning needs", Black et al. (2003, p. 2, emphasis original) go on to define the distinction between assessment for learning and formative assessment.

Some scholars have also introduced the concept of *assessment as learning (AaL)* (see e.g. Dann, 2002, 2014; Earl, 2003). Assessment as learning sees assessment as a learning activity and opportunity and, therefore, "assessment and learning become inextricably interlinked, so that their processes serve each other" (Dann, 2014, p. 164). However, several scholars, also those advocating assessment as learning, see it as an aspect of formative assessment or assessment for learning (see e.g. Dann, 2014, p. 149).

2.2 Design and collection of assessment evidence

Once the purpose of assessment is determined, the main questions of the design of assessment are *who* designs and controls the assessment methodology and process, *what* evidence is collected, as well as *how* and *when* that evidence is collected (see e.g. Pickford & Brown, 2006, p. 4). Thus, this section will start with *who* controls the assessment, with specific reference to internal and external assessment. After that, I will concentrate on *how* that evidence is collected. Does it involve small-scale or large-scale assessment? Is the evidence collected through traditional tests or more alternative assessment methods such as performance assessment? If tests are used to collect assessment evidence, are

they constrained or non-constrained tests? Although *what* evidence to collect and *when* to do it are important questions in the assessment process, they are not now discussed because of the focus and scope of this study.

2.2.1 External and internal assessment

The locus of control, in other words, *who* controls and designs the assessment, is one of the defining questions in assessment (e.g. Pickford & Brown, 2006, p. 4). Internal assessment is usually understood as assessment set and marked by the school – most often the teachers themselves in their own classrooms – whereas external assessment is designed and controlled by an external organisation or agency outside of the school (e.g. Bray, 1986; Harlen, 2007).

In many countries, internal assessments are mainly used to make decisions about instruction and are calibrated for the needs of a specific group of students (i.e. the class the teacher is teaching). Internal assessment is tied not only to the relevant curriculum but also to the instructional routines in the classroom. Accordingly, internal assessments are also called classroom(-based) (e.g. Hill, 2012; Popham, 2008; Rea-Dickins, 2007; Stoynoff, 2012) or teacher(-based) assessment (e.g. Davison & Leung, 2009; Gardner, 2012b) as it is the teachers who predominantly are the "agents of assessment" (Rea-Dickins, 2004, p. 249). Internal assessments may be quite frequent and they may entail, for instance, the teacher's observations in class or question and answer sessions, but they can also include more structured assessments, such as various kinds of tests, quizzes, journal writing, projects, presentations and reports (e.g. Marzano, 2010). Although the most common use of internal assessment is probably instructional and, thus, formative, internal assessment can also be used for summative purposes such as reporting students' progress to parents and it can result in marks and grades (Harlen, 2007; Popham, 2008). The common denominator is that "they are all under the control of the teacher and embedded in the curriculum" (Paris, Paris, & Carpenter, 2002, p. 142).

In contrast, external assessments are not under the control of the teacher. They are devised, controlled and often also marked by an external body, be it a commercial publisher, agency or organisation, educational administrators or national policymakers. External assessments do not occur as often as internal assessments but they "usually have greater importance, authority and stakes attached to them" (Paris et al., 2002, p. 142).

Yet again, the reality is not so clear-cut. Studies have shown that internal assessments are often influenced by external assessments. For instance, teachers can use external assessments as part of or as a model for their own assessments (Harlen, 2004, 2005; James & Lewis, 2012) or they can select tests or assessment tasks from external task pools or banks (Harlen, 2007). Also, some assessments may be externally designed and controlled but the actual student work is marked and graded by the students' own teachers, albeit on the basis of externally set criteria (see e.g. Bray, 1986; Marshall, 2011).

Students' own self-assessment or peer assessment could also be counted as internal assessment (Bray, 1986). However, as the terms internal and external

assessment are mostly used when discussing who controls the design of assessment procedures, students' self and peer assessment are discussed later, in the section dealing with the interpretation of assessment evidence.

2.2.2 Large-scale and small-scale assessment

Large-scale assessment or tests are forms of external assessment administered to large numbers of students for a variety of reasons (de Lange, 2007, p. 1114). Typically, but not necessarily, they are high-stakes tests intended to measure individual achievement (e.g. de Lange, 2007; Popham, 2001, p. 34). However, international surveys and tests that aim to evaluate programmes by measuring student achievement or attitude in various countries, such as PISA *(Programme for International Student Assessment),* TIMSS *(Trends in International Mathematics and Science Study)* and ICCS *(International Civic and Citizenship Education Study)*, do not – or at least should not – have direct high-stakes consequences for the participating students, teachers or schools (Volante, 2006). Nonetheless, they may have consequences for different educational programmes, curricula or participating countries. For instance, Finland's PISA success has won the Finnish school system international renown and has brought a great number of educators to see our schools (e.g. Rinne, Simola, Mäkinen-Streng, Silmäri-Salo, & Varjo, 2011, p. 35; Sahlberg, 2007). On the other hand, now that the results of PISA and TIMSS are widely seen "as criteria of good educational performance, reading, mathematical, and scientific literacy have now become the main determinants of perceived success or failure of pupils, teachers, schools, and entire education systems" (Sahlberg, 2007, pp. 177-178). In some countries, this has led to changes in curricula that shift teaching time to subjects and skills tested in PISA and TIMSS (Sahlberg, 2007). International surveys and their results have prompted or accelerated also other policy changes as well as public debate on education in several countries (e.g. Baird et al., 2011; Breakspear, 2012; Grek, 2009).

Small-scale assessments, in contrast, are assessments designed for smaller numbers of students, possibly even for just one student. They are typically internal assessments, designed and administered by teachers or possibly by students themselves.

2.2.3 Traditional or alternative assessment

Much of the assessment and evaluation literature has come from the United States (e.g. Takala, 1996). Therefore, their student assessment traditions, which have been strongly based on psychometrics and measuring distinct, decontextualised constructs, have dominated the field (Takala, 1996). As a consequence, standardised, large-scale paper-and-pencil tests, consisting mainly of select-answer items such as multiple-choice, have come to be considered traditional testing or assessment (see e.g. Gipps, 1994; Torrance, 1996).

The late 1980s and the 1990s witnessed a surge of alternative assessment methodology as a result of "growing dissatisfaction with traditional, multiple-choice forms of testing" (Herman, Aschbacher, & Winters, 1992, p. 1). First of all, traditional testing was considered too narrow and unable to capture the whole story of student learning and performance (e.g. Kohonen 1997, 1999; Shepard, 1989). Traditional assessments take place after learning and thus ignore the learning process altogether (Valencia 1990). They are also limited to a given time and place and the test-takers have only that opportunity to demonstrate all their knowledge and skills (e.g. Hamp-Lyons, 1996; Kohonen, 1997, 1999). Traditional testing, high-stakes testing in particular, has also been criticised for narrowing the curricula (see e.g. Shepard, 1989). Furthermore, multiple-choice test items often focus on measuring fragmentary, decontextualised skills and knowledge; such tests may not only distort learning outcomes but also turn learning processes into superficial rote learning (e.g. Shepard, 1989; Välijärvi, 1996). Memorising and memory retention leave little room for high-order learning, such as problem-solving, the integration of different skills and knowledge, critical thinking and creativity (e.g. Gipps, 1994; Harlen, 2012b).

Alternative assessment has been called by various names (see Kohonen, 1997, p. 13). However, more commonly, alternative assessment is also called performance(-based) assessment (e.g. Broadfoot, 1996a; Linn, 1994; Norris et al., 1998) or authentic assessment (e.g. Darling-Hammond, 1994; Darling-Hammond, Ancess, & Falk, 1995; Kohonen, 1997, 1999; O'Malley & Pierce, 1996; Torrance, 1996; Valencia, Hiebert, & Afflerbach, 1994; Wiggins, 1989, 1998). Many authors have also used these three terms "synonymously to mean variants of performance assessments that require students to generate rather than choose a response" (Herman et al., 1992, p. 2). For some scholars and teachers, a longer written answer to a question or an essay have qualified as performance assessment. For most, portfolios, presentations and research experiments, for example, meet the standards of alternative or authentic assessment:

> Rather than taking multiple choice tests in which students react to ideas or identify facts, these students engage in science experiments, conduct social science research, write essays and papers, read and interpret literature, and solve mathematical problems in real-world contexts. (Darling-Hammond et al., 1995, p. 2).

Establishing authenticity in the context and nature of the task does not necessarily require real-life tasks: "Assessment is authentic when we directly examine student performance on worthy intellectual tasks" (Wiggins, 1990, p. 2).

Traditional assessment has also focused on students' solo performance, i.e. what they can achieve alone, without guidance or scaffolding from a teacher or peer. Hence, group tests, where a small group of students take the same test together, in co-operation (either the whole test or parts of it) or pair/group assessment tasks (e.g. a pair dialogue or presentation or a co-written paper) can be considered alternative assessment methods.

Traditional assessment has been rather static in many ways: because of the grading traditions or criteria characteristics, for instance, the results of the test have remained unchanged even if the student has subsequently acquired more knowledge or skills in the area. Therefore more recent assessment developments such as *dynamic assessment* can also be regarded as alternative assessment methods. (For further information on dynamic assessment, see e.g. Oksanen, 2001, or *dynamic language assessment,* see e.g. Lantolf & Poehner, 2011; Leontjev, 2014, 2016; Poehner, 2007, 2008.)

2.2.4 Constrained or non-constrained assessment

Traditionally, taking a test or sitting an examination has taken place under tightly restricted conditions. The test time, its duration and place have been mandated and controlled and the testing situation rigorously invigilated (aka proctored). Testing aids or co-operation between test-takers have been regarded as cheating and have been strictly prohibited.

As mentioned above, a typical test has been *a closed-book test* where no books or testing aids are allowed. As an exception, mathematical tables, calculators or some dictionaries of ancient languages may sometimes be allowed in the testing situation itself. In this respect, its opposite is *an open-book test* where students can bring and consult their (course) books or other reference materials (Race, Brown, & Smith, 2005, p. 40). One example of an open-book test could be writing an essay in a foreign language with the help of a dictionary and a grammar book (Currie, 1986, pp. 125-126). Open-book tests may also be *open-web tests*, where students can consult the internet in addition to – or instead of – their books (Myyry & Joutsenvirta, 2015; Williams & Wong, 2009).

An open-notes test, where students are allowed to bring their course or lecture notes but no books, is a variation between closed-book and open-book tests. Another variation is *a cheat-sheet test*. It is a test where students are not only allowed but encouraged to bring some notes which are particularly constructed for the test (e.g. Erbe, 2007; Larwin, Gorman, & Larwin, 2013; Whitworth, 1990). This legitimate cheat sheet – aka crib notes – is often limited in size and also possibly in content and format (Larwin, 2012). These tests that allow memory aids are still normally taken by all students in the same place, at the same time and under teacher supervision (e.g. Race et al., 2005, pp. 40-43).

Computer-assisted assessment, also known as e-testing, e-assessment, distance or on-line testing, may give students some other aspects of freedom or agency as well (see e.g. Garrett, 2009; Myers, 2002; Van Maele, Baten, Beaven, & Rajagopal, 2013). For instance, many university e-exams can be taken at a time of the student's choice (e.g. Stowell & Bennett, 2010). Some of these computer-assisted tests are taken in a controlled environment, may be timed and allow no memory aids. Some may also be very high-stakes, and differ from traditional testing only in the format of taking the test with a computer, not paper and a pencil (see e.g. Dermo, 2009; Kalz & Ras, 2014). One such example in Finland is the ongoing transition from a paper-and-pen to a digital examination in the Matriculation Examination. Some e-assessments may even be virtually

invigilated (see e.g. Clarke, Dowland, & Furnell, 2013). On the other hand, some computer-assisted tests have no or few constraints: they can be sat at the time, place and also pace the students feel best themselves and co-operation as well as consulting materials are not prohibited (see e.g. Williams & Wong, 2009). DIALANG, a diagnostic foreign language on-line test is an example of one such test (Alderson, 2005; Alderson & Huhta, 2005). Today, most distance tests appear to be situated somewhere between these two extremes. Paper-and-pencil distance tests, or 'take-away tests' (Currie, 1986, pp. 126-127), still exist as well.

Probably the most common concern associated with *non-invigilated* or memory-aided tests is whether they are a reliable and also valid manner of testing students' knowledge and skills (e.g. Hollister, 2007). Views are divided. While many researchers as well as teachers and students trust their reliability and validity, some see non-invigilated, online tests as too open to dishonesty and to cheating and plagiarism (see e.g. Dermo, 2009; Hollister, 2007; Stowell & Bennett, 2010; Williams & Wong, 2009). Another concern associated with less constrained assessments is whether they enhance or decrease students' learning. Several scholars have concluded that less constrained and/or memory-aided testing methods have diminished students' test anxiety, enhanced their self-efficacy and improved and deepened their learning (e.g. Erbe, 2007; Gharib, Phillips, & Mathew, 2012; Larwin et al., 2013; Myyry & Joutsenvirta, 2015; Williams & Wong, 2009) but not everyone is convinced that these assessment methods benefit learning (e.g. Dickson & Miller 2005; Dickson & Bauer, 2008; Funk & Dickson, 2011).

2.3 Interpretation of the assessment evidence: Producing the assessment judgement

Designing the assessment methodology and collecting the assessment evidence are just part of the assessment process. The evidence gathered must be interpreted. The interpretation means "making a judgement about the quality of what is gathered", in other words, judging its worth and value (McMillan, 2000, p. 10). This interpretation "of what the results mean and how they can be used" (McMillan, 2000, p. 10) involves several phases. Although the two-dimensional figure (Figure 3) may seem to suggest that the whole interpretation stage follows the design and collection of assessment evidence, this is not meant to be the case. For instance, defining criteria and scoring guidelines are intrinsically linked to learning goals, and thus also assessment goals, so this phase should take place simultaneously with, or even precede, the designing of the assessment methodology and tasks (e.g. Currie, 1986). Chronologically, the phases following the gathering of assessment evidence include marking and scoring according to the criteria and turning the score into a grade.

This section concentrates on only two of the many aspects or decisions: *how* the evidence is scored and then the scores interpreted, and *by whom*.

2.3.1 Norm-referenced and criterion-referenced assessment

Assessments can be categorised as norm- or criterion-referenced assessment on the basis of how student performances are scored and how these scores are interpreted into marks and grades, (e.g. Bond, 1996; Heubert & Hauser, 1999). Their intended purposes are different and therefore content selection and, in particular, the ways in which the results are interpreted also differ (see e.g. Brown & Hudson, 2002, pp. 9-14).

The main function of a norm-referenced assessment is to rank students and their performances (Bond, 1996; Bray, 1986; Johnson, 1986; Notar, Herring, & Restauri, 2008; Yorke, 2007, p. 17). In a nutshell, norm-referencing means that a student's performance is compared with the performance of other students (Bond, 1996; Bray, 1986; Notar et al., 2008; Popham, 2008, pp. 122-123). In the case of standardised, large-scale achievement tests, individual performances are usually measured against a norm group (e.g. Bray, 1986; Johnson, 1986). The norm group means a representative group of students, sometimes a national sample, who were given the test prior to its use (Bond, 1996; Kubiszyn & Borich, 2013, p. 12). Subsequent test performances are then measured against the results of the norm group: if an examinee scores at the 86th percentile, it means that his or her performance "exceeded the performance of 85 percent of the test-takers in the *norm-group*" (Popham, 2008, p. 122, emphasis in the original).

Norm-referencing a test with a norm group is elaborate and expensive (Bond, 1996) and therefore cannot be done with all smaller tests or assessments. Another way of norm-referencing test results is to use the normal distribution of scores as the basis of assigning grades after the test has been taken (e.g. Brown & Hudson, 2002, p. 8), as has been the case in many school-leaving examinations such as the Matriculation Examination (see e.g. Mehtäläinen & Välijärvi, 2013; Juurakko-Paavola & Takala, 2013), although that is changing slightly with the Matriculation Examination.

In the absence of representative and sampled norm groups or large enough student populations, students' performance may still be compared with the performance of other students when interpreting the raw scores of the assessment into grades or marks. Sometimes student performance is compared with the performance of other students in the same course, or it may be measured against the Bell curve, for instance. This 'grading on the curve' may lead to unfair marking and grading as then a student's grade actually depends not only on their own skills but also on the skills and knowledge of their peers (Marzano, 2010, p. 17; Yorke, 2007, p. 17). Simply put, an average student will get worse grades in a group of excellent students and in a group of weaker students the student's grades will be better (see e.g. Ouakrim-Soivio, 2013, pp. 213-220).

Criterion-referenced assessments do not compare students' performances against one another but against predetermined performance levels, i.e.

standards, criteria or rubrics (Bond, 1996; Brown & Hudson, 2002; Kubiszyn & Borich, 2013, p. 12; Shrock & Coscarelli, 2010). The purpose is therefore not to rank students but to see how well they have learnt what they were supposed to learn (Brown & Hudson, 2002; Notar et al., 2008; Yorke, 2007, p. 18), and every student is marked and graded on the basis of their own merits, not on the basis of those of their peers (e.g. Johnson, 1986).

Although this may sound much fairer for individual students, criterion-referencing is not without problems. Despite using the same criteria, different teachers or markers may assess the performances differently (see e.g. Sadler, 2013). Variation may be caused by assessors understanding, interpreting, weighing or valuing the criteria differently. This, in turn, may result from the criteria themselves: the criteria may contain diverse sub-criteria (Bloxham, den-Outer, Hudson, & Price, 2016) or the language used may not be explicit enough (Sadler, 2013). Assessors may also disagree with the criteria and thus ignore them or adapt them to better suit their own expectations or preferences (Bloxham et al., 2016). Moreover, criterion-referenced testing is sometimes claimed to lead to grade inflation since the grade distribution is not predetermined or controlled (see e.g. Yorke, 2007, pp. 105-133).

Although norm- and criterion-referenced assessment may seem paradigmatically very different, several authors consider them to be a continuum and say that in assessment reality they coexist (Lok, McNaught, & Young, 2016; Miller et al., 2013, pp. 57-64; Tuokko, 2007, p. 116). In addition, the criteria and standards adopted may be implicitly norm-referenced (Lok et al., 2016; Yorke, 2007, p. 19; see also Brown & Hudson, 2002, pp. 13-14).

In second and foreign language teaching and testing, some authors use the terms *outcomes-based* (e.g. Brindley, 2001) and *standards-based* (e.g. Llosa, 2007, 2011) assessment. They are both closely linked with criterion-referenced assessment as they all compare the learner's performance against standards, criteria or benchmarks.

2.3.2 Assessed by whom?

Another question dealing with the locus of power and control is who assesses or marks students' work: is it an *external* assessor, the teacher or the students themselves? Assessing students' work, including marking tests and assigning grades, is probably most often carried out by the *teacher*. However, with large-scale or high-stakes examinations, teachers do not have the power to decide on the assessment criteria, even if they may do the preliminary marking (see e.g. Bray, 1986; Marshall, 2011, pp. 20-29). Power over the criteria as well as the final marking rests with external evaluators. There are exceptions to this, though. For instance, in large-scale national exams in Sweden, teachers assess and grade their own students (Gustafsson & Erickson, 2013).

Sometimes students themselves may have a role in assessing their work. According to one definition, *self-assessment* means that "students use criteria and apply standards to judge their own work" (De Grez, Valcke, & Roozen, 2012, p. 130; for several other definitions, see e.g. Noonan & Duncan, 2005).

According to this definition, students do not design the assessment themselves, or its criteria, but only apply given criteria to given pieces of work.

Self-assessments are mostly used for formative purposes, in other words, to enhance students' own learning and studying (e.g. Dochy, Segers, & Sluijsmans, 1999; Noonan & Duncan, 2005). Indeed, self-assessment is rather unanimously believed to be a necessary skill for effective learning: "Learning can only be effectively undertaken when the learner monitors what is known, what remains to be known and what is needed to bridge the gap between the two" (Boud, 1995, p. 13). Self-assessment is also a cornerstone of formative assessment (Black et al., 2003; Black & Wiliam, 2012). Self-assessment skills are also closely linked to self-directed learning and learner autonomy and are considered necessary for life-long learning (e.g. Boud, 1995, p. 14; Earl, 2003).

Nonetheless, self-assessment may be used summatively as well. For summative purposes such as grading, self-assessment is probably more widely used, or at least reported, in higher education, although some studies have dealt with summative peer and self-assessment or grading in basic or secondary education as well (e.g. Sadler & Good, 2006). When grading is involved, both teachers and students may question the reliability of self-assessment, also in higher education. In a study by Rodríguez-Gómez, Ibarra-Sáiz, Gallego-Noc, Gómez-Ruiz and Quesada-Serra (2012), both students and teaching staff suspected that self- or peer assessments carried out by students were subjective and biased because students did not "have sufficient mastery of the subject to carry out objective evaluations" (Rodríguez-Gómez et al., 2012, p. 12). Indeed, a wealth of studies in various fields show that self-assessment and its accuracy may depend on variables such as age, gender, race and academic achievement level (see e.g. Blatchford, 1997; Boud, Lawson, & Thompson, 2013; Chevalier, Gibbons, Thorpe, Snell, & Hoskins, 2009; Dunning, Heath, & Suls, 2004; Lew, Alwis, & Schmidt, 2010). Global judgements made on the basis of well-understood criteria (Falchikov & Goldfinch, 2000) as well as experience in self-assessment (Boud et al., 2013; Sadler & Good, 2006) have been indicated to improve accuracy (cf. Lew et al., 2010). Thus, self-assessment is usually regarded as a skill that needs to be explicitly practised and fostered in order to develop (see also Dochy et al., 1999; Rodríguez-Gómez et al., 2012). Nevertheless, some scholars remain rather sceptical about the accuracy of self-assessments (e.g. Eva & Regehr, 2008) and call for external assessment, standards and feedback to scaffold self-assessment (see e.g. Sargeant, 2008).

Despite various definitions (Noonan & Duncan, 2005), in an educational context, *peer assessment* usually means assessment carried out by students of the same 'status', i.e. students who are in the same course, class or group. Peer assessment is usually used as a supplementary assessment procedure, whose main function is to give additional, formative feedback and it does not substitute or overrule teacher assessment (Noonan & Duncan, 2005). Peer assessment, like self-assessment, is considered an integral part of assessment for learning (Black et al., 2003; Noonan & Duncan, 2005). Moreover, like self-assessment, peer assessment skills need to be developed. When the student role

changes from the traditional object of assessment to the assessor, students may have concerns about their ability to carry out assessment (Mok, 2011) or perhaps their peer-assessors' ability to do so (see e.g. Zhao, 2014). Some studies have, however, indicated reasonably high correlations between peer assessments and teacher assessments (see e.g. Matsuno, 2009; see also Sadler & Good, 2006). Many scholars believe that peer assessment can improve the learning of both the assessor and assessee: if "[o]rganized, delivered, and monitored with care, it can yield gains in the cognitive, social, affective, transferable skill, and systemic domains" (Topping, 1998, p. 269). However, a study by Sadler and Good (2006) with middle-school students, which saw significant improvement in learning when students were self-grading their test papers, did not find any learning gain when students were peer-grading test papers.

Co-assessment, i.e. a combination of self- and/or peer assessment with assessment carried out by the teacher, is a step closer to traditional assessment but still allows students an active role in assessment (Dochy et al., 1999). In their review of several studies, Dochy et al. (1999, p. 344) conclude that the combination of self-, peer and co-assessment has been found to be effective both for summative and formative purposes as it "makes tutors and students work together in a constructive way and as a result they come to higher levels of understanding by negotiation".

2.4 Communication of the assessment judgement

Price, Handley, Millar and O'Donovan (2010, p. 277) maintain that assessment feedback "is arguably the most important part of the assessment process". Feedback is therefore at the heart of the communication of the assessment judgement. First, *how* does the student get feedback from the assessment? Is the feedback in the form of a score or grade alone, or does the student get more detailed verbal feedback? Does the feedback feed *back*, or *forward*? Finally, the questions *to whom* and *by whom* this feedback is communicated are central in the communication of the assessment judgement.

2.4.1 Feedback: Grades and/or verbal feedback

Gardner (2012b, p. 109) calls giving marks and grades "the leitmotiv of summative assessment". Indeed, grading, whether expressed in letters, numbers, percentages or Latin words, is a phenomenon characteristic of education all over the world (e.g. Välijärvi, 1996, Wiggins, 2012). It is probably also the most prominent feature of assessment: Marzano (2010, p. 15) claims that at "the classroom level, any discussion of assessment ultimately ends up in a discussion of grading". Furthermore, grades, marks or percentages may often be the only feedback students get from assessments.

Nevertheless, many scholars underscore the problematic and complex nature of grading. One problem is the 'overall' nature of an "omnibus grade", as Marzano calls it (2010, p. 15). Harlen (2007, p. 27) elaborates on this overall nature as follows:

> Numerical scores from tests are a summation over a diverse set of questions and so have little meaning for what students actually know or can do for the same total can be made up in many ways. Scores also give a spurious impression of precision, which is very far from being the case.

In other words, the overall grade does not give any explicit information on what the student can or cannot do (see also Atjonen, 2014). However, as Harlen (2007) above points out, grades and marks are considered precise and objective indicators of learning, particularly by the general public (see also Gardner, 2012b) but in fact grades are far from being objective, precise or error-free. For instance, as teachers often design their own grading systems and philosophies for internal assessments, their grading practices and criteria may vary significantly (Marzano, 2010, pp. 15-19; McMillan, 2003), even within one school and school subject (see Guskey & Bailey, 2001, p. 1). External, large-scale assessments and their grades are not error free, either. Besides, grades are often used to serve several, even conflicting, purposes at the same time, such as ranking students, reporting results, providing feedback and motivating students (Brookhart, 2004, p. 23).

Grades have also been claimed to shift the students' focus from the learning to the 'self', i.e. the learners themselves (see e.g. Atjonen, 2014; Butler, 1987, 1988; Stobart, 2012), as grades are "interpreted in comparison to others", rather than as information on students' own learning (Stobart 2012, p. 240). Similarly, feedback comments given in addition to a grade or score may go unnoticed as students shift their attention from the learning task to the grade (Atjonen, 2014; Black et al., 2003, pp. 42-49; Butler, 1987; cf. Dlaska & Krekeler, 2013).

Feedback, on the other hand, can have a great influence on learning (e.g. Hattie, 2009, 2012; Hattie & Timperley, 2007; Wiggins, 2012). Hattie's syntheses (2009, 2012) of more than 800 meta-analyses, with over 200 million students at different ages and in different subjects, indicate that feedback has one of the most powerful impacts on student learning. Feedback is considered a vital element of assessment for learning (Black et al., 2003, pp. 42-49; Black & Wiliam, 1998a, 1998b, 2012).

While feedback is a complex issue and its effectiveness depends on several factors – such as the quality, timing and user-friendliness of feedback, and how the student receiving feedback reacts to it (see e.g. Brookhart, 2012; Hattie, 2009, 2012; Wiggins, 2012; Wiliam, 2012; Stobart, 2012) – it has the potential to improve learning (e.g. Stobart, 2012). However, in order to support and enhance learning, feedback should not only state or describe how things are at any given moment, but it should also *feed forward*, i.e. aim at improving future performance (e.g. Black & Wiliam, 1998b; Hattie & Timperley, 2007; Lizzio & Wilson, 2008; Wiggins, 2012).

Price et al. (2010), for instance, see correction as integral to a rather traditional and straightforward definition of feedback where "the role of feedback is to 'put things right' by taking a corrective action" (p. 278). In second and foreign language education, corrective feedback, in other words, marking and/or correcting students' errors, is probably the most common form of feedback. Recent studies on teacher feedback on second or foreign language writing have found that teachers primarily correct all student errors but, in addition to error correction, secondary school teachers in particular give rather little of any other feedback (e.g. Furneaux, Paran, & Fairfax, 2007; Guénette & Lyster, 2013; Lee, 2004). Also, although there is some recent second or foreign language literature that examines feedback in a broader sense, such as diagnostic feedback (e.g. Alderson et al., 2015; Jang & Wagner, 2013), most of the language education literature tends to narrow feedback down to corrective feedback, be it oral or written (Alderson et al., 2015; Jang & Wagner, 2013). Ergo, there has been a lively debate about the efficacy of corrective feedback in the second language writing and acquisition literature over the past couple of decades (see e.g. Bitchener & Ferris, 2012; Ferris, 2012; Guénette, 2007). However, despite numerous studies and analyses, no consensus on which corrective feedback method is the most effective – or even whether corrective feedback is beneficial at all – has been found (e.g. Guénette, 2007; Hyland & Hyland, 2006; Lee, 2005, 2008, 2014; Leontjev, 2016). (For further information on corrective feedback, see e.g. Bitchener, 2008; Ellis, Sheen, Murakami, & Takashima, 2008; Ferris, 1999, 2012; Hyland & Hyland, 2006; Leontjev, 2016; Lyster & Ranta, 2013; Simard, Guénette, & Bergeron, 2015; Truscott, 1996, 2007.)

2.4.2 By whom, to whom?

In an educational context, feedback is most often given by the teacher. Several studies in second or foreign language education have found that students prefer teacher feedback because they do not necessarily trust peer feedback to the same extent (e.g. Hyland & Hyland, 2006; Lee, 2008; Leki, 1991; Tarnanen & Huhta, 2011). Peer feedback is therefore usually given as supplementary feedback.

In upper secondary studies, feedback is primarily given to the students themselves. However, their parents or guardians may also receive feedback on the students' learning through grades, for instance (e.g. *National core curriculum for upper secondary schools* 2003, p. 224). Also teachers can get – or infer – feedback on their teaching through student assessment. For instance, the international comparative surveys such as PISA and ICCS inform the participating schools of their school's overall results (Linnakylä & Välijärvi, 2005, pp. 47-49). Schools also get some feedback from external examinations such as the Matriculation Examination. Sometimes that 'feedback' is also published in the media in the form of various ranking lists or league tables.

2.5 Use of the assessment judgement

In Figure 3, the last phase in the assessment process is the use of the assessment judgement. As already said many times in this study, the purpose of the assessment should define the use of its assessment judgements as well as its intended impact. However, that is not always the case. Particularly high-stakes test results are used for several additional purposes that they were not designed for, such as evaluating and comparing schools and teachers. These same results may also be used for various financial decisions ranging from allocating resources to schools to house prices (Jones et al., 2003; Koretz, 2008; Kubiszyn & Borich, 2013, pp. 30-33; Newton, 2007, 2012). In other words, the actual uses may differ from the intended use. That may also change the intended impact as well as the consequences of the assessment. Furthermore, tests, for instance language tests, may be 'repurposed' or 'retrofitted' to be used for new purposes that they were not originally designed for (see e.g. Fulcher & Davidson, 2009). However, because of the focus and scope of this study, I will now move on to focus on the consequences only.

2.5.1 Consequences: High-stakes and low-stakes assessment

If the consequences of an assessment are important for the learner, the assessment can be labelled as high-stakes assessment (Herbert & Hauser, 1999; Volante, 2006). What makes some assessment high-stakes is not the assessment itself, then, nor its contents or form, but primarily the way its results are used and what their impacts are on the student (Herbert & Hauser, 1999).

High and low stakes are therefore closely linked with pressure, i.e. high and low pressure (Nichols, Glass, & Berliner, 2006). An everyday example of a high-stakes test is a driving test: if the candidate does not pass, he or she will not get a driving licence. In an educational setting, high stakes normally refer to tests whose outcome has "high-stakes consequences for students – that is, when an individual student's score determines not just who needs help, but whether a student is allowed to take a certain program or class, or will be promoted to the next grade, or will graduate from high school" (Heubert & Hauser, 1999, p. 14).

The proponents of high-stakes testing, such as Bishop and Mane (2001), Cizek (2005) and Phelps (2005, 2012), to mention but a few, have argued that the high stakes attached to the test outcomes motivate students to study harder in order to gain rewards (e.g. better placement, or admission to further education) and to avoid punishing consequences such as retention or denial of graduation (see also e.g. Heubert & Hauser, 1999; Kornhaber & Orfield, 2001; Natriello & Pallas, 2001; Nichols et al., 2006). Along the same lines, high-stakes test scores have increasingly been used for other accountability purposes, such as evaluating an individual teacher's effectiveness or a school's performance, even though they were not designed for that purpose (e.g. Jones et al., 2003; Koretz, 2008; Stobart, 2008; Volante, 2006; see also Kuusela, 2003; Sahlberg, 2011). As

the rewards or threats are closely linked with money, job security and other significant factors (Kubiszyn & Borich, 2013, pp. 30-33), they are believed to act as highly effective incentives and thus improve the quality and effectiveness of the education (e.g. Cizek, 2005; cf. Amrein & Berliner, 2002.).

The opponents of high-stakes testing say that instead of improving teaching and learning, high-stakes tests lead to teaching to the test. As teachers devote more time to test revision and practice tests, not only the contents of teaching but also the selection of teaching and learning methodology become narrower (see e.g. Kornhaber & Orfield, 2001; Koretz, 2008; Mitchell & Salsbury, 2002, p. 118-119; Natriello & Pallas, 2001; Nichols & Berliner, 2007; Sahlberg, 2011). High-stakes tests are also believed to make students' learning shallower, as often students' primary purpose is to pass the test, not to learn the topics or skills per se (Harlen, 2005, 2012b; Natriello & Pallas, 2001; Sahlberg, 2011). All this easily leads to superficial rote learning instead of real conceptual understanding (e.g. Dietel, Herman, & Knuth, 1991; Harlen, 2005, 2012b; Volante, 2004).

Furthermore, some studies have found that high-stakes testing, instead of enhancing educational opportunities for disadvantaged student groups as was intended, in fact had a detrimental effect for minority, second-language, disabled or disadvantaged students (see e.g. Au, 2009; Madaus & Clarke, 2001; Mitchell & Salsbury, 2002; Natriello & Pallas, 2001; Rumberger, 2011). As high-stakes tests are often one single test with highly pressurised time and place constraints, they may also cause considerable stress and test anxiety (e.g. Aydın, 2009). Test anxiety can weaken memory and knowledge retention and, thus, test performance (Hembree, 1988). Underperforming in the test, in turn, can affect students' motivation, self-efficacy and self-esteem as learners (Harlen & Deakin Crick, 2003; Harlen, 2005).

A low-stakes test or assessment has no such serious consequences for the student (e.g. Sessoms & Finney, 2015; Wise & DeMars, 2005). As the defining factor is not the test itself but the use and perceived consequences of the results, what is a high-stakes test for one may not necessarily be that for someone else. There may also be 'medium stakes' (see e.g. Roever, 2001). A combination of low-stakes assessments may also eventually have high-stakes consequences when, as a sum total, they determine something of more importance, such as the final grade of the school-leaving certificate (see e.g. Thorsen, 2014).

2.6 Actual impact

"Assessment is never a neutral activity. It always has its impacts – both those intended and those not intended", writes Välijärvi (1996, p. 128). Broadly speaking, the intended impact of student assessment is to let students know how well they have reached the learning goals and then guide students to act upon this information to further their learning. Gardner (2012, p. 106-107),

when arguing that "assessment of any kind should ultimately improve learning", puts it as follows:

> Regardless of how 'learning' might be conceptualized, for example as the assimilation of new knowledge, the development of new understanding or the acquisition of new skills, and regardless of what theoretical position is taken on learning, be it socio-cultural, constructivist or behaviourist, it is difficult to contest the notion that assessment of the progress or outcome of that learning is beneficial to the learner. If this assessment pinpoints what we know, understand or can do, it affirms our learning. If it pinpoints difficulties or weaknesses, it enables us to focus our efforts, and the efforts of those who support our learning, on identifying how we might improve our learning. Assessment for the sake of assessment makes no sense but assessment for someone else's sake is an industry of epic proportions.

As we have seen above, assessment in the educational context is used for many other purposes as well. These purposes may differ greatly from the intended purpose or purposes. Furthermore, since assessment basically means judging the worth, value and importance of something (e.g. Atjonen, 2015, pp. 29-31; Linnakylä & Välijärvi, 2005), the worth, value or importance may vary between individuals.

The actual impact therefore depends on various factors but always also on the individual student. In other words, the actual impact that assessment has on the learner depends on how the learner experiences and reacts to the assessment and its outcome (cf. Hattie, 2009; Wiliam, 2012). This actual impact may help students to improve their learning and thus enhance their skills and knowledge, in other words, their resources. It may also motivate them. Hence, the actual impact may be empowering. However, the actual impact may also be disempowering and discourage students in their studies (see e.g. Rumberger, 2011).

The next chapter will therefore look at the concepts of empowerment and disempowerment as well as the potential link between assessment and (dis)empowerment in the light of research.

3 (DIS)EMPOWERMENT AND ASSESSMENT

This chapter focuses on the concepts of empowerment and disempowerment and their potential link with student assessment. However, having discussed empowerment, its history and varying definitions in several disciplines rather extensively in the Monograph part of this study (Pollari, 2000), I will not repeat all that discussion here. Instead, I will revisit the concept of empowerment in a more concise manner and I will also discuss and define the concept of disempowerment as understood in this study. I will then focus on some earlier research that has examined assessment and its impact on student empowerment.

3.1 Empowerment

In the late 1900s, empowerment became a common word in both academic and everyday discourse. However, empowerment was quite seldom explicitly defined (e.g. Karl, 1995, p. 14; Mondros & Wilson, 1994, p. 5; Perkins & Zimmerman, 1995). As it was, and still has been, used in different contexts and for several purposes, it has had varying meanings and connotations (Evans, 1992; Perkins & Zimmerman, 1995; Siitonen, 1999, p. 82). Empowerment can therefore be a problematic concept (Leach, Neutze, & Zepke, 2001) and quite elusive in its rather all-encompassing nature, which Robinson (1994, p. 12) describes as follows: "Empowerment is individual and collective; it is power and freedom; it is external and internal, political and personal, a means to an end and its own reward".

Nonetheless, in my opinion, at least three different aspects of empowerment could be discerned from the diverse use of the term: empowerment as giving power and resources, empowerment as taking power and/or resources and, finally, empowerment as taking charge of power and resources (for a more detailed discussion, see Pollari, 2000, pp. 51-57).

Historically, the roots of empowerment have been attributed to the Reformation and utopian socialism, to name but two possible sources; later, many others, including Gandhi, Freire, the Civil Rights movement as well as feminism and the sexual rights movements, have influenced notions of empowerment, particularly in the social and political sciences (Simon, 1994). Originally, empowerment was therefore mainly used in the emancipatory sense of giving the oppressed power (Freire, 1972) as well as resources, means and opportunities, through various forms of political, legislative, economic or social actions.

However, according to many scholars, empowerment is a process and thus cannot be simply given to people (Karl, 1995, p. 14). Adams (1991, p. 208) defines empowerment as "becoming powerful" and explains that it "embodies two dimensions: being given power and taking power." Cummins (1986; 2001, p. 653) sees empowerment as "the collaborative creation of power" that adds the power of both the empowered and the empowerer:

> Thus, power is created in the relationship and shared among participants. The power relationship is *additive* rather than *subtractive*. Power is *created with* others rather than being *imposed on* or *exercised over* others. (Cummins, 1996, p. 15; emphasis original.)

In addition, as empowerment is often seen as a process aiming towards greater participation and responsible autonomy, empowerment also entails a third dimension: actively taking charge of one's power and resources. Some scholars see empowerment as a process where power comes from within the person, not from outside (see e.g. Siitonen, 1999, p. 83).

Empowerment became an increasingly popular term also in education in the 1980s and 1990s (Perkins & Zimmerman, 1995). Several educational reforms or tools, ranging from critical pedagogy (e.g. Giroux, 1989, see also Darder, Baltodano, & Torres, 2003) and experiential learning (see e.g. Mulligan & Griffin, 1992; cf. Kolb, 1984) to computer networks and co-operative learning (Sapon-Shevin & Schniedewind, 1991) were considered empowering. Although the term was rather loosely defined also in education, the same three aspects – giving power and resources, accepting or taking power and resources as well as taking charge of the power and resources – could be read in the use of the term.

Nearly 20 years ago, I myself defined empowerment as "a process entailing the aspects of getting power, accepting and assuming it, and taking charge of it" (Pollari, 2000, p. 68). This definition of empowerment is presented in visual form in Figure 4, below. The empowered, although perhaps given power as a recipient or an object, is seen as an active agent who accepts power and "takes charge of it actively as a responsible subject": power includes here not only decision-making power but also "opportunities, resources and means to have both the readiness and willingness to take charge of one's actions and potentials actively and responsibly" (Pollari, 2000, p. 68). (For a longer discussion on empowerment, its definitions and roots in education, see Pollari, 2000, pp. 57-72 as well as Siitonen, 1999).

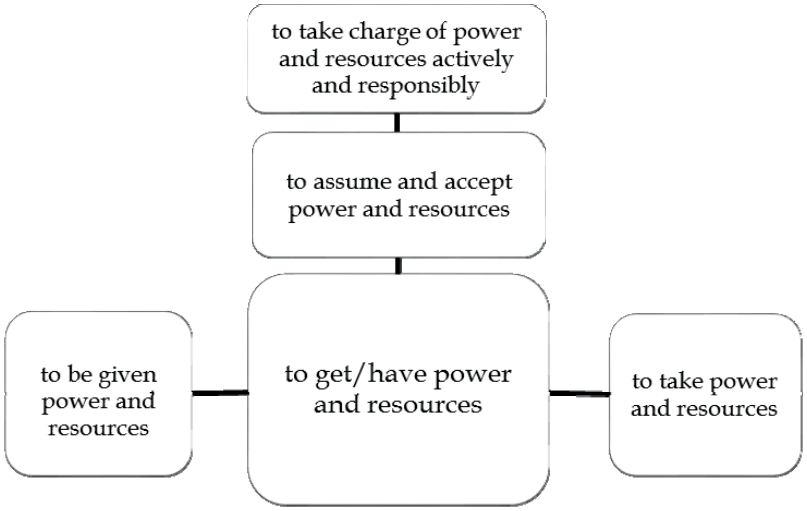

FIGURE 4 Empowerment as a concept but also as a process (based on Pollari, 2000, p. 68)

Around the same time, a theory of empowerment was being formulated within community psychology (see e.g. Perkins & Zimmerman, 1995; Rappaport, 1987; Zimmerman, 1995, 2000; Zimmerman & Rappaport, 1988; see also Schulz, Israel, Zimmerman, & Checkoway, 1995). The theory analyses empowerment at the levels of the individual, the organisation and the community, and it includes both processes and outcomes which may vary depending on the contexts and people involved (Zimmerman, 2000).

At the individual level of analysis, empowerment is referred to as psychological empowerment. Psychological empowerment has three components: intrapersonal, interactional and behavioural. The intrapersonal component is manifested not only by perceived control and self-efficacy, but also by competence and motivation (Zimmerman, 1995, 2000). The behavioural component entails "efforts to exert control" through active involvement (Zimmerman, 2000, p. 46). The interactional component provides a bridge between intrapersonal and behavioural components and it "suggests that people are aware of behavioral options or choices to act as they believe appropriate to achieve goals they set for themselves" (Zimmerman, 1995, p. 589). The theory of empowerment and its three levels are given visual form in Figure 5.

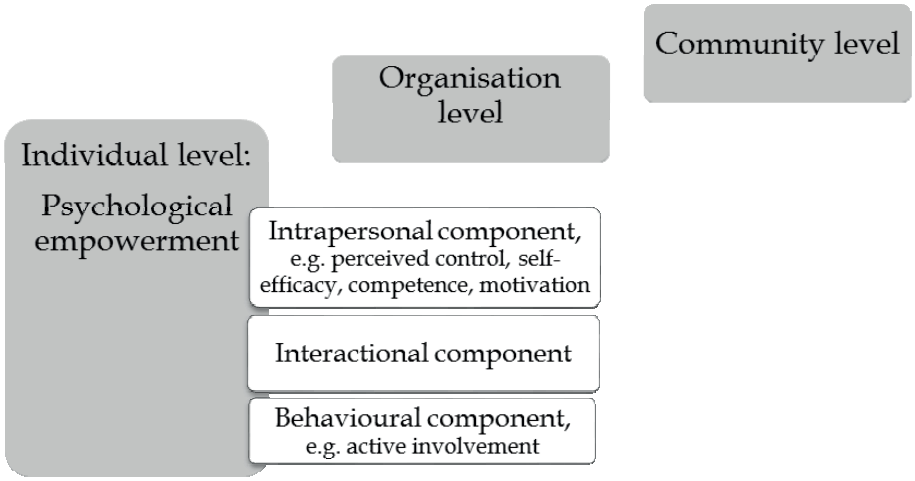

FIGURE 5 The theory of empowerment, its three levels of analysis and the components of psychological empowerment, as based on Zimmerman (1995, 2000).

In addition to these aspects or levels of empowerment, several writers have discussed different dimensions, contexts, domains or purposes of empowerment (see e.g. Perkins & Zimmerman, 1995). The focus can be, for instance, on legal, economic or cultural empowerment, or the purpose may be to empower different groups, ranging from women to minorities, from staff to patients or students. Some writers also define other levels of empowerment, such as psychological, social or political empowerment (see e.g. Francina & Joseph, 2013). Analysing some definitions of empowerment, Perkins and Zimmerman (1995, p. 570) conclude that "empowerment is more than the traditional psychological constructs with which it is sometimes compared or confused (e.g., self-esteem, self-efficacy, competency, locus of control)". They also conclude that psychological empowerment, sometimes also referred to as personal or individual empowerment (e.g. Bolaffi, Bracalenti, Braham, & Gindro, 2003), is "a goal common to all levels of intervention" (Perkins & Zimmerman, 1995, p. 574).

Although empowerment is more than the "traditional psychological constructs" of self-esteem, self-efficacy, competency and locus of control (Perkins & Zimmerman, 1995, p. 570), those constructs do play a major role in empowerment and in learning as well. So too do self-regulation and motivation: "In essence, highly self-regulated learners approach learning tasks in a mindful, confident manner, proactively set goals, and develop a plan for attaining those goals" (Cleary & Zimmerman, 2004, p. 538). Discussing minority students and their academic success or failure, Cummins (1986; 2001, p. 661) agrees:

> Students who are empowered by their school experiences develop the ability, confidence, and motivation to succeed academically. They participate competently in instruction as a result of having developed a confident cultural identity as well as school-based knowledge and interactional structures (Cummins 1983b, Tikunoff

> 1983). Students who are disempowered or "disabled" by their school experiences do not develop this type of cognitive/academic and social/emotional foundation.

As also the theory of empowerment recognises, both empowerment processes and their outcomes vary (Zimmerman, 2000). Referring to the works of Zimmerman (2000) and Schulz et al. (1995), Miller and Campbell (2006, p. 297) write that empowered outcomes "are evidenced by whether individuals or aggregate bodies of individuals engage in behaviors that permit effective pursuit of planned change and results in success." However, in some cases the actions meant to empower people "fail to foster the emancipatory potential that they make possible" (VanderPlaat, 1998, p. 87; see also Toomey, 2011). Individuals may also react differently to these actions and processes. As Leach et al. (2001, p. 294) put it: "Empowerment is not the same for everyone. A process that is empowering for some will be disempowering for others and will be resisted by them." Moreover, although the goal of empowerment is to foster a group's or an individual's agency and opportunities "to make effective choices, that is, to make choices and then to transform those choices into desired actions and outcomes" (Alsop, Bertelsen, & Holland, 2005, p. 10), some writers also highlight the right of those being empowered to decide not to use their power: "The choice is therefore with the individual, who, given the power, authority, skills and willingness to act, may choose to accept empowerment" (Rodwell, 1996, p. 309).

To summarise, I cite Miller and Campbell (2006, pp. 297-298), who, in my opinion, manage to incorporate many of the aspects and characteristics of personal empowerment in the following quotation:

> According to Schulz et al. (1995) and Zimmerman (2000), empowered individuals are critically aware and therefore able to analyze what must change, posses [sic] a sense of control and so feel capable of acting, and engage in participatory behaviors. An empowered person perceives their personal agency and acts in ways that reflect this perception.

Finally, in order to clarify the close ties of empowerment and numerous other related concepts, I attempt to visualise the conditions and processes, both external and internal, needed for successful outcomes of empowerment in the following figure.

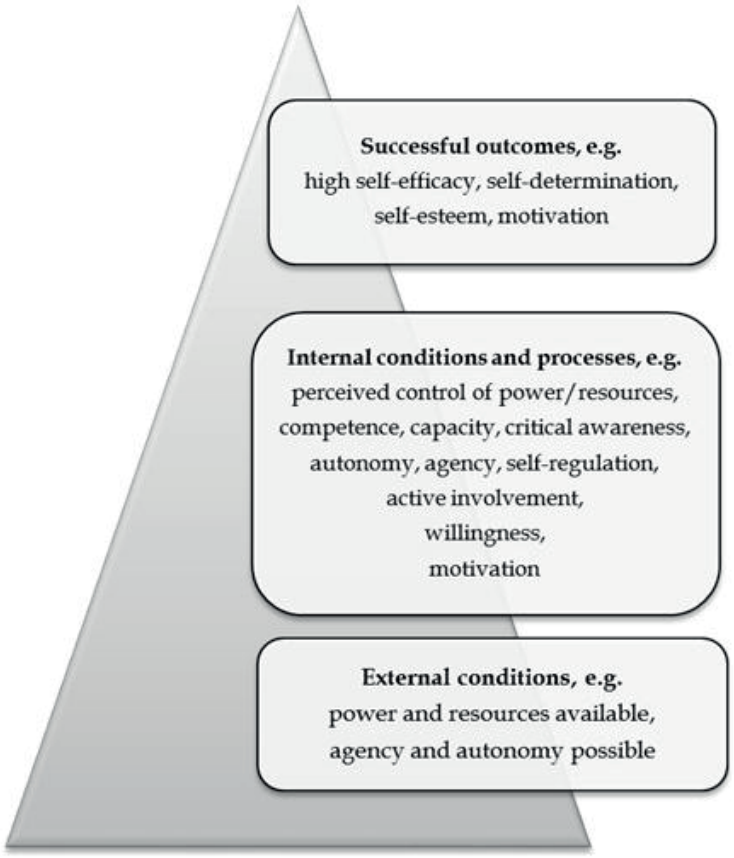

FIGURE 6 The external and internal conditions and processes of empowerment as well as the successful outcomes of empowerment.

3.2 Disempowerment

Like empowerment, so too disempowerment is used in different contexts with varying meanings. Rather often, disempowerment seems to be regarded as a term which requires no further definition (Kasturirangan, 2008, pp. 3-6; Toomey, 2011). For instance, Bolaffi et al. (2003) regard disempowerment and empowerment simply as opposites of each other. They define empowerment as "a process whereby people who are oppressed are enabled to gain some power and control over their lives": therefore, it is "the opposite of disempowerment – a process by which people are socially excluded because they are denied access to such power and control" (Bolaffi et al., 2003, p. 85).

Some authors also see power and resources as finite: if someone becomes empowered, somebody else has become disempowered (see e.g. Lorion & McMillan, 2008). These notions seem to regard empowerment and

disempowerment as a continuum of allocated power, with empowerment at one end and disempowerment at the other.

Disempowerment is also common in everyday discourse. The Merriam-Webster on-line Learner's dictionary defines *to disempower* as follows (http://learnersdictionary.com/definition/disempower, read 1.3.2016):

> To cause (a person or a group of people) to be less likely than others to succeed; to prevent (a person or a group) from having power, authority, or influence. To deprive of power, authority, or influence; to make weak, ineffectual, or unimportant.

Oxford Dictionaries gives a shorter definition: "Make (a person or a group) less powerful or confident" (https://en.oxforddictionaries.com/definition/disempower, read 1.3.2016). Both these dictionary definitions underline the aspects of confidence and self-efficacy, which are also important constituents of psychological empowerment (Zimmerman, 1995, 2000), even though Zimmerman does not use or define the term *psychological disempowerment* himself (Kasturirangan, 2008, p. 8-10). If people have been given power but they lack self-confidence, they are probably less likely to use their power. Disempowerment is therefore not simply a case of denying someone power and resources (cf. Bolaffi et al., 2003, p. 85).

In this study, disempowerment does not refer to students having or not having power, but it refers to students *experiencing* that they do not have power and/or resources to make decisions in order to fulfil their potential. In other words, disempowerment refers to a lack of *perceived* control and low self-efficacy (e.g. Zimmerman, 1995, 2000): students may actually have been given power but they do not either realise this or believe in their power and/or themselves. Therefore they do not, or cannot, take charge of their potential power, which may, in turn, lead to diminished motivation (Cleary & Zimmerman, 2004; Harlen, 2012b; Weber & Patterson, 2000).

3.3 Empowerment, disempowerment and student assessment[2]

Assessment is "very much an exercise of power" (Välijärvi, 1998, p. 13) and it has its impact on those assessed (see e.g. Välijärvi 1996; Shohamy, 2001). Assessment should therefore meet certain ethical requirements (e.g. Atjonen,

---

[2] At the organisational or community level, Fetterman (1996, 2001, 2002) has, with several colleagues (see e.g. Fetterman & Wandersman, 2005), discussed *empowerment evaluation* as a means for programme evaluation and improvement. Fetterman (2001, p. 14) characterises empowerment evaluations as follows: "Empowerment evaluations vary in size and scope. However, they all are shaped by a focus on self-determination, capacity building and helping others evaluate themselves." Empowerment evaluation shares many similarities with action research and is also informed by Zimmerman's theory of empowerment (Fetterman, 1996, 2001, pp. 13-14). However, as empowerment evaluation is not concerned with student assessment but with programme evaluation, it is outside the scope of the present study and will not be discussed here further.

2007; Välijärvi, 1996). Ethically speaking, in addition to being valid and reliable, assessment should also be fair, just and transparent (e.g. Atjonen, 2007; Race et al., 2006). Furthermore, assessment should avoid causing harm, and instead, aim at doing good, at promoting and motivating learning, for instance (e.g. Atjonen, 2007, pp. 34-51). Assessment should also respect students' autonomy, their right to make their own choices (Atjonen, 2007, pp. 37-39).

Nonetheless, from the students' point of view, assessment is often a rather disempowering experience. Students are the objects of assessment, with little, if any, say in the assessment process and its decisions (e.g. Aitken, 2012; Boud, 2007; Shohamy, 2001, 2007). Power over testing or assessment lies with their teachers or schools, or with external examination boards or testing agencies. However, the use of assessment information and decisions made on the basis of these assessments, such as graduation or access to further education, may sometimes have far-reaching consequences for students (e.g. Boud, 2007; Shohamy, 2001; Virta, 2002).

Assessment can also disempower students by affecting their learning, both the learning processes and learning outcomes. In the first place, as several studies have shown (see e.g. Darling-Hammond, Rustique-Forrester, & Pecheone, 2005; Madaus & Clarke, 2001; McNeil & Valenzuela, 2001), assessment can narrow the curriculum – both in the sense of content and methodology – and thus substantially limit students' learning. At worst, learning is not driven by students' learning needs or interests but rather imposed by high-stakes testing (Kornhaber & Orfield, 2001). This can disempower some student groups, particularly those belonging to a minority group (e.g. Darling-Hammond et al., 2005; Madaus & Clarke, 2001; Kornhaber & Orfield, 2001). Assessment can also impair students' learner role and their willingness and capacity for self-assessment, a skill necessary for life-long learning and any expertise (see e.g. Earl 2003):

> The capacity to make judgements is not well represented in many current assessment practices. Assessment items are often strongly knowledge-based, with criteria unilaterally set by teachers. The role of students tends to be to offer themselves to be assessed by others. This can create dependency on the authority of the teacher, rather than other sources of judgement, and can give rise to the implication that judgements are necessarily made by others. This is in contrast to the learner being positioned as an active agent in assessment decisions, as is advocated by many assessment theorists (e.g. Nicol and Macfarlane-Dick 2006; Nicol 2009). (Boud et al. 2013, 942-943)

In the school context, there is scant empirical evidence available of students' perceptions of the empowering or disempowering qualities of assessment. However, Aitken (2012) has studied Canadian students' anecdotes on assessment. The students, from primary school to university, specified several assessment practices that they found unfair. They mentioned, for example, lack of variety in the assessment methodology, too pressurised tests or insufficient test-taking time, secrecy over the test content, format or criteria, inadequate feedback and biased grading (Aitken, 2012). A European survey on foreign language assessment and its focus had rather similar results; in addition,

students mentioned irrelevant or too limited a focus as a feature of 'bad' assessment (Erickson & Gustafsson, 2005).

Although research on students' empowerment or disempowerment in terms of assessment is generally rather scarce, several authors have focussed on some particular assessment method as possibly empowering. For instance, many portfolio projects have aimed at empowering students both in their studies and assessment in several subjects at different school levels both internationally and in Finland (see e.g. Linnakylä, Kankaanranta, & Pollari, 1994; Pollari, Linnakylä, & Kankaanranta, 1996). These have also included foreign or second language portfolios (see e.g. Padilla et al., 1996; Permana, 2013; Pollari, 1996, 2000). Little and Erickson (2015; see also e.g. Little, 2005) highlight the possibilities of the Common European Framework of Reference for Languages (CEFR) and its European Language Portfolio (ELP) – based on the ideas of the learner autonomy movement (see e.g. Holec, 1979; Holec & Huttunen, 1997) – not only for integrating learning, teaching and assessment but also for promoting learner agency through self-assessment. However, Little and Erickson (2015, p. 125) note that although the CEFR reference levels are commonly used, the ELP has not been widely adopted in foreign language education, and therefore "the CEFR's underlying ethos has largely gone unrecognized or been ignored".

In addition to the ELP or its electronic version (Cummins & Davesne, 2009), shared assessment has been advocated as a way of empowering student writers in academic English at tertiary level (Pienaar, 2005). Peer assessment has also been used as a tool for engaging and empowering pupils in their EFL assessment at primary school level when preparing for high-stakes examinations (Bryant & Carless, 2010).

Other approaches are used to foster students' agency and autonomy in foreign language assessment as well (see e.g. Everhard & Murphy, 2015). They range from students' involvement in national language test development (Erickson & Åberg-Bengtsson, 2012) to formative assessment in EFL writing (Burner, 2015) and creating autonomy classrooms (Dam & Legenhausen, 2011). However, although agency and autonomy are closely linked with empowerment, these studies do not discuss the concept of empowerment or disempowerment as such. In addition, although outside the school and student assessment context, Shohamy (2001, 2007, 2014) has discussed the power of language testing and its potentially detrimental and undemocratic effects on test-takers extensively.

At tertiary level in particular, self-assessment has been widely implemented in order to foster students' empowerment and to enhance their learning and future professional skills (see e.g. Tan, Teo, & Ng, 2011; Tan, 2012, pp. 1-4). For this reason, most studies looking into assessment empowerment seem to have taken place in higher education and have focused on self- and peer assessment (see e.g. Kearney, Perkins, & Kennedy-Clark, 2016). Their results have been slightly mixed, ranging from quite positive to conflicting. For instance, in a study of 233 university students, Hanrahan and Isaacs (2001)

found that university students experienced self- and peer assessment as difficult and even uncomfortable but at the same time they felt that these methods enhanced their learning and understanding of the assessment and its criteria, for instance. Another study, by Patton (2012), explored 36 Australian undergraduates and their perceptions of peer assessment. The study found that although students supported peer assessment for formative assessment purposes, they "were highly critical of it as a summative practice" (Patton, 2012, p. 719).

In addition to using self-assessment, Leach, Neutze and Zepke (2000, 2001) decided to give adult learners also more power over assessment methods and criteria in the form of a choice: the students could either name their own tasks and the criteria to be used in the assessment, or take what the teachers suggested. Although the teachers had the final say, Leach et al. (2001, pp. 299-300) saw an opportunity in this:

> But this unequal power relationship can be used to create a context for learner empowerment; first to give learners insight into the academic discipline of assessment; and second, to create conditions for learners to empower themselves, to decide what evidence to present, what criteria to use, whether to self-assess or not, and whether to accept the judgement of authority, or to resist it.

The results showed that as students were different and had different experiences, they also responded differently to assessment empowerment: there were students who liked power-sharing, those who disliked it and those who disliked power-sharing at first but grew to appreciate it. Leach et al. (2001, p. 298) concluded that although the results were positive, "learners will vary in their desire and confidence to make judgements about their own work". This desire may also vary depending on how advanced and mature the students are: for example, a small-scale study by Francis (2008) found that third-year university students were more receptive to assessment empowerment than first-year students. In the study by Leach et al. (2001), in the name of empowerment the students could also decide to leave the assessment solely to the teachers. Tan (2012), however, disagrees with this choice: in his opinion giving students the right *not* to participate in assessment – self-assessment in his case – is not empowering. The students' decision not to participate "may be a sign of their docile and disciplined condition" and lack of self-confidence (Tan 2012, p. 140). It may also be due to low self-esteem (Tan et al., 2011). Moreover, if optional, it will not foster the learning and self-assessment skills of those who opt out (Tan, 2012).

To summarise the discussion above, traditional assessment does not appear to promote student empowerment in the sense of giving students decision-making power, autonomy or agency at any of the phases of the assessment process described in Figure 3. Most of the studies reported here have focused on self- and peer assessment as a potential vehicle for assessment empowerment in higher education. In other words, they have concentrated on *who produces the assessment judgement* as well as *who communicates* it. However, the study by Leach et al. (2000, 2001) offered students decision-making power,

agency and autonomy *in the design and collection of assessment evidence* as well as in its *interpretation*. Thus, the study by Leach et al. (2000, 2001) enabled more comprehensive assessment empowerment throughout the assessment process. They also permitted their students *not* to accept this empowerment and leave the assessment solely to their teachers.

4 STUDENT ASSESSMENT IN FINNISH UPPER SECONDARY SCHOOL

This chapter will address student assessment in Finnish upper secondary education. First, I will look at the national core curricula for upper secondary education: What do the core curricula, both past and present, say about student assessment? Do they state any particular requirements for assessment in English or in foreign languages in general? I will focus on the four core curricula in effect during the approximately 30 years that upper secondary education has been course-based, presented in Table 2 below. Although the title *Lukion opetussuunnitelman perusteet* has remained the same in Finnish, their English translations have varied. Hence, for consistency and reader-friendliness, I will refer to the core curricula with shorter English titles, also presented in Table 2, in the following sections.

TABLE 2 The national core curricula discussed in the present study.

Original Finnish name and publication information (English translation)	Henceforth in the present study
Lukion opetussuunnitelman perusteet 1985. Helsinki: Kouluhallitus/ Valtion painatuskeskus. (no English translation available)	*Core curriculum 1985*
Lukion opetussuunnitelman perusteet 1994. Helsinki: Opetushallitus. *(Framework curriculum for senior secondary school 1994*. Helsinki: National Board of Education.)	*Core curriculum 1994*
Lukion opetussuunnitelman perusteet 2003: Nuorille tarkoitetun lukiokoulutuksen opetussuunnitelman perusteet. Helsinki: Opetushallitus. *(National core curriculum for upper secondary schools 2003: National core curriculum for general upper secondary education intended for young people.* Helsinki: Finnish National Board of Education. Engl. translation 2004)	*Core curriculum 2003*
Lukion opetussuunnitelman perusteet 2015: Nuorille tarkoitetun lukiokoulutuksen opetussuunnitelman perusteet. Helsinki: Opetushallitus. *(National core curriculum for general upper secondary schools 2015: National core curriculum for general upper secondary education intended for young people.* Helsinki: Finnish National Board of Education. English translation in 2016)	*Core curriculum 2015*

After that, I will explore what earlier research has said about Finnish student assessment in upper secondary school and/or in foreign language studies.

I will conclude with an evaluative summary of Finnish upper secondary school student assessment and that of EFL on the basis of all of the above. In it I will refer to the concepts discussed in Chapter 2 and attempt to situate Finnish student assessment in the more general and international student assessment landscape with the help of these concepts.

4.1 Curricular guidelines for upper secondary student assessment

Teachers have strong autonomy in student assessment in Finland: not only do the teachers decide on the assessments, and thus design and organise them, but they also decide the assessment criteria, mark the assessments, draw conclusions on them and decide how to use the results (see e.g. Sahlberg, 2007; Vänttinen, 2011, p. 180). This autonomous role has gone more or less without question throughout the history of Finnish education (Vänttinen, 2011, p. 165). The Matriculation Examination, taken towards the end of upper secondary studies, has been the only external high-stake examination in the Finnish school context (Sahlberg, 2007; see also Atjonen, 2015, p. 34). There have been no compulsory school-leaving examinations at the completion of basic education, for instance (Sahlberg, 2007; see also Rinne et al., 2011, pp. 28-31).

Even though high-stakes testing has not been typical of assessment in Finnish education, student assessment – or evaluation (*arvostelu*) as it was called earlier – was long regarded as synonymous with grading (see e.g. Räisänen & Frisk, 1996), especially in upper secondary education (Välijärvi, 1996). Grading was usually based on evaluating the learning outcomes through tests. Thus, the assessment of learning through tests, but in the sense of teacher-made tests, has dominated Finnish educational assessment for a long time.

Despite their strong autonomy in student assessment, teachers have had to follow some regulations and guidelines. Traditionally, most of the regulations have been technical or bureaucratic in nature and focused mainly on grading (Vänttinen, 2011; see also Apajalahti, 1996). Nonetheless, the most important educational and pedagogical guidelines for student assessment have been given by the core curriculum of each era. I will therefore next discuss the core curricula for upper secondary education. In addition to the core curriculum that was in effect at the time of (most of) this study, I will also briefly explain some past and present curricula in order to throw light on the educational trends and traditions in Finnish student assessment.

4.1.1 Core curricula 1985 and 1994

In 1983, the Upper Secondary Schools Act (477/1983) changed upper secondary school education, its curriculum and structure significantly. Previously, upper secondary school education was quite strictly and centrally regulated (see e.g.

Core curriculum 1985; Välijärvi, 1996). The new national core curriculum for upper secondary education, *Core curriculum 1985*, still gave extensive guidelines and instructions to be followed, but also allowed local educational authorities some liberty in writing their own curricula. Furthermore, although the school structure was not yet non-graded, the upper secondary school curriculum became course-based and modularised, i.e. all subjects and their syllabi were divided into several courses of approximately 38 lessons. Each course was to be evaluated and graded separately and the final school-leaving grade in each subject was formed on the basis of the mean of all the previous grades in that subject.

Core curriculum 1985 added other general guidelines for student evaluation[3]. First of all, evaluation was to support both the upper secondary school studies and the attainment of its goals. Evaluation had to be as reliable and fair as possible. It was required to be encouraging and not too burdensome and attention was to be paid to "the quality of information, not only quantity" (*Core curriculum 1985*, p. 31). The criteria of each course were to be derived from the goals of each course and take the students' age and year into account so that the demands grew towards the end of the upper secondary school studies. Furthermore, students had to be informed about the criteria at the beginning of each course. The attainment of the goals could be evaluated through tests and continuous assessment, which included students' oral and written work, homework and participation in class work. The tests were also defined: "Tests can be summative tests that cover the goals of the course extensively or formative tasks that focus on some of the goals only" (*Core curriculum 1985*, p. 31). Summative tests were evaluated using grades 4-10, formative tasks could be evaluated "also in other manners" (*Core curriculum 1985*, p. 31). Even though the tests were so clearly determined, a course could also be evaluated on the basis of continuous assessment, without a summative test (*Core curriculum 1985*, p. 32). If summative tests were used, continuous assessment could either raise or lower the grade by one grade (*Core curriculum 1985*, p. 32).

However, in addition to the general guidelines for student evaluation, *Core curriculum 1985* did not give any additional guidelines for the assessment in second or foreign language education even though there are over 200 pages of text defining the course goals, foci, contexts and contents, including the grammatical structures that students were supposed to master (see *Core curriculum 1985*, pp. 61-282).

The next national core curriculum, *Core curriculum 1994*, changed and also further decentralised several things. First of all, the upper secondary school structure became non-graded, i.e. students did not have to follow their Year or class/group but could select courses more independently. Furthermore, the number of optional courses increased significantly, which enabled students to

---

[3] At that time, the word used in Finnish educational and curricular texts was *oppilasarvostelu*. To highlight the later change in terminology, I will use the word *evaluation* or even *grading* whenever the Finnish term used was *arvostelu*. Currently, the term *arvostelu* sounds rather judgemental.

have more individual study plans. Also schools and educational authorities were given a great deal of freedom in writing their own curricula. The core curriculum itself is an example of the decreased regulation: whereas the published *Core curriculum 1985* had over 400 pages, *Core curriculum 1994* had approximately one hundred.

In 1994 the term *oppilasarviointi*, student *assessment*, also replaced the earlier term *oppilasarvostelu* (student evaluation or grading) in the Finnish version of the core curriculum[4] for the first time (see *Lukion opetussuunnitelman perusteet 1994* in Finnish; see also Apajalahti 1996). The purpose of student assessment was to "give students feedback on the progress of their studies and on their learning achievements" both during upper secondary school and on completion of their studies: the purpose of that feedback was "to encourage and guide students in their studies" (*Core curriculum 1994*, p. 32). Grading, which was defined as one outcome of student assessment, was to be based on the objectives determined in the curriculum and "should aim at the best possible reliability and fairness" (*Core curriculum 1994*, p. 32). Students were therefore to be informed about the assessment and grading principles. Teachers were told to encourage their students to engage in self-assessment, which could be taken into account also in course grading. In addition to possible self-assessment, course grades were based on "possible written examinations, on continuous observation of the progress of studies and on the assessment of the student's products" (*Core curriculum 1994*, p. 32). No further regulations, instructions or advice were given on assessment in foreign or second language education in *Core curriculum 1994*[5].

Although there was a great change in both the physical and philosophical nature of the core curricula between 1985 and 1994, the main purposes of and requirements for assessment remained more or less the same. First of all, assessment had a dual function: it was to guide, support and encourage students' studies as well as to evaluate their attainment of the learning objectives. Secondly, the assessment and grading criteria were to be goal-referenced, i.e. based on the learning goals, and also transparent so that students knew them. Thirdly, evaluation and grading had to be fair and reliable, and finally, under both core curricula, grading could be based not only on written tests or exams but also on continuous assessment, including observation of students' participation in class and their oral or written work. Nevertheless, there were a few changes. Firstly, the term itself changed into student assessment (*arviointi*), most likely to de-emphasise the grade-oriented aspect and also avoid the judgemental and negative connotation of the term *arvostelu*.

---

4 Although the English translation of *Core curriculum 1994*, i.e. *Framework curriculum for senior secondary school 1994* uses the word *assessment* when discussing assessment in general, it uses the word *student evaluation* when discussing assessment that focuses on students' work and learning.

5 In total, *Core curriculum 1994* (in Finnish) has 25 pages of text concerning second or foreign language education in upper secondary school. In comparison, *Core curriculum 1985* had 220. For instance, the course descriptions in *Core curriculum 1994* are only a few lines long and the grammatical structures to be taught and mastered are no longer listed.

Secondly, tests, whether summative or formative, and their marking scales were no longer defined or regulated in *Core curriculum 1994*. Finally, self-assessment could be taken into account in assessment, also in grading.

4.1.2 Core curricula 2003 and 2015

The *National core curriculum for upper secondary schools 2003* (henceforth *Core curriculum 2003*), which was in place during the data gathering of Articles 1-5, states the objectives of assessment in upper secondary school as follows (p. 224):

> Student assessment aims to guide and encourage learning and to develop students' self-assessment skills. Students' learning and work shall be assessed diversely. (General Upper Secondary Schools Act, 629/1998, Section 17(1))
>
> The role of assessment of students' learning is to provide students with feedback on their progress and learning results both during and upon completion of upper secondary school studies. The purpose of such feedback is to encourage and guide students in their studies. In addition, assessment provides information for students' parents or guardians and for the needs of providers of further studies, representatives of working life and other similar groups. Assessment of students' learning will also help teachers and the school community as a whole to evaluate the effectiveness of education. Grading is one form of assessment.
>
> Assessment will encourage students in a positive way to set their own objectives and to readjust their working methods. (*Core curriculum 2003*, p. 224)

In addition, some guidelines are given for course assessment. Each course "will be assessed upon completion", and the purpose of the assessment is "to provide students with feedback on how well they have met the objectives of the course and on their progress in that subject" (*Core curriculum 2003*, p. 224). Assessment must be based on varied assessment practices and methods and may include self-assessment:

> Course assessment must be diverse and based not only on possible written tests, but also on continuous observation of students' progress in their studies and assessment of their skills and knowledge. Students' own self-assessment may also be taken into account, making use of methods such as course assessment discussions. (*Core curriculum 2003*, p. 224).

Assessment methods and practices will be determined in further detail by schools and local educational authorities in their local curricula (*Core curriculum 2003*, p. 224). In addition, students must know the learning goals as well as the assessment criteria of each course right from the start of the course, and these must be discussed with students:

> In addition to general assessment criteria, students must be informed of the criteria for assessment of each course at the beginning of the course, when these will be discussed with students. (*Core curriculum 2003*, p. 225).

Thus, the national core curriculum allows students some potential power in assessment: when discussing the assessment and its criteria, students can also have a say and perhaps give some suggestions. Students also have the right to

receive more detailed information about how the criteria are used in their particular case, i.e. to ask for clarification of their assessment (*Core curriculum 2003*, p. 225). Hence, although *Core curriculum 2003* does not use the words *empowerment* or *agency* directly, some traces of these concepts are present.

Core curriculum 2003 is the first of the four national core curricula discussed here to give additional guidelines or instructions concerning assessment in any particular subject. The guidelines in most subjects are clearly goal-referenced and also take the nature and the learning process of the particular subject well into account. For foreign or second language education, however, *Core curriculum 2003* mentions only one additional requirement:

> Assessment of the subject will take all areas of language proficiency into account in accordance with the priorities emphasised in the course descriptions (*Core curriculum 2003*, p. 102).

In 2010, when a so-called oral course was introduced, separate regulations were drawn up for its assessment. They stated, for instance, that an oral examination administered by the National Board of Education had to be used in assessment of the course. That requirement made the oral course the only course in the whole of the upper secondary school curriculum where an external test or examination was stipulated and the teacher did not have total autonomy in designing the course assessment.

In sum, *Core Curriculum 2003* (pp. 102, 224-225) states that assessment in English and all foreign or second language studies, as in all upper secondary education, must be diverse in both its forms and focus. Also, it must be transparent, in other words, students must know the goals and assessment criteria of each course, and these are to be discussed with students at the beginning of each course. Students are also entitled to know the rationale behind their assessment and grades. And, most importantly, the purpose of assessment is to guide and encourage learning and to develop students' self-assessment skills. Thus, assessment must give students feedback on their learning, both on its progress and on its results. Furthermore, when providing information for parents or potential employers, for instance, assessment also serves some external purposes.

The emerging ideas of *assessment of learning* and *assessment for learning* can thus be read in the national core curriculum already in 2003. However, as they are stated rather indirectly, assessment of learning has probably been the dominant function of assessment in upper secondary school studies. This may have been caused also by our rather grade- and test-oriented assessment culture (see e.g. Välijärvi, 1996) as well as by the modularised curriculum structure where each of the 75 courses are to be assessed separately. Besides, assessment is difficult and also slow to change. Steeped in the values, beliefs and attitudes of the surrounding society, assessment and particularly grading procedures are not only educational but also cultural practices with long traditions of stability and continuity and, hence, they are rather change-resistant all over the world (see e.g. Välijärvi, 1996; Suurtamm & Koch, 2014).

The latest core curriculum, *National Core Curriculum for general upper secondary schools 2015* (henceforth *Core curriculum 2015*), was still under construction during the time the data of this study was gathered, and it took effect in August 2016. *Core curriculum 2015* clarifies the two functions of assessment (i.e. assessment *of* learning and assessment *for* learning) and emphasises the importance of assessment for learning: "The purpose of assessment of learning is to promote the student's learning" (*Core curriculum 2015*, p. 240). The new curriculum also stipulates that students' learning is assessed also *during* each course; the purpose of this assessment is to enhance students' learning and give them feedback on their reaching of the course's objectives (*Core curriculum 2015*, p. 240).

In English, as in all foreign or second languages, assessment is defined and determined in a much more comprehensive way than previously:

> Assessment in foreign languages is based on the achievement of the general objectives of the instruction in foreign languages and special, syllabus- and language-specific objectives. Course-specific emphases and the closely related general and syllabus-specific objectives of foreign languages are taken into account for each course. Versatile feedback is provided on the student's progress at different stages of the learning process in all courses. Feedback is provided on the student's progress in the different areas of language proficiency as well as other objectives, such as language-learning skills and capabilities to act in target language environments. The students are guided in utilising self and peer assessment. Language portfolios can be utilised in all courses, also crossing the boundaries of individual subjects. (*Core curriculum 2015*, p. 115.)

Furthermore, *Core curriculum 2015* goes on to suggest that "where applicable, the Evolving Language Proficiency Scale, based on the European Framework of Reference, is used as a support for assessment, as a tool for the teacher, and an instrument for the student's self and peer assessment" (p. 115). *Core curriculum 2015* also determines that the assessment of the so-called oral course "is based on the grade awarded for the oral skills test set by the Finnish National Board of Education as well as other demonstration of knowledge and skills by the student during the course" (p. 116). Hence, the autonomy of foreign and second language teachers in designing and deciding on the assessment is more limited than the autonomy of teachers of any other subjects in upper secondary education.

As can be seen from the extracts above, *Core curriculum 2015* is leaning away from the test- or grade-oriented summative assessment tradition, with assessment of learning as its central purpose, towards assessment for learning[6]. However, its impact remains to be seen. The new core curriculum was not in place during the present study.

---

[6] Nonetheless, *Core curriculum 2015* for upper secondary school does not take this stance as clearly and strongly as the *National core curriculum for basic education 2014*, which clearly defines the roles and purposes of formative and summative assessment – and thus also assessment for learning and assessment of learning – in basic education.

4.2 Prior research on student assessment in Finnish upper secondary school

Research on student assessment in Finland is rather scarce. Although reports on learning outcomes in different subjects at the end of basic education (see e.g. Hildén et al., 2015; Ouakrim-Soivio, 2013; Tuokko, 2000, 2002, 2007) as well as research on some assessment experiments do exist, there is little detailed research on actual student assessment practices and procedures *in general* in Finland, such as tests that teachers have designed and/or used and their scoring and grading, so we actually cannot know for certain *how* teachers assess and grade their students (see e.g. Virta, 2002, p. 66-70; cf. e.g. Duncan & Noonan, 2007). In addition, only two studies dealing with students' experiences of assessment in any way have taken place during the past decade, that is, while the *National core curriculum for upper secondary schools 2003* was in place. Thus, I will have to look back in time.

Since the introduction of the current Finnish school system with comprehensive school/basic education and upper secondary school, there have been only a few studies that have dealt with student assessment in Finnish upper secondary school in any way at all. Some of the research was experimental in nature, i.e. the purpose was to experiment with some new features in assessment (e.g. Välijärvi, 1981, 1984; Syrjälä, 1989; see also Pollari, 1996, 1998). On the other hand, some of these studies were larger surveys or evaluation studies covering the whole spectrum of upper secondary school education, so student assessment played only a very minor part in them (e.g. Välijärvi, 1993; Välijärvi et al., 2009)[7]. Most of these studies were instigated or commissioned by educational officials, for instance, the National Board of Education or the Ministry of Education and Culture. As only one of these studies focused on assessment in English in upper secondary school (Pollari, 1998, 2000), I will also discuss a study by Tarnanen and Huhta (2011; see also Huhta & Tarnanen, 2009) as well as one by Härmälä, Huhtanen and Puukko (2014), both of which examined assessment in language education in basic education. The other studies reviewed here dealt with general assessment in upper secondary school. However, all the following studies asked students for their opinions or experiences of assessment. I will report the studies and their main findings in chronological order.

The first studies, by Välijärvi (1981, 1984), were based on a longitudinal study of a new way of student evaluation[8] that was carried out in an

---

7 The study on students' mathematical competence at the end of secondary education by Metsämuuronen (2016) also studied the relationship between mathematical competence and upper secondary school grades. However, this study did not explore how student assessment was carried out or how students experienced it.

8 Välijärvi (1981, 1984) uses the then commonly used Finnish word *arvostelu* (*kurssiarvostelu, päättöarvostelu*). As he translates the word as *evaluation* in his English abstracts, I will also use the words evaluation or grading when reporting his studies.

experimental upper secondary school, Alppila[9], in connection with a reform in the upper secondary school curriculum. The most tangible change introduced and experimented with in this study was the use of a different grading scale, 0-3, in students' course evaluation, instead of the traditional, and official, scale of 4-10. The participating students were asked to fill in a questionnaire in the three consecutive years of their upper secondary school studies (N=94/68/66). Their parents were also asked to answer a similar questionnaire.

After a quite positive start, students' – girls' in particular – attitudes towards the new scale became more negative as the experiment progressed. The scale was increasingly considered less accurate and just. One factor that caused friction was the fact that although course performance was assessed using the new scale, the final school-leaving grades would be given using the official 4-10 scale. However, students found positive aspects in the new grading scale as well: for instance, studying and evaluation had become less grade-oriented and grade-centred (Välijärvi, 1984, pp. 10-11).

The study by Välijärvi (1984) found that tests played a major role in assessing students' achievements and that students regarded tests as a good and reliable assessment method. Also, although students found studying for the tests taxing and stressful, they considered tests to be important: they motivated them to study, made the goals clearer and also gave quieter students a chance to show their knowledge and skills (Välijärvi, 1984). There were some gender differences in the responses. At the beginning of the experiment, girls had a more positive attitude towards the new grading scale than boys had. They also seemed to appreciate 'softer' assessment methods, such as continuous assessment, more than boys. Furthermore, girls suffered more from stress and anxiety caused by the tests (Välijärvi, 1984).

Syrjälä (1989) was the second scholar to study students' and teachers' views and experiences of student assessment as part of studying and teaching. This experiment in assessment, carried out in Alppila in 1982-1985, concentrated mainly on two things: another course grading scale (1-5, although again the official scale of 4-10 was to be used in the final school-leaving grade) and making assessment more varied. The assessment included continuous assessment of learning, verbal feedback, self- and peer assessment, as well as a wider range of types of test questions. Syrjälä's (1989) study consisted of some teacher and student interviews, written documents and also teacher and student questionnaires. Only third-year students were asked to respond. Although the number of student responses was quite small, 42, it represented 76% of the third-year students (Syrjälä, 1989, pp. 34-41). One of the six research questions focused on how upper secondary school students experienced student assessment, and another one on what tests, continuous assessment, performance assessment as well as self- and peer assessment meant to students (Syrjälä, 1989, pp. 40-41).

---

9 Alppila School was founded in 1959 as a national experimental school and it continued to function as a locus for several school experiments until the 1990s.

The students' reactions and experiences of assessment, tests and grading seemed slightly contradictory. Many students considered tests and assessment one of the most unpleasant features of upper secondary school studies, as they found assessment stressful and did not enjoy studying for the tests; yet most of the students found tests useful for learning because "you have had to revise for the tests" (Syrjälä, 1989, p. 77). Over 60% of the respondents also found the Matriculation Examination useful while 35% did not. Syrjälä (1989, p. 80) concludes that students seemed mostly concerned with the fairness of their grades and thus saw assessment in a rather limited way, as grading. However, 55% of the students did not think that grades could give enough information about their skills and knowledge.

In short, the student assessment research of the 1980s focused on experimenting with alternative grading scales, neither of which came to replace the traditional scale of 4-10 in upper secondary education.

In the 1990s, Välijärvi (1993) focused mainly on the new modularised, course-based curriculum and school structure but he also investigated its impact on student evaluation. In the modularised, course-based system, each course was evaluated and graded as a separate entity in which a student's previous grade in that subject played no role. The final school-leaving grade of each subject was then decided on the basis of the average of the course grades. Students (N=2,196), and female students in particular, mainly regarded the independent course-based evaluation system as positive and well-suited for the new study and curriculum structure. However, the way the school-leaving grade was decided divided opinions strongly and many students, male students in particular, considered the system unfair and demotivating. Välijärvi (1993, p. 125-134) concludes that although students' attitudes and opinions on student evaluation varied quite significantly between both students and schools, female students were more open to the new assessment system and thus less change-resistant.

The next stage was a large-scale student survey by Välijärvi and Tuomi (1995), which investigated upper secondary school as a learning environment and how it enabled students' individual study choices. One of the findings of their study was that students (N=2,850) experienced upper secondary school studies as demanding as well as strongly driven by tests. Half of the respondents felt that tests played too big a role in student assessment and grading, with 20% of the students saying that tests had a clearly negative effect on their studies; yet 49% of the students also felt that tests had a positive impact on their studying (Välijärvi & Tuomi, 1995, pp. 49-51). Furthermore, according to the students' experiences, "the Matriculation Examination casts a long shadow on the daily life of upper secondary schools" as nearly half of the teachers emphasised the importance of the Examination in their teaching (Välijärvi & Tuomi, 1995, p. 49).

The mid- and late-1990s could probably be characterised as the years of enthusiasm for authentic assessment, enabled and encouraged by the new *Core curriculum 1994*. There were several small-scale projects experimenting with

alternative, more authentic assessment methodology in Finnish schools but most of these experiments were not thoroughly documented or reported. One project that was reported was the *portfolio project*, instigated by the Institute for Educational Research (see e.g. Linnakylä, Pollari, & Takala, 1994; Pollari, Kankaanranta, & Linnakylä, 1996). The project also involved developing portfolio assessment in the teaching of English in upper secondary school (Pollari, 1996, 1998, 2000). In that study, portfolio assessment was used as a rather radical, alternative method of studying and assessment: for instance, no tests were taken during the portfolio course. The experiment allowed students a great deal of power and autonomy in deciding the topics, methods and also timetable of their pieces of work. In addition to a mandatory course grade, students also received a longer written assessment of their portfolios. The portfolio was mainly considered a nice and also empowering change by the participants (104 students and three teachers), enabling students' individual choices not only in their studies but also in assessment (e.g. Pollari, 2000). The turn of the millennium also witnessed some other upper secondary portfolio experiments, for instance those piloting the use of the European Language Portfolio (see e.g. Kohonen & Pajukanta, 2003; Lammi, 2002).

The largest study to investigate students' views on student assessment was the *Evaluation of pedagogy in Finnish upper secondary education* (Välijärvi et al., 2009). Its data consisted of a survey of third-year upper secondary school students (N=8,500) as well as interviews with students, teachers and heads of school. This evaluative study examined several features of Finnish upper secondary education, such as its objectives, students' flexible and individual study choices and teaching and working methods. A few questions on student assessment were included in the questionnaire section that dealt with teaching and working methods. These items showed that students considered assessment methods not to be very diverse but rather test-focused (Välijärvi et al., 2009, p. 54)[10]. Students nevertheless felt that assessment had given them a fairly good idea of their skills. Teachers had also discussed both the goals and the assessment criteria of each course with their students at the beginning of the course, as called for by the national framework curriculum (2003). However, self-assessment was not very widely used as part of course assessment. A large majority of the students, 75%, stated that good success in the Matriculation Examination was a goal directing their upper secondary school studies (Välijärvi et al., 2009, pp. 38-40). A teacher survey, part of another upper secondary school evaluation conducted two years later, corroborated earlier findings (Turunen et al., 2011, pp. 82-83). Both these studies therefore recommended that student assessment methodology should be made more varied, interactive and encouraging (Turunen et al., 2011, p. 88; Välijärvi et al., 2009, pp. 58-59). Also, assessment should focus on the whole learning process

10 This may partly be due to the upper secondary school structure where each of the approximately 35-lesson courses is assessed separately as an independent entity: also, the exam week system, where each 5-7-week period of the academic year ends with an exam week, is used in many schools and may have its effect on their assessment practices (Välijärvi et al., 2009, p. 54).

and students should be encouraged and trained to use self-assessment more (Välijärvi et al., 2009, p. 59).

In basic education, two recent large-scale studies have touched on assessment in language education. Although the curriculum and also the assessment guidelines for comprehensive school are different from those for upper secondary school, I will briefly report the main findings of these studies here as there have been no equivalent studies in the upper secondary school context.

The first of these studies was a project called *ToLP – Towards Future Literacy Pedagogies – Finnish 9th graders' and teachers' literacy practices in school and out-of-school contexts*, carried out in 2006-2009 (see Luukka et al., 2008). As part of that research project, both Huhta and Tarnanen (2009) and Tarnanen and Huhta (2011) examined foreign language assessment and feedback practices at the end of comprehensive school. The data of the study, reported in Huhta and Tarnanen (2009) as well as in Tarnanen and Huhta (2011), consisted of questionnaire surveys for students (N=1,720) and foreign or second language teachers (N=324, mainly English or Swedish), and teacher interviews. Tarnanen and Huhta found that "both the students and teachers agreed that the teacher carries out assessment far more frequently than any other actor in the assessment process" (Huhta & Tarnanen, 2009, p. 9). Most of the assessment, at least according to students, seemed to take place at the end of a course or a learning unit, and the teacher's role in grade-giving was dominant (Huhta & Tarnanen, 2009). Self- and peer assessment were used in the classrooms at least occasionally – teachers reported them taking place more often than students did – but they were "apparently used mostly for low-stakes, possibly formative, purposes, as the majority of both teachers and students said they do not play a significant role in determining students' grades in high-stakes final assessment" (Tarnanen & Huhta, 2011, p. 140). According to both teachers and students, not only test results but also effort, participation in class and attitude had an important role in grading. However, teachers and students had somewhat different views on what skills and content teachers considered important when assigning grades to their students, and therefore Tarnanen and Huhta (2011, pp. 140-141) concluded that many students did not appear to "fully know the criteria by which their performances are evaluated". Nonetheless, most students considered their foreign language grades accurate (Tarnanen & Huhta, 2011). All in all, the results draw a picture of fairly traditional assessment practices "that are partly consistent with the national curriculum" but do not seem to meet all its requirements (Huhta & Tarnanen, 2009, p. 17).

An evaluation of the learning outcomes in English at the end of basic education (Härmälä et al., 2014) was carried out in April, 2013 as part of an evaluation of the learning outcomes in most foreign/second languages studied in basic education (see Hildén et al., 2015). A total of 3,476 pupils (Year 9) and 220 teachers of English participated in the evaluation, which also involved a questionnaire survey including some questions on assessment and feedback practices. According to the participating teachers, written tests and

participation in class were the most important factors in course grading; doing homework and students' attitude also played a major role (Härmälä et al., 2014, p. 120). However, oral tests divided teachers' opinions: while a third considered them to be important in grading, another third regarded them as unimportant. The European Language Portfolio was not commonly used in language education or assessment. Furthermore, the teachers reported personalised feedback (e.g. discussing progress with the student or giving feedback on pair talk exercises) as well as self- and peer assessment much more commonly than the students did (Härmälä et al., 2014, p. 119), thus corroborating the findings of Tarnanen and Huhta (2011). The evaluations of the learning outcomes of other foreign/second languages had similar findings: written tests played a much more significant role in grading than oral skills, and self- and peer assessment did not appear to be a prominent feature in Finnish foreign or second language education (see e.g. Hildén & Rautopuro, 2014, pp. 111-130; Hildén et al., 2015).

In sum, previous research on student assessment in Finnish upper secondary schools shows that the types of student assessment used have been rather limited – with, perhaps, the exception of some assessment experiments in the 1980s and 1990s. The long tradition of considering student assessment, or evaluation, to be synonymous with grading (e.g. Apajalahti, 1996; Vänttinen, 2011) still seems to persist, at least to some extent. Nevertheless, students generally consider their grades to be fair and accurate. There has been little research into assessment in English or other foreign languages in Finnish upper secondary school, but it is not very likely that foreign language assessment differs greatly from the general assessment tendencies found in the studies introduced above.

4.3 Finnish student assessment in upper secondary school and in EFL: An evaluative summary

Finally, to funnel the discussion of the earlier chapters towards this study, I will briefly discuss student assessment in Finnish upper secondary schools and in EFL in particular. To do so, I will rely, firstly, on earlier research and literature. Secondly, I will rely on my own experience in the field of English. For over 20 years, I have not only taught English at comprehensive and upper secondary school but also worked as a teacher trainer. In that capacity, I have encountered hundreds of future teachers of English. Responsible for giving lectures and running workshops on assessment for these teacher trainees for at least the past ten years, I have had the opportunity to discuss the assessment practices used in their former schools. Furthermore, as an author of an upper secondary EFL course-book series and as a guest lecturer on assessment, I have met dozens of English teachers and discussed their assessment practices and concerns with them all over Finland. Thus, my insights are not based only on my views, opinions or classroom assessment practices – or those of my many colleagues –

but on those of the hundreds of EFL teacher trainees and dozens of EFL teachers I have encountered during my career.

In the following evaluative summary I will refer back to the concepts of students assessment discussed in Chapter 2. I will also refer back to Figure 3 (see Chapter 2).

First of all, the purpose of student assessment in Finnish upper secondary schools can be characterised as mainly *summative*. According to earlier research (e.g. Välijärvi & Tuomi, 1995; Välijärvi et al., 2009), assessment is rather *test-focused* and the summative test or exam at the end of the course usually carries considerable weight in assessment. Moreover, assessment appears to be somewhat *grade-centred*; some earlier research has found grading to be the dominant purpose and use of upper secondary school assessment (Syrjälä 1989). Grades appear to be the prevailing form of feedback as well, so there is not very much feedback, or feed-forward, that would help and guide learning forward. Nonetheless, students mostly feel that the assessment has given them a fairly good idea of their skills (Välijärvi et al., 2009, p. 54). Formative assessment does not seem to have gained much ground yet, or it is not regarded as assessment since it does not result in summative results and/or grades.[11] Indeed, some confusion over terminology may still persist: some teachers appear to consider formative assessment synonymous with continuous assessment, or with smaller tests and surprise quizzes, used also for summative purposes, as was the case earlier, in the 1980s, for instance.

Earlier research has also concluded that there is little variety in the methods used for collecting assessment evidence in Finnish upper secondary schools, and that the methods are not very interactive or participatory (Välijärvi et al. 2009; see also Turunen et al., 2011). The results of a study on teachers' views on their own assessment practices in comprehensive school (Atjonen, 2014) as well as one on the evaluation of pedagogy in Finnish basic education (Atjonen et al., 2008) and subsequent evaluations of language learning outcomes in basic education (e.g. Hildén et al., 2015; Härmälä et al., 2014) suggest similar conclusions. Thus, even though there are individual teachers and schools experimenting with alternative and innovative assessment methods, assessment can generally be considered quite *traditional* – although not necessarily in the American, multiple-choice testing sense. Nevertheless, Finnish foreign or second language assessment relies much more heavily on the select-answer approach than does the assessment in other subjects. For instance, nearly half of the EFL Matriculation Examination test score is currently based on multiple-choice items.

According to the national core curriculum, the focus and content of assessment should be determined by the objectives of each course, i.e. what is taught and studied in that particular course. The core curriculum also states

---

11 My guess is that because of the Finnish assessment/evaluation/grading tradition, the word assessment, or *arviointi* in Finnish, still carries the weight of the earlier form, *arvostelu*, and somehow has both a more formal and slightly judgemental tone. Therefore, *assessment for learning* (*AfL*) is not regarded as part of assessment, but part of teaching, by teachers and students alike.

that foreign language assessment "will take all areas of language proficiency into account in accordance with the priorities emphasised in the course descriptions" (*Core curriculum 2003*, p. 102). Although there is no research on the actual focus of foreign language assessment in Finland, assessment still appears to be more focused on correct language forms and vocabulary than language use, even though the teaching and studying in basic education cover communicative competencies (Tarnanen & Huhta, 2011, pp. 130-131; see also Wahlroos, 2012). Oral communication, i.e. speaking, appears to have been of less importance in both teaching and assessment (see e.g. Huhta & Hildén, 2013, p. 166). If this is the case also in upper secondary EFL studies, then assessment does not completely meet the requirements of instructional validity (Anderson, 2003, p. 11).

The design, collection and interpretation of the assessment evidence in Finland can primarily be regarded as *internal, classroom assessment*. However, although *teacher-controlled*, assessment is not necessarily teacher generated. Teachers most often compile the tests themselves, but out of testing exercises written by course-book authors and publishers. With new, digitalised testing exercises, the teacher cannot necessarily even edit or change the exercises. Sometimes teachers also opt for ready-made model tests, complete with grading instructions and criteria. In addition, especially towards the end of upper secondary EFL studies, teachers use past Matriculation Examination tests as testing material. Thus, although assessment and tests are not externally mandated or controlled, they are often externally designed, at least in parts. The interpretation of this assessment evidence may also be partly external when teachers use the criteria or grading instructions given by text-book authors or the Matriculation Examination Board. Similarly, EFL tests and other assessment tasks could be considered *small-scale* since teachers usually decide on their assessment methodologies quite individually. However, they often include larger-scale components, such as past Matriculation Examination test exercises or those provided by course-book publishers.

On the basis of both the present and previous national core curricula, Finnish upper secondary school student assessment can be regarded as *criterion-referenced* as each course is assessed on the basis of its objectives and how well the student has reached those objectives. However, the issue seems less clear-cut when one looks at the official guidelines[12], earlier research and classroom practice. For instance, Halonen (2007, p. 34) claims that "norm-referenced assessment is still used in school, for instance in the Matriculation Examination". Strictly speaking, the Matriculation Examination is not part of the upper secondary curriculum, but Halonen (2007) may have a point. The first assessment of students' test papers (e.g. an EFL essay) in the Matriculation Examination is criterion-referenced, but the final grading, i.e. the

12 The National Board of Education gives conflicting information on the matter in its guide for upper secondary school curriculum where Blom (2003, p. 70) claims that upper secondary school student assessment "is norm-based. It is based on the course objectives of each subject and their attainment, unlike the Matriculation Examination, which is based on the relative evaluation of knowledge and skills".

transformation of students' raw scores into the Matriculation Examination grades, is mainly based on the relative distribution of grades, or norm-referencing, in subjects with large test-taker populations (Mehtäläinen & Välijärvi, 2013, p. 81). Hence, the Matriculation Examination being a major goal for most students and thus also for many teachers, it may "cast a long shadow" (Välijärvi & Tuomi, 1995, p. 49) over assessment and marking practices. The ghosts of norm-referenced 'grading on the curve', which was the norm in Finland some decades ago (Apajalahti, 1996; Vänttinen, 2011), may therefore still haunt some classrooms (Ouakrim-Soivio, 2013, p. 216). However, there is no research on how foreign or second language teachers actually assess, score and grade their students in Finnish upper secondary education to date.

As is evident, teachers have, in principle, almost absolute power over the design, collection and interpretation of the assessment evidence in Finland. However, it appears that in practice, teachers have handed a great deal of that power over to external test makers. Teachers also have a great deal of power over the communication of the assessment judgement in the student assessment process. However, teachers have not shared the power with their students very much. Assessment procedures in which students can actually decide on the assessment methods, tools or criteria do not seem widespread. Neither do self- or peer assessments where students judge the quality of work against the given criteria for summative purposes: these may, however, be more commonly used for learning purposes than for grading (Tarnanen & Huhta, 2011). All in all, students' agency in the assessment process seems more or less limited to discussion of the course goals and assessment criteria at the beginning of each course, which is mandated by the national core curriculum: according to Välijärvi et al. (2009), approximately 80% of the students stated that the goals and criteria were indeed discussed with them. How much students can influence the actual design, collection or interpretation of the assessment evidence in these discussions is, however, not known. Given the lack of any research in this area, much of the above is therefore speculation based on tacit knowledge and anecdotal evidence.

Finally, student assessment in Finnish upper secondary education could mostly be regarded as relatively *low-stakes* assessment. However, as the final school-leaving grade is based on the average of the earlier course grades, their sum total has characteristics of high-stakes use (cf. Huhta & Tarnanen, 2009). Furthermore, since the stakes mostly depend on students and the impact assessment has on them (e.g. Herbert & Hauser, 1999), some students, in some circumstances, may consider upper secondary assessment more high-stakes assessment than others. Upper secondary school assessment may indeed occasionally also have high-stakes impacts.

In contrast, the Matriculation Examination, which is externally controlled, designed and marked, is a large-scale, high-stakes examination (Mehtäläinen & Välijärvi, 2013, p. 13; see also Lahtinen & Välijärvi, 2014). Even though the Matriculation Examination is not part of the upper secondary school

curriculum[13] but a separate examination system governed and organised by the Matriculation Examination Board (e.g. Lahtinen & Välijärvi, 2014, p. 88; Lindström, 1998), it is a major goal that has a significant impact on students' studies (Välijärvi et al., 2009, 38-40; see also Broadfoot, 1996b) and shapes upper secondary studies and teaching to a considerable extent (see Välijärvi et al., 2009, pp. 41-43). It therefore seems to have an impact on upper secondary school assessment, also in EFL, as the past Matriculation Examination tests are used for assessment purposes during upper secondary school studies.

Ultimately, the current assessment practices in upper secondary education do not seem to fully meet the general requirements set by the legislation and the core curriculum (see *Core curriculum 2003*, p. 224):

> Student assessment aims to guide and encourage learning and to develop students' self-assessment skills. Students' learning and work shall be assessed diversely. (General Upper Secondary Schools Act, 629/1998, Section 17(1))

From the perspective of empowerment, students' power and active agency is rather limited in assessment: students must be informed of the assessment criteria (*Core curriculum 2003*, p. 224), they may be given a say in assessment decisions (e.g. the design and collection of assessment evidence) when the assessment criteria are discussed at the beginning of the course, and their own self-assessment may also be taken into account in the course grades. However, whether students have any additional power in these decisions beyond being informed of the course criteria is decided in the local curricula (*Core curriculum 2003*, p. 224), or by individual teachers.

4.4 The niche of this study

As the above review shows, student assessment is a very prominent feature in education. It is also a highly complex, context-specific and powerful phenomenon which may have far-reaching consequences on students' lives. Yet while research on external, high-stakes testing abounds internationally, there has been little research on student assessment in Finland. Moreover, the impact of assessment on students' lives, and how students themselves experience assessment and its power, has not been studied widely, either internationally or in Finland. Thus, this study attempts to address some of these gaps in research.

---

13 The Matriculation Examination is considered the final examination of upper secondary education. It has two purposes: firstly, to assess whether students have acquired the skills and knowledge set in the upper secondary school curriculum and have the required maturity; and, secondly, to give eligibility for higher education (General Upper Secondary Schools Act, 629/1998, Section 18; see also e.g. Lahtinen & Välijärvi, 2014, p. 11). Nevertheless, students can acquire the upper secondary school certificate even if they do not sit or pass the Matriculation Examination. In order to get the Matriculation Examination certificate, however, students must have completed their studies and acquired a general upper secondary school certificate or a certificate of their vocational studies (Lahtinen & Välijärvi, 2014, p. 17).

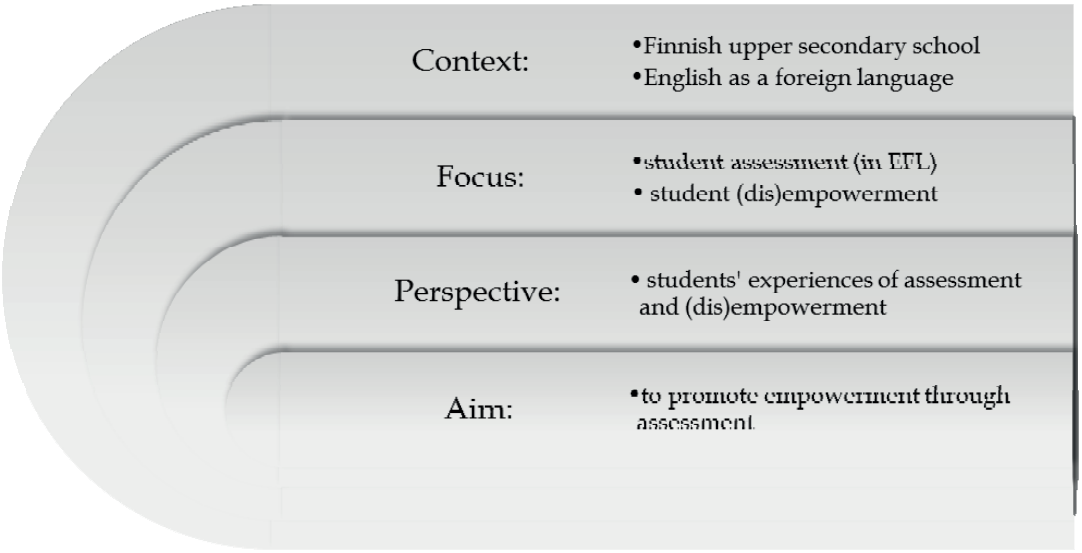

FIGURE 7 The context, focus, perspective and aim of the present study.

5 THE PRESENT STUDY

From the very start of my teaching career, assessment has troubled me. From very early on, it has also empowered me in the sense that assessment, and portfolio assessment in particular, offered me a new way of looking at my work, namely teacher-research (see e.g. Borg, 2013, p. 10). I have combined these two for most of my career by studying and experimenting with various assessment methods. This study is a culmination of that work.

The origins of the present study lie in an interest in finding out what students at our school thought of assessment in their upper secondary English studies. There being little research on students' experiences of assessment, the original purpose of this study was thus rather exploratory, aiming for better understanding of the matter (Borg, 2013, pp. 12-13; see also Dörnyei, 2007, p. 191). Exploratory investigations, as Dörnyei (2007, p. 308) phrases it, "help us to map the terrain first and fine-tune our specific research angle later in the project". This was also the case with the present study: with the experiences of disempowerment emerging from the survey data, I began to focus on them. Then, with my history with action/teacher research, i.e. developing and testing diverse assessment methodologies in order to empower students (see e.g. Borg, 2010, 2013), these two strands began to merge into the present study.

This study therefore has two aims. One is to delve into *the actual impact of assessment* as experienced by students themselves, and in this case, into assessment empowerment or disempowerment, *to find out what predicts disempowerment in assessment and how assessment disempowerment and empowerment manifest themselves*. The other aim is linked to *the intended impact of assessment*: the aim is *to experiment with diverse assessment methods in order to see whether they could foster students' empowerment in assessment*. Given this dual intent, it is difficult to summarise the aim of this study in one single research purpose or question.

Like the two aims of the study, the reporting of the present study is also divided into two parts, each of which focuses on different aspects and relies on different data and methodology. The first part, i.e. Part 1, comprises three sub-studies, all of which are articles. They focus on finding out what students'

experiences of and reactions to assessment and feedback are in their EFL studies in a Finnish upper secondary school. These articles aim to answer the following research questions: *Do students experience assessment in their upper secondary EFL studies as (dis)empowering? What predicts disempowerment? How do assessment disempowerment as well as empowerment manifest themselves?*

The second part of the present study, Part 2, also comprises three sub-studies. These studies, which are two articles and a monograph, aim to experiment with some assessment methods as a possible way of empowering students. Hence, the second research question is: *Could some assessment methods foster student empowerment in EFL studies?* However, when discussing whether these methods could empower students or not, I will also look into *how assessment empowerment and disempowerment manifest themselves in these experiments*.

Table 3 summarises the foci, participants, research design and main methods of data collection and analysis of the sub-studies. All of these, along with the main results of each sub-study, will be discussed in further detail in the following sections. My aim is to give the reader a comprehensive picture of each sub-study. The results will be summarised once more in the discussion chapter of this study, but only from the perspective of the present compilation study.

TABLE 3 The present study and its sub-studies with their central characteristics (see Dörnyei, 2007, p. 169, for typological classifications of data collection and analysis)

Sub-study	Main research focus	Participants My role	Research design Data and data collection method(s)	PRINCIPAL and additional data analyses	Time
Part 1: (Dis)empowering assessment?					
Pollari, P. (2017a). The power of assessment: What (dis)empowers students in their EFL assessment in a Finnish upper secondary school? *Apples – Journal of Applied Language Studies, 11*(2), 147–175.	Assessment (dis)empowerment, its predictors and manifestations	N=146 upper secondary school students My role: a teacher-researcher who (had) taught many of the participants	Survey Questionnaire (the whole questionnaire) Data: Concurrent QUAN+qual	Mixed methods, QUAN → qual: QUAN: descriptive statistics, Pearson's correlation coefficient, PCA, stepwise RA Qual: close reading of open-ended answers	Questionnaire administered in March 2014
Pollari, P. (2017b). To feed back or to feed forward? Students' experiences of and responses to feedback in a Finnish EFL classroom. *Apples – Journal of Applied Language Studies, 11*(4), 11–33.	Feedback and its role in assessment (dis)empowerment	As above (n=140) My role: as above	Survey The questionnaire as above, but focusing on the 15 Likert-scale and one open-ended feedback questions Data: QUAN+qual	Mixed methods QUAN + qual: QUAN: descriptive statistics, Pearson's correlation coefficient, PCA; Qual: content analysis of open-ended answers	As above
Pollari, P. (2016). Daunting, reliable, important or "trivial nitpicking"? Upper secondary students' expectations and experiences of the English test in the	Stress, test anxiety and the Matriculation examination test	As above (n=142) My role: as above	Survey The questionnaire as above, but focusing on the 9 Likert-scale and two open-ended questions on the Matriculation Examination test	Mixed methods QUAN → qual: QUAN: descriptive statistics, Pearson's correlation coefficient Qual: content analysis of open-ended answers	As above

Matriculation Examination. *AFinLA-e. Soveltavan kielitieteen tutkimuksia 2016/n:o 9*, 184-211.			Data: QUAN+qual		
Part 2: In search of empowering assessment methodology					
Pollari, P. (2015). Can a cheat sheet in an EFL test engage and empower students? *AFinLAn vuosikirja–AFinLA Yearbook*, (73), 208-225.	Cheat-sheet test	N=101 upper secondary school students on four ENA5 courses My role: primarily the teacher but also the researcher of the study	Teacher-research: Teaching/assessment experiment, (action research) Data: QUAL+ quan Students' cheat sheets; students' comments (questionnaire); students' test results	Mixed methods: QUAL → quan: QUAL: inductive qualitative content analysis Quan: descriptive statistics, Pearson's correlation coefficient; t-test, ANOVA, ANCOVA	January-February in 2013 and 2014 (47 and 54 students, respectively).
Pollari, P. (submitted) How to make corrective feedback more learner-centred? A feedback experiment in upper secondary EFL studies in Finland	(Corrective) feedback experiment	N= 46 upper secondary school students My role: as above	Teacher-research: Teaching experiment on CF/action research Data: QUAL: Students' choices and comments (on essay papers); students' comments (questionnaire);	Mixed methods QUAL →QUAN QUAL: content analysis of student comments QUAN: descriptive statistics,	May 2014, Jan-Feb. 2016 (30 and 16 students respectively)
Pollari, P. (2000). "*This is my portfolio": Portfolios in upper secondary school English studies*. Jyväskylä: Institute for Educational Research.	Portfolios as a vehicle for student empowerment	108 from four ENA6 courses, two schools, their three teachers My role: role varied in different groups but mainly a participating observer	Illuminative evaluation case study/action research Data: QUAL Primary: Students' portfolios Secondary: students' comments on e.g. questionnaires or in interviews; teachers' comments and assessments; my observations	QUAL: close reading, qualitative content analysis and the four-field mapping of student cases	1994

5.1 Part 1: (Dis)empowering assessment?

Part 1 is composed of three articles, all of which are based on a questionnaire survey conducted in our school in March 2014. Since the data collection questionnaire and participants are the same in these three articles, I will report them first. Then, I will present some of the preliminary results of the survey that prompted all these three articles and also the whole of the present study. Only after that will I move on to the aims, research questions and results of each article.

5.1.1 Part 1 and its background

5.1.1.1 The data collection questionnaire

The data collection method of the first three articles of this study was an extensive web-based (MrInterview) questionnaire that was sent to the second- and third-year upper secondary school students of Jyväskylä Teacher Training School (*Jyväskylän normaalikoulu*) in March 2014. The questionnaire had both Likert-scale items and open-ended questions (see Appendix 1). In addition to gender and year, the students were also asked to report their English grades (the previous grade, the grade they would give themselves as well as their final grade in basic education), how many English courses they had completed, and how many different English teachers they had had during those courses. They were also asked when they would take or when they had taken their English test in the Matriculation Examination.

The first Likert-scale section (with a four-point scale) dealt with students' goal-orientation in their upper secondary school studies in general. This section was based on the questionnaire items used in the evaluation of pedagogy in Finnish upper secondary education which, in turn, were directly based on the *National Core Curriculum for upper secondary schools 2003* (see Välijärvi et al., 2009, p. 39).

The following sections concentrated on assessment practices and methodology as well as students' experiences of them in their English studies in upper secondary school. Also, I wanted to know whether assessment met the requirements set by the *Core curriculum 2003*.

Most of these questions were five-point Likert-scale items (125 items) but there were also 11 open-ended questions. The open-ended questions were optional, in other words, students could continue to answer the questionnaire even if they left all or some of them unanswered. Apart from the first Likert-scale section concentrating on students' goal-orientation, all the other Likert-scale items used a five-point scale. Even though some research experts advocate omitting the middle option and thus making the respondents take a stand, I considered it fairer and more empowering for students to let them have the right to express uncertainty (see e.g. Cohen, Manion, & Morrison, 2013, pp. 386-390). Thus, the five-point scale was also a philosophical choice. Moreover, from a research point of view, I find that if respondents are more or less coerced to

take a stand on a matter where they do not have an opinion, it can diminish the trustworthiness of the results (Cohen et al., 2013, pp. 389-390).

To make the admittedly long questionnaire more student-friendly, the questions were divided into themed sections, with instructions at the beginning of each section. This meant that students could concentrate on one topic area at a time. In addition to the background and goal-orientation questions, there were altogether seven sections: empowerment and agency in assessment processes; the frequency of different assessment methods; the usefulness of different assessment methods; the accuracy and guidance of assessment; students' experiences of and views on assessment; the Matriculation Examination; and feedback.

The questionnaire and its items drew theoretical inspiration from the extensive literature on assessment, empowerment and foreign/second language education. Several studies and their questionnaires, for instance those of Välijärvi et al. (2009), Välijärvi (1981, 1984), Syrjälä (1989) and ToLP – Towards Future Literacy Pedagogies (see Luukka et al., 2008; Tarnanen & Huhta, 2011), offered both theoretical and methodological consolidation and invaluable ideas for specific questions for this study. However, as there was no previous research on most of the topic areas of the questionnaire in this context, the nature of the questionnaire was quite exploratory and it had to be particularly designed for this study (Cohen et al., 2013, pp. 256-259; Creswell, 2014, pp. 155-160; see also Patton, 2002, pp. 192-193).

Consequently, the questionnaire was highly contextualised and tailor-made. Several items were based on the *National core curriculum for upper secondary schools 2003* and current assessment practices both in Finland and at our school. Students' ideas and comments on assessment, gathered throughout my teaching career, had shaped the questionnaire considerably, and the open-ended questions were designed so that students could elaborate upon their ideas and express them more freely.

Four research experts on educational assessment and/or foreign language education as well as three colleagues at school commented on the evolving versions of the questionnaire. In the construction of the internet questionnaire itself, careful attention was paid to the student-friendliness of its instructions, wording, order and layout, for instance. The questionnaire was tested and re-tested and commented on by a senior researcher with established expertise in student surveys and in research on upper secondary education. Finally, the internet questionnaire was piloted by four upper secondary students. Each round of testing and comments led to further refinements. All these measures were taken to ensure the content validity and reliability of the questionnaire (e.g. Cohen et al., 2013, p. 188-209; see also Messick, 1989).

5.1.1.2 Participants

Out of 199 students, 146 filled in the questionnaire (response rate 73.4%). The second-year students (79 students, i.e. 54.1% of the respondents) answered the questionnaire during one of their English lessons. The third-year students responded in their own time (67 students, 45.9% of the respondents). This data

gathering method had to be adopted with the third-year students since most of them were preparing for the Matriculation Examination and no longer had any lessons. The missing third-year students are thus students who did not volunteer to participate, or did not access the letter or the follow-up request at all (six students). Most of the missing second-year students were absent from those lessons.

Eighty-six of the respondents were female (58.9%), 60 male (41.1%). The average of their previous English grades was 8.6 (min. 6, max.10, with 4 being the lowest and 10 the highest grade in the Finnish system). So far in upper secondary school, they had studied, on average, 6.7 courses (min. 4, max.11) and had had 3.7 different English teachers (min. 2, max. 7). The first-year students were excluded from this survey as I wanted students to have adequate experience on English studies and assessment in upper secondary school.

The number of participants in the studies reported in Article 2 and Article 3 is slightly smaller (140 and 142 respectively) as only those students who answered all of the Likert-scale items in the relevant sections of the questionnaire, namely those on feedback or on the Matriculation Examination, were included. (For further information, see Articles 1 and 2.)

Although the results cannot be generalised to other schools, they give quite an accurate picture of the situation in one school at the time of the study since the respondents represent the total student population of our school well, in terms of both gender and grades[14].

5.1.1.3 Preliminary results prompting Articles 1-3

Initially, I had aimed to find out what our students' overall experience of assessment during their upper secondary school English studies was. In their opinion, did the assessment meet the requirements set by *Core curriculum 2003*? In other words, did the assessment encourage and guide their learning? Did the assessment give students feedback on their progress and develop their self-assessment skills? Were the assessment methods varied, accurate and fair? Did students know the goals and criteria? I also wanted to know whether students felt that the assessment practices allowed them any power or agency in the

---

[14] The missing students did not distort the gender ratio in any way. However, the missing students may have had slightly lower previous grades than the respondents or the total population. This is, however, difficult to verify for two reasons. Firstly, the students answered the questionnaire completely anonymously, which means that it is impossible to check which course actually was the previous course of each respondent. Secondly, the previous grade is self-reported. Nonetheless, if compared to the 'best guesses', the means of the respondents' self-reported previous grades seemed slightly higher. For instance, the mean of the grades of course ENA5 (second-year students) as reported by the school was 8.42; the respondents' mean was 8.61. The difference is somewhat similar with third-year students: the mean of their self-reported previous grade was 8.55, whereas the mean of the grades of the last compulsory course (ENA6) was 8.27. However, the mean of the final, school-leaving grade of the third-year students, as reported by the school, was higher (m=8.66) than their self-reported previous grade. In sum, it would be quite warranted to say that the self-reported previous grades of the respondents correspond with the grades of the total population quite well.

assessment processes. If they did, did the students take advantage of that power and agency and use it?

The frequency of different assessment methods used in the classrooms showed that the assessment in EFL studies at our school seemed to include the usual range: the most commonly used methods were vocabulary quizzes, course tests, listening comprehension tests, and essays, written either at school or at home. Oral tests or other oral evidence were taken into account at least in some courses. Self-assessments were used quite regularly, and occasionally they had had some effect on the grade as well. Peer assessments were rarer but not totally absent. Some 'more alternative' methods such as cheat-sheet tests or co-produced pieces (e.g. presentations, pair discussions) had been used, but not very often. In general, formative and diagnostic assessment seemed rare.

Overall, the students seemed quite happy with the EFL assessment. Nearly all the students agreed that assessment had been based on the criteria discussed at the beginning of the course and they knew why they received the grade they did, which they considered accurate and fair. A large majority also felt that the assessment had given them a good overall picture of their skills. On the other hand, over half of the students felt that assessment only stated how things were, but did not guide them forward. In short, assessment seemed to work well, even very well, as assessment *of* learning, but not as assessment *for* learning.

However, there were critical voices as well. A total of 15-20% of the students said that the assessment methods had discouraged them and undermined their willingness to learn English.

With conflicting findings on feelings of power and agency emerging from the data, I started to focus on the experiences of empowerment and particularly disempowerment in assessment: Why do students experience assessment so differently at the same school, with the same teachers? Does the data reveal any explanations for the conflicting experiences? This led to Article 1.

5.1.2 Article 1: Assessment (dis)empowerment

Pollari, P. (2017a). The power of assessment: What (dis)empowers students in their EFL assessment in a Finnish upper secondary school? *Apples – Journal of Applied Language Studies, 11*(2), 147–175.

5.1.2.1 Aims and research questions

As mentioned earlier, the aim of Part 1 of the present study is *to find out what predicts disempowerment in assessment and how assessment disempowerment and empowerment manifest themselves*. The aims and research questions of Article 1 are thus almost identical with those of Part 1:

Do students experience assessment in their upper secondary EFL studies as empowering or disempowering?
What predicts (dis)empowerment?
How do assessment disempowerment and empowerment manifest themselves?

5.1.2.2 Data analysis

Principally, all the data provided by the questionnaire was analysed quantitatively. Firstly, the descriptive statistics (e.g. frequencies, means and standard deviations) were calculated. Then, in order to reduce the dimensionality of the rather large pool of data, a varimax-rotated principal component analysis (e.g. Brown, 2009; Jokivuori & Hietala, 2007, pp. 89-95; Metsämuuronen, 2009) was conducted to summarise the variance of each section/topic area of the questionnaire into a few principal components. These analyses resulted in principal components that were transformed into a total of 28 sum variables (see Appendix 1 in Article 1). Next, to have a general view of which sum variables might correlate with one another, a correlation matrix of these 28 sum variables as well as the background variables of gender, year and the previous grade was calculated.

One of the sum variables was named *Disempowerment* as its items involved the central features or results of disempowerment: assessment is not seen as a good, beneficial factor facilitating learning, but as something that drains the students' power, resources and motivation. In other words, it refers to a lack of perceived control as well as low self-efficacy and motivation, which are features of the intrapersonal component of psychological empowerment (Zimmerman, 1995, 2000).

As the first step, students' differing experiences of assessment (dis)empowerment were analysed and grouped with the help of means and standard deviations. Secondly, a stepwise regression analysis (e.g. Jokivuori & Hietala, 2007, pp. 39-55; Metsämuuronen, 2009) was run to find out which variables were the strongest predictors of disempowerment.

In order to add depth and to illustrate "what the individual variation means" (Patton, 2002, p. 15), qualitative data and analysis was also used in the third approach, i.e. in the illuminative close-ups of three information-rich student cases (see Patton, 2002, p. 242). Methodologically, the case analyses are based on mixed methods that complement each other (Lund, 2012): the qualitative data is used both to check the accuracy and validity of the quantitative findings and to further explain them in order to provide as comprehensive an analysis as possible (Creswell, 2014, pp. 215-225). First of all, the cases had to qualify in their category (disempowered/non-disempowered/empowered) on the basis of the quantitative analysis of their numerical answers. Next, the open-ended answers of each of these students who qualified were carefully read, analysed and compared with one another through close reading, which Brummett (2010, p. 25) characterises as follows: "Close reading is a mindful, disciplined reading of an object with a view to deeper understanding of its meanings" (see also Thomas, 2006). Then, the most information-rich cases – "those from which one can learn a great deal about issues of central importance" (Patton, 2002, p. 46) – were purposefully selected. Students' open-ended answers are cited extensively to allow the reader a sort of *thick description*, i.e. giving enough authentic data and 'voice' for the reader to

be able to judge the trustworthiness of the description and get "a sense of the cognitive and emotive state" of each student (Ponterotto, 2006, p. 547).

5.1.2.3 Findings

The findings of Article 1 showed that students were, on average, quite happy with their assessment. However, the study also showed that different students had very different experiences of assessment. Consequently, two opposing groups of students were formed on the basis of the *Disempowerment* sum variable. These were *the Disempowered* (n=21) and *the Non-disempowered* (n=18). The name *Non-disempowered* is admittedly very clumsy, but as the Disempowerment sum variable did not include any items concerning power given to students or students taking active charge of their decision-making power, i.e. empowerment, I could not call them *empowered*.

When comparing the means of the sum variables of the Disempowered and the Non-disempowered groups with the means of *the whole respondent group*, a few sum variables or topic areas showed clear differences. For instance, an individual variable, "*Assessment causes me anxiety and stress*", as well as the sum variable of *Stressful and discouraging assessment* divided opinions between these three groups. Also, students' responses to feedback, its utility, importance and role in learning seemed to differentiate between the groups. The groups seemed rather different in their sense of their ability to analyse their own strengths and weaknesses. Of these three groups, the Disempowered also considered the assessment methods that had been used to be the least varied and good, thought that course tests had had too much weight, and regarded assessment as the least accurate or fair. They also wanted to have more influence on the assessment methodology and criteria than the other two groups.

However, when comparing the sum variable concerning *Given empowerment* (e.g. whether the goals and assessment methodology had been discussed at the beginning of the course, and whether students had been given a chance to influence them), the difference became noticeably smaller.

Next, to see which of these factors or sum variables might predict Disempowerment best, a step-wise regression analysis was run. It resulted in an eight-step model, which altogether accounted for 59.3% of the variance. The five most significant predictors of disempowerment, accounting together for over 50% of the variance, were *Stressful and discouraging assessment*, *No pressurised or high-stakes tests, Grades over feedback, Good and versatile assessment*, which related negatively with disempowerment, and *Inadequate feedback*. In other words, disempowered students felt both stressed and demotivated by the assessment. Test anxiety was a clear predictor: no high-stakes tests but softer, i.e. less pressurised and more formative assessment was called for. The current assessments and their methodology were not considered good and varied enough, and they did not give students a fair chance to show all their skills or knowledge. Furthermore, feedback had failed to serve its purpose of facilitating learning. Feedback was either overshadowed by grades and therefore feedback was not given much attention and was considered less important than grades or

scores, or students had not received enough feedback to guide and encourage their learning.

In addition, *Success-oriented goals* as well as *English for life, not for the Exam* both predicted *Disempowerment* negatively. The last predictor in this model was *Personality affects assessment*. Thus, students' ownership of their English studies as well as their goal-orientation played a role in assessment (dis)empowerment.

These findings are shown in visual form in Figure 8 below.

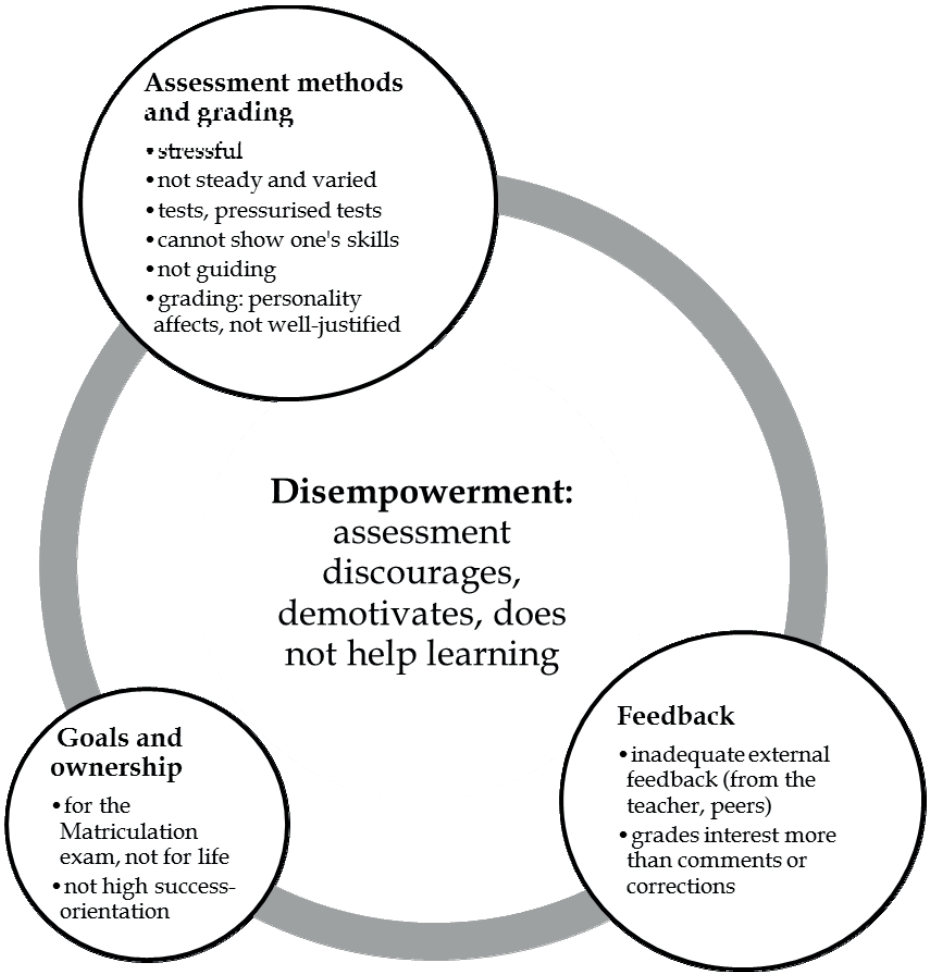

FIGURE 8 Assessment disempowerment and student experiences predicting it.

Finally, as the third approach to the research questions, three student cases were selected and reported. The disempowered student, a second-year female student, exhibited strong assessment anxiety, which also seemed to have led to diminished self-efficacy and motivation to study English. She would have liked to have more power to decide on the assessment methodology and she wanted

softer, more formative assessment methods that would not cause so much stress and pressure. In contrast, the non-disempowered student, a second-year male student, manifested no assessment stress or anxiety: he was very self-confident and believed in his English skills. He therefore did not care what assessment methods were used and did not want to have a say on assessment. The final case was deemed an empowered student as he had used the power given to him to influence assessment. He was a third-year male student, who had quite clear opinions on assessment, its role and function as well as different assessment methods.

All the three ways of analysis used in this study resulted in the same conclusions regarding disempowerment. First of all, assessment seemed to cause the disempowered students a great deal of anxiety and stress. The disempowered students feared high-stakes testing, such as the Matriculation Examination, but also course tests or exams. Thus, test anxiety (see e.g. Cassady, 2010; Hembree, 1988) seemed to have a connection with assessment disempowerment. All in all, the current assessment methodology was not considered good and diverse enough, and it did not give the students a fair chance to show all their English skills or knowledge. That could, in turn, contribute to the loss of self-efficacy and motivation in their English studies. Therefore, the disempowered students would have liked to have more power to influence the assessment methodology as they hoped for 'softer', i.e. less pressurised and more formative, assessment methodology – in other words, they wanted *assessment for learning* in addition to assessment of learning. Nevertheless, they did not seem to use the power they had already been given to influence the assessment methodology. In other words, they either did not perceive that they had any power, or they did not believe in that power or their own ability to use it. Furthermore, the descriptive statistics showed differences between the previous English grades – as well as gender and year – of the disempowered and non-disempowered student groups even though none of these background variables predicted disempowerment in the stepwise regression analysis.

Secondly, feedback and how it was experienced played a significant role. The feedback that the disempowered students had received had not met their expectations and needs. Either they had not had enough feedback, or it had not been helpful. In some cases, the dissatisfaction had resulted in students rejecting teacher comments and concentrating on the grades only. Thirdly, the disempowered students did not seem to feel ownership of their English studies: they seemed to study English rather for the sake of the grades, or the Matriculation Examination, than for their own goals. On the other hand, they did not seem to have a strong success-orientation, either. In general, they exhibited lower scores in all goal-orientation sum variables on average than other students.

The empowered and non-disempowered students seemed to acknowledge the power they had been given in the assessment process. The difference between the two student groups was in their active involvement (Zimmerman,

1995, 2000) in that process: whereas the empowered students had actively used the power they had, the non-disempowered students had decided not to use it. Nonetheless, they had both made their own decisions on the matter, based on their willingness to use – or not to use – that power.

5.1.3 Article 2: Feedback and assessment (dis)empowerment

Pollari, P. (2017b). To feed back or to feed forward? Students' experiences of and responses to feedback in a Finnish EFL classroom. *Apples – Journal of Applied Language Studies, 11*(4), 11–33.

The preliminary findings of the assessment survey had indicated that half of the students felt that assessment did not really help or guide their learning forward. Also, Article 1 indicated that feedback played a significant role in students' assessment (dis)empowerment. Hence, Article 2 focuses on *the communication of the assessment judgement*, and *feedback*, in particular, and students' responses to feedback.

5.1.3.1 Aims and research questions

One of the aims of the article was to explore how students experienced feedback in their EFL studies and assessment. Within the focus of the present study, the most pertinent research questions of Article 2 are:
What kinds of responses did students have to feedback?
How did these responses relate to assessment empowerment and disempowerment?

5.1.3.2 Data analysis

Article 2 used the data collected by the questionnaire discussed above. However, instead of the whole questionnaire, this study concentrated on 15 Likert-scale items and one open-ended question dealing with feedback.

First of all, descriptive statistics of the 15 Likert-scale items were calculated. Independent samples T-tests were also conducted to test the statistical significance of the differences of means by gender and year.

As described above (see Chapter 5.1.2.2), the varimax-rotated principal component analysis was initially run to summarise each topic area into a few principal components, which, in turn, were transformed into sum variables. This data and these variables were used in further analysis. Pearson correlation coefficients were calculated to analyse the correlations between variables. Students' gender, year and previous English grade were used as independent variables. In addition, qualitative content analysis (Tuomi & Sarajärvi, 2009, pp. 103-119) was used to analyse the answers to the open-ended question, which offered additional, illuminative data for this sub-study.

5.1.3.3 Findings

In the context of the present study, the aim of Article 2 was to discover what kinds of responses to feedback students had and whether their responses were related to empowerment.

According to Article 2, a vast majority of students wanted more feedback on both their language skills and learning skills. At the same time, most of the students seemed content with the feedback they had received and found it helpful and motivating. There was, nonetheless, also a significant minority of students who were not completely happy with the existing feedback and gave several suggestions on how to improve it.

Firstly, the students wanted to have feedback that would not only refer to the present state of their skills but would improve their future performance and learning. Secondly, the students wanted feedback that was individual and personalised, and so clear, concrete and specific that they knew what it meant and what they should do. They also wished to have more feedback during the course, not only at the end of it, so that they could act upon it. They wanted constructive and balanced feedback, and more varied methods of giving feedback. In sum, they wanted feedback that would function as *assessment for learning*.

However, as pointed out in earlier research, (e.g. Hattie & Timperley, 2007; Wiliam, 2012), the effectiveness of feedback did not seem to depend only on feedback itself, but also on students' different responses to it. The principal component analysis extracted four principal components, which were turned into sum variables. They were *Guiding feedback*, *Self-feedback*, *Inadequate feedback* and *Grades over feedback*. Feedback could be highly appreciated and work very well, as was the case in *Guiding feedback*. Alternatively, feedback could work well, but the feedback given by teachers or peers was unnecessary because of the students' good self-assessment skills, as seen in *Self-feedback*. On the other hand, feedback could, for one reason or another, also fail. *Inadequate feedback* did not meet all students' needs for external feedback, which is what they valued and craved. Alternatively, as was the case with *Grades over feedback*, feedback in the form of teacher comments or corrections was not much valued or welcomed.

These responses to feedback manifested clear differences in the experiences of empowerment and disempowerment related to assessment. With *Guiding feedback* and *Self-feedback*, assessment in general was considered empowering. Assessment was seen as varied, appropriate and just, and it seemed to serve students well. Therefore, assessment empowered students in their learning process: it gave them power and useful resources to conduct their studies. By contrast, with *Inadequate feedback* and *Grades over feedback* assessment was experienced as a disempowering factor that had not succeeded in motivating, guiding and helping students in their learning, nor had it given them a chance to show all their skills in English.

Whereas previous success in English studies did not correlate with any of these four feedback responses, gender may have had an influence on *Inadequate feedback* and *Grades over feedback*, both of which also correlated with experiencing assessment disempowerment. Female students manifested a stronger tendency towards *Inadequate feedback*. Male students, on the other hand, showed a stronger preference for *Grades over feedback* than female

students, as also did second-year students. Then again, third-year students seemed to be more capable of *Self-feedback* or more willing to give it, and they also experienced *Guiding feedback* more than second-year students.

In sum, how students experienced feedback and reacted to it seemed to have a clear connection with assessment (dis)empowerment.

5.1.4 Article 3: Stress, anxiety and the Matriculation Examination

Pollari, P. (2016). Daunting, reliable, important or "trivial nitpicking"? Upper secondary students' expectations and experiences of the English test in the Matriculation Examination. In A. Huhta & R. Hildén (eds.) *Kielitaidon arviointitutkimus 2000-luvun Suomessa. AFinLA-e. Soveltavan kielitieteen tutkimuksia 2016/n:o 9*, 184-211.

As we have seen, Article 1 showed that assessment seemed to cause the disempowered students stress and anxiety. Test anxiety (see e.g. Cassady 2010; Hembree, 1988) seemed to play a role in assessment disempowerment: the disempowered students disliked pressurised testing, and particularly high-stakes testing, such as the Matriculation Examination. Furthermore, the disempowered students seemed to study English more for the Matriculation exam than for their own future goals. The disempowered students also felt that assessment methods did not give them a fair chance to show all their English skills or knowledge and did not take all areas of language skills into account. Thus, in order to shed more light on this correlation between disempowerment with test anxiety and somewhat extrinsic goal-orientation, Article 3 focuses on students' experiences and expectations of the Matriculation Examination and its English test, its importance, validity, reliability as well as the anxiety caused by it.

5.1.4.1 Aims and research questions

For the purposes of the present study, the aim of Article 3 is to answer the following research questions:

How important a goal is the English test in the Matriculation Examination for upper secondary English studies?
Does the English test in the Matriculation Examination bring on test anxiety?
Do students consider the Matriculation exam test a valid and reliable way of showing their English skills?

5.1.4.2 Data analysis

The data explored in this article comes primarily from the Matriculation Examination section of the questionnaire and its Likert-scale items as well as from the goal-orientation questions of the same questionnaire (see Appendix 1). That data was analysed quantitatively using descriptive statistics. Students' gender, previous (self-reported) grade as well as whether or not they had taken at least a part of the English Matriculation Examination test were used as independent variables. Independent samples T-tests were conducted to test the

statistical significance of the differences of means of gender and the test-taking. Pearson correlation coefficients were calculated to analyse the correlations between variables.

There were also two open-ended questions dealing with the Matriculation exam in the questionnaire. The answers to these questions offered qualitative data which was analysed through content analysis (e.g. Patton, 2002, pp. 452-455; Tuomi & Sarajärvi, 2009, pp. 103-119). The content analysis started as inductive analysis "discovering patterns, themes, and categories in one's data" (Patton, 2002, p. 453). However, as the emerging categories and themes, particularly with the open-ended answers to Question 9, seemed to match the quality criteria for assessment presented in the literature, the second round of content analysis turned into deductive content analysis (e.g. Patton, 2002, pp. 452-455; Tuomi & Sarajärvi, 2009, pp. 113-117). At that stage, the data was re-categorised according to the already existing quality characteristics of validity, reliability and fairness.

5.1.4.3 Findings

At the beginning of the questionnaire, the students were asked how much some goals had influenced their studies in upper secondary school. Over 85% of all the respondents said that getting a good mark in the Matriculation Examination had been a goal that had affected their studies either very much or quite a lot. The Matriculation Examination thus seemed to be an important goal – and even more important than a good upper secondary school certificate. However, when asked why they were studying English, the results changed. While about 30% of the respondents said that a good grade in the Matriculation Examination was the most important goal of their upper secondary English studies, approximately 55% of the respondents disagreed. Around 90% of all the respondents said that they were studying English primarily for their own future and not for the Matriculation Examination.

The second research question dealt with test anxiety, and the results of this study seem clear: the Matriculation Examination caused some fear or anxiety in about 60% of the respondents. However, the actual experience of taking the test did not seem quite as bad as students had anticipated or expected. Female students were clearly more susceptible to Matriculation Examination anxiety, and among them, the anxiety was significantly higher (m=3.84) than among male students (m= 2.70)**.** Students' previous grades did not correlate with anxiety (r= -.125).

Also, approximately one student in four mentioned either stress or anxiety in their open-ended answers to Question 9: *"What do you think of the Matriculation Examination in Advanced English? What kind of thoughts/emotions does the examination give rise to?"* Eight of the students who mentioned stress or anxiety had already taken the test. Their stress or anxiety was mostly linked to the test-taking situation or with the high stakes of the exam. The students who had not yet taken the test mentioned anxiety or apprehension more often. Their anxiety or fear ranged from slightly anxious excitement to strong fear that had affected their study plans. On the other hand, ten students were confident of

their skills and not worried or anxious about the test. In sum, the expectations seemed somehow stronger, either more anxiety-ridden or more relaxed and confident, than the actual experiences. Nevertheless, approximately 60% of all the respondents, and over 70% of the female respondents, said that the English test of the Matriculation Examination frightened them at least to some extent.

The third research question focused on students' views of the validity and reliability of the Matriculation Examination English test. There were two Likert-scale items that dealt with its reliability. According to these items, over half of the students who had already taken the Matriculation exam English test thought that it was not a reliable way to show their skills and considered teacher-based assessments a more accurate assessment of their skills. Not everybody agreed with them, though, and, with 30% of these students undecided, students did not seem totally convinced that teacher-based assessment would necessarily be much better. Male students in general seemed to trust the reliability of the Matriculation exam more than did female students. Students' previous English grade did not correlate with these two items. However, students' scepticism about the reliability of the Matriculation exam English test correlated with Matriculation Examination anxiety.

Many students seemed quite critical of the validity of the Matriculation test in their open-ended answers. Its content validity, or content relevance and coverage, was not regarded as particularly good because speaking was not tested. Furthermore, detailed knowledge related to grammatical exceptions or rare vocabulary was considered irrelevant to real-life communication skills. The level was also seen as too demanding when compared with the goals and syllabi of Advanced English courses.

The reliability of the test was not considered to be very high, either. Students who had already taken the test mentioned several threats to reliability. Deliberately tricky questions were considered the greatest threat (see also Anckar, 2011; Huhta, Kalaja, & Pitkänen-Huhta, 2006). The pressurised test-taking situation and luck were also regarded as threats to the reliability of the test. For these reasons, students did not seem very convinced that they could show their English skills very reliably in the Matriculation exam English test. However, although not necessarily happy with the test and its format, the students seemed to consider the scoring and grading of their test papers quite fair.

To summarise, even though students said that they were studying English primarily as a skill for life, the English test in the Matriculation Examination seemed to be an important goal in their English studies. The Matriculation examination English test also appeared to cause considerable stress and anxiety to a great number of students, and particularly to female students, mostly because of the pressurised test-taking situation and the high-stakes nature of the examination. Furthermore, students seemed rather critical of the validity and reliability of the test as a test of their English skills because of its limited content coverage and relevance as well as its construct-irrelevant variance.

However, as seen above in Article 1, unlike the majority of students, the disempowered students seemed to be studying English more for the Matriculation examination than for their own future. The fact that the Matriculation examination certificate or grade was very important for them *per se,* combined with both test anxiety in pressurised test situations and frustration that the assessment methodology did not allow them to exhibit what they considered their true skills, all seemed to have caused them severe stress and thus also contributed to assessment disempowerment in general. However, fear or anxiety about the Matriculation Examination did not alone predict assessment disempowerment.

5.2 Part 2: In search of empowering assessment methodology

As the previous articles in Part 1 indicated, *the actual impact* of assessment could range from empowering to disempowering. Part 2 focuses on *the intended impact* of assessment. The disempowered students felt that the existing assessment methodology was too limited and did not enable them or allow them to show all their English skills. The aim of Part 2 of the present study was therefore *to experiment with some alternative assessment methods to explore whether they could empower students in the assessment process*. The assessment methodologies experimented with in these studies are a cheat-sheet test, more personalised corrective feedback and, finally, portfolio assessment.

5.2.1 Article 4: A cheat-sheet test

Pollari, P. (2015). Can a cheat sheet in an EFL test engage and empower students? *AFinLAn vuosikirja–AFinLA Yearbook*, (73), 208-225.

5.2.1.1 Aims and research question

This article reports a teaching/assessment experiment focusing on cheat-sheet tests. The experiment aimed to empower students in the sense of giving them more agency and power in the design and collection of assessment evidence (see Figure 3 in Chapter 2), and more specifically, in *how* evidence is collected. The research question of this article is:
Could a cheat-sheet test empower students in their EFL studies?

5.2.1.2 Participants

A total of 101 students (61 female and 40 male students, aged 17-18) took part in this study in 2013 and 2014 (47 and 54, respectively). They were second-year students, on their penultimate compulsory English course (ENA5, the culture course).

5.2.1.3 Data and data analysis

Article 4 relies on three different sources of data: the students' cheat sheets, their comments written on a questionnaire as well as their test results.

The contents of the cheat sheet were limited to grammatical elements of the course. The size of the cheat sheet was restricted (one side of A4) so that students would have to process and summarise the information they selected. They made the cheat sheets out of class and they could make them as they wanted (hand-written or typed, with colours, coding and pictures, or not, for instance). The only requirement was that each student made their cheat sheets themselves, i.e. they were not to copy anybody else's cheat sheet. After the test, the students' cheat sheets were collected for analysis.

Students' comments were collected as data by using an open-ended questionnaire in Finnish (see Appendix in Article 4). Comments were collected both before the test and immediately after it. A final round of student comments was collected after I had handed the marked and graded tests as well as the cheat sheets back to the students.

The students' test results were the third source of data. The test comprised both grammar and reading comprehension but the cheat sheet was made only for the grammar part of the test. The grammar exercise was a traditional multiple-choice exercise with 40 items scoring 40 points. Although rather behaviouristic, some of its items required processing two grammatical constructs at once (e.g. articles and capital letters) and, admittedly, the exercise was quite detailed and challenging. The maximum score on the reading comprehension (RC) part was also 40 points. It consisted of a traditional multiple-choice reading comprehension exercise that was based on an authentic film review (20 points), an interpretation exercise (translate/explain five out of the six underlined sentences in the film review in Finnish, 15 points) as well as a short written response to the film review (in English, 5 points).

Inductive qualitative content analysis was used to analyse the students' cheat sheets and comments: after several readings, the cheat sheets as well as the comments were placed in categories that emerged from the data (e.g. Tuomi & Sarajärvi, 2009, pp. 108-113). However, not all of the students volunteered answers on every question in the questionnaire, and some answers were also rather vague. When a comment proved difficult to categorise, I consulted a second reader, an experienced educational researcher.

The students' test results and previous grades were also used for additional quantitative analyses. In order to try to establish whether the cheat sheet had had any measurable effect on test results, the quality of the cheat sheet was first compared with the grammar results (see Table 2 in Article 4). The comparison of means and Pearson's correlation coefficient were used to analyse the test results; then, a t-test, one-way analysis of variance as well as analysis of covariance were used to analyse the statistical significance of the differences of the means (see e.g. Jokivuori & Hietala, 2007; Metsämuuronen, 2009).

5.2.1.4 Findings

The great majority of students had a positive attitude towards the cheat-sheet test when the idea was introduced to them. Only five students expressed dislike of the idea.

After the test, the cheat sheets were collected and analysed. Out of 101 students, a total of 92 students had made a cheat sheet. Nine students, seven male and two female students, had decided not to make a cheat sheet for the test. With one exception, these students formed quite a homogenous group on the basis of their prior grades: the mean of their previous grades was clearly higher than that of those who made the cheat sheet.

Immediately after the test, the students were asked if the cheat sheet had been beneficial in the test. Although feeling almost unanimously that the cheat sheet had been helpful in one way or another, for instance by decreasing their test anxiety, the students did not believe its impact on their actual test results to be strong. Overall, the students felt that the cheat sheet had rather helped them to study and learn better than offered them the right answers in the test. For instance, over a third of the students said they had used the cheat sheet mainly for checking some of their answers and some did not use their cheat sheet at all in the test situation even though they had one with them.

All in all, a thorough, well-prepared cheat sheet seemed to result in a small gain in test results on average. The bigger gain, however, was in students' experiences. By far the majority of students still said that the cheat sheet had been beneficial because it had improved their studying and learning as well as their recollection. Some also said that it had made them prepare for the test better than usual. Fourteen students, all girls, mentioned feeling less insecure or stressed because of the cheat sheet. And finally, most of the students *liked* the cheat sheet, and quite a few would have liked to use it more often in assessment at school.

Nonetheless, cheat sheets did not suit every student. Out of the 92 students who had made a cheat sheet, six said the impact of the cheat sheet had been non-existent or negative, mainly because of the difficulty of the test, lack of time or lack of preparation. Two of those six students also considered that the cheat sheet had a negative effect on learning.

In sum, a cheat-sheet test was not a universal panacea, and it did not suit every student's learning or testing preferences. Nevertheless, according to the findings of this limited study, a cheat-sheet test proved to be *one* learner-friendly assessment method that most students found beneficial for learning and studying. It also reduced test anxiety considerably. Furthermore, a cheat-sheet test allowed students several opportunities for agency when preparing for the test as well as when taking the test. They could decide, for instance, whether to make a cheat sheet or not, what the content of their cheat sheet would be as well as whether to use it in the test situation or not. In sum, a cheat-sheet test can empower students far more than traditional closed-book tests do. In addition, the cheat-sheet test brought together studying, learning and assessment and also enabled the formative use of summative assessment (Black et al., 2003, pp. 53-57). Thus, the cheat-sheet test could be characterised as a method suitable for both *assessment for learning* and *assessment as learning*.

5.2.2 Article 5: Individual choice on corrective feedback

Pollari, P. (submitted for review). How to make corrective feedback more learner-centred? A feedback experiment in upper secondary EFL studies in Finland.

The following article reports two small-scale teaching experiments that focus on the *communication of the assessment judgement* and, more precisely, on corrective feedback. Corrective feedback is one of the most frequent forms of feedback that foreign language teachers give to their students. Broadly speaking, it can be divided into two major categories in terms of both its medium and its treatment of errors. Firstly, corrective feedback can be *oral or written*. Secondly, *direct* corrective feedback means that the teacher corrects the students' mistakes, whereas *indirect* corrective feedback means that the teacher indicates the mistakes but does not correct them.

5.2.2.1 Aims and research questions

Giving students the chance to choose an individual way of receiving (corrective) feedback on their essays, these experiments were designed to find out what kind of feedback and error correction would serve students best in their own opinion, and why. The research questions of this article are:
Could individual choice on (corrective) feedback make feedback more learner-centred?
Could individual choice enhance student empowerment?

5.2.2.2 Participants

The first experiment took place with 12 female and 18 male second-year upper secondary students, aged 17-18, in May 2014. The course was an advanced, non-compulsory course (ENA7). The second experiment also took place in an advanced, non-compulsory EFL course (ENA9) with 16 third-year students (13 female and three male students, aged 18-19) in January and February 2016. Altogether, the number of students who participated was 46. However, as the students wrote two essays in the second experiment, the number of feedback choices was bigger than the actual number of students.

5.2.2.3 Data and data analysis

The data of this article consists of students' (corrective) feedback choices and the reasons the students gave for their choices. In addition, the students gave feedback on the experiment with the help of a simple questionnaire.

The individual choices were calculated and presented quantitatively. Students' comments and reasons for their choices offered qualitative data and were analysed through content analysis. However, as the experiments were not intended as rigorous research experiments, the data analysis methodology is quite rudimentary.

5.2.2.4 Findings

The quantitative findings of the experiments showed clearly that students had individual needs and preferences. Even though students were given a rather

limited agency, i.e. the choice between written and oral feedback and direct or indirect error treatment, they came up with a variety of feedback options. Altogether, students chose written feedback 44 times, and oral 16 times. Direct corrective feedback was chosen 39 times and indirect 20 times. Some students left their preferences unknown or wanted both oral and written feedback. Furthermore, the qualitative analysis of the students' reasons for choosing their preferred methods indicated that the reasons were highly individual.

At the end of the experiments, students were asked to give feedback on the experiment. Unfortunately, I did not get comments from every student. However, none of the 36 students who did respond considered the personalised feedback they had got in the experiment to be less useful than prior feedback practices, with the teacher deciding the one and only way of giving corrective feedback to all students. Seventeen students said that the feedback had been as useful as the feedback they usually received. What is worth noting is that these seventeen students had all chosen written feedback, eleven with direct and six with indirect correction. Nineteen students, with differing feedback choices, said that feedback in the experiment had been more beneficial for them.

Hence, although limited only to agency in choosing the method of corrective feedback, the experiments attest that students find even a small increase in their power beneficial, as they can choose the method that they experience as suitable for themselves. Choosing for themselves the most suitable feedback method may also enhance students' engagement with feedback and help them to see feedback as assessment meant to improve their learning *(assessment for learning)* rather than just as assessment stating their results *(assessment of learning*) or their shortcomings (*error* correction).

5.2.3 Monograph: Portfolio

Pollari, P. (2000). "*This is my portfolio*"*: Portfolios in upper secondary school English studies.* Jyväskylä: Institute for Educational Research.

Generally speaking, the last sub-study can be regarded as an illuminative evaluation case study. Although the monograph was written a long time ago, I have decided to include it in the present study for a number of reasons. First of all, the portfolio study introduced and used portfolios as an innovative, learner-centred tool for learning and assessment in Finnish EFL education at a practical, grassroots level. Secondly, it introduced the concept of empowerment into the Finnish discussion on foreign and second language education. It sparked several other portfolio programmes in language education, also outside upper secondary school education. Thirdly, it influenced my conceptions of assessment, its purposes and possibilities and thus also the assessment practices in my work. And, most vitally, the portfolio approach gave students agency and decision-making power at *every phase of the assessment process* (see Figure 3), from defining the purpose of the portfolio to its actual impact. In sum, it finds its logical place in this study, and in Part 2, both from the perspectives of its

theoretical background and its goal as an assessment experiment aiming to foster student empowerment. Hence, it deserves to be revisited at this point.

5.2.3.1 Aims and research questions

Originally, the portfolio study had two purposes. Firstly, as a pedagogical innovation, it aimed to try out and develop portfolios in Finnish foreign language education. Its first research goal was therefore to see how the portfolio programme proceeded and progressed in the classrooms. Secondly, and more pertinently for the present study, the aim of this study was to foster students' active and responsible role both in learning and in assessment. Hence, its research question was: *Did the portfolio programme foster students' empowerment?*

5.2.3.2 Participants

The portfolio courses were carried out in two schools in Jyväskylä, the Teacher Training School of the University of Jyväskylä and Kesy Upper Secondary School (Kesyn lukio) in the spring term of 1994. Three teachers of English, two from the Teacher Training School and one from Kesy, participated in the study with one or two of their second-year-student groups (ENA6 course). The pilot group had consisted of eight students, and the actual portfolio experiments started out with 108 students (58 female, and 50 male; 68 of the students were from Kesy and 40 from the Teacher Training School).

The participating groups varied considerably in size. Both teachers from the Teacher Training School had only one second-year A-English (English as the first foreign language) group each. Their groups consisted of 22 and 18 students. The teacher from Kesy participated in the study with two large groups of 30 and 38 students, both of which had their portfolio course in the same period of the school year. The decision to experiment with groups of different sizes as well as with one teacher having a large number of students at the same time was, however, deliberate: we wanted to test whether class size and the number of students or groups would have an effect on the introduction and implementation of the portfolio approach.

5.2.3.3 Data and data analysis

The portfolio study relied on several sources of data. The primary source of data was the students' portfolios. Each portfolio included not only the student's pieces of work but also a prologue (i.e. an introduction to the portfolio), the student's self-assessments of the selected pieces, and an epilogue (i.e. a summative reflection and evaluation of the portfolio project). Furthermore, a working log documenting their working process was required.

Secondary sources of data included other sources of students' own comments (e.g. questionnaires and interviews with selected students) as well as the teachers' final assessments, comments and observations. My observations as a researcher were also used as secondary data.

All these sets of data were used with a slightly different focus for different purposes. For instance, when explaining the portfolio process, a heavier focus was on the teachers' and my observations during the process as well as on the

students' comments on the process written in their working logs or elsewhere in their portfolios. When analysing and mapping all the portfolio cases (i.e. students) on the four-field map, the main focus was on the students' portfolios but attention was also paid to the teachers' assessments and comments. In the portrayal of selected student cases, the focus was on students' portfolios but other sources of their comments were used as well.

However, the main data analysis method remained basically the same throughout the whole research project. The data was analysed qualitatively. The students' portfolios were read through, holistically, several times in order to know the data in thorough and "intimate ways" (Marshall & Rossman, 1989, p. 11). Then, in order to place each case on the analysis map, each portfolio was read once more. All the other information gathered during the programme was used in the analysis as well. Thus, data triangulation (e.g. Denzin, 1978; Janesick, 1994) was applied throughout the whole analysis process. (For more information on the data analyses, see Pollari, 2000, pp. 93-94; 165-176; 183-188.)

5.2.3.4 Findings

The research interest of this sub-study was to see whether the portfolio programme could foster students' empowerment. The answer to this research question is, in brief, yes: as a pedagogical innovation, the portfolio programme proved interesting, rewarding and also empowering, for various reasons.

Firstly, compared to other, more traditional courses, the portfolio approach empowered students by giving them a great deal of *power* and *agency* at several stages of the assessment process (see Figure 3). First of all, although the overall purpose of the portfolio was to combine studying and assessment in this course, and therefore it had both *a formative and a summative purpose* (although these terms were not widely used in the 1990s), students could give their portfolios other, more personal purposes such as showcasing their own interests, expressing themselves or exploring new issues. Also, at the phase of designing and collecting assessment evidence, students could rather freely decide on the specific topics and forms of their own pieces of work, so long as they were within the guidelines requiring that the pieces exhibited a variety of both topics and formats. In addition to the content and format of their pieces of work, students also had a great deal of power in directing their working processes. Of the approximately 35 lessons of the course, students were expected to attend about 15 lessons. Otherwise, they were free to decide on their own working pace, schedule and also place. Furthermore, the students decided which pieces to select for their final portfolio to be assessed for their course grade. They were also asked to give their own self-assessment, where they could state their own criteria, and they were invited to suggest a course grade for themselves even though they did not have any legitimate power over their grades. Therefore students were empowered also at *the interpretation and production of the assessment judgement*.

Secondly, in terms of empowering students, the students were given *resources* to support their self-directed and empowered learner role. At the very beginning of the course, they were offered ideas, examples and materials to

help them to come up with the topics of the work. Moreover, in addition to hearing about and discussing all the requirements and assessment criteria of the portfolio course in class, the students were also given this information in writing. In other words, they knew from the very beginning how their work would be assessed and they could go back to the criteria or requirements whenever needed. Students could also consult their teacher whenever they wanted to. Additional support was given during class conference lessons. There students were supposed to give each other peer feedback: this, however, did not always work as some students seldom brought any work to these lessons.

Thirdly, students had several opportunities for *self-efficacy*. In addition to choosing their own topics and working methods, and giving their own criteria and self-assessment for their work, they also presented their work to the whole class at the end of the course. During this portfolio celebration, which took place during the time allotted to the course exam, students could choose whether to present their work in Finnish or English and whether to present just one or two pieces of work, or the whole portfolio. Thus, at the phase of communication they could express their ideas and also get feedback from other students. Finally, portfolios offered opportunities for self-efficacy also at the level of the intended and actual use of the assessment. For several students, portfolios offered an opportunity to pursue their own interests or to try out self-directed and independent studies, and for quite a few, portfolios offered a vehicle for self-expression and creativity. To some students, their own use and purpose of the portfolio surpassed the intended use of the portfolio as simply a vehicle for formative and summative assessment.

However, the portfolio programme was not problem-free. Both the students' and teachers' new roles demanded effort and adjustment and thus also caused some friction at times. Perhaps the greatest problem was that some students were not ready or willing to assume such an active but also responsible role as was expected of them. Although the teachers offered a great deal of support, the change was quite abrupt and radical. As the course lasted only approximately six weeks, there was no opportunity for any longer-term learner training and a more gradual power-shift. However, the majority of students embraced their decision-making power and freedom; students also clearly differed from one another in their preparedness and willingness to assume a more empowered learner role.

In order to analyse the individual portfolios as a vehicle for the students' empowerment, a total of 101 portfolios were analysed with the help of a four-field map, below (see Figure 8). With 108 students starting the course, there were thus seven missing portfolios. Two students discontinued their upper secondary studies during the course, never submitting their portfolios; three male students handed in their portfolio so late the following term that they were not included in the study; and an additional two portfolios were submitted on time but unfortunately misplaced before the four-field analysis of the portfolios.

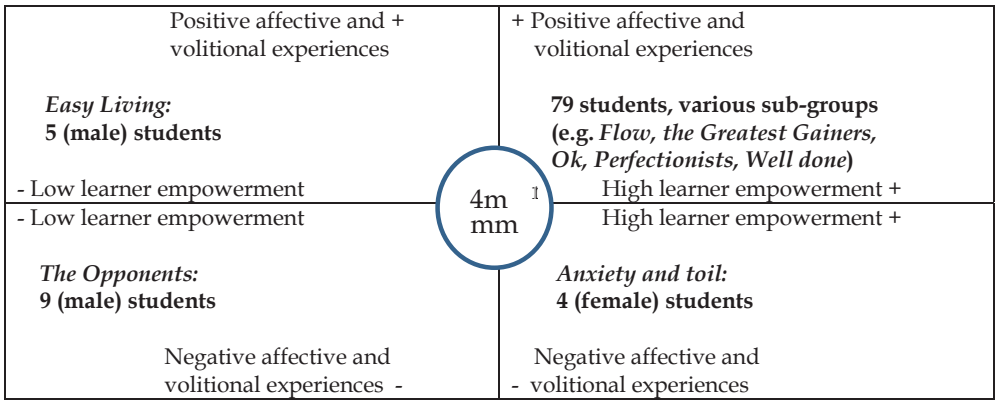

FIGURE 9 The portfolio students situated on the map on the basis of their learner empowerment

According to the analysis, 83 students took an active and responsible learner role. Their working processes were not necessarily easy, but at the end of the course both the students themselves and their teachers agreed that they had worked hard, some students probably much harder than during an ordinary, more teacher-directed course. Thus, it could be claimed that the portfolio programme somehow either fostered or provided opportunities for these students' empowerment.

Nevertheless, among these 83 students there were four students who very responsibly fulfilled all the requirements of the course, but because their working processes were full of toil and anxiety, they themselves did not really consider the experience to have been empowering. Full of self-doubt and distrust in their English skills, the students struggled through the course in a strained and stressed manner. In other words, they took charge of the power and agency responsibly, but not willingly. Although I did not call them disempowered at the time, I would now claim that the portfolio programme disempowered them in that it caused them excessive anxiety and stress.

In contrast, there were 14 students whose working processes during the course did not appear very active. These students produced very little work in their portfolios: "the less work, the better" seemed to be their motto. They were all male students. However, five of them liked the portfolio approach because it was easy, while nine disliked it because they considered the approach too demanding and laborious.

All these categories naturally accommodated variety. Several portfolio portraits were therefore presented to illustrate the individual variety in students' portfolio and empowerment processes.

In sum, the portfolio assessment in this course offered students an opportunity for empowerment, either as power, resources, agency and/or self-efficacy, at all levels of the assessment process. However, not all students

embraced that opportunity. Students' portfolio processes were highly individual, and so were their processes of empowerment, or disempowerment. Nevertheless, the portfolio experiment was the most empowering assessment approach that I am aware of in that it gave students the opportunity and power to decide on their own studying and assessment at all phases of the assessment process, from the purpose to the actual impact of assessment (see Figure 3).

6 DISCUSSION AND CONCLUSIONS

In this final chapter, I will summarise and discuss the main findings of the present study by answering the main research questions one more time, but this time on the basis of the whole study, not its separate sub-studies.

As this study, and all its sub-studies, started from assessment issues and questions arising from classroom work, it was very much practically oriented and motivated. Therefore, although this is quite unorthodox, I will present the practical implications before the scientific contributions of the study. After that, I will discuss the limitations of the study as well as give suggestions for further research. Last of all, I will present my final conclusions and closing remarks.

6.1 Summary and discussion of the results

The research interest of this study was to find out what students' experiences of assessment and (dis)empowerment were in their upper secondary EFL studies in one Finnish upper secondary school, namely the school where I teach. The Finnish core curricula clearly state that assessment should, first and foremost, encourage and guide students' learning and promote their self-assessment skills (see e.g. *Core curriculum 2003*, p. 224-225). Also, assessment should somehow involve students in the assessment process and give them agency, even if limited. In other words, assessment should also empower students. Furthermore, the current objective of education is not only to teach students certain facts or skills but also to teach them transferable skills and the desire and ability to learn throughout their lives. This entails also learning to learn and self-assessment skills. Education should thus empower students in the sense of giving them resources and skills for the future.

Consequently, the first research question in this study was: *Do students experience assessment and feedback in their upper secondary EFL studies as empowering or disempowering?* I will summarise the answers to this research question, as well as to the others, briefly in the following paragraphs.

This study has indicated that while most students were generally quite happy with assessment, there was a significant minority of students who found assessment disempowering. For them, assessment caused stress and anxiety and they felt that the assessment was not good enough or varied enough to allow them to show all their skills and knowledge. The assessment also discouraged them and reduced their desire to study English.

In addition, the feedback given during their EFL studies had not been completely successful with all students. Sometimes students felt that they had not received enough feedback, or that the feedback had not guided and enhanced their learning. Also, students had at times ignored the feedback and concentrated on their grades instead of the teachers' comments and corrections. However, feedback was also considered to be beneficial and to give guidance. Students could also do self-feedback, i.e. use the feedback information extracted from the learning situations to self-assess their work. In these cases, feedback had achieved its objective as stated in the national *Core Curriculum 2003* and had both guided and encouraged students' learning and developed their self-assessment skills. Thus, it had been empowering in the sense of giving the students skills and resources.

This study has shown, then, that students' reactions to assessment and feedback vary considerably. In its conclusion that the actual impact of feedback, or assessment, depends on students' individual responses to it, these results strongly support the findings of Hattie (2009) and Wiliam (2012). The study has also indicated clearly that students' experiences of empowerment vary: as Leach et al. (2001, p. 294) put it: "Empowerment is not the same for everyone. A process that is empowering for some will be disempowering for others and will be resisted by them."

All in all, according to the present study, assessment and feedback are in themselves neither empowering nor disempowering, but potentially both. Approximately 15% of the students in Part 1 of the present study seemed to feel disempowered by assessment. Then again, a good 10% of the students did not feel disempowered by assessment at all: they were labelled as non-disempowered in this study. In addition, there were students who seemed to experience assessment as empowering and who had embraced the given agency in assessment and used their power, although limited, to influence the assessment processes. However, the majority of students were situated somewhere between these groups. In other words, most students had quite a neutral attitude to assessment and feedback: assessment and feedback had served their purpose as part of the studying and learning cycles and as part of school life well but probably had not had any particularly strong effect on students' empowerment or disempowerment processes.

Nonetheless, the approximately 15% of students who felt disempowered translates into approximately four students in each group of 25 students. In my opinion, that is quite a few. Therefore another research question was asked to find out what might explain or predict assessment disempowerment. The disempowered students felt both stressed and demotivated by assessment, for a variety of reasons. First of all, test anxiety was a clear predictor of disempowerment: no high-stakes tests but more formative and less pressurised assessment, *assessment for learning*, was called for. The current assessment methods were not considered to be good and varied enough, and they did not give students a fair chance to show all their skills or knowledge in English.

Secondly, feedback had failed to serve its purpose of facilitating learning. Feedback was either overshadowed by grades, and therefore students had paid little attention to it and had considered it to be less important than their grade or score, or students had not received enough feedback to guide and foster their learning. However, assessment empowerment (or non-disempowerment) was linked to the experience of learning guided by feedback: the students felt either that feedback was beneficial and had fostered their learning or that they did not need more external feedback because they themselves could infer feedback from the learning situations. In addition, students' ownership of their English studies as well as their goal-orientation played a role in assessment (dis)empowerment. Students' personality was also seen as a factor that influences assessment. The disempowered students hoped for some additional, alternative assessment methodology.

To summarise, the five most significant predictors of disempowerment, accounting together for over 50% of the variance in this study, were *Stressful and discouraging assessment*, *No pressurised or high-stakes tests, Grades over feedback, Good and versatile assessment*, which related negatively with disempowerment, and *Inadequate feedback*.

As two of the predictors of disempowerment concerned feedback, and another two were linked to stress, anxiety and pressurised tests, they both deserved a closer look. Article 3 therefore dealt with the Matriculation Examination. As it is the only high-stakes test in the Finnish school system, it stands in marked contrast to ordinary Finnish upper secondary school assessment, i.e. internal teacher-based assessment, and that could be seen in the results: six students out of ten suffered from Matriculation Examination anxiety at least to some extent. As has been found in earlier research (Hembree, 1988; see also Cassady, 2010), female students appeared much more vulnerable to test anxiety, with more than two-thirds of them saying that they were afraid of the Matriculation Examination. Their anxiety or fear ranged from slightly anxious excitement or apprehension to strong, disempowering fear that affected their upper secondary school study plans. In contrast, half of the male students did not seem to suffer from any Matriculation exam anxiety. Students' previous

grades did not correlate with anxiety; thus, the stress and anxiety caused by the Matriculation exam test did not appear to be linked to their English skills.

In addition to the pressurised, time-constrained testing situation, one reason why the Matriculation Examination seemed to cause anxiety was its impact on students' future and their further studies. Nonetheless, Matriculation exam anxiety as such did not predict assessment disempowerment in the regression analysis.

Could some assessment methods foster empowerment? If yes, how?

The aim of Part 2 of the present study was to try out some assessment methods to see if they could encourage students' empowerment. This had a clear pedagogical goal: in addition to studying the effects of some assessment methods on students' empowerment, its aim was to develop a new kind of assessment methodology that could foster empowerment. Perkins and Zimmerman (1995, p. 570) maintain that "empowerment-oriented interventions enhance wellness while they also aim to ameliorate problems, provide opportunities for participants to develop knowledge and skills, and engage professionals as collaborators instead of authoritative experts". Although I would not venture to call these teaching experiments interventions as such, they too aimed at students' well-being as well as the creation of new, collaborative opportunities for students to develop and show their knowledge and skills.

One assessment approach that was experimented with in the present study was the cheat-sheet test (Article 4). In terms of student empowerment in the assessment process, the cheat-sheet test increased the agency and power of students at one phase only, namely in the *How?* phase of designing and collecting assessment evidence. More precisely, the students had more agency and power when studying for the test as well as in the test-taking situation itself. Nonetheless, even though limited to that phase alone, the cheat-sheet test seemed to empower many students. The students could develop and exercise their agency as well as self-assessment skills throughout the cheat-sheet process. In the first place, they could each decide for themselves whether to make a cheat sheet or not. Constructing the cheat sheet developed their self-assessment skills as well as their learning-to-learn skills in a very tangible manner and introduced a new study method to those who had never written revision sheets. Moreover, because the cheat sheet reduced some students' test anxiety, it allowed them to focus on both studying and taking the test without excessive, disruptive stress. Finally, the students could decide when to use the cheat sheet in the test – if at all. So, compared to traditional closed-book tests, the cheat-sheet test allowed students several additional opportunities to act as active agents. In addition, the cheat-sheet test could be considered *assessment for/as learning* as it combined studying, learning and both formative and summative assessment. However, the change in the assessment process was small.

The second experiment, which was reported in Article 5, dealt with more individualised (corrective) feedback. Once again, the change in students' agency with regards to the whole assessment process was small: they could choose the way they wanted feedback. Thus, they could affect the communication phase of the assessment process, nothing else, and even in that, their choices were limited to selecting between written or oral feedback and between the direct and indirect treatment of errors. Nevertheless, although it was such a small change, students found it positive: a good half of the students considered the more individualised feedback better for their learning than the traditional teacher choice, and nearly half found it just as good. Thus, for most students the individualised choice seemed to enhance the potential of corrective feedback to be *assessment for learning*.

The third experiment, the portfolio, which was reported in the monograph, was a much more radical approach. It changed the students' role and agency in the whole assessment process. While the majority of students welcomed, even embraced, that change, approximately 15% of the students did not like it. Some of them did not like the amount of work the portfolio required or its topic area. Some of them disliked the student-centred and self-directed working method. Also, three students did not manage to hand in their portfolios at the end of the course. Yet in my opinion only the four Anxiety students could be considered truly disempowered by the portfolio approach: the 'intervention' certainly did not enhance their well-being.

All these experiments showed that even though the alternative assessment methods proved promising, they were not a panacea that could empower all of the students. Furthermore, both the cheat-sheet test and the portfolio approach were disliked by some students. So, did the assessment methods that were experimented with in these sub-studies improve the situation in any way? Yes, in my opinion, they did. Although students' empowerment may not appear radically better in terms of numbers, the students who were empowered or disempowered were most likely different students in different studies. For instance, the students who seemed to clearly benefit from the cheat-sheet test were students who ordinarily suffered from test anxiety. Even if the cheat sheet did not improve their test results very much, they could prepare for and take the test feeling less insecure and anxious or stressed. On the other hand, most of the students who did not like the idea of the cheat-sheet test were students who normally did very well in tests and did not want any additional testing aids. Also, the majority of students who disliked the portfolio approach did not want to invest much work or effort in their studies. Thus, taking a test 'cold' pleased them more than creating a portfolio. However, even though the number of students who disliked the alternative approach (the portfolio or the cheat sheet) was approximately the same as the number of disempowered students in Part 1, one has to remember that disliking an assessment method does not mean that the students feel disempowered by it.

How do assessment disempowerment and empowerment manifest themselves? The empowered, the disempowered and the non-disempowered of these sub-studies.

All the sub-studies of the present study indicated that students' experiences of and reactions to the assessment methodology vary considerably. This was the case with each of the assessment methods that were tried out, namely the cheat sheet, personalised corrective feedback and the portfolio. All these methods were met with a range of responses. All these experiments also had their empowered and non-disempowered students, and the portfolio approach as well as the cheat-sheet test also had their 'victims', the disempowered students.

In sub-study 6 (see Pollari, 2000, p. 68) I defined the empowered as follows:

> In other words, the empowered is here regarded as an active agent who may be given power as an object or a recipient, but who accepts it and takes charge of it actively as a responsible subject. Power refers here, for instance to having power to make decisions concerning oneself, but also to having enough opportunities, resources and means to have both the readiness and willingness to take charge of one's actions and potentials actively and responsibly

Following this definition, to which I still subscribe, the empowered students in these studies were, for instance, the empowered student case presented in Article 1 as well as the Constant Flow and the Greatest Gainers of the portfolio study. They clearly benefited from either assessment in general or from the particular assessment method that was the object of the experiment. So too did the students who found the cheat sheet very helpful both for their learning and their test anxiety. They also actively, responsibly and willingly took charge of the agency and power given to them. As a result, assessment enabled them to customise either their studies, the contents of their portfolio or the cheat sheet so that they would best benefit from them. Furthermore, they could adjust the difficulty level to fit their goals and skills. Being empowered also means gaining something positive in the process, such as better learning, self-efficacy or diminished anxiety. In this data, a willingness to try out and take charge of the opportunities given to them seems to be the common denominator between the empowered students; they were both female and male students, with a whole range of previous English grades.

Then again, as Leach et al. (2001, p. 294) put it, a "process that is empowering for some will be disempowering for others". In Chapter 3.2 I defined disempowerment as follows:

> Thus, in this study, disempowerment does not refer to students having or not having power, but it refers to students experiencing that they do not have power and/or resources so that they could make decisions in order to fulfil their potential. In other words, disempowerment refers to the lack of perceived control and low self-efficacy (e.g. Zimmerman, 1995, 2000): students may actually have been given power but they do not either realise it or believe in their power and/or themselves. Therefore they do not, or cannot, take charge of their potential power, which may, in turn, lead to diminished motivation (Cleary & Zimmerman, 2004; Harlen, 2012b; Weber & Patterson, 2000).

The disempowered student of Article 1 and *Misery* of the portfolio study were undoubtedly discouraged and disempowered by either assessment in general

or by the portfolio approach. Assessment, or the portfolio, caused them considerable anxiety and stress and they did not trust their language skills at all. They also lacked self-efficacy as learners of English and seemed to be studying English for external reasons, such as the grades, the Matriculation Examination or because they had to, but not for themselves. All in all, they did not appear to have either the resources or the confidence to take charge of the power and agency they had been given. Although to a much smaller extent, perhaps also the one or two students who had invested in their cheat sheets but were badly disappointed with the cheat-sheet test could be considered the disempowered of that assessment experiment. The three students mentioned above, as well as the majority of those considered disempowered in Article 1, were female students.

Leach et al. (2001, p. 294) claim that some students will resist a process that is meant to be empowering. In these studies, the clearest cases of those resisting a particular assessment methodology were the nine male students in the portfolio approach who were named *the Opponents*. Their previous grades ranged between poor (5) and very good (9). They neither liked the course and its method, nor attempted to work actively or responsibly. However, they took a clear stance that the portfolio was not a proper or suitable teaching or assessment method in upper secondary school, and they trusted themselves and their point of view enough to actively resist it. Therefore, although they disliked the portfolio, they did not seem to be disempowered by it.

All the alternative assessment methods had their *non-disempowered* students as well. For instance, the student case named *Easy Living* in the portfolio study reminds one to a considerable extent of the non-disempowered student of Article 1. In terms of the theory of empowerment (Zimmerman, 1995, 2000), they both manifested a clear intrapersonal component of psychological empowerment as they trusted their language skills and themselves. So did most of the students who decided not to make a cheat sheet for the test. Their self-efficacy appeared strong. All these students also used their empowerment in the sense that they decided *not* to invest very much in the proffered agency (Leach et al., 2001; Rodwell, 1996), either in terms of investing much effort in the making of the portfolio, or the cheat sheet, or using their potential decision-making power in the assessment process. Thus, they manifested an interactional component of psychological empowerment as they appeared to "act as they believe appropriate to achieve goals they set for themselves" (Zimmerman, 1995, p. 589). In short, the non-disempowered students in this study could be defined as students who have high self-efficacy but who may decide *not* to take charge of the power, agency or learner role given to them because they do not consider it relevant or useful for achieving the goals they have set for themselves. In this data, the majority of these students were male students with either very good or excellent previous grades (9 or 10).

In brief, the manifestations of assessment (dis)empowerment presented above illuminate and corroborate in a clear, tangible way the findings of other studies and the empowerment literature: empowerment is not the same for

everyone (e.g. Leach et al., 2000, 2001; Zimmerman, 1995, 2000), nor is everyone equally confident, ready or willing to assume the power, resources or the role given to them.

6.2 Practical implications

The practical implications of this study are quite clear: there should be as much variety as possible in the assessment and feedback methods employed in EFL studies in upper secondary school as a whole. Naturally, the selection of each assessment method depends first and foremost on the purpose of that assessment, but the methodology employed all through upper secondary EFL studies should be as diverse as possible. As this study has shown, students are individuals and they have different skills and aims as well as different strengths and weaknesses. Therefore, they also experience assessment and its methods differently and react to them in individual ways. Also, assessment should not be only *summative* and focus only on learning outcomes *(assessment of learning)* but more attention should be paid to *assessment for learning*, i.e. *formative assessment*. Furthermore, diversifying the assessment methodology as well as assessment purposes could mean that assessment would not discourage and disempower the same students every time. A broader assessment methodology might offer those usually discouraged or even disempowered by traditional assessment the chance of positive experiences that might, for a change, foster their self-efficacy. However, using more diverse assessment methodology will not make assessment problem free and it will not please all the students, either, as the assessment experiments of this study have indicated.

This study highlighted four aspects of assessment that call for more attention: variety in assessment methods and purposes, the anxiety that assessment may cause, feedback and its role, and agency in the assessment process.

The current national core curriculum, *Core curriculum 2015*, emphasises the role of assessment in promoting and enhancing learning. In other words, the core curriculum advocates *assessment for learning*, and not only *assessment of learning*. This seems to call for some change in the traditional ways of assessment in Finnish upper secondary school: although assessment has always been teacher based, its role has been to report the learning outcomes through grading, rather than to support the learning process. In a nutshell, assessment has mainly been seen as summative testing and grading which take place after the learning process is over. Sometimes assessment has taken place only during the so-called exam week. More assessment for learning is required. This includes formative assessment whose sole purpose is to enhance and help learning. Formative assessment could also gradually build the self-confidence of those students who now seem to lack it as learners of English, as it would allow them less pressurised assessment situations and give them evidence of their learning. However, formative assessment may not be very familiar to

many teachers and we have little evidence of its systematic use in Finland. Further in-service as well as pre-service teacher training is therefore needed on assessment, and particularly on formative assessment.

One crucial factor in assessment for learning, and in learning in general, is feedback (Black & Wiliam, 1998a, 1998b, 2012; Hattie & Timperley, 2007; Hattie, 2009). Good feedback that meets learners' needs and individual reactions in the right way can have a great impact on learning, but in foreign language education feedback has mainly been regarded as corrective feedback. In other words, feedback in secondary schools has concentrated on correcting language errors and spelling mistakes (e.g. Furneaux, Paran, & Fairfax, 2007; Guénette & Lyster, 2013; Lee, 2004). This rather narrow and mechanistic view of feedback can have many limitations. First of all, how does it accord with the new core curriculum and its objectives? C*ore curriculum 2015* (p. 114) states, for instance, that foreign language "instruction strengthens the students' confidence in their own abilities in learning languages and using them confidently, and provides possibilities for experiencing the joy of learning". *Core curriculum 2015* goes on to say that one of the general objectives of language studies is that the student "gains confidence to utilise his or her language proficiency creatively in studies, at work, and during leisure time" (p. 114). If feedback is limited to correcting all errors but nothing else, it hardly builds the confidence of all learners. Moreover, language is for communication. Should that communication not be included in the feedback? One can communicate meaning, emotions and experiences, sometimes very powerfully, even if the language used is not error free.

Feedback and its role should therefore be reconsidered. As students in this study said, they would like to have more feed forward, in other words, more feedback that guides them and their studies and helps them to improve their future performance. In addition, more feedback is needed during the learning process, not only after it, to enhance learning. Also, feedback should be individual and personalised, and so clear, concrete and specific that students know what it means and what they should do: as Price et al. (2010, p. 279) quite rightly say "feedback can only be effective when the learner understands the feedback and is willing and able to act on it". Students also hoped for constructive and balanced feedback, not just error correction, and they wanted more varied methods of giving feedback. All this, coupled with the findings that students have individual, different reactions to feedback, as also suggested by earlier research literature (Hattie, 2009; Wiliam, 2012), is a tall order for any teacher. Further professional training for us teachers in how to give effective feedback and how to rework student writing so that students benefit more from corrective feedback would therefore be welcomed.

Another form of feedback, the formative use of summative assessments, such as course tests, is one of the cornerstones of assessment for learning (Black et al., 2003, pp. 31, 53-57). This clearly needs more attention. Earlier, I cited Välijärvi et al. (2009, p. 54), who speculated that the modular curriculum and upper secondary school structure with the widely used exam week system may be one reason for test-focused assessment in upper secondary school. Also, with

the courses ending with the exam week and new courses starting immediately after it, there may not be sufficient time for students and teachers to analyse and make use of the information given by course-week assessments in order to improve further learning. For instance, my own teaching experience has shown that surprisingly many students never come to collect their test papers – all they want to know is the course grade (which they can see in the digital assessment system). This means, of course, that they do not get *any* information that could pinpoint what they knew and what they did not know (see Gardner, 2012, p. 107). This is more or less the case also when students get back separate answer sheets, or mere scores, but not the actual test papers with the test tasks: when this happens, students can see in which test exercises they scored well and where not, but they do not get any relevant information which would enable them to focus their attention on identifying how they might improve their learning (Gardner, 2012, p. 107). Even if there is a follow-up lesson where the test papers are handed back and/or test tasks discussed – which is not the case in every school – this discussion may be very hasty, with too much information, and it does not much benefit further learning. And finally, alas, some students are too interested only in marks and grades to pay much attention to further elaboration or comments, as this study has also indicated. Consequently, in my view, we should find ways to utilise the summative assessment information better to support learning, and not only to report it.

Clear differences between empowered and disempowered students were found in self-feedback, i.e. in getting feedback from the learning situation itself, as well as in knowing one's own strengths and weaknesses. Disempowered students seemed rather dependent on external feedback, given by the teacher, and did not, or could not, engage in assessing their own work. Students need to be guided more in the process of self-assessment and self-feedback. In order to be able to assess their own work, students should, first and foremost, know the learning goals of that particular task as well as the criteria for good work. That requirement did not necessarily seem to be met with the students in this study (see also Black & Wiliam, 1998b). Thus, goal-setting, opening up and explaining the criteria for good work as well as self- and peer-assessment methods should be the focus of both learner and teacher training.

Core curriculum 2015 also calls for diversity in the assessment evidence and in the methodology used for course assessments and grades. Although it is quite safe to assume that assessment in Finnish upper secondary EFL studies takes into account at least several domains of language skills (such as reading and listening comprehension and writing), the assessment methodology is not necessarily varied. As this study has also shown, certain kinds of tests seem to dominate assessment in upper secondary school. Very often these tests and their exercise types are influenced by the Matriculation Examination. The result is that multiple-choice comprehension questions as well as essays of 150-250 words appear to be the staple of assessment methodology. They do prepare students well for their Matriculation Examination and, in that sense, for their future, but multiple-choice reading or listening comprehension is a rare skill in

real life, and real life is the reason given by most of the students in this study for studying English.

What would diverse assessment methodology mean in classroom reality? Portfolios offered a total break from the test-oriented assessment culture and allowed students unprecedented opportunities for agency and empowerment. This does not mean that I would advocate portfolios for every upper secondary school course. The portfolio can mean a great deal of work, both for the teacher and the students, so perhaps one or two courses during the upper secondary English courses would suffice. Semi-portfolios, i.e. a combination of portfolio work and more traditional teaching and assessment, might be quite viable as well. Self-assessment and peer assessment should probably get a stronger role in assessment, as required by *Core curriculum 2015*, and also in summative assessment. Naturally, I am not denying the need for tests, not even for practice tests for the Matriculation exam, but advocating a repertoire of more diverse tests. For instance, cheat-sheet tests proved feasible, useful and empowering as an alternative method. Foreign and second language education has long employed oral tests where students communicate in pairs or small groups – could we not start using pair or group work also in other modes of assessment such as assignments written in pairs or tests taken in small groups? In such cases the testing situation could genuinely be also a learning situation where students could learn from one another. Besides, if essays can be written at home, why not take-away tests or assessment assignments, either digital or otherwise, that are taken without the constraints of time or place? Many of these methods, of course, require trust and as well as a change from seeing assessment as a form of control to seeing it as a form of learning.

In addition to diversity in assessment methodology, I would like to promote more choice and agency for students in the assessment process by involving them in assessment design and interpretation beyond just being informed of the course objectives and criteria. Such involvement does not have to be as radically learner-centred as the portfolio approach; even a small change in agency can go a long way, as the individualised corrective feedback experiment indicated. Perhaps one scenario might be making a school-based assessment blueprint that covers all upper secondary English courses but yet leaves room for some individual choice; this could ensure that students get a balanced array of different assessment methods during their EFL studies while allowing students a say in assessment decisions.

At the same time, I would also like to empower teachers in assessment procedures by offering them further pre-service and in-service training in assessment, or in *assessment literacy*, as several authors call "an understanding of the principles of sound assessment" (Volante & Fazio, 2007, p. 750). As Härmälä and Hildén (2012) very aptly summarise the dilemma, the teacher's assessment task is "demanding but assessment training scarce" (see also Härmälä, 2012). This appears to be the case not only in Finland but also internationally (see e.g. Popham, 2009; Volante & Fazio, 2007; see also Fulcher, 2012). Student assessment is such an important and powerful part of school life

that it is high time it was seen as an area of expertise in its own right and not just as an automatic part of teaching that "anyone who can teach can do". I therefore wholeheartedly agree with Takala (1996), who emphasised the importance of teachers' expertise in assessment 20 years ago and concluded that "the teacher who knows the local circumstances and students and who masters assessment methodology well" is best equipped to assess the students (Takala, 1996, p. 221). But as Takala (1996) also suggested, teachers need good pre- and in-service training for assessment. Therefore, in my view, assessment, or assessment literacy (see e.g. Stiggins, 1991, 1995; Popham, 2009), should be allocated much more time and attention in pre-service teacher training than it is now. As both practice and research show, assessment-related activities constitute a significant part of teachers' work (e.g. Mertler, 2004, 2009; Brookhart, 2011). If assessment and its foundations are not properly taught, questioned, conceptualised and reflected on, teachers tend to pick up assessment and grading practices as "on-the-job experience", potentially without proper reflection or justification of their assessment practices: as McMillan (2003, p. 38) claimed already over ten years ago, "Teachers evidently fill the void in various and numerous ways, ways that are difficult even for them to identify".

Furthermore, we in-service foreign or second language teachers at upper secondary level rather depend on at least some external testing materials, as it is practically impossible for us to create reading and listening comprehension materials and exercises because of the time it would take us to do this from scratch. Consequently, we have practically handed over a great deal of our assessment agency to text-book publishers who also produce testing materials. We pick and use materials written by authors who do not necessarily have any training in assessment literacy or test development. If we base our assessment design on such materials without careful, educated reflection on their validity or construct-irrelevant variance, for instance, it is no wonder if our assessment practices raise some concerns. Sound assessment literacy is a prerequisite for sound assessment practices (Stiggins, 1995; Popham, 2009; Volante & Fazio, 2007).

Although the national core curricula do not appear to offer enough guidance or clear enough criteria to support teacher assessment and grading (see e.g. Ouakrim-Soivio, 2013, p. 109) and despite my concern over the inadequate training in assessment available to teachers, it is, nevertheless, my firm belief that we do not need more external, high-stakes testing. If we want to safeguard uniformity in teacher assessment and grading, proper assessment training as well as more easily understandable objectives and criteria in the core curricula would be much more effective tools than any external test. There is also ample evidence of the many harmful effects of high-stakes testing on teaching and learning all around the world. Moreover, as Koretz (2008, p. 316) so aptly puts it, "A test, even a very good one, is always just a test: a valuable source of information, but still only a limited and particular view of student performance".

As this study also attests, pressurised testing is disempowering for the significant number of students who suffer from test anxiety. One of the most pressurised situations in the Finnish school system is the Matriculation Examination. One clear reason for anxiety was linked with the high stakes of the Matriculation Examination. Many students feared that they would underperform under pressure and therefore not have such good opportunities for further education as their skills really merit. In such cases, the actual impact of assessment would truly be disempowering. If the Matriculation Examination once again becomes what its name and original use suggests, an entrance exam to further education, in addition to being the final examination of upper secondary school (Kaarninen & Kaarninen, 2002; Lindström, 1998), what will happen to students who suffer from test anxiety? Will it compromise their equal chances for further education? I do not advocate entrance exams as such, but at least they offer a second – and perhaps different – chance for students who, for one reason or another, underperform in the Matriculation Examination. Also, if students who do not gain access to their preferred place of further education start repeating the syllabi of upper secondary education that they have already completed, and resitting the different parts of the Matriculation Examination several times in order to improve their grades, are they not somehow going to go backwards in their life and learning?

If the Matriculation Examination results are increasingly used for selection for further education, then the examination itself should probably undergo significant changes. Digitalisation is expected to increase the range of exercise types, but the biggest issue, as the students in this study said, is the validity and reliability of the tests that make up the Matriculation Examination, not its medium. The format of the test should therefore perhaps be carefully reconsidered. Cost-efficient and seemingly reliable (at least in the sense of rater reliability) as the multiple-choice questions are, is there over-reliance on them in the foreign/second language tests? Moreover, is the variance that multiple-choice items create too much based on construct-irrelevant variance (Black & Wiliam, 2012; Messick, 1996) at the moment? According to earlier research, test anxiety may manifest itself as anxiety blockage and retrieval failure, heightened by a test situation that the student considers threatening and therefore "the student begins to lose confidence in her knowledge base, and continue to question the accuracy of the responses she has offered", which "becomes particularly salient in multiple choice tests with distracters that 'look good'" (Cassady, 2010, p. 13). Then again, has the *impact of digitalisation on test anxiety* been considered? It would seem quite plausible that students who are anxious about the testing situation itself will envisage more risks involved with the computer than with paper and pen. In other words, will the tool, the computer, itself not heighten the anxiety of those who fear that everything will go wrong in the test?

Another point worth reconsidering is whether it would compromise the purpose of the examination if each test did not create as much variance as now, in other words, if more students did well in the test? According to the *General*

Upper Secondary School Decree (629/1998, 18 §), the purpose of the Matriculation Examination is to examine whether students have reached the goals of upper secondary education, i.e. the skills, knowledge and maturity as defined in the curriculum (see also Lahtinen & Välijärvi, 2014, p. 11). Should the test not be criterion-referenced as, for instance, are the so called YKI tests, which test candidates' language proficiency (see e.g. Huhta & Hildén, 2016)? Then the Matriculation Examination could truly examine students' knowledge against criteria based on the objectives of upper secondary education, instead of being norm-referenced and thus ranking students against one another. Finally, if the upper secondary school certificate loses all its importance when applying for further education, for what purpose do we need course grades – would a pass/fail for each course not suffice, as a colleague suggested? Then teachers could perhaps concentrate more on teaching, and *assessment for learning*, instead of assessment of learning.

6.3 Scientific contributions

Although the practical implications of this study may seem more salient, the present study has some scientific contributions and implications as well. First of all, the present study contributes to fields of research that are under-researched not only in Finland but also elsewhere. It has given new, more detailed information on student assessment in Finnish upper secondary school and also in foreign or second language education. This study agrees with the findings of earlier studies on upper secondary education (Välijärvi & Tuomi, 1995; Välijärvi et al., 2009) and language education (Hildén et al., 2015; Hildén & Rautopuro, 2014; Härmälä et al., 2014; Tarnanen & Huhta, 2011), that assessment seems rather test-based and not very varied in its methodology or agency (see also Atjonen, 2014). However, in an attempt to stay true to the emancipatory origins of empowerment (e.g. Freire, 1972, see also e.g. Cummins, 1986, 1996, 2001), the present study has also aimed to change the situation by introducing and experimenting with alternative assessment methodology that might not only broaden the methodological repertoire but also foster students' power, agency and self-efficacy, in other words, their empowerment. To my knowledge, all three teaching experiments, i.e. the cheat sheet, the personalised corrective feedback and the portfolio approach, were the first reported research studies on their topics in foreign language education in Finland. They shed light on previously unexplored territory.

Moreover, the portfolio study (i.e. sub-study 6) first introduced and further defined the concept of empowerment in the context of foreign or second language education in Finland some 20 years ago. The present study, in its entirety, also corroborates the findings of earlier empowerment studies and literature, such as those of Leach et al. (2000, 2001) and Zimmerman (1995, 2000), in indicating that empowerment is not the same for everyone, nor is everyone equally willing or ready to assume the power and resources that are made

available for them. The present study also shows this in a concrete and tangible way by presenting some real cases of empowered, disempowered and non-disempowered students. Furthermore, in terms of the theory of empowerment (Zimmerman, 1995, 2000), the *non-disempowered students* are also conceptually a new, interesting group. Therefore, this study makes its contribution to the existing empowerment literature.

Most importantly, the present study has asked students themselves for their own experiences and opinions on assessment and thus given them a legitimate voice in assessment. The findings of this research thus complement as well as support earlier research by Aitken (2012) and by Erickson and Gustafsson (2005). The results have proven that students experience assessment and (dis)empowerment very differently and thus also react to it in individual ways. This study indicates that test anxiety is linked to assessment disempowerment and that particularly female students suffer from stress and anxiety related to pressurised test situations such as the Matriculation Examination. These results therefore support earlier studies that have found that test anxiety is a significant factor in study motivation and success, and that most of test anxiety sufferers are female students (e.g. Hembree, 1988, Cassady, 2010).

This study also corroborates the views of Hattie (2009, 2012) and Wiliam (2012), for instance, in saying that students react to feedback in highly individual ways. Although not uncovering all the eight responses to feedback reported by Wiliam (2012), this study revealed four individual responses to feedback: *Guiding feedback, Self-feedback, Inadequate feedback* and *Grades over feedback*. The last two of these were linked to assessment disempowerment. Thus, in addition to exploring feedback and individual feedback responses in foreign language education in Finland for the first time, this study has made individual feedback responses more visible and shown that feedback plays a crucial role in assessment (dis)empowerment. The results of this study also support and complement the findings of earlier studies by demonstrating that, in general, students appreciate teacher feedback (e.g. Leki, 1991) but that not all students are interested in it (e.g. Cohen, 1987). This study also agrees to some extent with the results of the studies of Butler (1987, 1988), Kohn (1999, 2011), Pulfrey, Buchs and Butera (2011) as well as Pulfrey, Darnon and Butera (2013) that grades can, indeed, overshadow teacher feedback and corrections with some students but not necessarily with all students (see Dlaska & Krekeler, 2013). The study further suggests that even a small change towards more personalised feedback might motivate and engage students much more in the feedback process and thus corroborates Guénette's (2007) ideas about the importance of students' motivation in response to corrective feedback.

Last but not least, although not part of the empirical results of this study, one theoretical contribution that this study makes is Figure 3, depicting the assessment process. To my knowledge, the figure is the first attempt to visualise the whole assessment process, from its purpose to its actual impact, in one figure. It also shows that the assessment process does not end with the design

and collection of assessment evidence, as some earlier figures might suggest (see e.g. Pickford & Brown, 2006, p. 4). The figure might also function as a starting point for constructing a practical theory of classroom assessment for teachers, something that now appears to be lacking in both the pedagogical and the assessment literature.

6.4 Limitations of the present study

This study is not without its limitations. Firstly, most of this study was limited to just one Finnish upper secondary school: the sub-studies reported in Articles 1-3 as well as in Articles 4 and 5 took place in one school, i.e. the Teacher Training School of the University of Jyväskylä. As they stand, the findings cannot be generalised to other Finnish upper secondary schools. Furthermore, the academic achievement of the student population in our school is above the national average, so there were not many participants in this study, especially in the data in Articles 1-5, who were struggling with their upper secondary studies. With larger and more varied student samples, students' experiences of empowerment and disempowerment as well as their responses to feedback or to the English test in the Matriculation Examination might look different, as they might also in other contexts and assessment cultures. Then again, the purpose of the present study was not to give a statistically generalisable picture of all upper secondary students in Finland but to focus on analysing and understanding students' experiences of assessment and (dis)empowerment as well as possibly trying to change the situation within this selected context and with these participants. Studying, as it was, matters central to my teaching in my own professional context, it can be claimed that the present study meets the requirements set for teacher research (see Borg, 2010, pp. 392-393; Borg, 2013).

Secondly, this study is limited in its time frame. On the one hand, as much of the data of this study was gathered at the same time (in March 2014), Part 1 of this study (i.e. Articles 1-3) cannot exhibit potential changes in empowerment over time. Furthermore, the timing of the data collection was not ideal, as the majority of third-year students had already left school to prepare for the Matriculation Examination. On the other hand, the data of the portfolio sub-study was collected over 20 years ago. In addition, the portfolio study took place under a different core curriculum, *Core curriculum 1994*. Its educational setting is therefore not completely comparable with the setting in Articles 1-5.

This study relies on mixed methods in its data collection and analysis. However, all the data for Articles 1-3 was collected at the same time by means of one internet-based questionnaire. Several of the findings therefore rely on one questionnaire. Because of the absence or scarcity of prior research on most of the topics of this study, the questionnaire was tailor-made for this study. As a result, it did not have many points of reference or apply previously tested models, and therefore it is open to criticism. Although the students seemed to answer the questionnaire quite carefully, the questionnaire was extensive and

would have benefited from further pruning. In hindsight, I could, perhaps, have omitted the part dealing with the frequency of different assessment methods altogether, to shorten the questionnaire. Naturally, other data collection instruments, for instance a different questionnaire, or alternative methods, such as student interviews or narratives, might have yielded additional perspectives and further information.

The remaining sub-studies, i.e. Articles 4 and 5 as well as the monograph, were teaching experiments, the aim of which was to try out, and even develop, alternative assessment methodology. That being the case, they were not highly rigorous, controlled research experiments, but part of ordinary classroom work. Consequently, there were many 'intervening variables' that were not part of the original plan but unexpectedly became part of the work, and also often required some *ad hoc* intervention or action. These intervening variables affected, for instance, the data collection, particularly in Article 5. Furthermore, as the portfolio study was a total novelty in many respects, its whole process was affected by new, unexpected situations. The data analysis methodologies used in these teaching experiments might also be open to question. For instance, the four-field map used as an instrument of analysis in the portfolio experiment was tailor-made and therefore had no legitimisation from prior research, as such. The data analysis methods used in Article 5 were also rather simple.

My role in this study has been complex, and could be seen as a potential limitation as well. In Articles 1-3 I was a teacher-researcher conducting research in my own professional context, i.e. in the school where I teach. Also, at some point or another I had taught many, and probably most, of the students who answered the questionnaire, so they knew me. In the following sub-studies, reported in Articles 4 and 5, I was experimenting with my own teaching and therefore all the participating students were my students at that time. The portfolio experiment (Monograph) was the only one where I was not involved as an actual teacher but as a researcher. Nonetheless, one of the two participating schools was my former employer, some of the participating students were my former students and all the three teachers were my colleagues and friends. One might therefore claim that I was too much of an insider and perhaps had too much invested in these experiments to be totally objective.

On the other hand, many of these limitations can also be considered strengths of this study. As the instruments of data collection and analysis were specially designed for these sub-studies, they were also highly contextualised and matched the purposes, contexts and also participants of the studies to as large an extent as possible. For instance, although some students said that the questionnaire had been rather long, none of them said that they had not understood any of the questions. Furthermore, although not having any prior research to which to anchor them could be seen as a limitation, many of these sub-studies were firsts of their kind and broke new ground in Finnish foreign or second language education research. Also, even though the results of this study cannot be generalised to other schools, the results give an accurate picture of

the situation in our school at the time of the study. And, finally, my role as an insider guaranteed that I knew the contexts, participants and all the realities of these sub-studies truly inside out. Hence, this study seems to fulfil Borg's (2010, p. 395) definition of teacher research:

> I thus define teacher research as systematic inquiry, qualitative and/or quantitative, conducted by teachers in their own professional contexts, individually or collaboratively (with other teachers and/or external collaborators), which aims to enhance teachers' understandings of some aspect of their work, is made public, has the potential to contribute to better quality teaching and learning in individual classrooms, and which may also inform institutional improvement and educational policy more broadly.

6.5 Future research

Perhaps one of the most important findings of this study is the need for more research on student assessment in the Finnish school context. Research is needed on both student assessment in foreign/second language education in Finland and student assessment in Finnish upper secondary education in general. Moreover, although basic education is not within the scope of this study, I would venture to say that more research is also needed on student assessment in Finnish basic education. After all, "assessment is the field where the battle for real renewal of teaching and curricula is either lost or won" (see Mehtäläinen, 1994, p. 103).

The specific areas that would, in my opinion, merit further research are numerous. First of all, as teachers or schools have a great deal of power over their assessment procedures, it would be important to find out how teachers actually construct their assessment processes: How do they design and collect assessment evidence? What kind of assessment evidence do they collect and how varied is it? What areas or skills do they assess? How do they interpret that assessment evidence into assessment judgements, in other words, what are their criteria and scoring procedures? How do they translate their assessment evidence and judgements into summative grades? Simply put: how, and on what basis, do teachers assign grades in upper secondary school language studies? Are their grades comparable, or is 'inter-grader' reliability as low as it is sometimes claimed to be? Research has shown that foreign language teachers rely heavily on course books and materials provided by course book authors (Luukka et al., 2008; Taalas, Tarnanen, Kauppinen & Pöyhönen, 2008). It is therefore more than reasonable to assume that foreign language teachers rely also on ready-made testing and assessment materials to a considerable degree. We therefore need to know how publishers construct their assessment material packages. These materials, the result of several "pre-packaged sets of decisions" (see Sheldon, 1988, p. 238), have a considerable influence on assessment procedures in Finnish schools, and possibly will have even more influence following digitalisation, so it is vital to analyse what the theoretical bases and

practical philosophies behind the materials are. So far, however, I have not found any research on this topic in Finland.

This study also highlighted two particular areas, feedback and pressurised testing, that both deserve further research. What kind of feedback do foreign/second teachers give in Finland? What kind of feedback would students find beneficial? What are the effects of pressurised high-stakes testing, such as the Matriculation Examination, on students' studies, or on their well-being, or future choices? What kind of a washback effect does the Matriculation Examination have on upper secondary education and how strong is it? How does the potential washback effect on foreign/second studies and assessment accord with the objectives and requirements set out in the new *Core curriculum 2015?*

With the (re)introduction of the concepts of formative and summative assessment, or assessment for learning and assessment of learning, in the new core curricula for basic education and for upper secondary education, research on the methodology and effects of formative assessment is also called for. Research to test and develop various kinds of methods for formative assessment in the Finnish school context (cf. Black et al., 2003) would greatly benefit both teachers and their students – without it, there is a risk that formative assessment will remain a concept in the national core curricula but will never become truly alive in classrooms.

In the same way, as a teacher, I would also urge much more classroom-based research and many more teaching experiments trying out and developing different assessment methods for both formative and summative assessment, as well as feedback methods. That would be a good way to develop the new, innovative methodology that is clearly needed for the purpose of both summative and formative assessment, and for teacher assessment as well as self- and peer assessment and feedback.

As a teacher-researcher, I would naturally welcome more teacher research. The general consensus seems to be that teacher-research and teachers' research engagement "has the potential to be a powerful transformative force in the work and professional development of language teachers" but, alas, "such engagement remains a minority activity in our field" (Borg, 2010, p. 391). As a rather sad illustration of this, Dörnyei said, in 2007 (p. 191) that he was "still to meet a teacher who has been voluntarily involved in an action research project". As the benefits of teacher-research on language education seem quite extensive, I hope that new ways of enabling teacher research engagement can be found in the future, despite the wide range of barriers that teacher research tends to meet. (For an extensive review of language teacher research engagement, see Borg, 2010, 2013.)

As a teacher trainer, I would also be interested in research into teacher trainees' conceptions of assessment and how they construct their own practical philosophies and procedures of assessment during their training. What kind of skills and knowledge do they get during their teacher training, and do they find it adequate and/or relevant? How well prepared are they with respect to

assessment when they start work? If they do not feel adequately prepared for assessment as part of their work, how do they "fill the void" at work (McMillan, 2003, p. 38)?

Finally, it is vital to have further research that focuses on the students' perspectives on assessment, both on classroom-based assessment and high-stakes assessment. Research involving students in the development of assessment and investigating participatory forms of assessment and assessment development, whether in the classroom, locally or even nationally, as has been the case in Sweden (Erickson & Åberg-Bengtsson, 2012), could also enable, even legitimise, students' role and agency.

It is high time that both student assessment and assessment research gave students a legitimate voice: they are the ones who ultimately carry the consequences of student assessment. I will therefore end by quoting a student from this study:

> *On hienoa, että tällaista asiaa kartoitetaan näin laajasti. Arviointi ei ole yhdentekevä asia. Arviointi ei ole vain numero Wilmassa. Arvioinnilla voi olla suuri vaikutus siihen, millaiseksi opiskelija itsensä tuntee sekä tämän motivaatioon tulevina oppimisen hetkinä. (3F110)*

> *It's great that this topic is being investigated so widely. Assessment is no trivial matter. Assessment is not just a grade in Wilma (=internal electronic communication and record-keeping system). Assessment can have a strong impact on how students see themselves and on their motivation to learn in the future.*

TIIVISTELMÄ (FINNISH SUMMARY)

Opiskelija-arvioinnilla on merkittävä asema kaikkialla koulumaailmassa, niin myös suomalaisessa lukiokoulutuksessa. Jokainen lukiolainen saa vuosittain useita kurssiarvosanoja, ja niiden taustalla on kymmeniä erilaisia kokeita ja muita arviointinäyttöjä. Vaikutusvaltaisesta roolistaan huolimatta arviointia ja sen käytännön toteutusta suomalaisessa lukiossa on tutkittu varsin vähän. Myöskään sitä, miten opettajat arvioivat opiskelijoidensa suoriutumista kieltenopetuksessa, ei ole maassamme laajalti tutkittu. Opiskelijoiden omiin kokemuksiin arvioinnista ja sen vaikutusvallasta on perehdytty vieläkin vähemmän.

Tämän tutkimuksen tavoitteena on tutkia, kuinka lukiolaiset itse kokevat arvioinnin osana lukion englannin opintojaan. Tavoitteena on selvittää, voimaannuttaako arviointi opiskelijoita vai aiheuttaako arviointi ennemminkin voimattomuuden kokemuksia: toisin sanoen, auttaako arviointi opiskelijoiden mielestä heitä opinnoissaan vai lannistaako ja latistaako se heitä ja heidän opiskeluaan. Lisäksi pyrkimyksenä on tutkia, voivatko erilaiset arviointimenetelmät edistää opiskelijoiden positiivisia kokemuksia arvioinnin roolista ja sen vaikutusvallasta.

Tutkimuksen teoreettinen tausta

Opiskelija-arviointi on käsitteenä hyvin monitahoinen ja laaja, sillä opiskelija-arviointi on eri koulukulttuureissa varsin erilaista. Taustoittaakseni suomalaista opiskelija-arviointia ja sen ominaispiirteitä lähestyn tässä tutkimuksessa opiskelija-arviointia ensin sen keskeisimpien yleisten käsitteiden kautta (Luku 2). Nivon nämä käsitteet eräänlaiseen arviointiprosessikaavioon, joka sisältää mm. arvioinnin tarkoituksen, arvioinnin suunnittelun ja toteutuksen sekä arviointituloksen hyödyntämisen ja vaikutuksen. Lisäksi kartoitan, mitä lukion opetussuunnitelmat sekä aiemmat lukion ja kieltenopiskelun opiskelija-arviointia käsittelevät suomalaistutkimukset sanovat arvioinnista (Luku 4). Näiden yhteenvetona totean, että lukion opiskelija-arviointi – niin yleisesti kuin kieltenopetuksessakin – on varsin summatiivista ja numerokeskeistä arviointia, jonka keskeisenä tavoitteena on vaikuttanut olevan kurssiarvosanan antaminen. Lukion kurssiarviointi perustuu suurelta osin erilaisiin kokeisiin, eli arviointi ei ole menetelmiltään kovin monipuolista eikä itse- tai pariarvioinneilla ole merkittävää sijaa (ks. mm. Välijärvi et al., 2009; Härmälä et al., 2014). Sen sijaan opiskelijat ovat hyvin tietoisia kurssien tavoitteista ja arviointiperusteista ja kokevat arvioinnin antaneen melko hyvän kuvan osaamisestaan (Välijärvi et al., 2009). Toisin kuin monissa muissa maissa, opettajalla on arvioinnin suhteen täysi valta, joten arviointi on – ainakin periaatteessa – opettajien suunnittelemaa ja toteuttamaa pienimuotoista luokkahuonearviointia *(small-scale internal classroom assessment)*, jossa yksittäisillä kokeilla tai kurssiarvosanoilla ei ole kovin kriittistä painoarvoa *(low-stakes assessment)*. Käytännössä kieltenopettajat kuitenkin turvautuvat arvioinnissa suurelta osin oppikirjasarjojen valmiisiin arviointimateriaaleihin tai esimerkiksi aiempien ylioppilaskirjoitusten tehtäviin: ylioppilastutkintohan on koululaitoksemme ainoa laajamittainen ulkoinen tutkinto, ja sillä on suuri merkitys

opiskelijoiden tulevien opintojen kannalta *(external large-scale high-stakes examination)*. Tutkimusta, joka kattavasti kartoittaisi lukion tai lukion kieltenopetuksen arviointikäytänteitä, ei kuitenkaan Suomessa ole vielä tehty.

Opiskelija-arvioinnin lisäksi tämän tutkimuksen keskiössä ovat käsitteet *empowerment* ja *disempowerment* (Luku 3). *Empowerment*, jonka käännöksinä suomessa käytetään mm termejä *valtautuminen* ja *valtauttaminen* sekä *voimaantuminen* ja *voimaannuttaminen* (Siitonen, 1999, pp. 82-90), on ollut tutkimuksen kohteena useilla eri aloilla, joten sen määritelmät vaihtelevat. Vallan ja voimavarojen saamisen tai haltuunoton lisäksi läheisiä käsitteitä ovat mm. autonomia, itsemääräämisoikeus, toimijuus, motivaatio, itsetunto sekä käsitys omasta toimintakyvystä ja mahdollisuudesta toteuttaa omia tavoitteitaan. Zimmermanin (1995, 2000) yhteisöpsykologian alalla kehittämän teorian mukaan *empowerment* jakautuu yksilö-, organisaatio- ja yhteisötasoon. Yksilötasolla kyse on psykologisesta voimaantumisesta, joka sisältää paitsi yksilön oman (*intrapersonal*) kokemuksen siitä, että hänellä on valtaa ja kykyä toimia, myös halun ja pyrkimyksen käyttää tätä toimintavaltaa ja -kykyä. Osa tutkijoista kuitenkin on sitä mieltä, että yksilöllä on myös oikeus kieltäytyä käyttämästä toimintavaltaansa. Termille *disempowerment* on vieläkin vaikeampaa löytää suomennosta, sillä kyse ei ole vain vajaavaltaisuudesta tai rajoitetusta itsemääräämisoikeudesta vaan eräänlaisesta voimattomuudesta tai *voimaantumattomuudesta* (Siitonen, 1999, p. 83). Vaikka *voimattomuus* ja *voimaannuttomuus* ovat mielestäni sanoina varsin kömpelöjä, käytän niitä tässä tiivistelmässä parempien termien puuttuessa. Tässä työssä *disempowerment* tarkoittaa, että opiskelija kokee, ettei hänellä ole valtaa tai voimavaroja tehdä sellaisia päätöksiä, jotka voisivat mahdollistaa hänen potentiaalinsa toteutumisen: opiskelijalla voi itse asiassa olla valtaa, mutta hän ei koe omaavansa sitä, tai hän ei usko omaan valtaansa tai kykyynsä käyttää sitä, eikä siten kykene ottamaan valtaa ja/tai voimavaroja haltuunsa, mikä saattaa vuorostaan mm. vähentää opiskelumotivaatiota.

Tutkimuksen tavoitteet ja toteutus

Väitöstutkimukseni koostuu viidestä artikkelista, aiempaan lisensiaatintyöhöni pohjautuvasta monografiasta ja kokoomaosasta. Tämän kokonaisuuden tarkoituksena oli selvittää, miten lukiolaiset kokevat arvioinnin osana lukion englannin opintoja.

Ensimmäiset kolme artikkelia perustuvat keväällä 2014 toteutettuun laajaalaiseen kyselytutkimukseen, johon osallistui yhteensä 146 Jyväskylän normaalikoulun lukion toisen ja kolmannen vuosikurssin opiskelijaa. Kyselyaineistosta nousseiden teemojen pohjalta halusin ennen kaikkea selvittää, onko arviointi opiskelijoiden kokemuksen mukaan opintoja auttava ja opiskelijaa voimaannuttava tekijä vai päinvastoin, eli aiheuttaako arviointi opiskelijoille lähinnä ahdistusta ja voimattomuuden tunnetta ja siten mahdollisesti jopa haittaa opintoja. Aineistoa analysoitiin monimenetelmäisesti, vaikkakin pääpaino oli kvantitatiivisissa menetelmissä (mm. pääkomponenttianalyysi ja askeltava regressioanalyysi). Seuraavat osatutkimukset ovat opetuskokeiluja, joissa opettaja-tutki-

jana kokeilin erilaisia arviointimenetelmiä tai -käytänteitä selvittääkseni, voivatko erilaiset arviointimenetelmät edistää opiskelijoiden voimaantumista. Nämä aineistot on kerätty eri aikoina ja niitä on analysoitu eri menetelmin, joskin niiden analyysissa pääpaino on ollut kvalitatiivisilla menetelmillä.

Osatutkimusten tulokset

Vaikka ensimmäisissä kolmessa artikkelissa raportoituun kyselytutkimukseen osallistuneet opiskelijat olivat pääsääntöisesti tyytyväisiä arviointiin, n. 15-20% opiskelijoista koki arvioinnin aiheuttavan ahdistusta ja voimattomuuden tunnetta. Siten ensimmäisen artikkelin tavoite oli selvittää, mikä lukion englannin arvioinnissa aiheuttaa voimattomuuden tunnetta näille opiskelijoille. Ko. opiskelijat kokivat, että arviointi aiheutti heille liiallista stressiä ja ahdistusta: koeahdistus ja varsinkin ns. paineistetut kokeet (ylioppilaskirjoitukset ja kurssikokeet) näyttivät liittyvän voimattomuuden tunteeseen. Lisäksi nykyistä arviointia ei pidetty tarpeeksi monipuolisena eikä sen koettu antavan opiskelijalle mahdollisuutta osoittaa kaikkea osaamistaan. Toiseksi keskeiseksi tekijäksi nousi palaute: joko palautetta ei ollut saatu riittävästi tai sitä ei koettu opintojen kannalta hyödylliseksi. Arvioinnin aiheuttamaa voimattomuutta kokeneet opiskelijat tuntuivat lisäksi opiskelevan englantia lähinnä arvosanan tai ylioppilastutkinnon vuoksi eivätkä niinkään omien päämääriensä toteuttamiseksi. Toisena ääripäänä aineistosta nousivat esiin opiskelijat, joita arviointi tuntui voimaannuttavan: opiskelijat tunnistivat oman toimintavaltansa ja myös käyttivät sitä, ja arviointi auttoi heitä opinnoissaan. Lisäksi aineistosta löytyi opiskelijoita, joille arviointi ei aiheuttanut minkäänlaista ahdistusta, stressiä tai voimattomuutta, mutta he eivät myöskään halunneet käyttää omaa toimintavaltaansa, sillä heille arviointi oli melko merkityksetön asia.

Toinen artikkeli paneutui tarkemmin palautteeseen ja sen kytköksiin opiskelijakokemuksiin arvioinnista ja sen voimaannuttavuudesta. Aineistosta nousi esiin neljä erilaista reaktiota palautteeseen. Niistä *ohjaava palaute*, eli palaute, jonka opiskelijat kokivat ohjaavan ja auttavan heitä opinnoissaan, samoin kuin *itsepalaute*, eli se, että opiskelija osasi itse arvioida osaamistaan ja työtään esim. opiskelutilanteista saamansa informaation avulla ilman, että opettaja tai toinen opiskelija antoi hänelle palautetta, liittyivät arvioinnin voimaannuttaviin kokemuksiin. Sen sijaan *riittämätön palaute* ja *arvosanojen arvostaminen palautetta enemmän* korreloivat voimattomuuskokemusten kanssa.

Kolmas artikkeli paneutui opiskelijoiden kokemuksiin sekä ennakko-odotuksiin englannin ylioppilaskokeesta. Ylioppilaskoe aiheutti jonkinasteista ahdistusta opiskelijoiden enemmistössä, ja varsinkin tytöt ilmaisivat pelkäävänsä ylioppilaskoetta. Lisäksi opiskelijat kritisoivan mm. puhumista testaavan kokeen puuttumista sekä joidenkin osioiden liiallista vaikeutta tai vaikeaselkoisuutta. Toisaalta koetta pidettiin tärkeänä, ja monille hyvä ylioppilastutkintotodistus oli tärkeämpi kuin lukion päättötodistus.

Tämän tutkimuskokonaisuuden loppuosa keskittyi erilaisten arviointimenetelmien opetuskokeiluihin. Neljännessä artikkelissa raportoin ns. lunttilappu-

koetta. Opiskelijat kokivat lunttilapun pääsääntöisesti opiskelua ja oppimista tukevana menetelmänä, joka myös selkeästi vähensi koejännitystä ja stressiä. Lunttilapun ei kuitenkaan koettu vaikuttaneen koetulokseen suurestikaan.

Viidennessä artikkelissa kokeilin henkilökohtaisempaa korjaavaa palautetta *(corrective feedback, CF)* lukiolaisten kirjoitelmien palautteessa. Opiskelija itse sai valita joko suullisen tai kirjallisen palautteen. Lisäksi hän sai valita, halusiko kirjoitelman virheet valmiiksi korjattuina *(direct CF)* vai merkittyinä *(indirect CF)*. Tulokset osoittivat, että opiskelijat arvostivat valinnan mahdollisuutta ja kokivat, että täten palaute oli heille itselleen sopivampaa ja siten myös monen mielestä tehokkaampaa.

Työn viimeisenä osana on portfoliokokeilusta kertova monografia, joka perustuu aiempaan lisensiaatintyöhöni. Portfoliokokeilu osoitti, että näin tavallisesta kurssista poikkeava ja opiskelijalähtöinen arviointi- ja opiskelumenetelmä vaati opiskelijoilta paljon mm. itseohjautuvuutta ja omaehtoista työtä, mutta toisaalta se myös mahdollisti omien kiinnostusten ja tavoitteiden toteuttamisen oman kielitaidon tasolla. Täten portfolio parhaimmillaan voimaannutti lukuisia opiskelijoita niin englannin opiskelijoina kuin kielenkäyttäjinä. Vaikka portfoliokokeilusta on kulunut aikaa jo noin 20 vuotta, kokeilu antaa vertailupohjaa muille tässä työssä raportoiduille arviointikokeiluille.

Lopuksi

Tämän väitöstutkimuksen osana olleen kyselyn perusteella noin 15% koulumme opiskelijoista koki arvioinnin tekijänä, joka ahdisti ja lannisti heitä huomattavasti ja siten haittasi opintoja ja opiskelumotivaatiota. Ahdistusta heille aiheuttivat varsinkin painoarvoltaan tärkeät kokeet ja koetilanteet. He myös kokivat, että käytetyt arviointimenetelmät eivät antaneet heille mahdollisuutta osoittaa kaikkea osaamistaan. Opiskelijat kaipasivat myös enemmän palautetta, joka auttaa opintoja eteenpäin eikä vain totea nykyisiä taitoja tai niiden puutteita. Toisaalta noin 10 % opiskelijoista ei kokenut arvioinnin stressaavan, ahdistavan tai lannistavan yhtään. Suuri osa opiskelijoista sijoittui tähän väliin: useimmat olivat varsin tyytyväisiä arviointiin osana englannin opintojaan, mutta he ottivat sen koulutyöhön olennaisesti kuuluvana osana, joka ei todennäköisesti vaikuttanut heidän henkilökohtaiseen voimaantumiseensa sen suuremmin. Voimaantumisprosessit ovat hyvin yksilöllisiä, joten se, mikä voimaannuttaa yhtä opiskelijaa, ei välttämättä voimaannuta jotakin toista opiskelijaa.

Arvioinnin ahdistamien opiskelijoiden kokemuksista on kuitenkin opiksi otettavaa yleisemminkin. He toivoivat lisää formatiivista ja vähemmän stressiä aiheuttavaa arviointia, joka tukee oppimista eikä pelkästään totea sitä. Niinpä uusien opetussuunnitelman perusteiden mainitsema formatiivinen, oppimista edistävä arviointi sai näiltä opiskelijoilta kannatusta, vaikka se ei heille vielä tuttua ollutkaan. Lisäksi he kaipasivat monipuolisempia arviointimenetelmiä, jotka toisivat esiin tietoja tai taitoja erilaisin tavoin.

Tässä väitöstutkimuksessa kokeilin lisäksi kolmea erilaista arviointitapaa, portfolioarviointia, lunttilappukoetta sekä henkilökohtaisempaa korjaavaa pa-

lautetta, joilla kaikilla pyrittiin opiskelijoiden voimaannuttamiseen. Kaikki kokeillut arviointimenetelmät lisäsivät opiskelijoiden mahdollisuutta aktiiviseen toimijuuteen myös arvioinnissa. Lisäksi monet opiskelijat kokivat nämä arviointimenetelmät muutenkin voimaannuttavina: ne antoivat opiskelijoille mahdollisuuksia mm. omien kiinnostuksenkohteiden, oppimistarpeiden ja tavoitteiden mukaiseen opiskeluun ja arviointiin sekä koejännityksen vähentämiseen. Mikään mainituista arviointimenetelmistä ei kuitenkaan osoittautunut arviointimenetelmäksi, joka voimaannuttaisi kaikkia opiskelijoita: osa opiskelijoista selkeästi karsasti kokeiltuja menetelmiä ja kannatti 'tavallista' kokeisiin pohjautuvaa arviointia sen tuttuuden tai vaivattomuuden takia.

Tämä väitöstutkimuskokonaisuus toteutettiin pääsääntöisesti koulussa, jonka opiskelija-aines on tiedoiltaan ja taidoiltaan selkeästi valtakunnallisen keskiarvon yläpuolella. Niinpä tämän tutkimuksen tulokset eivät ole yleistettävissä muihin suomalaisiin lukioihin. Siitä huolimatta tämän tutkimuksen pohjalta suosittelen, että arvioinnin tulisi olla mahdollisimman monipuolista niin menetelmiltään kuin sisällöiltään ja että formatiiviseen, oppimista edistävään arviointiin tulisi kiinnittää entistä enemmän huomiota opintojen aikana. Lisäksi palautteen tulisi auttaa opiskelijaa opinnoissaan eteenpäin eikä vain todeta tämänhetkistä tilannetta. Erilaisia palautteenantomenetelmiä samoin kuin toimivia itsearviointimenetelmiä tulisi kehittää. Yksi vaihtoehto olisi, että lukio-opintojen aikana eri kursseilla painotettaisiin opetussuunnitelman kurssikuvausten ja -tavoitteiden sekä käytännön mahdollisuuksien mukaan erilaisia arviointitapoja ja -kohteita, jotta lukioarvioinnin kokonaisuus antaisi monipuolisen ja kokonaisvaltaisen kuvan opiskelijan tiedoista ja taidoista. Opiskelijoiden mahdollisuutta osallistua arviointiin ja sen suunnitteluun tulisi myös lisätä. Lisäksi opettajien arviointiosaamiseen ja sen päivittämiseen tulisi kiinnittää suurempaa huomiota niin opettajien perus- kuin täydennyskoulutuksessa. Lisäksi toivon, että arviointi kiinnostaisi tutkimuksentekijöitä jatkossa nykyistä enemmän, sillä se on oppimisen, opintojen ja kouluelämän keskeinen osa-alue, jota on tutkittu varsin vähän. Tämän tutkimuksen jälkeen jää paljon tutkittavaa, sillä tämä tutkimus herättää mielestäni enemmän uusia kysymyksiä kuin antaa vastauksia.

REFERENCES

Adams, R. (1991). *Protests by pupils: Empowerment, schooling and the state*. London: Falmer.

Aitken, N. (2012). Student voice in fair assessment practice. In C. F. Webber & J. L. Lupart (Eds.), *Leading student assessment*. Studies in educational leadership; ISSN 15 (pp. 175-200). Dordrecht: Springer Netherlands.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., Haapakangas, E., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York: Routledge.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301-320. doi:10.1191/0265532205lt310oa

Alsop, R., Bertelsen, M. F., & Holland, J. (2005). *Empowerment in practice: From analysis to implementation*. Washington D.C.: World Bank.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education Policy Analysis Archives*, 10, 18. DOI: http://dx.doi.org/10.14507/epaa.v10n18.2002

Anckar, J. (2011). *Assessing foreign language listening comprehension by means of the multiple-choice format: Processes and products*. Jyväskylä: University of Jyväskylä.

Anderson, L. W. (2003). *Classroom assessment: Enhancing the quality of teacher decision making*. Mahwah N.J.: Erlbaum.

Apajalahti, M. (1996). Peruskoulun oppilasarvioinnin arviointi. In A. Räisänen & T. Frisk (Eds.), *Silta uuteen opiskelija-arviointiin*: *Arviointia opiskelija-arvioinnista* (pp. 83-93). Helsinki: Opetushallitus [National Board of Education].

Atjonen, P. (2007). *Hyvä, paha arviointi*. Helsinki: Tammi.

Atjonen, P. (2014). Teachers' views of their assessment practice. *The Curriculum Journal, 25*(2), 238-259.

Atjonen, P. (2015). *Kehittävä arviointi kasvatusalalla*. Joensuu: Kirjokansi.

Atjonen, P., Halinen, I., Hämäläinen, S., Korkeakoski, E., Knubb-Manninen, G., Kupari, P., ... & Wikman, T. (2008). *Tavoitteista vuorovaikutukseen. Perusopetuksen pedagogiikan arviointi.* Koulutuksen arviointineuvoston julkaisuja, 30. Jyväskylä: Koulutuksen arviointineuvosto.

Au, W. (2009). *Unequal by design: High-stakes testing and the standardization of inequality*. New York: Routledge.

Aydın, S. (2009). Test anxiety among foreign language learners: A review of literature. *Journal of Language and Linguistic Studies, 5*(1), 127-137.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Baird, J., Hopfenbeck, T. N., Newton, P., Stobart, G., & Steen-Utheim, A. T. (2014). *State of the field review: Assessment and learning*. Oslo: Norwegian Knowledge Centre for Education.

Baird, J.A., Issacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *Policy effects of PISA*. Oxford, England: Oxford University Centre for Educational Assessment.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25.

Biggs, J. (1998). Assessment and classroom learning: A role for summative assessment? *Assessment in Education: Principles, Policy & Practice, 5*(1), 103-110. doi:10.1080/0969595980050106

Bishop, J. H., & Mane, F. (2001). The impacts of minimum competency exam graduation requirements on college attendance and early labor market success of disadvantaged students. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 51-84). New York: The Century Foundation Press.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*(2), 102-118.

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. New York: Routledge.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead UK: Open University Press.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-73.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Black, P., & Wiliam, D. (2012a). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 11-32). London: Sage.

Black, P., & Wiliam, D. (2012b). The reliability of assessments. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 243-263). London: Sage.

Blatchford, P. (1997). Pupils' self assessments of academic attainment at 7, 11 and 16 years: Effects of sex and ethnic group. *The British Journal of Educational Psychology, 67*(2), 169-84.

Blom, H. (2003). Opiskelijan oppimisen arviointi. In J. Kauppinen, L. Jääskeläinen, M. Montonen, & A. Tella (Eds.*), Lukion opetussuunnitelmaopas* (pp. 70-75). Helsinki: Opetushallitus.

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Taylor (Ed.), *Educational evaluation: New roles, new means: The 68th yearbook of the National Society for the Study of Evaluation, Part II, Vol. 68*(2) (pp. 26-50). Chicago: University of Chicago Press.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971*). Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466-481.

Bolaffi, G., Bracalenti, R., Braham, P., & Gindro, S. (2003). *Dictionary of race, ethnicity and culture.* London: Sage.

Bond, L. A. (1996). Norm-and criterion-referenced testing. ERIC/AE digest. http://ericae.net/edo/ed410316.htm

Borg, S. (2010). Language teacher research engagement. *Language Teaching, 43*(4), 391-429.

Borg, S. (2013). *Teacher research in language teaching: A critical analysis*. Cambridge: Cambridge University Press.

Boud, D. (1995). *Enhancing learning through self assessment*. London: Kogan Page.

Boud, D. (2007). Reframing assessment as if learning were important. In D. Boud, & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 14-25). London: Routledge.

Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education, 38*(8), 941-956.

Bray, E. (1986). Fitness for purpose. In R. Lloyd-Jones, E. Bray, G. Johnson & R. Currie (Eds.) *Assessment: From principles to action* (Repr. ed.) (pp. 17-34). Basingstoke: Macmillan.

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance*. OECD Education Working Papers, (71). OECD Publishing. http://dx.doi.org/10.1787/5k9fdfqffr28-en

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing, 18*(4), 393-407. doi:10.1177/026553220101800405

Broadfoot, P. (1996a). Performance assessment in perspective: International trends and current English experience. In A. Croft (Ed.), *Primary education: Assessing and planning learning* (pp. 35-65). London: Routledge/The Open University.

Broadfoot, P. (1996b). Liberating the learner through assessment. In G. Claxton, T. Atkinson, M. Osborn & M. Wallace (Eds.), *Liberating the learner - Lessons for professional development in education* (pp. 32-44). London: Routledge.

Brookhart, S. (2004). *Grading*. Upper Saddle River, NJ: Pearson Education.

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

Brookhart, S. (2012). Preventing feedback fizzle. *Educational Leadership, 70*(1), 24-29.

Brown, G. T., & Hirschfeld, G. H. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy & Practice, 15*(1), 3-17.

Brown, J. D. (2009). Principal components analysis and exploratory factor analysis–definitions, differences, and choices. *JALT Testing & Evaluation SIG Newsletter, 13*(1), 26-30.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Brummett, B. (2010). *Techniques of close reading*. Thousand Oaks, CA: Sage.

Bryant, D. A., & Carless, D. R. (2010). Peer assessment in a test-dominated setting: Empowering, boring or facilitating examination preparation? *Educational Research for Policy and Practice, 9*(1), 3-15.

Burner, T. (2015). Formative assessment of writing in English as a foreign language. *Scandinavian Journal of Educational Research,* 59, 1-23. doi:10.1080/00313831.2015.1066430

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*(4), 474-482. doi:http://dx.doi.org.ezproxy.jyu.fi/10.1037/0022-0663.79.4.474

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*(1), 1-14. doi:10.1111/j.2044-8279.1988.tb00874.x

Cassady, J. C. (2010). Test anxiety: Comtemporary theories and implications for learning. In J. C. Cassady (Ed.), *Anxiety in schools: The causes, consequences, and solutions for academic anxieties*, (pp. 5-26). New York: Peter Lang.

Chevalier, A., Gibbons, S., Thorpe, A., Snell, M., & Hoskins, S. (2009). Students' academic self-perception. *Economics of Education Review, 28*(6), 716-727.

Cizek, G. J. (2005). High-stakes testing: Contexts, characteristics, critiques, and consequences. In R. Phelps (ed.), *Defending Standardized Testing*, (pp. 23-54). Mahwah, NJ: Lawrence Erlbaum.

Clarke, N. L., Dowland, P., & Furnell, S. M. (2013). E-invigilator: A biometric-based supervision system for e-assessments. *International Conference on Information Society, June 2013,* (pp. 238-242). Piscataway, NJ: IEEE.

Cleary, T. J., & Zimmerman, B. J. (2004). Self-regulation empowerment program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools, 41*(5), 537-550. doi:10.1002/pits.10177

Cohen, A. (1987). Student processing of feedback on their compositions. In A. Wenden, & J. Rubin (Eds.), *Learner strategies in language learning. Language teaching methodology series* (pp. 57-83). Englewood Cliffs N.J.: Prentice-Hall.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education* (7th ed.). Abingdon: Routledge.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Los Angeles: Sage.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.

Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review, 56*(1), 18-37.

Cummins, J. (1996). *Negotiating identities: Education for empowerment in a diverse society*. Ontario, CA: California Association for Bilingual Education.

Cummins, J. (2001). HER classic reprint: Empowering minority students: A framework for intervention. *Harvard Educational Review*, *71*(4), 649-676.

Cummins, P. W., & Davesne, C. (2009). Using electronic portfolios for second language assessment. *The Modern Language Journal, 93*(1), 848-867.

Currie, R. (1986). Examinations in schools. In R. Lloyd-Jones, E. Bray, G. Johnson & R. Currie (Eds.) *Assessment: From principles to action* (Repr. ed.) (pp. 119-139). Basingstoke: Macmillan.

Dam, L., & Legenhausen, L. (2011). Explicit reflection, evaluation, and assessment in the autonomy classroom. *Innovation in Language Learning and Teaching, 5*(2), 177-189.

Dann, R. (2002). *Promoting assessment as learning: Improving the learning process.* London: Routledge/Falmer.

Dann, R. (2014). Assessment as learning: Blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice, 21*(2), 149-166.

Darder, A., Baltodano, M., & Torres, R. D. (2003). Critical pedagogy: An introduction. In A. Darder, M. Baltodano & R. D. Torres (Eds.), *The critical pedagogy reader* (pp. 1-21). New York, NY: RoutledgeFalmer.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64*(1), 5-30.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work.* New York: Teachers College Press.

Darling-Hammond, L., Rustique-Forrester, E., & Pecheone, R. L. (2005). *Multiple measures approaches to high school graduation*. Stanford, CA: Stanford University School Redesign Network.

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly, 43*(3), 393-415.

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self-and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education, 13*(2), 129-142.

de Lange, J. (2007). Large-scale assessment and mathematics education. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 1111-1142). Charlotte, NC: National Council of Teachers of Mathematics.

Denzin, N. (1978). *The research act: A theoretical introduction to sosiological methods* (2nd ed.). New York: McGraw-Hill.

Dermo, J. (2009). E-assessment and the student learning experience: A survey of student perceptions of e-assessment. *British Journal of Educational Technology, 40*(2), 203-214. doi:10.1111/j.1467-8535.2008.00915.x

Dickson, K. L., & Bauer, J. J. (2008). Do students learn course material during crib sheet construction? *Teaching of Psychology, 35*(2), 117-120. doi:10.1177/009862830803500211

Dickson, K. L., & Miller, M. D. (2005). Authorized crib cards do not improve exam performance. *Teaching of Psychology, 32*(4), 230-233. doi:10.1207/s15328023top3204_6

Dietel, R., Herman, J., & Knuth, R. (1991). *What does research say about assessment?* Oak Brook: NCREL. http://www.education.umd.edu/EDMS/MARCES/mdarch/pdf/msde000013.pdf

Dlaska, A., & Krekeler, C. (2013). Does grading undermine feedback? The influence of grades on the effectiveness of corrective feedback on L2 writing. *The Language Learning Journal*, 1-17. doi: 10.1080/09571736.2013.848226

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher education, 24*(3), 331-350.

Duncan, C. R., & Noonan, B. (2007). Factors affecting teachers' grading and assessment practices. *Alberta Journal of Educational Research, 53*(1), 1-21.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69-106. doi:10.1111/j.1529-1006.2004.00018.x

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

Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*(3), 353-371.

Erbe, B. (2007). Reducing test anxiety while increasing learning: The cheat sheet. *College Teaching, 55*(3), 96-98.

Erickson, G., & Åberg-Bengtsson, L. (2012). A collaborative approach to national test development. In D. Tsagari & I. Csépes (Eds.), *Collaboration in language testing and assessment* (pp. 93–108). Frankfurt am Main: Peter Lang.

Eva, K. W., & Regehr, G. (2008). "I'll never play professional football" and other fallacies of self-assessment. *Journal of Continuing Education in the Health Professions, 28*(1), 14-19.

Evans, N. (1992). Linking personal learning and public recognition. In J. Mulligan & C. Griffin (Eds.), *Empowerment through experiential learning: Explorations of good practice* (pp. 85-93). London: Kogan Page.

Everhard, C. J., & Murphy, L. (Eds.). (2015). *Assessment and autonomy in language learning*. New York: Palgrave Macmillan.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322. doi:10.3102/00346543070003287

Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing, 8*(1), 1-11. doi:http://dx.doi.org.ezproxy.jyu.fi/10.1016/S1060-3743(99)80110-6

Ferris, D. R. (2012). Written corrective feedback in second language acquisition and writing studies. *Language Teaching, 45*(4), 446-459. doi:http://dx.doi.org.ezproxy.jyu.fi/10.1017/S0261444812000250

Fetterman, D. M. (1996). Empowerment evaluation: An introduction to theory and practice. In D. M. Fetterman, S. J. Kaftarian & A. Wandersman (Eds.), *Empowerment evaluation: Knowledge and tools for self-assessment and accountability* (pp. 3-46). Thousand Oaks, CA: Sage.

Fetterman, D. M. (2001). *Foundations of empowerment evaluation*. Thousand Oaks, CA: Sage.

Fetterman, D. M. (2002). 2001 Invited address: Empowerment Evaluation: Building Communities of Practice and a Culture of Learning. *American Journal of Community Psychology, 30*(1), 89-102.

Fetterman, D. M., & Wandersman, A. (Eds.). (2005). *Empowerment evaluation principles in practice.* New York, NY: Guilford Press.

Framework curriculum for the senior secondary school 1994. Helsinki: National Board of Education.

Francina, P. X., & Joseph, M. V. (2013). Women empowerment: The psychological dimension. *Rajagiri Journal of Social Development, 5*(2), 163-176.

Francis, R. A. (2008). An investigation into the receptivity of undergraduate students to assessment empowerment. *Assessment & Evaluation in Higher Education, 33*(5), 547-557. doi:10.1080/02602930701698991

Freire, P. (1972). *Pedagogy of the oppressed*. London: Penguin Books.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123-144.

Funk, S. C., & Dickson, K. L. (2011). Crib card use during tests: Helpful or a crutch? *Teaching of Psychology, 38*(2), 114-117.

Furneaux, C., Paran, A., & Fairfax, B. (2007). Teacher stance as reflected in feedback on student writing: An empirical study of secondary school teachers in five countries. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(1), 69-94.

Gardner, J. (2012a). Assessment and learning: Introduction. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 1-8). London: Sage.

Gardner, J. (2012b). Quality assessment practice. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 103-121). London: Sage.

Garrett, N. (2009). Computer-assisted language learning trends and issues revisited: Integrating innovation. *Modern Language Journal, 93*(1), 719-740.

General Upper Secondary Schools Act, 629/1998 see *Lukiolaki 629/1998* http://www.finlex.fi/fi/laki/ajantasa/1998/19980629

Gharib, A., Phillips, W., & Mathew, N. (2012). Cheat sheet or open-book? A comparison of the effects of exam types on performance, retention, and anxiety. *Psychology Research, 2*(8), 469-478.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer Press.

Grek, S. (2009). Governing by numbers: The PISA 'effect'in Europe. *Journal of Education Policy, 24*(1), 23-37.

Guénette, D. (2007). Is feedback pedagogically correct: Research design issues in studies of feedback on writing. *Journal of Second Language Writing, 16*(1), 40-53. doi:http://dx.doi.org.ezproxy.jyu.fi/10.1016/j.jslw.2007.01.001

Guénette, D., & Lyster, R. (2013). Written corrective feedback and its challenges for pre-service ESL teachers. *Canadian Modern Language Review, 69*(2), 129-153.

Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning.* Thousand Oaks, CA: Corwin Press.

Gustafsson, J. E., & Erickson, G. (2013). To trust or not to trust? Teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability, 25*(1), 69-87.

Halonen, M. (2007). Monikielinen Suomi – maahanmuuttajataustaisten koululaisten suomen kielen taito. *Nuorisotutkimus, 25*(4), 33-49.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research & Development, 20*(1), 53-70.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*(3), 317-336.

Harjanne, P., & Tella, S. (2009). Investigating methodological reality in Finnish foreign language classrooms: Revisiting the KIELO project's rationale and research. In R. Kantelinen & P. Pollari (Eds.), *Language Education and Lifelong Learning* (pp. 135-154). Joensuu: University of Eastern Finland.

Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In *Research Evidence in Education Library,* Issue 1. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tension and synergies. *Curriculum Journal, 16*(2), 207-223.

Harlen, W. (2012a). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 87-102). London: Sage.

Harlen, W. (2012b). The role of assessment in developing motivation for learning. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 171-183). London: Sage.

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education, 10*(2), 169-208.

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice, 4*(3), 365-379. doi:10.1080/0969594970040304

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hattie, J. A. C. (2009). *Visible learning : A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hattie, J. (2012). Know thy impact. *Educational Leadership, 70*(1), 18-23.

Hayward, L. (2012). Assessment and learning: The learner's perspective. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 125-139). London: Sage.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: ASCD. http://files.eric.ed.gov/fulltext/ED352389.pdf

Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington D.C.: National Academy Press,.

Hildén, R., & Rautopuro, J. (2014). *Ruotsin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013*. [Learning outcomes for syllabus A in Swedish at the end of basic education in 2013]. Helsinki: Finnish Education Evaluation Centre/Finnish National Board of Education.

Hildén, R. Härmälä, M. Rautopuro, J., Huhtanen, M., Puukko, M., & Silverström, C. (2015). *Outcomes of language learning at the end of basic education in 2013*. Helsinki: Finnish Education Evaluation Centre/Finnish National Board of Education.

Hill, K. (2012). *Classroom-based assessment in the school foreign language classroom*. Frankfurt am Main: Peter Lang.

Holec, H. (1979). *Autonomy and foreign language learning*. Strasbourg: Council for Cultural Co-operation.

Holec, H., & Huttunen, I. (Eds.). (1997). *Learner autonomy in modern languages: Research and development*. Strasbourg: Council of Europe.

Hollister, K. K. (2007). Proctored vs. un-proctored exams in a hybrid course: A brief comparison of student results. *Journal of Educational Technology, 4*(2), 63-68.

Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky & F. M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp. 469-482). Malden, MA: Blackwell.

Huhta, A., & Hildén, R. (2013). Kielitaidon arvioinnin metodologiset vaihtoehdot. In A. Räisänen (Ed.), *Oppimisen arvioinnin kontekstit ja käytännöt* (pp. 159-186). Helsinki: Opetushallitus.

Huhta, A., & Hildén, R. (2016). Kielitutkinnot ja muu laajamittainen kielitaidon arviointi Suomessa. *AFinLA-E: Soveltavan Kielitieteen Tutkimuksia, (9)*, 3-26.

Huhta, A., Kalaja, P., & Pitkänen-Huhta, A. (2006). Discursive construction of a high-stakes test: The many faces of a test-taker. *Language Testing, 23*(3), 326-350.

Huhta, A., & Tarnanen, M. (2009). Assessment practices in the Finnish comprehensive school–what is the students' role in them. In S. May (Ed.), *LED2007: Refereed conference proceedings of the 2nd International Conference*

on Language, Education and Diversity. Hamilton, New Zealand: Wilf Malcolm Institute of Educational Research (WMIER), University of Waikato.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83-101.

Härmälä, M., & Hildén, R. (2012). Moniulotteinen, monitasoinen arviointi. *Kieli, koulutus ja yhteiskunta: Kielikoulutuspolitiikan verkoston verkkolehti*. Retrieved from http://www.kieliverkosto.fi/journals/kieli-koulutus-ja-yhteiskunta-lokakuu-2012/

Härmälä, M., Huhtanen, M., & Puukko, M. (2014). *Englannin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013*. Helsinki: Finnish Education Evaluation Centre/Finnish National Board of Education.

James, M., & Lewis, J. (2012). Assessment in harmony with our understanding of learning: Problems and possibilities. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 187-205). London: SAGE.

Janesick, V. (1994). The dance of qualitative research design: Metaphor, methodolatry, and meaning. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (pp. 209-219). Thousand Oaks, CA: Sage.

Jang, E., & Wagner, M. (2013). Diagnostic feedback in the classroom. In A. Kunnan (Ed.), *The companion to language assessment* (pp. II:6:42:693-711). Hoboken, NJ: John Wiley & Sons.

Johnson, G. (1986). Criterion-referenced assessment. In R. Lloyd-Jones, E. Bray, G. Johnson & R. Currie (Eds.) *Assessment: From principles to action* (Repr. ed.) (pp. 95-118). Basingstoke: Macmillan.

Jokivuori, P., & Hietala, R. (2007). *Määrällisiä tarinoita: Monimuuttujamenetelmien käyttö ja tulkinta*. Porvoo: WSOY.

Jones, G. M., Jones, B. D., & Hargrove, T. (2003). *The unintended consequences of high-stakes testing*. Lanham, USA: Rowman & Littlefield Publishers.

Juurakko-Paavola, T. & Takala, S. (2013). *Ylioppilastutkinnon kielikokeiden tulosten sijoittaminen Lukion opetussuunnitelman perusteiden taitotasoille.* Helsinki: Ylioppilastutkintolautakunta.
http://www.ylioppilastutkinto.fi/images/sivuston_tiedostot/Raportit_tutkimukset/FI_2013_kielikokeet_taitotasot.pdfJuurakko

Kaarninen, M., & Kaarninen, P. (2002). *Sivistyksen portti: Ylioppilastutkinnon historia*. Helsinki: Otava.

Kalz, M., & Ras, E. (Eds.). (2014). *Computer assisted assessment: Research into e-assessment*. International Conference, CAA 2014, Zeist, The Netherlands, June 30--July 1, 2014. Proceedings. Cham, Switzerland: Springer.

Karl, M. (1995). *Women and empowerment: Participation and decision making*. London: Zed Books.

Kasturirangan, A. (2008). *The balance of psychological empowerment and disempowerment for survivors of domestic violence.* (Unpublished doctoral dissertation). University of Illinois, Chicago, IL.

Kearney, S., Perkins, T., & Kennedy-Clark, S. (2016). Using self- and peer-assessments for summative purposes: Analysing the relative validity of

the AASL (authentic assessment for sustainable learning) model. *Assessment & Evaluation in Higher Education, 41*(6), 840-853. doi:10.1080/02602938.2015.1039484

Kohn, A. (1999). *Punished by rewards: The trouble with gold stars, incentive plans, A's, praise, and other bribes*. Boston: Houghton Mifflin.

Kohn, A. (2011). The case against grades. *Educational Leadership, 69*(3), 28-33.

Kohonen, V. (1997). Authentic assessment as an integration of language learning, teaching, evaluation and the teacher's professional growth. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment. Proceedings of LTRC 96* (pp. 7-22). Jyväskylä: University of Jyväskylä.

Kohonen, V. (1999). Authentic assessment in affective foreign language education. In J. Arnold (Ed.), *Affect in language learning* (pp. 279-296). Cambridge: Cambridge University Press.

Kohonen, V., & Pajukanta, U. (Eds.). (2003). *Eurooppalainen kielisalkku 2–EKS-projektin päätösvaiheen tuloksia*. Tampereen yliopiston opettajankoulutuslaitoksen julkaisuja, 28.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, N.J.: Prentice-Hall.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us.* Cambridge, Mass.: Harvard University Press.

Kornhaber, M. L., & Orfield, G. (2001). High-stakes testing policies: Examining their assumptions and consequences. In G. Orfield, & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 1--18). New York: The Century Foundation Press.

Kubiszyn, T., & Borich, G. D. (2013). *Educational testing and measurement: Classroom application and practice* (10th ed.). Hoboken, NJ: Wiley.

Kuusela, J. (2003). *Koulujen paremmuusjärjestyksestä*. Helsinki: Opetushallitus.

Lahtinen, A., & Välijärvi, J. (2014). *Ylioppilastutkinto.* Suomalaisen tiedeakatemian kannanottoja, 5. Helsinki: Suomalainen Tiedeakatemia.

Lammi, K. (2002). *Kielisalkku lukion ruotsin kielen opiskeluun motivoinnissa: Opiskelijoiden ja opettajan kokemuksia.* Jyväskylä: Jyväskylän yliopisto, soveltavan kielentutkimuksen keskus.

Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research, 15*(1), 11-33. doi:10.1177/1362168810383328

Larwin, K. (2012). Student prepared testing aids: A low-tech method of encouraging student engagement. *Journal of Instructional Psychology, 39*(2), 105-111.

Larwin, K. H., Gorman, J., & Larwin, D. A. (2013). Assessing the impact of testing aids on post-secondary student performance: A meta-analytic investigation. *Educational Psychology Review, 25*(3), 429-443. doi:http://dx.doi.org.ezproxy.jyu.fi/10.1007/s10648-013-9227-1

Leach, L., Neutze, G., & Zepke, N. (2000). Learners' perceptions of assessment: Tensions between philosophy and practice. *Studies in the Education of Adults, 32*(1), 107-119.

Leach, L., Neutze, G., & Zepke, N. (2001). Assessment and empowerment: Some critical questions. *Assessment & Evaluation in Higher Education, 26*(4), 293-305. doi:10.1080/02602930120063457

Leahy, S., & Wiliam, D. (2012). From teachers to schools: Scaling up professional development for formative assessment. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 49-71). London: SAGE.

Lee, I. (2004). Error correction in L2 secondary writing classrooms: The case of Hong Kong. *Journal of Second Language Writing, 13*(4), 285-312. doi:http://dx.doi.org.ezproxy.jyu.fi/10.1016/j.jslw.2004.08.001

Lee, I. (2005). Error correction in the L2 writing classroom: What do students think? *TESL Canada Journal, 22*(2), 1-16.

Lee, I. (2008). Student reactions to teacher feedback in two Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*(3), 144-164. doi:http://dx.doi.org.ezproxy.jyu.fi/10.1016/j.jslw.2007.12.001

Lee, I. (2014). Feedback in writing: Issues and challenges. *Assessing Writing, 19*, 1-5. doi:http://dx.doi.org.ezproxy.jyu.fi/10.1016/j.asw.2013.11.009

Lee, Y. (2015). Diagnosing diagnostic language assessment. *Language Testing, 32*(3), 299-316. doi:10.1177/0265532214565387

Leki, I. (1991). The preferences of ESL students for error correction in college-level writing classes. *Foreign Language Annals, 24*(3), 203-218. doi:10.1111/j.1944-9720.1991.tb00464.x

Leontjev, D. (2014). The effect of automated adaptive corrective feedback: L2 English questions. *Apples: Journal of Applied Language Studies, 8*(2), 43-66.

Leontjev, D. (2016). *ICAnDoiT: The impact of computerised adaptive corrective feedback on L2 English learners*. Jyväskylä: University of Jyväskylä.

Lew, M. D., Alwis, W. A. M., & Schmidt, H. G. (2010). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education, 35*(2), 135-156.

Lindström, A. (1998). *Ylioppilastutkinnon muotoutuminen autonomian aikana*. Jyväskylä: Koulutuksen tutkimuslaitos.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

Linnakylä, P., Pollari, P., & Takala, S. (Eds.). (1994). *Portfolio: Arvioinnin ja oppimisen tukena*. Jyväskylä: Kasvatustieteiden tutkimuslaitos.

Linnakylä, P., & Välijärvi, J. (2005). *Arvon mekin ansaitsemme: Kansainvälinen arviointi suomalaisen koulun kehittämiseksi*. Jyväskylä: PS-Kustannus.

Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgements in the assessment process. *Language Testing, 22*(3), 321-336.

Little, D., & Erickson, G. (2015). Learner identity, learner agency, and the assessment of language proficiency: Some reflections prompted by the

Common European Framework of Reference for Languages. *Annual Review of Applied Linguistics, 35*, 120-139.

Lizzio, A., & Wilson, K. (2008). Feedback on assessment: Students' perceptions of quality and effectiveness. *Assessment & Evaluation in Higher Education, 33*(3), 263-275.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*(4), 489-515. doi:10.1177/0265532207080770

Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing, 28*(3), 367-382. doi:10.1177/0265532211404188

Lloyd-Jones, R. (1986). An overview of assessment. In R. Lloyd-Jones, E. Bray, G. Johnson, & R. Currie (Eds.) *Assessment: From principles to action* (Repr. ed.) (pp. 1-13). Basingstoke: Macmillan.

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. *Assessment & Evaluation in Higher Education, 41*(3), 450-465.

Lorion, R. P., & McMillan, D. W. (2008). Does empowerment require disempowerment? Reflections on psychopolitical validity. *Journal of Community Psychology, 36*(2), 254-260.

Lukion opetussuunnitelman perusteet 1985. Helsinki: Kouluhallitus.

Lukion opetussuunnitelman perusteet 1994. Helsinki: Opetushallitus.

Lukion opetussuunnitelman perusteet 2003: Nuorille tarkoitetun lukiokoulutuksen opetussuunnitelman perusteet. Helsinki: Opetushallitus.

Lukion opetussuunnitelman perusteet 2015: Nuorille tarkoitetun lukiokoulutuksen opetussuunnitelman perusteet. Helsinki: Opetushallitus.

Lund, T. (2012). Combining qualitative and quantitative approaches: Some arguments for mixed methods research. *Scandinavian Journal of Educational Research, 56*(2), 155-165. doi:10.1080/00313831.2011.568674

Luukka, M., Pöyhönen, S., Huhta, A., Taalas, P., Tarnanen, M., & Keränen, A. (2008). *Maailma muuttuu - mitä tekee koulu? Äidinkielen ja vieraiden kielten tekstikäytänteet koulussa ja vapaa-ajalla* [The world changes - how does the school respond? Mother tongue and foreign language literacy practices at school and in free-time]. Jyväskylä: University of Jyväskylä, Centre for Applied Language Studies.

Lyster, R., & Ranta, L. (2013). Counterpoint piece: The case for variety in corrective feedback research. *Studies in Second Language Acquisition, 35*(01), 167-184.

Madaus, G., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In G. Orfield, & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 85-106). New York: The Century Foundation Press.

Marshall, B. (2011). *Testing English: Formative and summative approaches to English assessment*. London: Continuum.

Marshall, C., & Rossman, G. (1989). *Designing qualitative research*. Newbury Park, CA: Sage.

Marzano, R. J. (2010). *Formative assessment & standards-based grading*. Bloomington, IN: Marzano Research Laboratory.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75-100.

McMillan, J. H. (2000). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Press.

McMillan, J.H. (2003). Understanding and improving teachers' classroom assessment decision-making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 22*(4), 34-43.

McNeil, L., & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 127-150). New York: The Century Foundation Press.

Mehtäläinen, J. (1994). *Elämää akvaariossa. Kokemuksia koulukohtaisen opetussuunnitelmatyön ensivaiheista.* Kasvatustieteiden tutkimuslaitoksen julkaisusarja B. Teoriaa ja kaytäntöjä, 88. Jyväskylä: Jyväskylän yliopisto, Kasvatustieteiden tutkimuslaitos.

Mehtäläinen, J., & Välijärvi, J. (2013). *Ylioppilaskokeiden arvosanojen vertailtavuus eri aineissa vuosina 2007–2011*. Jyväskylä: Koulutuksen tutkimuslaitos.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education, 33*(1), 49-64.

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*(2), 101-113.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series, 1996*(1), i-18.

Metsämuuronen, J. (2009). *Tutkimuksen tekemisen perusteet ihmistieteissä: Tutkijalaitos [The fundamentals of research in human sciences: The researcher edition]* (4th ed.). Helsinki: International Methelp.

Metsämuuronen, J. (2016). *Oppia ikä kaikki. Matemaattinen osaaminen toisen asteen koulutuksen lopulla*. Helsinki: Kansallinen koulutuksen arviointikeskus.

Miller, R. L., & Campbell, R. (2006). Taking Stock of Empowerment Evaluation An Empirical Review. *American Journal of Evaluation, 27*(3), 296-319.

Miller, M. D., Linn, R. L., & Gronlund, N. (2013*). Measurement and assesment in teaching* (11th ed.). Singapore: Pearson.

Mitchell, B. M., & Salsbury, R. E. (2002). *Unequal opportunity: A crisis in America's schools?* Westport, CT: Bergin & Garvey.

Mok, J. (2011). A case study of students' perceptions of peer assessment in Hong Kong. *ELT Journal, 65*(3), 230-239.

Mondros, J., & Wilson, S. (1994). *Organizing for power and empowerment.* New York: Columbia University Press.

Mulligan, J., & Griffin, C. (Eds.) (1992). *Empowerment through experiential learning: Explorations of good practice*. London: Kogan Page.

Myers, M. J. (2002). Computer-assisted second language assessment: To the top of the pyramid. *ReCALL, 14*(1), 167-181.

Myyry, L., & Joutsenvirta, T. (2015). Open-book, open-web online examinations: Developing examination practices to support university students' learning and self-efficacy. *Active Learning in Higher Education, 2015, 16*(2), 119-132.

National core curriculum for basic education 2014. (2016). Helsinki: Finnish National Board of Education.

National core curriculum for upper secondary schools 2003: National core curriculum for general upper secondary education intended for young people (2004). Helsinki: Finnish National Board of Education.

National core curriculum for upper secondary schools 2015: National core curriculum for general upper secondary education intended for young people (2016). Helsinki: Finnish National Board of Education.

Natriello, G., & Pallas, A. M. (2001). The development and impact of high-stakes testing. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Iinequality and high-stakes testing in public education* (pp. 19-38). New York: The Century Foundation Press.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 149-170. doi:10.1080/09695940701478321

Newton, P. E. (2012). Validity, purpose and the recycling of results from educational assessments. In J. Gardner (Ed.), *Assessment and Learning* (2nd ed., pp. 264-276). London: Sage.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*, 1. DOI: http://dx.doi.org/10.14507/epaa.v14n1.2006

Nichols, S., & Berliner, D. (2007). *Collateral damage: How high-stakes testing corrupts America's schools.* Cambridge, MA: Harvard Education Press.

Nicol, D. (2009). Assessment for learner self-regulation: Enhancing achievement in the first year using learning technologies. *Assessment & Evaluation in Higher Education, 34*(3), 335-352.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

Noonan, B., & Duncan, C. R. (2005). Peer and self-assessment in high schools. *Practical assessment, research and evaluation, 10*(17), 1-8.

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. K. (1998). *Designing second language performance assessment*. Honolulu, HI: University of Hawai'i Press.

Notar, C. E., Herring, D. F., & Restauri, S. L. (2008). A web-based teaching aid for presenting the concepts of norm referenced and criterion referenced testing. *Education, 129*(1), p.119-124.

Oksanen, E. (2001). *Arvioinnin kehittäminen erityisopetuksessa: Diagnosoinnista oppimisen ohjaukseen laadullisena tapaustutkimuksena.* Jyväskylä: University of Jyväskylä.

O'Malley, J. M., & Pierce, L. V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Reading, MA: Addison-Wesley.

Ouakrim-Soivio, N. (2013). *Toimivatko päättöarvioinnin kriteerit? Oppilaiden saamat arvosanat ja Opetushallituksen oppimistulosten seuranta-arviointi koulujen välisten osaamiserojen mittareina*. Helsinki: Opetushallitus.

Padilla, A. M., Aninao, J. C., & Sung, H. (1996). Development and implementation of student portfolios in foreign language programs. *Foreign Language Annals, 29*(3), 429-438.

Paris, S. G., Paris, A. H., & Carpenter, R. D. (2002). Effective practices for assessing young readers. In B. M. Taylor & P. D. Pearson (Eds.), *Teaching Reading: Effective Schools, Accomplished Teachers*, (pp. 141-160). Mahwah, NJ: Lawrence Erlbaum.

Patton, C. (2012). "Some kind of weird, evil experiment": Student perceptions of peer assessment. *Assessment & Evaluation in Higher Education, 37*(6), 719-731.

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Perkins, D. D., & Zimmerman, M. A. (1995). Empowerment theory, research, and application. *American Journal of Community Psychology, 23*(5), 569-79.

Permana, S. (2013). *Empowering EFL students in writing through portfolio-based instruction.* (Doctoral dissertation, Universitas Pendidikan Indonesia).

Perusopetuksen opetussuunnitelman perusteet 2014. Helsinki: Opetushallitus.

Phelps, R. P. (2005). The rich, robust research literature on testing's achievement benefits. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 55–90). Mahwah, NJ: Psychology Press.

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing, 12*(1), 21-43.

Pickford, R., & Brown, S. (2006). *Assessing skills and practice*. Abingdon: Routledge.

Pienaar, C. (2005). Shared assessment: Empowering student writers. *Language Matters, 36*(2), 193-204.

Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *Modern Language Journal, 91*(3), 323-340. doi:10.1111/j.1540-4781.2007.00583.x

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. Philadephia: Springer Science and Business Media.

Pollari, P. (1996). ”Greetings from the CultureWorld!” Portfolio-kokeilu lukion englannin opetuksessa. In P. Pollari, M. Kankaanranta & P. Linnakylä

(Eds.), *Portfolion monet mahdollisuudet* (pp. 137-156). Jyväskylä: Kasvatustieteiden tutkimuslaitos.

Pollari, P. (1998). *"This is my portfolio": Portfolios as a vehicle for students' empowerment in upper secondary school English studies*. Unpublished Licentiate thesis. Jyväskylä: University of Jyväskylä.

Pollari, P. (2000). *"This is my portfolio": Portfolios in upper secondary school English studies*. Jyväskylä: Institute for Educational Research.

Pollari, P., Kankaanranta, M., & Linnakylä, P. (Eds.). (1996). *Portfolion monet mahdollisuudet*. Jyväskylä: Kasvatustieteiden tutkimuslaitos.

Ponterotto, J. G. (2006). Brief note on the origins, evolution, and meaning of the qualitative research concept thick description. *The Qualitative Report, 11*(3), 538-549.

Popham, W. J. (2001). Teaching to the test? *Educational Leadership, 58*(6), 16-20.

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into practice, 48*(1), 4-11.

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education, 35*(3), 277-289. doi:10.1080/02602930903541007

Pulfrey, C., Buchs, C., & Butera, F. (2011). Why grades engender performance-avoidance goals: The mediating role of autonomous motivation. *Journal of Educational Psychology, 10*3(3), 683-700.

Pulfrey, C., Darnon, C., & Butera, F. (2013). Autonomy and task performance: Explaining the impact of grades on intrinsic motivation. *Journal of Educational Psychology, 105*(1), 39-57.

Race, P., Brown, S., & Smith, B. (2005). *500 tips on assessment*. (2nd ed.). Abington: RoutledgeFalmer.

Räisänen, A., & Frisk, T. (1996). Oppilas- ja opiskelija-arvioinnin taustaa. In A. Räisänen & T. Frisk (Eds.), *Silta uuteen opiskelija-arviointiin: Arviointia opiskelija-arvioinnista* (pp. 9-26). Helsinki: Opetushallitus

Rappaport, J. (1987). Terms of empowerment/exemplars of prevention: Toward a theory for community psychology. *American Journal of Community Psychology, 15*, 121-148.

Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing, 21*(3), 249-258. doi:10.1191/0265532204lt283ed

Rea-Dickins, P. (2007). Classroom-based assessment: Possibilities and pitfalls. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 505-520). New York: Springer.

Rinne, R., Simola, H., Mäkinen-Streng, M., Silmäri-Salo, S., & Varjo, J. (2011). *Arvioinnin arvo: Suomalaisen perusopetuksen laadunarviointi rehtoreiden ja opettajien kokemana.* Jyväskylä: Suomen kasvatustieteellinen seura.

Robinson, H. A. (1994). *The ethnography of empowerment: The transformative power of classroom interaction*. Washington, DC: Falmer Press.

Rodríguez-Gómez, G., Ibarra-Sáiz, M. S., Gallego-Noche, B., Gómez-Ruiz, M., & Quesada-Serra, V. (2012). Student voice in learning assessment: A pathway not yet developed at university. *Revista Electrónica de Investigación y Evaluación Educativa, 18*(2), art.2. doi: 10.7203/relieve.18.2.1991

Rodwell, C. M. (1996). An analysis of the concept of empowerment. *Journal of Advanced Nursing, 23*(2), 305-313. doi:10.1111/1365-2648.ep8542315

Roever, C. (2001). Web-based language testing. *Language Learning & Technology, 5*(2), 84-94.

Rumberger, R. W. (2011). *Dropping out: Why students drop out of high school and what can be done about it*. Cambridge, MA: Harvard University Press.

Rust, C., O'Donovan, B., & Price, M. (2005). A social constructivist assessment process model: How the research literature shows us this could be best practice. *Assessment & Evaluation in Higher Education, 30*(3), 231-240. doi:10.1080/02602930500063819

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*(1), 77-84. doi:10.1080/0969595980050104

Sadler, D. R. (2013). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy & Practice, 20*(1), 5-19. doi:10.1080/0969594X.2012.714742

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment, 11*(1), 1-31.

Sahlberg, P. (2007). Education policies for raising student learning: The Finnish approach. *Journal of Education Policy, 22*(2), 147-171.

Sahlberg, P. (2011). The fourth way of Finland. *Journal of Educational Change, 12*(2), 173-185.

Sapon-Shevin, M., & Schniedewind, N. (1991). Cooperative learning as empowering pedagogy. In C. Sleeter (Ed.), *Empowerment through multicultural education* (pp. 159-178). Albany, N.Y.: SUNY Press.

Sargeant, J. (2008). Toward a common understanding of self-assessment. *Journal of Continuing Education in the Health Professions, 28*(1), 1-4.

Schulz, A. J., Israel, B. A., Zimmerman, M. A., & Checkoway, B. N. (1995). Empowerment as a multi-level construct: perceived control at the individual, organizational and community levels. *Health Education Research, 10*(3), 309-327.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné & M. Scriven (Eds.), *Perspectives of Curriculum Evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Sessoms, J., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing, 15*(4), 356-388. doi:10.1080/15305058.2015.1034866

Sheldon, L. E. (1988). Evaluating ELT textbooks and materials. *ELT Journal, 42*(4), 237-246.

Shepard, L. (1989). Why we need better assessments. *Educational Leadership, 46*(7), 4-9.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-391.

Shohamy, E. (2007). Tests as power tools: Looking back, looking forward. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner & C. Doe (Eds.), *Language testing reconsidered* (pp. 141-152). Ottawa: University of Ottawa Press/Les Presses de l'Université d'Ottawa.

Shohamy, E. (2014). *The power of tests: A critical perspective on the uses of language tests.* Abington: Routledge.

Shrock, S. A., & Coscarelli, W. C. (2010). Criterion-referenced measurement. In P. Peterson, E. Baker & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 31-35). Elsevier Science.

Siitonen, J. (1999). *Voimaantumisteorian perusteiden hahmottelua*. Oulu: Oulun yliopisto.

Simard, D., Guénette, D., & Bergeron, A. (2015). L2 learners' interpretation and understanding of written corrective feedback: Insights from their metalinguistic reflections. *Language Awareness, 24*(3), 233-254. doi:10.1080/09658416.2015.1076432

Simon, B. L. (1994). *The empowerment tradition in American social work: A history*. New York: Columbia University Press.

Stake, R. E. (2004). *Standards-based & responsive evaluation*. Thousand Oaks, CA: Sage.

Stiggins, R. J. (1991). Assessment Literacy. *Phi Delta Kappan, 72*(7), 534-539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238-245.

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

Stobart, G. (2012). Validity in formative assessment. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 233-242). London: Sage.

Stowell, J. R., & Bennett, D. (2010). Effects of online testing on student exam performance and test anxiety. *Journal of Educational Computing Research, 2010, 42*(2), 161-171.

Stoynoff, S. (2012). Looking backward and forward at classroom-based language assessment. *ELT Journal, 66*(4), 523-532.

Suurtamm, C., & Koch, M. J. (2014). Navigating dilemmas in transforming assessment practices: Experiences of mathematics teachers in Ontario, Canada. *Educational Assessment, Evaluation and Accountability, 26*(3), 263-287.

Syrjälä, L. (1989). *Oppilasarviointi osana lukio-opiskelua ja opetusta: Oppilaiden ja opettajien näkemyksiä ja kokemuksia Alppilan lukiossa* Oulu: University of Oulu.

Taalas, P., Tarnanen, M., Kauppinen, M., & Pöyhönen, S. (2008). Media landscapes in school and in free time–two parallel realities. *Digital Kompetanse, 4*(3), 240-256.

Takala, S. (1994). Arviointi – ongelma ja mahdollisuus. In P. Linnakylä, P. Pollari & S. Takala (Eds.), *Portfolio arvioinnin ja oppimisen tukena* (pp. 1-8). Jyväskylä: Kasvatustieteiden tutkimuslaitos

Takala, S. (1996). Suoritusarviointi puntarissa: Mahdollisuuksia ja ongelmia valtakunnallisissa kokeissa. In P. Pollari, M. Kankaanranta & P. Linnakylä (Eds.), *Portfolion monet mahdollisuudet* (pp. 207-222). Jyväskylä: Kasvatustieteen tutkimuslaitos.

Takala, S. (1997). Vieraan kielen kehittymisen arviointiperusteita. In Jakku-Sihvonen, R. (Ed.), *Onnistuuko oppiminen – oppimistulosten ja opetuksen laadun arviointiperusteita peruskoulussa ja lukiossa* (pp. 87-114). Helsinki: Opetushallitus.

Tan, K. H. K. (2012). *Student self-assessment: Assessment, learning and empowerment*. Singapore: Research Publishing.

Tan, H. K. K., Teo, C. T., & Ng, C. S. (2011). Variation in students' conceptions of self-assessment and standards. *Education Research International, Vol. 2011*, Article ID 487130. doi:10.1155/2011/487130

Taras, M. (2005). Assessment – summative and formative – some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478.

Tarnanen, M., & Huhta, A. (2011). Foreign language assessment and feedback practices in Finland. In D. Tsagari & I. Csépes (Eds.), *Classroom-based language assessment. Language testing and evaluation Vol. 25* (pp. 129-146). Frankfurt am Main: Peter Lang.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation, 27*(2), 237-246.

Thorsen, C. (2014). Dimensions of norm-referenced compulsory school grades and their relative importance for the prediction of upper secondary school grades. *Scandinavian Journal of Educational Research, 5*8(2), 127-146. doi:10.1080/00313831.2012.705322

Toomey, A. (2011). Empowerment and disempowerment in community development practice: Eight roles practitioners play. *Community Development Journal, 46*(2), 181-195.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276. doi:10.3102/00346543068003249

Torrance, H. (1996). The role of assessment in educational reform. In H. Torrance (Ed.), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment.* (Repr. ed., pp. 144-156). Buckingham: Open University Press.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327-369.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*(4), 255-272. doi:http://dx.doi.org.ezproxy.jyu.fi/10.1016/j.jslw.2007.06.003

Tuokko, E. (2000). *Peruskoulun 9. vuosiluokan englannin (A1-kieli) oppimistulosten kansallinen arviointi 1999.* Helsinki: Opetushallitus.

Tuokko, E. (2002). *Perusopetuksen päättövaiheen ruotsin kielen oppimistulosten kansallinen arviointi 2001*. Helsinki: Opetushallitus.

Tuokko, E. (2007). *Mille tasolle perusopetuksen englannin opiskelussa päästään? Perusopetuksen päättövaiheen kansallisen arvioinnin 1999 Eurooppalaisen viitekehyksen taitotasoihin linkitetyt tulokset.* Jyväskylä: Jyväskylän yliopisto.

Tuomi, J., & Sarajärvi, A. (2009). *Laadullinen tutkimus ja sisällönanalyysi* (6., uud. laitos). Helsinki: Tammi.

Turunen, H., Herajärvi, S., Kupiainen, S., Pirkkalainen, L., Syyrakki, S., Virtanen, V., ... & Ohranen, S. (2011). *Lukiokoulutuksen opetussuunnitelman perusteiden ja tuntijaon toimivuuden arviointi.* Koulutuksen arviointineuvoston julkaisuja, 55. Jyväskylä: Koulutuksen arviointineuvosto.

Van Maele, J., Baten, L., Beaven, A., & Rajagopal, K. (2013). E-Assessment for Learning: Gaining insight in language learning with online assessment environments. In B. Zou, M. Xing, C. Xiang, Y. Wang & M. Sun (Eds.), *Computer-assisted foreign language teaching and learning: Technological advances* (pp. 246-261). Hershey, PA: IGI Global.

Valencia, S. (1990). Assessment: A portfolio approach to classroom reading assessment: The whys, whats, and hows. *The Reading Teacher, 43*(4), 338-340.

Valencia, S. W., Hiebert, E. H., & Afflerbach, P. (1994). *Authentic reading assessment: Practices and possibilities*. Newark, DE: International Reading Association.

VanderPlaat, M. (1998). Empowerment, emancipation and health promotion policy. *Canadian Journal of Sociology/Cahiers canadiens de sociologie, 23*(1), 71-90.

Virta, A. (2002). Arviointi oppimisen ja opetuksen punaisena lankana. In E. Lehtinen & T. Hiltunen (Eds.), *Oppiminen ja opettajuus* (pp. 63-86). Turku: Turun yliopisto. Kasvatustieteiden tiedekunnan julkaisuja B: 71.

Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy, (35).* http://www.umanitoba.ca/publicat ions/cjeap/articles/volante.html

Volante, L. (2006). An alternative vision for large-scale assessment in Canada. *Journal of Teaching and Learning, 4*(1). http://137.207.184.83/ojs/leddy/index.php/JTL/article/view/89

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

Välijärvi, J. (1981*). Lukion oppilasarvioinnin kehittäminen osana opetuksellista kokonaisuudistusta.* Jyväskylä: Kasvatustieteiden tutkimuslaitos.

Välijärvi, J. (1985). *Oppilasarvionti lukion opetuksellisen kehittämisen kohteena* Jyväskylä: Kasvatustieteiden tutkimuslaitos.

Välijärvi, J. (1993). *Kurssimuotoisuus opetussuunnitelman moduulirakenteen sovelluksena lukiossa*. Jyväskylä: Kasvatustieteiden tutkimuslaitos.

Välijärvi, J. (1996). Oppilasarviointi opiskelun uudistumisen tukena ja tukahduttajana lukiossa. In A. Räisänen & T. Frisk (Eds.), *Silta uuteen opiskelija-arviointiin* (pp. 123-142). Helsinki: Opetushallitus.

Välijärvi, J. (1998). National assessment and the criteria for evaluation. In H. Jokinen, & J. Rushton (Eds.), *Changing contexts of school development – the challenges to evaluation and assessment. Report on the international SBD seminar April 21-26, 1997 Jyväskylä*. Jyväskylä: Institute for Educational Research.

Välijärvi, J., & Tuomi, P. (1995). *Lukio nuorten valintojen ja oppimisen ympäristönä.* Jyväskylä: Kasvatustieteiden tutkimuslaitos.

Välijärvi, J., Huotari, N., Iivonen, P., Kulp, M., Lehtonen, T., Rönnholm, H., . . . Ohranen, S. (2009). *Lukiopedagogiikka*. Koulutuksen arviointineuvoston julkaisuja 40. Jyväskylä: Koulutuksen arviointineuvosto.

Vänttinen, M. (2011). *Oikeasti hyvä numero: Oppilaiden arvioinnin totuudet ja totuustuotanto rinnakkaiskoulusta yhtenäiskouluun*. Joensuu: Itä-Suomen yliopisto.

Wahlroos, S. (2012). Monipuolisemmalla arvioinnilla kohti yhteismitallisempia arvosanoja. *Kieli, koulutus ja yhteiskunta*. http://www.kieliverkosto.fi/journals/kieli-koulutus-ja-yhteiskunta-lokakuu-2012/

Weber, K., & Patterson, B. R. (2000). Student interest, empowerment and motivation. *Communication Research Reports, 17*(1), 22-29. doi:10.1080/08824090009388747

Whitworth, R. (1990). Using crib notes as a learning device. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 64*(1), 23-24. doi:10.1080/00098655.1990.9955797

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.

Wiggins, G. (1990). The case for authentic assessment. *ERIC digest.* http://eric.ed.gov/?id=ed328611

Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Wiggins, G. (2012). Seven keys to effective feedback. *Educational Leadership, 70*(1), 10-16.

Wilcox, B. (1992). *Time-constrained evaluation: A practical approach for LEAs and schools.* London: Routledge.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3-14.

Wiliam, D. (2012). Feedback: Part of a system. *Educational Leadership, 70*(1), 31-34.

Williams, J. B., & Wong, A. (2009). The efficacy of final examinations: A comparative study of closed-book, invigilated exams and open-book,

open-web exams. *British Journal Of Educational Technology, 40*(2), 227-236. doi:10.1111/j.1467-8535.2008.00929.x

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17. doi:10.1207/s15326977ea1001_1

Yorke, M. (2007). *Grading student achievement in higher education: Signals and shortcomings*. Abington: Routledge.

Zhao, H. (2014) Investigating teacher-supported peer assessment for EFL writing. *ELT Journal 68*(2), 155-168.

Zimmerman, M. (1995). Psychological empowerment: Issues and illustrations. *American Journal of Community Psychology, 23*(5), 581-599.

Zimmerman, M. A. (2000). Empowerment theory: Psychological, organizational and community levels of analysis. In J. Rappaport & E. Seidman (Eds.), *Handbook of community psychology* (pp. 43-63). New York: Kluwer Academic/Plenum.

Zimmerman, M., & Rappaport, J. (1988). Citizen participation, perceived control, and psychological empowerment. *American Journal of Community Psychology, 16*(5), 725-750.

APPENDIX 1

Hyvä lukiolaisemme

Teen tutkimusta arvioinnista englannin kielessä. Tutkimus tulee perustumaan tähän kyselyyn ja opiskelijahaastatteluihin.
Kyselyssä tiedustellaan näkemystäsi ja kokemustasi englannin kielen opetuksesta ja arvioinnista lukiossa.

Vastatessasi ajattele lukioaikaasi ja lukion englannin opetusta kokonaisuutena.

Vastauksesi ovat ehdottoman luottamuksellisia. Julkaistavasta raportista ei yksittäisiä vastauksia voida jäljittää.

Pirjo Pollari
Normaalikoulu
Jyväskylän yliopisto
pirjo.pollari@norssi.jyu.fi

Seuraava

A1 Sukupuolesi

- ○ Nainen
- ○ Mies

A2 Vuosikurssisi

- ○ 2. vsk
- ○ 3. vsk
- ○ 4. vsk

Edellinen | Seuraava

A3 Oletko jo kirjoittanut englannin (ylioppilaskokeen)?

- ○ Olen, enkä aio kirjoittaa sitä uudestaan.
- ○ Olen, mutta kirjoitan sen uudestaan korottaakseni arvosanaa.
- ○ Kirjoitan sen tänä keväänä.
- ○ En ole vielä, mutta kirjoitan sen ensi syksynä.
- ○ En ole vielä, kirjoitan sen myöhemmin.

Edellinen | Seuraava

A4 Minkä kouluarvosanan itse antaisit englannin kielen taidoistasi tällä hetkellä?

A5 Mikä oli edellinen englannin kurssiarvosanasi?

A6 Mikä oli peruskoulun päättötodistuksen englannin arvosanasi?

A7 Kuinka monta englannin kurssia olet lukiossa suorittanut?

A8 Montako englannin opettajaa sinulla on ollut lukiossa?

Edellinen Seuraava

B Missä määrin seuraavat tavoitteet ovat ohjanneet koko lukio-opiskeluasi?

	Erittäin paljon	Melko paljon	Jonkin verran	Ei lainkaan
Saada hyvä yleissivistys	◉	○	○	○
Opiskella mahdollisimman paljon kiinnostavia kursseja	○	○	○	○
Oppia tekemään päätöksiä ja valintoja	○	○	○	○
Hyvä menestyminen ylioppilaskirjoituksissa	○	○	○	○
Opetella itse ottamaan vastuuta asioista	○	○	○	○
Oppia suunnittelemaan opintojani ja tulevaisuuttani	○	○	○	○
Hyvä päättötodistus	○	○	○	○
Selvittää itselleni, mitä isona oikeastaan haluan tehdä	○	○	○	○
Mennä samoille kursseille kuin kaverinikin	○	○	○	○
Oppia tulemaan toimeen erilaisissa ryhmissä ja erilaisten ihmisten kanssa	○	○	○	○
Oppia tuntemaan itseni, vahvuuteni ja heikkouteni	○	○	○	○
Oppia ilmaisemaan itseäni	○	○	○	○
Päästä lukion jälkeen opiskelemaan tavoittelemaani ammattiin	○	○	○	○

Edellinen | Seuraava

C Seuraavat väittämät liittyvät englannin kurssien arvioinnin toteutukseen opiskelemillasi kursseilla.

	Jokaisella kurssilla	Lähes jokaisella kurssilla	Joillakin kursseilla	1-2 kurssilla	Ei yhdelläkään kurssilla
Opettaja on kertonut kurssin tavoitteet kurssin alussa.	○	○	○	○	○
Opettaja on kertonut kurssin alussa, mistä kurssin arvosana koostuu.	○	○	○	○	○
Olen tiennyt kurssin alusta asti, miten eri suoritukset (esim. sanakokeet, koe) painottuvat kurssiarvosanassa.	○	○	○	○	○
Opettaja on kertonut, millainen koe tulee olemaan.	○	○	○	○	○
Olen tiennyt etukäteen, jos sanakokeissa ei hyväksytä kuin ko. tekstin sanoja.	○	○	○	○	○
Opettaja on selittänyt kokeiden (myös sanakokeiden, tms.) palautuksen yhteydessä miten ne on arvioitu.	○	○	○	○	○
Opettaja on päättänyt yksin kaiken arviointiin liittyvän.	○	○	○	○	○

Edellinen Seuraava

Kysymys C jatkuu

	Jokaisella kurssilla	Lähes jokaisella kurssilla	Joillakin kursseilla	1-2 kurssilla	Ei yhdelläkään kurssilla
Me opiskelijat olemme voineet ehdottaa tai valita arviointitapoja, aiheita tms.	○	○	○	○	○
Olemme halutessamme voineet vaikuttaa siihen, mistä osatekijöistä kurssin arvosana koostuu.	○	○	○	○	○
Olen ottanut kantaa arviointitapoihin tai arvioinnin perusteisiin.	○	○	○	○	○
Meillä on ollut mahdollisuus itse ehdottaa kurssiarvosanaa ja perustella se.	○	○	○	○	○
Olen ehdottanut ja perustellut arvosanaa itselleni.	○	○	○	○	○
Koen, että olen voinut oikeasti vaikuttaa arviointiin ja kurssiarvosanaani.	○	○	○	○	○

Edellinen | Seuraava

D Kuinka usein opettajasi ovat käyttäneet seuraavia arviointimenetelmiä?

	Jokaisella kurssilla	Lähes jokaisella kurssilla	Joillakin kursseilla	1-2 kurssilla	Ei yhdelläkään kurssilla
Kurssikoe	○	○	○	○	○
Sanakokeita	○	○	○	○	○
Kuuntelukoe tai kokeita	○	○	○	○	○
Suullinen koe	○	○	○	○	○
Muu suullinen näyttö (esim. esitelmä)	○	○	○	○	○
Kirjoitelma	○	○	○	○	○
Kotona tehtävä kirjoitelma	○	○	○	○	○
Esitelmä, tutkielma tai muu laajempi työ	○	○	○	○	○
Tiivistelmä	○	○	○	○	○
Pistokoe (esim. sanakoe josta ei ilmoiteta etukäteen)	○	○	○	○	○
Alkutesti, joka testaa kuinka hyvin osaatte asiat jo entuudestaan	○	○	○	○	○
Formatiiviset testit (=eivät vaikuta lopulliseen arvosanaan mutta kertovat kuinka hyvin asia on opittu)	○	○	○	○	○
Portfolio (arviointi perustuu kokonaan tai osittain omiin töihin ja mahdollisesti niiden itsearvioihin)	○	○	○	○	○
Väittely tai ryhmäkeskustelu	○	○	○	○	○
Näytelmä tms. esitys; video- tai äänitallenne	○	○	○	○	○

Edellinen | Seuraava

Kysymys D jatkuu

	Jokaisella kurssilla	Lähes jokaisella kurssilla	Joillakin kursseilla	1-2 kurssilla	Ei yhdelläkään kurssilla
Posteri	○	○	○	○	○
Lunttilappukoe (rajatut, omatekemät muistiinpanot mukana kokeessa)	○	○	○	○	○
Koe, johon saa tuoda mukanaan kirjan ja/tai kaikki muistiinpanot	○	○	○	○	○
Koe/tehtävä, joka tehdään parin tai ryhmän kanssa mutta kullakin on oma osuutensa (esim. suullinen koe)	○	○	○	○	○
Koe/tehtävä, joka tehdään parin tai ryhmän kanssa eikä kenenkään omaa osuutta voi erottaa (esim. yhteinen kirjallinen esitelmä)	○	○	○	○	○
Itsearviointi	○	○	○	○	○
Itsearviointi, joka vaikuttaa kurssiarvosanaan	○	○	○	○	○
Etänä (esim. netin välityksellä) tehty koe	○	○	○	○	○
Tietokoneen avulla tehty koe (kurssikoe, sanakoe, tms.)	○	○	○	○	○
iPadin tai tietokoneen avulla tehty tai tallennettu suullinen koe	○	○	○	○	○
Pari- tai kaveriarvioinnit	○	○	○	○	○
Pari- tai kaveriarvioinnit, jotka vaikuttavat arvosanaan	○	○	○	○	○

E Millaisia arviointimenetelmiä haluaisit käytettävän nykyistä enemmän? Miksi?

F Millaisia arviointimenetelmiä haluaisit käytettävän nykyistä vähemmän? Miksi?

Edellinen Seuraava

G Nämä väittämät käsittelevät englannin kurssien arvioinnin, esimerkiksi saamasi arvosanan, osuvuutta ja oikeudenmukaisuutta. Ota niihin kaikkiin kantaa ajatellen kaikkia lukiossa suorittamiasi englannin kursseja.

	Jokaisella kurssilla	Lähes jokaisella kurssilla	Joillakin kursseilla	1-2 kurssilla	Ei yhdelläkään kurssilla
Arviointi on antanut minulle hyvän kuvan osaamisestani.	○	○	○	○	○
Tiedän, miksi olen saanut sen arvosanan kuin olen saanut.	○	○	○	○	○
Koen saaneeni sen mitä olen ansainnutkin.	○	○	○	○	○
Kurssien arvioinnissa ei ole riittävästi otettu huomioon kaikkea osaamistani.	○	○	○	○	○
Olen voinut kysyä halutessani perusteluja koe- tai kurssiarvosanoille.	○	○	○	○	○
Arviointi ei ole ollut reilua.	○	○	○	○	○
Arviointi on lannistanut tai vähentänyt haluani opiskella.	○	○	○	○	○
Arviointi on ollut kannustavaa.	○	○	○	○	○
Arviointinäyttöjä on ollut tasaisesti kurssin aikana.	○	○	○	○	○
Arviointimenetelmät ovat olleet monipuolisia.	○	○	○	○	○
Kurssien arvostelussa kurssikokeiden osuus on painottunut liikaa.	○	○	○	○	○
Kaikki kielitaidon osa-alueet on arvioinnissa otettu huomioon.	○	○	○	○	○
Arviointi on painottunut liikaa kurssin loppuun.	○	○	○	○	○

Kysymys G jatkuu

	Jokaisella kurssilla	Lähes jokaisella kurssilla	Joillakin kursseilla	1-2 kurssilla	Ei yhdelläkään kurssilla
Arviointi (kokeet, kirjoitelmat, jne.) on aiheuttanut minulle liikaa stressiä.	○	○	○	○	○
Arviointi on suosinut joitakin opiskelijoita tai opiskelijatyyppejä.	○	○	○	○	○
Tuntiaktiivisuus on vaikuttanut liikaa arvosanaan.	○	○	○	○	○
Kurssiarvosanasta on jätetty pois heikoimmat suoritukset (esim. heikoin sanakoe).	○	○	○	○	○
Kaikkein heikoimmin menneet suoritukseni ovat vaikuttaneet arvosanaan liikaa.	○	○	○	○	○
En tiedä miksi olen saamani arvosanan saanut.	○	○	○	○	○
Kurssiarvosanani ovat perustuneet niihin asioihin, joista kurssin alussa sovimme.	○	○	○	○	○
Opiskelijan persoonallisuus on vaikuttanut arvosanaan.	○	○	○	○	○
Olen saanut alhaisempia arvosanoja kuin olisin mielestäni ansainnut.	○	○	○	○	○
Olen saanut korkeampia arvosanoja kuin olisin mielestäni ansainnut.	○	○	○	○	○

Edellinen Seuraava

H Jos olet saanut alhaisempia arvosanoja kuin olisit mielestäsi ansainnut, minkä arvelet olevan siihen syynä?

I Jos olet saanut korkeampia arvosanoja kuin olisit mielestäsi ansainnut, minkä arvelet olevan siihen syynä?

J Jos olet jo kirjoittanut englannin (ylioppilaskokeen), saitko mielestäsi ansaitsemasi arvosanan? Miksi/miksi et?

Edellinen | Seuraava

K Kuinka hyödyllisenä englannin kielen oppimiselle pidät seuraavia arviointitapoja?

	Erittäin hyödyllinen	Jokseenkin hyödyllinen	En osaa sanoa	Jokseenkin hyödytön	Täysin hyödytön
Arvosana perustuu pääasiassa kurssikokeeseen	○	○	○	○	○
Arvosana perustuu monenlaiseen arviointinäyttöön (sanakokeet, kuuntelut, koe, jne.)	○	○	○	○	○
Portfolioarviointi	○	○	○	○	○
Lunttilappukoe	○	○	○	○	○
Kirja/muistiinpanot mukaan kokeeseen	○	○	○	○	○
Ei kurssikoetta ollenkaan	○	○	○	○	○
Sanakokeet	○	○	○	○	○
Kotona tehtävä kirjoitelma, esitelmä tms.	○	○	○	○	○
Alkutestit (=aiempien tietojen testaaminen)	○	○	○	○	○
Pistokoe (esim. sanakoe josta ei ilmoiteta etukäteen)	○	○	○	○	○

Edellinen Seuraava

Kysymys K jatkuu

	Erittäin hyödyllinen	Jokseenkin hyödyllinen	En osaa sanoa	Jokseenkin hyödytön	Täysin hyödytön
Ns. formatiiviset kokeet ja testit, jotka eivät vaikuta lopulliseen arvosanaan ollenkaan	○	○	○	○	○
Erilaiset itsearviot	○	○	○	○	○
Arvosana perustuu lähinnä tuntityöskentelyyn	○	○	○	○	○
Arvosana perustuu lähinnä itsearvioon	○	○	○	○	○
Pari- tai kaveriarvioinnit	○	○	○	○	○
Tietokoneavusteinen koe	○	○	○	○	○
Suullisen kielitaidon koe tai muu suullinen näyttö	○	○	○	○	○
Suullisen kielitaidon valtakunnallinen, virallinen koe (=suullisen kurssin koe)	○	○	○	○	○
Ylioppilaskoe	○	○	○	○	○

Edellinen Seuraava

L Jos pidät joitakin arviointitapoja oppimisen kannalta erittäin hyödyllisinä, niin miksi?

M Jos pidät joitakin arviointitapoja oppimisen kannalta täysin hyödyttöminä, niin miksi?

Edellinen Seuraava

N Seuraavat väittämät liittyvät henkilökohtaiseen näkemykseesi/kokemukseesi arviointimenetelmistä ja arvioinnin toteutuksesta. Ota niihin kantaa ajatellen englannin kursseja kokonaisuutena.

	Täysin samaa mieltä	Jokseenkin samaa mieltä	En osaa sanoa	Jokseenkin eri mieltä	Täysin eri mieltä
Kurssien arviointimenetelmät ovat minulle yhdentekeviä.	○	○	○	○	○
Arviointi ahdistaa ja stressaa minua.	○	○	○	○	○
Arviointimenetelmät antavat minulle mahdollisuuden osoittaa miten paljon osaan.	○	○	○	○	○
Käytetyt arviointimenetelmät lannistavat minua.	○	○	○	○	○
Arviointi vain toteaa, se ei ohjaa tai auta oppimaan paremmin.	○	○	○	○	○
Arvosanaan vaikuttavia töitä, kokeita tms. pitäisi olla nykyistä vähemmän.	○	○	○	○	○
Opettajat päättääköot kaiken arviointiin liittyvän.	○	○	○	○	○
Arvosanan pitäisi perustua vain taitoihin, ei esim. ahkeruuteen tai aktiivisuuteen.	○	○	○	○	○
On hyvä, jos heikoimmat suoritukset voidaan jättää arvosanasta pois.	○	○	○	○	○

Edellinen | Seuraava

Kysymys N jatkuu

	Täysin samaa mieltä	Jokseenkin samaa mieltä	En osaa sanoa	Jokseenkin eri mieltä	Täysin eri mieltä
Jos kokeita tehdään ilman opettajan valvontaa, opiskelijat huijaavat.	○	○	○	○	○
Eri opettajat painottavat arvioinnissa ihan eri asioita.	○	○	○	○	○
Arviointi on vähentänyt haluani oppia.	○	○	○	○	○
Jos opettaja pyytää itsearviota ja/tai kurssiarvosanaehdotustani, teen sen aina.	○	○	○	○	○
Arvioinnin tulee olla kaikille sama.	○	○	○	○	○
Opettaja päättää arvosanani yksin - en minä voi siihen vaikuttaa.	○	○	○	○	○
En pidä itsearvioinneista.	○	○	○	○	○
Haluan enemmän valtaa päättää siitä, kuinka minua arvioidaan.	○	○	○	○	○
Jos saan itse valita arviointitapoja, se motivoi minua opiskelemaan enemmän.	○	○	○	○	○
En ole kiinnostunut kurssien arviointiperusteista.	○	○	○	○	○

Edellinen | Seuraava

O Haluatko enemmän valtaa vaikuttaa arviointiin? Miksi? Miten! Miksi et?

Edellinen Seuraava

P Seuraavat väittämät käsittelevät ylioppilaskirjoituksia (ylioppilaskoetta). Ota niihin kaikkiin kantaa, vaikka sinulla ei vielä olisi henkilökohtaista kokemusta kirjoituksista.

	Täysin samaa mieltä	Jokseenkin samaa mieltä	En osaa sanoa	Jokseenkin eri mieltä	Täysin eri mieltä
Lukiossa pitäisi käyttää vain samoja arviointitapoja kuin yo-kirjoituksissakin.	○	○	○	○	○
Opettajat opettavat vain ylioppilaskirjoituksia varten.	○	○	○	○	○
Opiskelen englantia elämää ja tulevaisuuttani enkä yo-kirjoituksia varten.	○	○	○	○	○
Yo-kirjoitusten arviointi ei vastaa opettajien arviointikäytänteitä.	○	○	○	○	○
Yo-kirjoitukset pelottavat minua.	○	○	○	○	○
Opettajani ovat opastaneet minua liian vähän yo-kirjoituksia varten.	○	○	○	○	○
Englannin opintojeni tärkein tavoite minulle on hyvä arvosana yo-kirjoituksissa.	○	○	○	○	○
Voin yo-kirjoituksissa luotettavasti osoittaa, kuinka hyvin englantia osaan.	○	○	○	○	○
Opettajien antama arviointi antaa oikeamman kuvan osaamisestani kuin yo-koe.	○	○	○	○	○

Edellinen | Seuraava

Q Mitä mieltä olet A-englannin yo-kokeesta? Millaisia tunteita/ajatuksia koe sinussa herättää? Miksi?

Edellinen Seuraava

R Seuraavat väittämät liittyvät saamaasi palautteeseen. Ota niihin kaikkiin kantaa ajatellen englannin opetusta ja opiskeluasi lukiossa kokonaisuutena.

	Täysin samaa mieltä	Jokseenkin samaa mieltä	En osaa sanoa	Jokseenkin eri mieltä	Täysin eri mieltä
Saan kurssien aikana tarpeeksi palautetta osaamisestani, jotta voin vaikuttaa opiskeluuni ko. kurssin aikana.	○	○	○	○	○
Opettajani kirjoittaa tarpeeksi palautetta esim. kirjoitelman loppuun.	○	○	○	○	○
En tiedä, mitkä ovat heikkouteni ja /tai vahvuuteni englannissa.	○	○	○	○	○
Saan tarpeeksi palautetta myös muilta opiskelijoilta.	○	○	○	○	○
Katson aina tarkasti virheeni ja korjaukset palautetuista kokeista ja kirjoitelmista.	○	○	○	○	○
Muilta opiskelijoilta saamani palaute on hyödyllistä.	○	○	○	○	○
Haluaisin opettajalta enemmän palautetta taidoistani.	○	○	○	○	○
Haluaisin opettajalta enemmän palautetta siitä, kuinka minun tulisi kehittää opiskeluani.	○	○	○	○	○

Edellinen Seuraava

Kysymys R jatkuu

	Täysin samaa mieltä	Jokseenkin samaa mieltä	En osaa sanoa	Jokseenkin eri mieltä	Täysin eri mieltä
Opettajan antama palaute on auttanut minua korjaamaan kielitaitoni puutteita.	○	○	○	○	○
Saamani kurssiarvosana ohjaa seuraavan kurssin opiskeluani.	○	○	○	○	○
Saamani koenumero kiinnostaa minua enemmän kuin opettajan korjaukset tai kommentit.	○	○	○	○	○
Saamani arviointi ja palaute on auttanut ja ohjannut opiskeluani.	○	○	○	○	○
Saamani arviointi ja palaute on motivoinut minua.	○	○	○	○	○
Arvioin itse osaamistani, kun tarkastamme tunneilla (koti)tehtäviä.	○	○	○	○	○
Saan tarpeeksi tietoa osaamisestani mm. tehtävien tekemisen ja tarkastuksen avulla.	○	○	○	○	○

Edellinen Seuraava

S Jos et ole saanut riittävästi palautetta osaamisestasi, niin millä tavoin annettua palautetta toivoisit?

T Mikä on mielestäsi arvioinnin tärkein tehtävä? Siis miksi koulussa tarvitaan arviointia? Vai tarvitaanko sitä?

U Kysely on nyt lopussa. Olisiko Sinulla vielä muita arviointiin liittyviä terveisiä opettajillesi?

Edellinen Seuraava

ORIGINAL PAPERS

I

THE POWER OF ASSESSMENT: WHAT (DIS)EMPOWERS STUDENTS IN THEIR EFL ASSESSMENT IN A FINNISH UPPER SECONDARY SCHOOL?

by

Pollari, Pirjo (2017a)

Apples – Journal of Applied Language Studies, 11(2), 147–175.

Apples – Journal of Applied Language Studies
Vol. 11, 2, 2017, 147–175

The power of assessment: What (dis)empowers students in their EFL assessment in a Finnish upper secondary school?

Pirjo Pollari, University of Jyväskylä

Assessment wields a great deal of power over students. Yet, there is little research on how students, either in general or as individuals, experience assessment. Therefore, this study aimed to explore what disempowers or empowers students in EFL assessment. A total of 146 students from one Finnish upper secondary school answered a questionnaire on assessment and feedback in their EFL studies. The study utilises mixed methods: primarily, the questionnaire data was analysed quantitatively (principal component analysis, step-wise regression analysis), secondarily, qualitative data and analysis were also used. The analyses showed that students reacted to assessment in highly individual ways. While many students appreciated assessment, a significant minority found assessment disempowering. Assessment caused them considerable anxiety and they did not consider assessment methods good and versatile enough. Furthermore, feedback played a role in assessment disempowerment. Therefore, EFL assessment and feedback methods should be more versatile in order to also cater for those students who currently may feel disempowered by assessment.

Keywords: assessment, students' experiences, empowerment, disempowerment, upper secondary education, EFL

1 Introduction

Assessment plays a powerful role in education. It determines whether students succeed or not; in other words, it defines value (see e.g. Atjonen, 2007, p. 19; Linnakylä & Välijärvi, 2005, p. 16) and worth (see Shohamy, 2001) of their work, and thus affects them significantly. It may motivate students externally but may also cause them stress and anxiety. Yet, there is little research on students' experiences of the power of assessment internationally (Aitken, 2012), and hardly any in Finland. Furthermore, in the context of foreign language (FL) education in Finnish upper secondary schools, there is none so far. So, how do upper secondary students actually experience assessment as part of their EFL studies? In their opinion, does it guide and improve their learning or does it

Corresponding author's email: pirjo.pollari@norssi.jyu.fi
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies
University of Jyväskylä

http://apples.jyu.fi
http://dx.doi.org/10.17011/apples/urn.201708233543

cause them stress and dishearten them? Do students feel that they have power over assessment, and if they do not, would they like to have some?

To find that out, students at one Finnish school answered a web-based questionnaire dealing with assessment and feedback during their upper secondary English studies. Even though the first overall results showed that most students were quite satisfied with assessment and its methods, content and timing, for instance, there were also those who felt that assessment had rendered them powerless and distressed. Subsequently, some of them had lost their motivation to study English. With the majority of students considering assessment good, accurate and fair, why did these students feel so differently? What disempowered them in assessment?

Firstly, I will define the concept of assessment briefly and then discuss empowerment and disempowerment and their role in assessment. Next, I will introduce the present study, its participants as well as data collection and analysis methodology. The main findings of the entire survey will be presented in a nutshell, but the key focus of this article is centred upon what the data revealed about students' empowerment and, in particular, disempowerment in assessment, and their possible predictors. Moreover, to illuminate students' experiences at an individual level, I will present three student cases. Finally, I will discuss the findings, their limitations and possible implications.

2 Conceptual framework

2.1 Assessment as defined in this article

Assessment is a broad concept, with various definitions for different contexts and purposes (e.g. Wiliam, 2011). In the school context, assessment has often been divided into diagnostic, formative and summative assessment, with formative assessment primarily supporting learning and summative reporting the results of learning. Currently, assessment at school is increasingly defined as assessment *for* learning and assessment *of* learning (e.g. Black & Wiliam, 1998, 2012; Gardner, 2012).

In this article, the term *assessment* refers to assessment as it is generally understood in Finnish schools and also defined by the *National core curriculum for upper secondary schools 2003*, which was in force at the time of this study. Accordingly, assessment here entails all aspects of classroom assessment, from various forms of formative assessment and feedback to a variety of student work, quizzes and tests, and, finally, to the assigning of summative course grades.

There is little research on assessment in upper secondary or foreign language education in Finland, but the little there is suggests that assessment in upper secondary school focuses on grading, which, in turn, is mostly based on teacher-controlled tests, and is neither very versatile nor interactive (Välijärvi et al., 2009). Self- and peer-assessments do not appear very common for summative purposes in FL education (Tarnanen & Huhta, 2011). Furthermore, the Matriculation Examination, the only high-stakes examination in the Finnish school context taken towards the end of upper secondary education, seems to affect teaching, studying and assessment practices in upper secondary education (e.g. Atjonen, 2007).

As students receive at least approximately 60 course grades (and at least six English grades) during their upper secondary education in Finland, it is safe to

say that assessment and grading, although part of upper secondary pedagogy in general, are a prominent phenomenon also *per se*. Grades are probably the most tangible recognition that students receive of their work. Moreover, according to extensive research, assessment has a crucial impact on students' studying and learning as well as on their motivation, self-concept and self-efficacy (e.g. Atjonen, 2007; Crooks, 1988; Harlen, 2012; Herman & Linn, 2014; Reay & Wiliam, 1999; Takala, 1994; Välijärvi, 1996).

2.2 Empowerment

The roots of empowerment have been attributed to various origins, ranging from Enlightenment to Marxism, from Civil Rights to feminist theories (e.g. Simon, 1994; Traynor, 2003). Thus, depending on contexts and purposes, it has had varying meanings (Francis, 2008; Perkins & Zimmerman, 1995).

First, empowerment was mainly used in an emancipatory sense of giving power to the oppressed (Freire, 1972). However, several scholars started to regard empowerment as a process that cannot simply be given to people (e.g. Karl, 1995; Rappaport, 1987; Zimmerman, 1995). Hence, Adams (1991, p. 208) defined empowerment as "becoming powerful" and explained that it "embodies two dimensions: being given power and taking power".

Furthermore, empowerment was seen as a collaborative process aiming towards greater power, participation and responsible autonomy (e.g. Cummins, 1986). Therefore, empowerment also entails a third dimension: actively taking charge of one's power and resources (Pollari, 2000).

In the 1980s and 1990s, a theory of empowerment was formulated within community psychology (see e.g. Perkins & Zimmerman, 1995; Rappaport, 1987; Zimmerman, 1995, 2000; Zimmerman & Rappaport, 1988). The theory analyses empowerment at individual, organisation and community levels and it includes both processes and outcomes, which may vary depending on the contexts and people involved (Zimmerman, 2000).

At the individual level of analysis, empowerment is referred to as *psychological* empowerment. Psychological empowerment has three components: intrapersonal, interactional and behavioural. The *intrapersonal* component is manifested by perceived control and self-efficacy, but also by competence and motivation (Zimmerman, 1995, 2000). The *behavioural* component entails "efforts to exert control" through active involvement (Zimmerman, 2000, p. 46). The interactional component provides a bridge between intrapersonal and behavioural components and it "suggests that people are aware of behavioural options or choices to act as they believe appropriate to achieve goals they set for themselves" (Zimmerman, 1995, p. 589).

As the theory of empowerment recognises, both empowerment processes and their outcomes vary (Zimmerman, 2000). In some cases, the actions meant to empower people "fail to foster the emancipatory potential that they make possible" (VanderPlaat, 1998, p. 87; see also Toomey, 2011). Moreover, although the goal of empowerment is to foster a group's or individual's agency and opportunities "to make effective choices, that is, to make choices and then to transform those choices into desired actions and outcomes" (Alsop et al., 2005, p. 10), some writers also highlight the right of those being empowered to decide *not* to use their power: "The choice is therefore with the individual, who, given the power, authority, skills and willingness to act, may choose to accept empowerment" (Rodwell, 1996, p. 309).

2.3 Disempowerment

Disempowerment is usually regarded as the opposite of empowerment (e.g. Bolaffi et al., 2003) and thus a term which seems to require no further definition (Kasturirangan, 2008; Toomey, 2011). Yet, like empowerment, disempowerment is used in different contexts with varying meanings. For instance, power and resources are sometimes seen finite: if someone becomes empowered, then someone else becomes disempowered (e.g. Lorion & McMillan, 2008). This notion seems to regard empowerment and disempowerment as the polar ends of allocated power.

However, many everyday definitions, such as dictionary definitions, of disempowerment include aspects of confidence and self-efficacy, which are important constituents of psychological empowerment (Zimmerman, 1995, 2000). Accordingly, even if people have been given power, but they lack self-confidence, they are probably less likely to use their power. Disempowerment is therefore not simply a case of denying someone power and resources.

Thus, in this article, disempowerment refers to students *experiencing* a lack of power and/or resources to make decisions in order to fulfil their potential. In other words, disempowerment refers to the lack of *perceived* control and low self-efficacy (e.g. Zimmerman, 1995, 2000): students may actually have been given power but they either do not realise it or believe in their power and/or themselves. Therefore, they do not, or cannot, take charge of their potential power, which may, in turn, lead to diminished motivation (Harlen, 2012; Weber & Patterson, 2000).

2.4 Empowerment and disempowerment in assessment

Assessment, from the students' point of view, is often a rather disempowering endeavour: as objects of assessment, students do not have much say in the assessment decisions (e.g. Aitken, 2012; Boud, 2007). Yet, decisions made on the basis of these assessments may have far-reaching consequences for students.

In the school context, empirical evidence of students' perceptions of the empowering or disempowering qualities of assessment is rather scarce. However, Aitken (2012) has studied Canadian students' anecdotes on assessment. The students, from primary school to university, mentioned several assessment practices that they found unfair. These included a lack of variety in assessment methodology, too pressurised tests or insufficient test-taking time, secrecy over test content, format or criteria, inadequate feedback and biased grading (Aitken, 2012). A European survey on FL assessment and its focus had rather similar results; in addition, students mentioned irrelevant or too limited a focus as a feature of 'bad' assessment (Erickson & Gustafsson, 2005).

Foreign or second language learning literature has discussed particular assessment approaches that could enhance learners' empowerment. For instance, Little (2005) and Little and Erickson (2015) highlight the possibilities of the *Common European Framework of Reference for Languages* (CEFR) and its European Language Portfolio (ELP) not only in integrating learning, teaching and assessment but in promoting learner *agency* through self-assessment. In addition to the ELP and its electronic version (Cummins & Davesne, 2009), course-based portfolios have been studied as a vehicle for student empowerment in upper secondary EFL studies in Finland (Pollari, 2000). Likewise, shared assessment

has been advocated as a way of empowering student writers in academic English at tertiary level (Pienaar, 2005). In primary school EFL, Bryant and Carless (2010) have investigated whether peer-assessment might empower pupils when preparing for examinations in Hong Kong. There have also been other approaches to foster students' agency and *autonomy* in FL assessment (see e.g. Dam & Legenhausen, 2011; Erickson & Åberg-Bengtsson, 2012) but these studies do not discuss the concept of (dis)empowerment as such.

Most research looking into assessment as a vehicle for empowerment has taken place in higher education and has focused on self- and peer-assessment. These studies have included several disciplinary areas such as health psychology, the humanities and social sciences. Their results have been somewhat mixed. For instance, in a study of 233 university students, Hanrahan and Isaacs (2001) found that university students experienced self- and peer-assessment difficult and even uncomfortable, but at the same time they felt that these methods enhanced their learning and understanding of the assessment and its criteria. Another study, by Patton (2012), explored 36 Australian undergraduates and their perceptions towards peer-assessment. The study found that although students supported peer-assessment for formative assessment purposes, they "were highly critical of it as a summative practice" (Patton, 2012, p. 719).

One of the most comprehensive assessment experiments attempting to empower students was reported by Leach et al. (2000, 2001). In addition to self-assessment, they decided to give adult education students more power over both assessment methods and criteria by offering choice: the students could name their own tasks and criteria to be used in assessment, or take what the teachers suggested. Their results showed that students had differing responses to assessment empowerment: there were students who liked power-sharing, those who disliked it and those who disliked power-sharing first but grew to appreciate it. Accordingly, Leach et al. (2001) conclude that although the results were mainly positive, "learners will vary in their desire and confidence to make judgements about their own work" (p. 298).

This desire and confidence may also vary depending on how advanced and mature students are (Francis, 2008). Thus, in the name of empowerment, the students in the study by Leach et al. (2000, 2001) could also decide to leave the assessment solely to the teachers. Tan (2012), however, disagrees with this choice: in his opinion giving students the right *not* to participate in assessment – self-assessment in his case – is not empowering. Moreover, if optional, it will not foster the learning and self-assessment skills of those who opt out (Tan, 2012).

The Finnish school system has only one high-stakes test, the Matriculation Examination. Otherwise, teachers decide on assessment and its methodology, within the boundaries of the National core curriculum for upper secondary schools. Although the core curriculum does not use the word empowerment as such, some traits of the concept are present. Firstly, assessment must aim at guiding and encouraging learning and it must be diverse. Secondly, the course goals and assessment criteria are to be discussed with students at the beginning of each course. Furthermore, students may be given a say in determining their course grades, but that is left for schools and teachers to decide (for further information, see *National core curriculum for upper secondary schools, 2003)*.

Thus, Finnish students should have at least some power in the assessment process so why do some students still feel disempowered in assessment?

3 The present study

3.1 Aims

This article is part of a larger study the aim of which was to find out how students at one school experienced assessment during their upper secondary EFL studies. For instance, did assessment encourage and guide students' learning, as required by the National core curriculum? Furthermore, were the assessment methods considered versatile, accurate and fair? Did they allow students any power or agency in assessment?

With conflicting findings of power and agency emerging from the data, I began to focus on the students' experiences of empowerment and, particularly, of disempowerment in assessment. Therefore, the research questions of this article are:

1. Do the students who found assessment disempowering differ from other students in any clear respect? If yes, how?
2. What predicts disempowerment in assessment?
3. How are assessment disempowerment and empowerment manifested at an individual level?

3.2 Data collection

To get a comprehensive view on students' experiences of EFL assessment in this upper secondary school in a practical and economical manner, its second- and third-year students were asked to answer a web-based questionnaire anonymously. In addition to background questions, the questionnaire had eight sections with 139 Likert-scale items and 11 open-ended questions. Each section covered one topic area: students' goal orientation; empowerment and agency in assessment processes; the usefulness of different assessment methods; the frequency of different methods; the accuracy and guidance of assessment; students' personal experiences of and views on assessment; the Matriculation Examination; and feedback.

The questionnaire drew theoretical inspiration from extensive literature on assessment, empowerment and FL education. Studies such as the evaluation of pedagogy in Finnish upper secondary education (Välijärvi et al., 2009) and *Towards Future Literacy Pedagogies* (Luukka et al., 2008; Tarnanen & Huhta, 2011) offered invaluable ideas for specific questions. However, with no previous research on most of the topic areas of the questionnaire in this context, the questionnaire was quite exploratory in its nature and had to be specifically designed for this study (Cohen et al., 2013; Creswell, 2014).

Most items on the questionnaire were based on the *National core curriculum for upper secondary schools 2003* and on the current assessment practices both in Finland and at this school. Four research experts on educational assessment and/or FL education as well as three colleagues at school (the upper secondary school head teacher, a student counsellor and another English teacher) commented on the evolving versions of the questionnaire. These experts were consulted to ensure that the content of the questionnaire was valid from practical, legislative and theoretical perspectives. Student voice was also included in the questionnaire as students' ideas and comments on assessment,

gathered during my teaching career of over 20 years, shaped the questionnaire considerably. Furthermore, the open-ended questions were placed at the end of each topic area, after the Likert-scale items, and were designed so that they would enable students to elaborate and express their ideas more freely (see Appendix 6).

The questionnaire was repeatedly tested and commented on by a senior researcher with expertise in both student surveys and in research on upper secondary education. Finally, the internet questionnaire was piloted by four upper secondary students. Each round of testing and comments contributed to further refinements. All these measures were taken to ensure the content validity and reliability of the questionnaire (e.g. Cohen et al., 2013; Messick, 1989).

3.3 Participants

Out of 199 students, 146 answered (response rate 73.4%). The second-year students (79 students, i.e. 54.1% of the respondents) answered the questionnaire during one of their English lessons in March 2014 and the third-year students, already preparing for the Matriculation Examination, in their own time (67 students, 45.9% of the respondents). Eighty-six respondents were female (58.9%), 60 male (41.1%). The average of their previous English grade (self-reported) was 8.58 (range 6–10, with 4 being the lowest and 10 the highest grade in the Finnish system). So far in upper secondary school, they had studied, on average, 6.7 courses (range 4–11) and had had 3.7 different English teachers (range 2–7). The first-year students were excluded from this survey as I wanted students to have had adequate experience of English studies and assessment at upper secondary school. Regarding gender and grades, the respondents are a good representation of the total student population of the school at the time of the study.

3.4 Data analysis

Principally, the data was analysed quantitatively. Originally, in order to reduce the dimensionality of the whole data, a varimax-rotated principal component analysis (e.g. Brown, 2009; Metsämuuronen, 2009) was conducted to summarise the variance of each section of the questionnaire into a few principal components. This analysis revealed a strong (dis)empowering component in assessment. On the basis of the resulting principal components, altogether 28 sum variables were formed[1] (see Appendix 1). The SPSS software was used for the statistical analyses.

Firstly, to address the research questions of this article, students' differing experiences of assessment (dis)empowerment were analysed and grouped with the help of means and standard deviations. Secondly, a stepwise regression analysis (e.g. Jokivuori & Hietala, 2007; Metsämuuronen, 2009) was run to find out which variables might predict disempowerment the strongest.

In order to add depth and to illustrate "what the individual variation means" (Patton, 2002, p. 15), qualitative data and analysis were also used in the third approach, i.e. in the illuminative close-ups of three individual students. Methodologically, these case analyses are based on mixed methods that complement each other: the qualitative data is used to both check the accuracy and validity of the quantitative findings and further explain them, and vice versa, in order to provide as comprehensive analysis as possible (Creswell, 2014). Firstly, the cases had to qualify in their category (disempowered/non-

disempowered/empowered) on the basis of the quantitative analysis of their responses to the Likert-scale items. Secondly, the open-ended answers of each of these qualified students were carefully read, analysed and compared with one another through close reading, which Brummett (2010, p. 25) characterises as follows: "Close reading is a mindful, disciplined reading of an object with a view to deeper understanding of its meanings" (see also Thomas, 2006). Then, the most information-rich cases – "those from which one can learn a great deal about issues of central importance" (Patton, 2002, p. 46) – were purposefully selected.

4 Findings

One section of the questionnaire dealt with students' personal experiences of and views on assessment and its agency and power. The principal component analysis of that section extracted six components with Eigenvalues bigger than 1. The most effective component (17.43% of variance) was transformed into a sum variable which consisted of the four items that had the strongest loadings in this component (see Table 1). Henceforth, the resulting sum variable is called *Disempowerment* as its items cover central features or results of disempowerment: assessment is not seen as a factor facilitating learning, but rather as something that drains the students' power, resources and motivation. In other words, it refers to the lack of perceived control, self-efficacy and motivation, which are the features of the intrapersonal component of psychological empowerment (Zimmerman, 1995, 2000).

Table 1. The items and their loadings in the sum variable of *Disempowerment* (Cronbach's alpha .76).

Item	Loading
Assessment methods give me an opportunity to show how much I know.	-.788
The assessment methods (that are used) discourage me.	.771
Assessment has diminished my willingness to learn.	.749
Assessment just states, it does not guide or help me to learn better.	.615

The *Disempowerment* sum variable was the main starting point for all the following analyses. However, the analyses and findings deal with that sum variable from different angles. Firstly, I will present the 'big picture' of all the data using the *Disempowerment* sum variable as a dividing point which divides students into different groups. Then I will focus on the predictors of disempowerment with the help of a stepwise regression analysis. Finally, I will introduce three individual student cases which rely also on the students' open-ended answers.

4.1 Disempowered and non-disempowered student groups

To see the general trends of the data, the means of each of the 28 sum variables, as well as two individual variables (see Appendix 1), were calculated for the whole respondent group. Then, to see how the students who felt disempowered differed from the whole respondent group of this study, these means were

calculated also for the group that can be considered disempowered. The means are presented as graphs in Figure 1.

The disempowered group was defined on the basis of the sum variable named *Disempowerment* mentioned above. The mean of the whole respondent group for this sum variable was 2.48, with the minimum value of 1 and maximum 4.5 (*SD* .79). The cut-off point for including a student in the disempowered group was one SD above the mean (*M* + 1 *SD*, i.e. 2.48+0.79=3.27). This resulted in a group comprising 21 students (14.4%), most of whom were girls (see Table 2).

Also, I wanted to explore the students who, according to their questionnaire responses, did not appear disempowered at all. Calculating the cut point on the same principle (2.48–0.79=1.69), the resulting group had altogether 18 students (12.3%). However, I could not call these students empowered on the basis of this sum variable since the sum variable did not entail any items concerning power given to students or students actively taking charge of their decision-making power. Hence, they are rather clumsily called *non-disempowered*. The three student groups (i.e. the disempowered, the non-disempowered and the whole respondent group) differed from one another both in their gender ratio as well as in their grades: the disempowered students had the lowest previous grades (*M*=7.86) and the non-disempowered the highest (*M*=8.83) of these three groups. Furthermore, the disempowered students showed the biggest difference between the grade they would have given themselves and the one received (see Table 2).

Table 2. Descriptive statistics of the whole respondent group as well as the disempowered and the non-disempowered student groups.

	All respondents *n*=146	**The disempowered *n*=21**	**The non-disempowered *n*=18**
Number of female and male students / ratio	86 females, 60 males 58.9% / 41.1%	14 females, 7 males 66.7% / 33.3%	8 females, 10 males 44.4% / 55.6%
Second-year/third-year students ratio	79 / 67 54.1% / 45.9%	13 / 8 61.9% / 38.1%	11 / 7 61.1% / 38.9%
Mean of previous English grade	8.58	7.86	8.83
Mean of own estimate/i.e. self-grade	8.64	8.05	8.83
Mean of final English grade in basic education	9.06	8.57	9.11

When comparing the means of the sum variables of the Disempowered and the Non-disempowered with the means of the whole respondent group, a few sum variables or topic areas showed clear differences. For instance, the individual variable *Assessment causes me anxiety and stress* as well as the sum variable of *Stressful and discouraging assessment* divided opinions between these three groups (see Figure 1). Also, students' responses to feedback, its usefulness, importance and role in learning seemed to set these groups apart. The groups seemed rather different in their experienced ability to analyse their strengths and weaknesses. The Disempowered also considered the assessment methodology the least versatile and good, thought that course tests had had too much weight and also regarded assessment as the least accurate or just out of these three

groups. They also wanted to have more influence on the assessment methodology and criteria than the other two groups.

However, when comparing the sum variable concerning *Given empowerment* (e.g. whether the goals and assessment methodology were discussed at the beginning of the course, and whether students were given a chance to influence them), the difference became noticeably smaller. Furthermore, all the student groups seemed rather unanimous in their views on the degree of usefulness of some assessment methods, such as self-assessment or other 'softer', i.e. more formative, and versatile methods. At first glance, it looked as if the disempowered students also felt that they had been given power to participate in the decision-making process, but somehow they had not quite embraced it or it had not resulted in assessment methodologies of their choice.

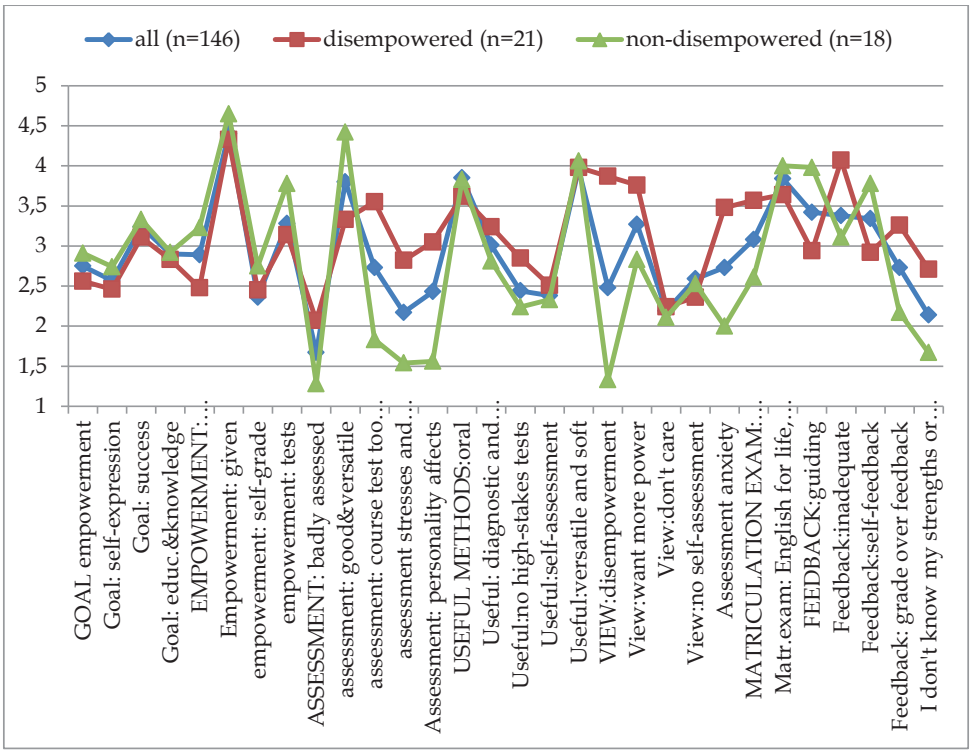

Figure 1. The line chart depicting the sum variable means of all respondents as well as the disempowered and the non-disempowered student groups (see Appendices 1 and 3 for more information on the sum variables).

In summary, several factors seemed to contribute to students feeling disempowered or not in assessment. Yet, the mere means of the sum variables did not adequately explain what might best predict disempowerment.

4.2 Predictors of disempowerment

To find out which sum variables or background factors such as grade, gender or year (as dummy variables) might best predict *Disempowerment*, a stepwise regression analysis was run. The analysis produced a model with eight predictors,

which altogether accounted for 59.3% of the variance. The distribution of the residuals was evaluated following the normality assumption. The normal probability plot of the residuals was approximately linear and the histogram of the residuals was almost normal. Also, the scatterplots of residuals indicated homoscedasticity, confirming the constant variance. Furthermore, as the tolerance (.59–.88) and VIF indexes (1.1–1.7) indicated that multicollinearity was quite low (see also correlation matrix, Appendix 5), this model was accepted.

The most significant predictor of *Disempowerment* was the sum variable of *Stressful and discouraging assessment.* It explained 34.3% of the variance in *Disempowerment* (see Appendix 2; the beta weights and standardised betas in the last model are presented in Appendix 3). Students felt that assessment caused them too much stress and discouraged and demotivated them. When compared with the sum variable of *Disempowerment*, this sum variable had one item (*Assessment has discouraged me or diminished my willingness to study*) which overlapped with some of those of Disempowerment, which may explain its high explanatory power to some extent. However, the two sum variables and their items were by no means identical (see Appendix 4).

In the next step, a sum variable indicating that students did not consider pressurised tests useful for their learning, *No pressurised or high-stakes tests*, was added to the model[2]. Thus, it was the second most significant predictor of *Disempowerment*. In other words, these students regarded tests with aids – e.g. cheat-sheet or open-book tests – as beneficial for learning, whereas more pressurised assessments such as course tests or the Matriculation Examination were not considered good or useful for learning purposes. This sum variable accounted for an additional 7.9% of the variance. Alone, as the only predictor in the linear regression analysis, it would have explained 12.1% of the variance.

The next step added a feedback sum variable, *Grades over feedback,* which accounted for an additional 6.4% of the variance. Alone, it would have accounted for 11.3% of the variance. *Grades over feedback* meant that students were more interested in their grades and scores than in teacher comments or corrections, which they did not necessarily even consult carefully. They may even have rejected feedback.

The sum variable of *Good and versatile assessment* was the fourth most significant predictor of *Disempowerment,* accounting for an additional 3.9% of the variance in this model. As it was negatively related to *Disempowerment,* it means that disempowered students felt that assessment had *not* been good and versatile. Alone, this sum variable would have accounted for 27.1% of the variance of *Disempowerment*, which was caused by their high mutual correlation (r = -.52, p < .01), but its high correlation with *Stressful and discouraging assessment* (r = -.57, p < .01) reduces its additional explanatory power (see Appendix 5).

The following step in the regression model added another feedback sum variable, *Inadequate feedback*. Inadequate feedback refers to students wanting more feedback both from their teachers and peers. *Inadequate feedback* accounted additionally for 2.1% of the variance – as a single predictor, it would have accounted for 12.6% of the variance.

The sixth step added the sum variable of *Success-oriented goals*: students stated a good school-leaving certificate and good grades in the Matriculation Examination as well as a study place in the field of their choice after graduation as the main objectives of their studies in upper secondary school. This sum variable and *Disempowerment* had a negative relationship, i.e. success-oriented goals predicted *Disempowerment* negatively: the higher the success-orientation,

the less disempowered those students felt. It accounted for an additional 1.7% of the total variance (alone: 2.8%).

Slightly contradictorily, the sum variable of *English for life, not for the Exam* also related negatively to *Disempowerment*, and was the penultimate predictor of *Disempowerment* (an additional 1.4% of the variance: alone, 3.1%). In other words, the more the students considered that they were studying English for themselves, not for the Matriculation Examination, the less disempowered they felt.

Finally, one more sum variable improved the explanatory power of this model, namely the sum variable of *Personality affects assessment*: students felt that assessment favours some student and personality types. It accounted for an additional 1.5% of the variance. However, alone it would have predicted as much as 20.0% of the variance.

All in all, according to this stepwise regression analysis, the five most significant predictors of disempowerment, accounting together for over 50% of the variance, were *Stressful and discouraging assessment, No pressurised or high-stakes tests, Grades over feedback, Good and versatile assessment*, which related negatively with disempowerment, and *Inadequate feedback*. In other words, disempowered students felt both stressed and demotivated by assessment. Test anxiety was a clear predictor: no high-stakes tests but 'softer', i.e. more formative and less pressurised assessment was called for. The current assessment methods were not considered good and versatile enough, and they did not give students a fair chance to show all their skills or knowledge. Furthermore, feedback had failed to serve its purpose of facilitating learning. Feedback was either overshadowed by grades, and therefore insufficient attention was paid to feedback and it was considered less important than grades or scores, or students had not received enough feedback to guide and enhance their learning. In addition, students' ownership of their English studies as well as their goal-orientation played a role in assessment (dis)empowerment. Students' personality was also seen as a factor that influences assessment.

4.3 Focus on individuals: three student cases

To illustrate how students as individuals behind these means and quantitative analyses *experienced* assessment, I will present three student cases. The cases, a disempowered, a non-disempowered and an empowered student, were selected on the basis of two main criteria: they represent their category in a clear and illuminative manner, and they had answered a sufficient number of the open-ended questions so that there was enough data in their own voices to "provide depth, detail, and individual meaning" (Patton, 2002, p. 16). Accordingly, the following account primarily relies on the students' open-ended answers. The answers were originally written in Finnish but I have attempted to maintain both their meanings and style as faithfully as possible. The students' quantitative answers are presented in Figure 2 below.

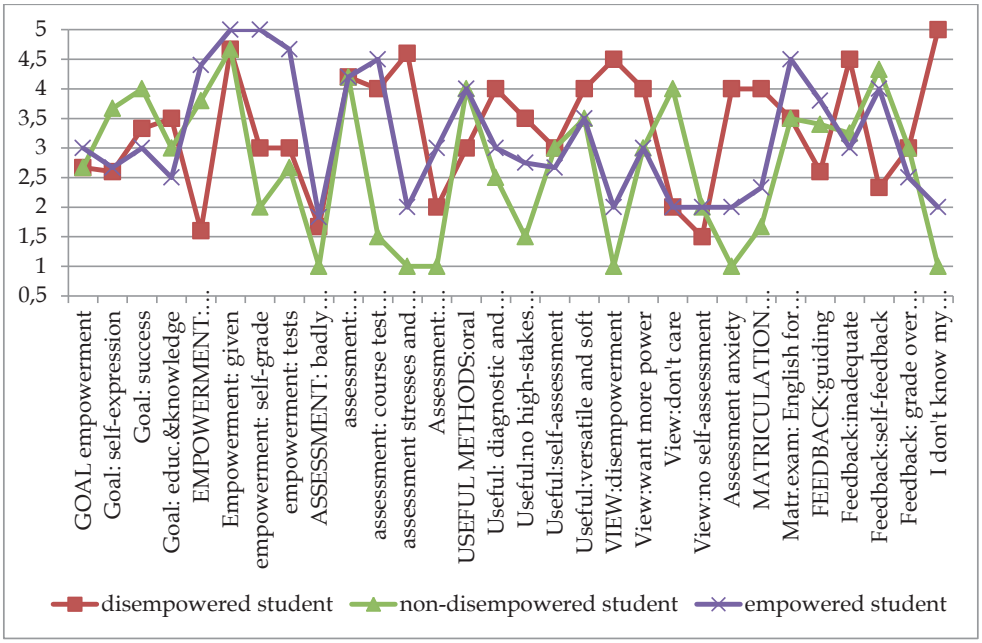

Figure 2. The sum variable means of the three individual students.

4.3.1 "I'm beside myself with fear"

The highest value for the sum variable of *Disempowerment*, 4.5, was by a second-year female student who had studied five English courses with four different teachers in upper secondary school. Her final English grade in basic education two years earlier had been 9, but now her English grade was 7. She seemed to consider the grade quite fair since she would have given herself the same grade. She also regarded the assessment methodology as quite versatile and fair (see Figure 2). Yet, assessment caused her stress and anxiety to such an extent that she seemed to have lost trust in her ability to learn English as well as her willingness to study it: "I am crap at English" she wrote twice in her answers, and "I hate English" were her final words in the questionnaire. She had answered all the questions in a detailed and thorough manner, so I do not think the comments above were mere bursts of teenage rant but sincere comments.

Why did she consider herself so poor at English? Why had she lost her self-efficacy as a learner of English? One explanation might lie in pressurised test situations and high-stakes tests. Although not considering herself unfairly or badly assessed, she felt that the course test influenced the final course grade too heavily and thus caused too much stress. She would have preferred less stressful assessment methods. She also hoped for more formative assessment:

> There could be grammar tests that don't affect the grade. They would be excellent groundwork/practice for the course test. Assessment methods in English have to be versatile so that vocabulary, grammar, listening comprehension and pronunciation are all assessed. I'd like to have two grammar tests in each course. This way, things would still be fresh in your mind and you wouldn't face a horrible excess of grammar that is hard to digest and learn in the test week. Cheat-sheet tests are also good and could be used more because you learn well when you write down notes. In my opinion, in assessment, more

> attention should be paid to whether you have taken part actively in group or pair discussions because they assess how actively and confidently you speak English and what your attitude is to it in other respects as well. (Q1)

She also regarded chances to compensate for some weaknesses with extra effort as useful for learning:

> Some assessment methods motivate you more to work harder. Motivating 'tips' like vocab tests that improve your grade are good. Perhaps there could be some extra tasks etc. you could do to improve your grade as well in the course? (Q6)

However, these compensatory methods, or any assessment methods, should not significantly increase the student's workload at home, and therefore, she did not consider home assignments useful for learning. She also wanted to have more power to influence assessment so that she could organise her use of time more rationally and efficiently:

> I want to influence how many vocabulary tests we have and which ones of them affect the grade. This way, I can plan my own timetables with regards to my studies at least a little and also concentrate on other languages I study. Self-assessment method is good, it may help open the teacher's eyes, too. (Q8)

She did not appear very self-regulated on the basis of her answers in goal-orientation sum variables nor in the sum variable of *Self-feedback*, which refers to students seeking feedback themselves from various teaching and learning situations (e.g. checking homework) without being given feedback explicitly by the teacher or peers. Consequently, she also considered feedback inadequate and would have liked to have feedback "Orally and in writing as often as possible" (Q10).

All these answers give a picture of a student for whom languages were not her forte and who needed to work hard at them. She was probably busy outside school, and thus did not like to work at home very much. On the basis of her answers, her ambition to do quite well at school as well as her lack of self-efficacy as a learner of English had probably started prior to her upper secondary school studies. She had had a very good English grade (9) in her final report of basic education, but yet the idea of her poor English had affected her study choices for upper secondary school at that time. She explained her choices when asked about her thoughts about the Matriculation Examination:

> In upper secondary school, the thing I am most afraid of is that Matriculation exam. I chose Advanced Maths so that I won't have to take the Advanced English exam. I'm beside myself with fear because I don't believe I'll pass it with dignity. I think my English Matriculation exam grade will be the tarnish of my diploma. But what can you do if you are crap at something. (Q9)

Nonetheless, despite her negative and anxious comments, she considered assessment needed: "It tells the student about the level of their skills and knowledge. So, yes, it is needed." (Q11)

Yet, she saw assessment and its function in a rather static and summative way: its purpose is to tell the students the level of their skills.

4.3.2 "I'll manage, no matter what method"

Next, the opposite of the disempowered student is portrayed by a totally non-disempowered student. He was one of the three male students whose value in the sum variable of *Disempowerment* was the lowest possible (1). He was a second-year student, with 9 as his previous grade. Although he would have given himself a 10, he did not feel that assessment had been unfair. In his opinion, the assessment methodology had been versatile and good: the course test did not carry too much weight and personality did not affect assessment. Furthermore, assessment caused him no stress, anxiety or disempowerment at all, not even the forthcoming Matriculation Examination he was planning to sit the following autumn: "I'll pass it even if I have my eyes shut and hands tied behind my back." (Q9)

There appeared to be a clear reason for his extreme non-disempowerment. He trusted his English skills so much that he felt convinced he would manage well no matter what methods were used in assessment. Therefore, he did not want more power to influence assessment methodology:

> No, I personally just don't care how a course is assessed. It makes no difference what methods are used, my English is so good that I'll manage with them all. Often even without studying/reading. And yes, I am a little arrogant. (Q8)

As assessment methodology did not matter to him, he did not offer his opinions on what methods should be used more, or what would be useful for learning. However, he had an opinion on what *not* to use:

> Cheat-sheet tests and tests with your book and/or notes. They don't assess any other skills than perhaps how to find information and if you can bring your notes, then also how well you can write notes. The main thing is to test your ENGLISH SKILLS, right? I just can't see how they could be useful for anything or anyone. (Q7)

As could be seen, this student did not appear disempowered by assessment in the slightest. Assessment did not seem to matter to him, and, accordingly, he did not want to have or use any power to influence the assessment, either. Although not answering the question on the need and function of assessment at school, the student appeared to perceive the purpose of assessment at school as assessment *of* learning rather than assessment *for* learning.

4.3.3 "It's good to listen to us, too."

The final case depicts an empowered student. He was a third-year student and he had very high means in all sum variables dealing with empowerment (see Figure 2). His English grades, both the final grade of basic education and the previous grade as well as his own suggestion, were all 9. He had already taken the English Matriculation Examination the previous autumn and was relatively satisfied with its result – "Yes, totally fair considering how much I studied for it" (Q5) – but not quite happy with the examination itself: "There's no oral part. Yet it's one of the basic elements of language skills. Anyways, the exam has become "too" difficult over the years, doesn't require real English skills anymore." (Q9)

Oral skills seemed very important for him, and he emphasised the importance of assessing them in general as well: "Discussion, or talking in front of the class

to be precise! Pronunciation and speaking need to be focused on more as they are extremely important things." (Q1)

Furthermore, he criticised the course test as a testing method, basically because of its reliance on memory-retention and recall:

> The course test begins to be a pretty old format. Memorising things by heart is altogether a bit outdated (you can find everything real quick on the net). I'm not saying that remembering everything by heart is a bad thing, on the contrary it is good to remember! but as I said, a bad format. (Q2)

In his opinion, another useless assessment method would be "a course grade based on self-assessment" (Q7).

He scored 2 in the Disempowerment sum variable, so although his score was lower than the average (2.48), it was not low enough to include him in the group of non-disempowered students. What made him different from the non-disempowered student above was his attitude towards power and agency in his English studies. He had clearly taken charge of his chances to influence assessment procedures as well as the knowledge of assessment goals, criteria and methodology. He also felt empowered by this, as can be seen in the sum variables dealing with agency and empowerment (see Figure 2). Hence, he had opinions on assessment methodology and their usefulness, and he welcomed the chance to have a say on assessment: "At the end of the day, it's the teacher who decides. However, everybody's a different learner so it's good to listen to our opinions on assessment." (Q8)

Moreover, he considered assessment useful and it had a clear purpose for him: "To tell us what should be improved, for instance things that I haven't paid any attention to myself. It is really needed!"

Thus, the empowered student considered assessment necessary and he saw the role of assessment as improving and guiding learning, in other words as assessment *for* learning, and not only as stating the level of skills (i.e. assessment *of* learning).

5 Discussion and conclusions

The first research question of this article was to find out, in rather general terms, if the students who found assessment disempowering differed from the whole group in any clear respect. Next, this article aimed to focus on factors that could best predict disempowerment. Finally, the aim was to explore how assessment empowerment and disempowerment manifested themselves on an individual level.

According to the descriptive statistics in the first round of analysis, most students in the disempowered student group were female. Compared to the non-disempowered student group as well as to all respondents, the average of their English grades was also slightly lower. Also, the means of the sum variables indicated several other factors where these student groups differed from one another. Yet, the different means of the sum variables did not adequately explain what might best predict disempowerment. Therefore, a stepwise regression analysis was run and it produced a model with eight predictors. The five most significant predictors of disempowerment, accounting together for over 50% of the variance, were *Stressful and discouraging assessment*, *No pressurised or high-stakes tests, Grades over feedback, Good and versatile*

assessment, which related negatively with disempowerment, and *Inadequate feedback*. However, even though the descriptive statistics showed differences between the previous grades, gender and year of the disempowered and non-disempowered student groups, none of these background variables predicted disempowerment in the stepwise regression analysis. Finally, three student cases were presented to illuminate how individual students experienced assessment disempowerment, non-disempowerment and empowerment.

All the analyses of this study resulted in the same conclusions on disempowerment. First of all, assessment seemed to cause the disempowered students a great deal of anxiety and stress. The disempowered students feared high-stakes testing, such as the Matriculation Examination, but even course exams had too much weight or pressure for their comfort. Thus, test anxiety (see e.g. Cassady, 2010; Hembree, 1988; Knekta, 2017) had a clear connection with assessment disempowerment. In line with earlier studies (e.g. Hembree, 1988; Knekta, 2017), test anxiety and stress was more prominent with female students. Students also felt that their personalities could play too strong a role in the grading process. All in all, the current assessment methodology was not considered either good or diverse enough, and the students felt that they were not given a fair chance to show all their English skills or knowledge. That could, in turn, contribute to the loss of self-efficacy and motivation in their English studies (e.g. Harlen, 2012). Therefore, the disempowered students would have liked more power to influence the assessment methodology as they hoped for more formative and less pressurised assessment methodology.

Secondly, feedback and how it was experienced played a significant role. Feedback had not met students' expectations and needs: either they had not had enough feedback, or it had not been helpful. In some cases, the dissatisfaction had resulted in students ignoring teacher comments and concentrating on grades only. Focusing on grades which had not always met their expectations may have, in turn, decreased students' intrinsic motivation as well as self-efficacy and self-confidence (Butler, 1988; Kohn, 2011; Pulfrey et al., 2013).

Thirdly, the disempowered students did not seem to feel *ownership* of their English studies: they seemed to study English more for the sake of the grades, or the Matriculation Examination, rather than for their own goals. Yet, they did not seem to have a strong success-orientation, either. In general, they exhibited lower scores in all goal-orientation sum variables on average than other students.

However, the disempowered students also acknowledged the given empowerment. They had been informed of the goals as well as the assessment processes and criteria at the beginning of the courses and they had had a fair chance to discuss and to influence them if willing to do so. Yet, even though they wanted to have more power to influence assessment, they had probably not experienced or assumed that power even when possible. One possible reason for this might be that, in their own opinion, their self-assessment skills were lacking as they did not know their strengths and weaknesses in English. Thus, they did not engage in self-feedback as much as some other students. Some seemed to have very low self-confidence as learners of English. However, although the disempowered student group had the lowest previous grade in comparison with all respondents or the non-disempowered student group in the descriptive statistics, the grade as a background variable did not predict disempowerment in the stepwise regression analysis.

Compared to the disempowered students, the non-disempowered students scored slightly higher in the goal-orientation and the empowerment sum

variables. Nonetheless, the clearest differences between the non-disempowered and the disempowered students were in the personal experiences of assessment anxiety and stress as well as feedback. In other words, the non-disempowered students seemed happier with assessment and they got more benefit from assessment and feedback. Their self-assessment skills seemed better and they knew their strengths and weaknesses in English.

The non-disempowered students were conceptually an interesting group. In terms of the theory of empowerment (Zimmerman, 1995, 2000), they manifested a clear intrapersonal component of psychological empowerment as they trusted their skills and themselves. However, some of the non-disempowered students did not exhibit the behavioural component of active involvement. They were happy to be passive objects of assessment and did not wish to have any active agency in assessment. Yet, their self-efficacy seemed strong. They also manifested an interactional component of psychological empowerment as they appeared to "act as they believe appropriate to achieve goals they set for themselves" (Zimmerman, 1995, p. 589). Hence, if empowerment is considered to entail the right to choose whether to use their power or not (Leach et al., 2001; Rodwell, 1996), then they, too, were empowered.

As often maintained in empowerment literature, empowerment is not the same for everyone (e.g. Leach et al., 2000, 2001; Zimmerman, 1995, 2000), nor is everyone equally willing or ready to assume the given power and resources. On the basis of this study, I cannot but agree with Leach et al. (2001, p. 298): "Similarly in assessment, learners will vary in their desire and confidence to make judgements about their own work." In the case of the non-disempowered students, they did not all necessarily have a *desire* to take charge of their power, while in the case of the disempowered students, they probably did not have the tools and, moreover, *confidence* to take charge of their given power. The empowered students, however, had desire, tools and confidence to participate actively in the assessment process.

Practically speaking, if the objective of education is to educate learners who will all have high levels of self-regulation and autonomy, then perhaps all the non-disempowered students should somehow be motivated to assume a more active decision-making role. However, in my opinion, the truly disempowered students need attention first. Decreasing their anxiety and enhancing their confidence, ownership and feelings of self-efficacy in learning and studying would be vital. Being such a prominent phenomenon at school, assessment inevitably plays a crucial role in the empowerment process. For example, introducing less pressurised testing situations such as cheat-sheet tests or home exams occasionally, or as an alternative method, might ease some of their anxiety. Smaller tests as well as more formatively-oriented assessment might also help to decrease their stress. Formative assessment could gradually build their confidence and self-efficacy as they could see that they do learn all the time. It would also give them a chance to 'fill the gap' between the desired outcome and their performance during the learning process (Sadler, 1989), instead of just stating the shortcomings afterwards (Black & Wiliam, 1998). It would be important to foster their ownership of their English skills, to make them see that even if they do not get full marks in tests, their English skills are useful and worthwhile.

The disempowered students would most likely benefit from more personalised feedback. Furthermore, feedback should feed forward, help them to improve their future performance instead of just scrutinising their present or past performance (Hattie, 2009). Giving feedback without grades might help

them to focus on their skills and not only on (possibly disappointing) grades (Butler, 1988; Kohn, 2011).

In addition, students should be both invited and trained to engage in self-assessment. Small, clearly defined self-assessment tasks, against clear, tangible goals and criteria might foster their trust both in their self-assessment skills and in their English skills. Making concrete choices, such as choosing how many vocabulary tests to take, might safely train them in using their decision-making power but also make them aware that they do have some power.

This study was limited to one school only, and thus the findings cannot be generalised as such. Furthermore, since the academic achievement of the student population in this school is above the national average, this study did not have many respondents who struggled with their upper secondary studies. With larger and more varied student groups, students' (dis)empowerment experiences might look different, as they might also in other contexts and cultures. Moreover, other data collection instruments, for instance a different questionnaire, might have altered the findings. Although the students seemed to have answered the questionnaire quite attentively, it was extensive and would have benefitted from further pruning. As all the data was collected simultaneously, this study cannot exhibit potential changes in empowerment over time or in different situations, either. Hence, alternative methods, such as student interviews or narratives, might have yielded additional information.

There is plenty of room for further research regarding students' views on assessment both in foreign language education and in education in general. Also, students' experiences of what disempowers or empowers them in assessment should be examined further, and with more varied student samples and methods. A longitudinal study could indicate how and whether students' assessment experiences change over time. The questionnaire of the present study could also be retested and refined further. Nevertheless, this study allows some insight into students' own experiences of assessment and the factors that may empower or disempower them in assessment. Moreover, it shows tangibly that behind all the means and averages, individual students react to assessment in highly individual ways. It is thus a new opening in important but under-researched areas of both FL and upper secondary school assessment. I hope this study shows that assessment should be versatile and it should take students' perceptions and ideas into consideration during the whole assessment process in order to also cater for those students who currently may feel disempowered by assessment. After all, "assessment of any kind should ultimately improve learning" – of all students (Gardner, 2012, p. 106).

Endnotes

[1] The principal component analysis does not explicitly assume normal distribution (Chatfield & Collins, 1980, p. 58). However, as the components were used in a further statistical analysis, it is worth mentioning that most variables used in the PCA were slightly skewed to the right.

[2] Although this sum variable had a rather low internal consistency (Cronbach's alpha .50), it was kept in the analysis because its content was considered relevant for the analysis. This was the case with the sum variables of *Grade over feedback* and *English for life*. With those two, the reason for a rather low internal consistency was a small number of items in the sum variable; with *No pressurised or high-stakes* tests it was the low inter-item correlation. Nonetheless, with no explicitly determined cut-off value for Cronbach's alpha, some researchers have suggested values of .70, .60 or even .50 (see Jokivuori & Hietala, 2007, p. 104). In this study, I have chosen the value of .60. However, the most crucial reason for including or excluding some sum variable has been the relevance of its content.

References

Adams, R. (1991). *Protests by pupils: Empowerment, schooling and the state*. London: Falmer.

Aitken, E. N. (2012). Student voice in fair assessment practice. In C. F. Webber & J. L. Lupart (Eds.), *Leading student assessment. Studies in educational leadership* (pp. 175–200). Dordrecht: Springer Netherlands.

Alsop, R., Bertelsen, M. F., & Holland, J. (2005). *Empowerment in practice: From analysis to implementation*. Washington, DC: World Bank.

Atjonen, P. (2007). *Hyvä, paha arviointi* [Good, bad assessment]. Helsinki: Tammi.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Black, P., & Wiliam, D. (2012). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 11–32). London: Sage.

Bolaffi, G., Bracalenti, R., Braham, P., & Gindro, S. (2003). *Dictionary of race, ethnicity and culture*. London: Sage.

Boud, D. (2007). Reframing assessment as if learning were important. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 14–25). London: Routledge.

Brown, J. D. (2009). Principal components analysis and exploratory factor analysis–Definitions, differences, and choices. *JALT Testing & Evaluation SIG Newsletter, 13*(1), 26–30.

Brummett, B. (2010). *Techniques of close reading*. Thousand Oaks, CA: Sage.

Bryant, D. A., & Carless, D. R. (2010). Peer assessment in a test-dominated setting: Empowering, boring or facilitating examination preparation? *Educational Research for Policy and Practice, 9*(1), 3–15.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*(1), 1–14.

Cassady, J. C. (2010). Test anxiety: Contemporary theories and implications for learning. In J. C. Cassady (Ed.), *Anxiety in schools: The causes, consequences, and solutions for academic anxieties* (pp. 5–26). New York: Peter Lang.

Chatfield, C., & Collins, A. (1980). *Introduction to multivariate analysis*. London: Chapman and Hall.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education* (7th ed.). Abingdon: Routledge.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Los Angeles: Sage

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481.

Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review, 56*(1), 18–37.

Cummins, P. W., & Davesne, C. (2009). Using electronic portfolios for second language assessment. *Modern Language Journal, 93*(1), 848–867.

Dam, L., & Legenhausen, L. (2011). Explicit reflection, evaluation, and assessment in the autonomy classroom. *Innovation in Language Learning and Teaching, 5*(2), 177–189.

Erickson, G., & Gustafsson, J.-E. (2005). *Some European students' and teachers' views on language testing and assessment. A report on a questionnaire survey*. European Association for Language Testing and Assessment. Retrieved from http://www.ealta.eu.org/resources.htm

Erickson, G., & Åberg-Bengtsson, L. (2012). A collaborative approach to national test development. In D. Tsagari & I. Csépes (Eds.), *Collaboration in language testing and assessment* (pp. 93–108). Frankfurt am Main: Peter Lang.

Francis, R. A. (2008). An investigation into the receptivity of undergraduate students to assessment empowerment. *Assessment & Evaluation in Higher Education, 33*(5), 547–557.

Freire, P. (1972). *Pedagogy of the oppressed*. London: Penguin Books.

Gardner, J. (2012). Quality assessment practice. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 103–121). London: Sage.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research & Development, 20*(1), 53–70.

Harlen, W. (2012). The role of assessment in developing motivation for learning. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 171–183). London: Sage.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77.

Herman, J., & Linn, R. (2014). New assessments, new rigor. *Educational Leadership, 71*(6), 34–37.

Jokivuori, P., & Hietala, R. (2007). *Määrällisiä tarinoita: Monimuuttujamenetelmien käyttö ja tulkinta* [Quantitative stories: The use and interpretation of multivariate methods]. Porvoo: WSOY.

Karl, M. (1995). *Women and empowerment: Participation and decision making*. London: Zed Books.

Kasturirangan, A. (2008*). The balance of psychological empowerment and disempowerment for survivors of domestic violence.* Unpublished doctoral dissertation. University of Illinois, Chicago, IL.

Knekta, E. (2017). Are all pupils equally motivated to do their best on all tests? Differences in reported test-taking motivation within and between tests with different stakes. *Scandinavian Journal of Educational Research*, *61*(1), 95–111.

Kohn, A. (2011). The case against grades. *Educational Leadership, 69*(3), 28-33.

Leach, L., Neutze, G., & Zepke, N. (2000). Learners' perceptions of assessment: Tensions between philosophy and practice. *Studies in the Education of Adults, 32*(1), 107–119.

Leach, L., Neutze, G., & Zepke, N. (2001). Assessment and empowerment: Some critical questions. *Assessment & Evaluation in Higher Education, 26*(4), 293–305.

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]. Jyväskylä: PS-Kustannus.

Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgements in the assessment process. *Language Testing, 22*(3), 321–336.

Little, D., & Erickson, G. (2015). Learner identity, learner agency, and the assessment of language proficiency: Some reflections prompted by the Common European Framework of Reference for Languages. *Annual Review of Applied Linguistics, 35*, 120–139.

Lorion, R. P., & McMillan, D. W. (2008). Does empowerment require disempowerment? Reflections on psychopolitical validity. *Journal of Community Psychology, 36*(2), 254–260.

Luukka, M., Pöyhönen, S., Huhta, A., Taalas, P., Tarnanen, M., & Keränen, A. (2008). *Maailma muuttuu – mitä tekee koulu? Äidinkielen ja vieraiden kielten tekstikäytänteet koulussa ja vapaa-ajalla* [The world changes – how does the school respond? Mother tongue and foreign language literacy practices at school and in free-time]. Jyväskylä: University of Jyväskylä, Centre for Applied Language Studies.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Metsämuuronen, J. (2009). *Tutkimuksen tekemisen perusteet ihmistieteissä: Tutkijalaitos* [The essentials of research methods in human sciences: The researcher edition] (4th ed.). Helsinki: International Methelp.

National core curriculum for upper secondary schools 2003 (English translation printed in 2004). Helsinki: Finnish National Board of Education.

Patton, C. (2012). "Some kind of weird, evil experiment": Student perceptions of peer assessment. *Assessment & Evaluation in Higher Education, 37*(6), 719–731.

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Perkins, D. D., & Zimmerman, M. A. (1995). Empowerment theory, research, and application. *American Journal of Community Psychology, 23*(5), 569–579.

Pienaar, C. (2005). Shared assessment: Empowering student writers. *Language Matters, 36*(2), 193–204.

Pollari, P. (2000). *"This is my portfolio": Portfolios in upper secondary school English studies*. Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Pulfrey, C., Darnon, C., & Butera, F. (2013). Autonomy and task performance: Explaining the impact of grades on intrinsic motivation. *Journal of Educational Psychology, 105*(1), 39–57.

Rappaport, J. (1987). Terms of empowerment/exemplars of prevention: Toward a theory for community psychology. *American Journal of Community Psychology*, *15*(2), 121–148.

Reay, D., & Wiliam, D. (1999). 'I'll be a nothing': Structure, agency and the construction of identity through assessment. *British Educational Research Journal, 25*(3), 343–354.

Rodwell, C. M. (1996). An analysis of the concept of empowerment. *Journal of Advanced Nursing, 23*(2), 305–313.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–391.

Simon, B. L. (1994). *The empowerment tradition in American social work: A history*. New York: Columbia University Press.

Takala, S. (1994). Arviointi – ongelma ja mahdollisuus [Assessment – a problem and a possibility]. In P. Linnakylä, P. Pollari & S. Takala (Eds.), *Portfolio arvioinnin ja oppimisen tukena* [Portfolio supporting assessment and learning] (pp. 1–8). Jyväskylä: Institute for Educational Research.

Tan, K. H. K. (2012). *Student self-assessment: Assessment, learning and empowerment.* Singapore: Research Publishing.

Tarnanen, M., & Huhta, A. (2011). Foreign language assessment and feedback practices in Finland. In D. Tsagari & I. Csépes (Eds.), *Classroom-based language assessment. Language testing and evaluation, Vol. 25* (pp. 129–146). Frankfurt am Main: Peter Lang.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation, 27*(2), 237–246.

Toomey, A. (2011). Empowerment and disempowerment in community development practice: Eight roles practitioners play. *Community Development Journal, 46*(2), 181–195.

Traynor, M. (2003). A brief history of empowerment: Response to discussion with Julianne Cheek. *Primary Health Care Research & Development, 4*(2), 129–136.

Välijärvi, J. (1996). Oppilasarviointi opiskelun uudistumisen tukena ja tukahduttajana lukiossa [Student assessment supporting or suppressing the renewal of studying in the upper secondary school]. In A. Räisänen & T. Frisk (Eds.), *Silta uuteen opiskelija-arviointiin* [Bridge to new student assessment] (pp. 123–142). Helsinki: National Board of Education.

Välijärvi, J., Huotari, N., Iivonen, P., Kulp, M., Lehtonen, T., Rönnholm, H., Knubb-Manninen, G., Mehtäläinen, J., & Ohranen, S. (2009). *Lukiopedagogiikka* [Evaluation of pedagogy in Finnish upper secondary education]. Jyväskylä: Finnish Education Evaluation Council.

VanderPlaat, M. (1998). Empowerment, emancipation and health promotion policy. *Canadian Journal of Sociology/Cahiers canadiens de sociologie, 23*(1), 71–90.

Weber, K., & Patterson, B. R. (2000). Student interest, empowerment and motivation. *Communication Research Reports, 17*(1), 22–29.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3–14.

Zimmerman, M. A. (1995). Psychological empowerment: Issues and illustrations. *American Journal of Community Psychology, 23*(5), 581–599.

Zimmerman, M. A. (2000). Empowerment theory: Psychological, organizational and community levels of analysis. In J. Rappaport & E. Seidman (Eds.), *Handbook of community psychology* (pp. 43–63). New York: Kluwer Academic/Plenum.

Zimmerman, M. A., & Rappaport, J. (1988). Citizen participation, perceived control, and psychological empowerment. *American Journal of Community Psychology, 16*(5), 725–750.

Appendices

Appendix 1

The 28 sum variables based on a varimax-rotated principal component analysis of each topic area of the questionnaire (each topic area is mentioned at the beginning of the name of the sum variable) as well as two additional variables (in italics)

GOAL: empowerment as goal
Goal: self-expression as goal
Goal: success-oriented goals
Goal: education and knowledge as goal

EMPOWERMENT: experienced empowerment
Empowerment: given empowerment
Empowerment: self-grade empowerment
Empowerment: test empowerment

ASSESSMENT: badly assessed
Assessment: good and versatile assessment
Assessment: course test too weighted
Assessment: stressful and discouraging assessment
Assessment: personality affects assessment

USEFUL METHODS: oral
Useful: diagnostic and formative
Useful: no high-stakes tests at all
Useful: self-assessment
Useful: versatile and soft

VIEW: disempowerment
View: want more power
View: don't care
View: no to self-assessment
View: Assessment anxiety: "Assessment causes me anxiety and stress"

MATRICULATION EXAM: fear
Matriculation exam: English for life, not for exam

FEEDBACK: guiding feedback
Feedback: inadequate feedback
Feedback: self-feedback
Feedback: grade over feedback
Feedback: "I don't know my strengths or weaknesses in English"

Appendix 2: Model Summary

Model	R	R^2	Adj R^2	R^2 Change	F Change	df 1	df 2	Sig. F Change
1	.586	.343	.339	.343	72.2	1	138	<.001
2	.650	.422	.414	.079	18.7	1	137	<.001
3	.697	.486	.475	.064	17.0	1	136	<.001
4	.725	.526	.512	.039	11.2	1	135	.001
5	.739	.546	.530	.021	6.1	1	134	.014
6	.751	.564	.544	.017	5.3	1	133	.022
7	.760	.578	.555	.014	4.3	1	132	.040
8	.770	.593	.568	.015	4.9	1	131	.028

Predictors:

M1: Stressful and discouraging assessment
M2: Stressful and discouraging assessment, No pressurised or high-stakes tests
M3: Stressful and discouraging assessment, No pressurised or high-stakes tests, Grade over feedback
M4: Stressful and discouraging assessment, No pressurised or high-stakes tests, Grade over feedback, Good and versatile assessment
M5: Stressful and discouraging assessment, No pressurised or high-stakes tests, Grade over feedback, Good and versatile assessment, Inadequate feedback
M6: Stressful and discouraging assessment, No pressurised or high-stakes tests, Grade over feedback, Good and versatile assessment, Inadequate feedback, Success-oriented goals
M7: Stressful and discouraging assessment, No pressurised or high-stakes tests, Grade over feedback, Good and versatile assessment, Inadequate feedback, Success-oriented goals, English for life, not for the Matriculation exam
M8: Stressful and discouraging assessment, No pressurised or high-stakes tests, Grade over feedback, Good and versatile assessment, Inadequate feedback, Success-oriented goals, English for life, not for the Matriculation exam, Personality affects assessment

Appendix 3: The beta weights and standardised betas in the last model

	Unstandardised Coefficients		**Standardised Coefficients**		
	B	Std. Error	Beta	t	Sig.
(Constant)	1.587	.684		2.320	.022
Stressful and discouraging assessment	.291	.080	.266	3.650	.000
No pressurised or high-stakes tests	.355	.074	.289	4.826	.000
Grade over feedback	.209	.052	.247	3.992	.000
Good and versatile assessment	-.197	.097	-.147	-2.038	.044
Inadequate feedback	.167	.066	.157	2.506	.013
Success-oriented goals	-.244	.079	-.184	-3.104	.002
English for life, not for the Matriculation exam	-.130	.058	-.135	-2.241	.027
Personality affects assessment	.126	.056	.150	2.223	.028

Appendix 4: The eight predictors of disempowerment

The sum variables, their items and their loadings

Stressful and discouraging assessment (Cronbach's alpha .68):

- Assessment (tests, essays, etc.) has caused me too much stress. .654
- Participation in class has affected the grade too much. .566
- Assessment has discouraged or diminished my willingness to study. .561

No pressurised or high-stakes tests (Cronbach's alpha .50):

- Matriculation Exam -.666
- The grade is mainly based on the course exam/test -.646
- No course test at all .518
- Book/notes allowed in the test .486

Grade over feedback (Cronbach's alpha .55):

- The test mark or score interests me more than the teacher's comments or corrections. .770
- I always check my mistakes and corrections carefully when I get my tests or essays back. -.600

Good and versatile assessment (Cronbach's alpha .75):

- There have been assessments steadily and evenly throughout the course. .673
- Assessment methods have been versatile .653
- All parts of language proficiency have been taken into account in assessment. .588
- I know why I have received the grade I have received. .550
- Assessment has given me a good overall picture of my skills. .534

Inadequate feedback (Cronbach's alpha .72):

- I would like to have more teacher feedback on my skills. .894
- I would like to have more teacher feedback on how to develop my studying. .825
- I get enough feedback from other students. -.590
- My teacher writes enough feedback at the end of the essay, for instance. -.539

Success-oriented goals (Cronbach's alpha .66):

- Good results in the Matriculation Exam. .856
- To gain access to study for the career I want after upper secondary school. .701
- Good final upper secondary school diploma. .695

English for life, not for the Matriculation Exam (Cronbach's alpha .55):

- I study English for life and for my future, not for the Matriculation Exam -.857
- For me, the most important goal of my English studies is a good grade in the Matriculation Exam. .760

Personality affects assessment (Cronbach's alpha .61):

- The student's personality has affected the grade. .745
- Assessment has favoured some students or student types. .708

Appendix 5

The correlation matrix of the eight predictor sum variables and disempowerment

	Disempowerment	**stress**	**No pressurised test**	**Grade over feedback**	**Good & versatile assessment**	**Inadeq. feedback**	**success**	**Engl.for life**	**Personal. affects assessment**
Disempowerment									
Stress	.586**								
No pressurised test	.349**	.109							
Grade over feedback	.337**	.248**	-.149						
Good & versatile assessment	-.521**	-.574**	-.123	-.211*					
Inadeq. feedback	.355**	.273**	.252**	-.113*	-.323**				
Success	-.166*	-.055	-.075	.069	.166*	.040			
Engl for life	-.175*	-.174*	.186*	-.243**	.026	-.038	-.180*		
Personality affects assessment	.447**	.467**	.131	.225**	-.394**	.236**	.121	-.075	

Appendix 6

Open-ended questions (originally in Finnish in the questionnaire)

Q1: What kinds of assessment methods would you like to have used more than what are used at the moment?

Q2: What kinds of assessment methods would you like to have used less than what are used at the moment?

Q3: If you have received a lower grade than you think you would have deserved, what do you think was the reason for that?

Q4: If you have received a higher grade than you think you would have deserved, what do you think was the reason for that?

Q5: If you have already taken the Matriculation exam in English, did you get the grade you deserved in your opinion? Why/why not?

Q6: If you consider some assessment method(s) really useful for learning, why do you think so?

Q7: If you consider some assessment method(s) totally useless for learning, why (do you think so)?

Q8: Do you want more power to influence assessment? Why? How? Why not?

Q9: What do you think of the Matriculation Examination in Advanced English? What kinds of thoughts/emotions does the examination evoke?

Q10: If you haven't received enough feedback, how and what kind of feedback would you like to get?

Q11: In your opinion, what is the most important function of assessment? In other words, why is assessment needed at schools? Or is it needed?

Received December 20, 2016
Revision received April 14, 2017
Accepted June 1, 2017

II

TO FEED BACK OR TO FEED FORWARD? STUDENTS' EXPERIENCES OF AND RESPONSES TO FEEDBACK IN A FINNISH EFL CLASSROOM

by

Pollari, Pirjo (in press b)

Apples – Journal of Applied Language Studies, 11(4), 11–33.

.

Apples – Journal of Applied Language Studies
Vol. 11, 4, 2017, 11–33

To feed back or to feed forward? Students' experiences of and responses to feedback in a Finnish EFL classroom

Pirjo Pollari, University of Jyväskylä

Good feedback is a powerful element in learning. Ultimately, however, the impact feedback has on learning depends on how the learner responds to that feedback. So far, foreign or second language studies on feedback have mainly concentrated on different methods of error correction, not on students' responses to feedback in general. This study aims to find out what students thought of the feedback they had received in their EFL studies. Furthermore, the study seeks to discover students' different responses to that feedback. The data was gathered using a web-based questionnaire filled out by 140 students. The students, aged 17–19, were all from a single Finnish upper secondary school. The data was analysed mainly quantitatively. The results show that although students were primarily content with their feedback, they wanted more guiding feedback, i.e. more feed forward. They also wanted more personalised feedback as well as feedback that takes place during the learning process, and not only after it. In addition, the varimax-rotated principal component analysis brought out four different responses to feedback. The results indicate that feedback should be more differentiated to support and empower students in their EFL learning better.

Keywords: feedback, students' responses to feedback, EFL teaching, empowerment

1 Introduction

Feedback can have a strong influence on learning (e.g. Hattie, 2009, 2012; Hattie & Timperley, 2007; Wiggins, 2012) and, thus, good feedback lies at the heart of good pedagogy (see e.g. Black & Wiliam, 1998; Sadler, 1998). Accordingly, feedback is considered a vital element of *formative assessment*, or *assessment for learning* (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Black & Wiliam, 1998, 2012; Taras, 2005). However, even if feedback itself is good, informative and balanced, it does not always work since its impact on learning depends on the response which the feedback triggers in the learner (Hattie & Timperley, 2007; Wiliam, 2012). For instance some students pay little attention to received comments (e.g. Black et al., 2003), or do not notice feedback at all. Also, several studies show that comments and corrections in students' foreign or second language (FL/L2) writing do not improve their writing or its grammatical

Corresponding author's email: pirjo.pollari@norssi.jyu.fi
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies
University of Jyväskylä

http://apples.jyu.fi
http://dx.doi.org/10.17011/apples/urn.201708073429

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

accuracy significantly (see e.g. Bitchener & Ferris, 2012; Ferris, 2012; Guénette, 2007; Semke, 1984; Truscott, 1996, 2007). Why not? Do students not find the feedback they receive beneficial?

This study seeks to find out whether students think the feedback they receive during their upper secondary school studies of English as a foreign language (EFL) is good enough to guide and facilitate their learning. Moreover, it aims to discover what students' responses to feedback are. The data was gathered using a web-based questionnaire answered by 146 students (aged 17–19) in one Finnish upper secondary school.

This teacher-research study focuses on an under-researched but practically very relevant topic and context. Although it is widely accepted that feedback impacts learning greatly, there still is not much detailed classroom research on *how* feedback actually works (Murtagh, 2014). Therefore, several researchers have called for more feedback research, for instance teacher-research, in "naturalistic classroom contexts to explore the real needs of teachers and students" (Lee, 2014, p. 1; see also Bitchener & Ferris, 2012; Hyland, 2010; Jakobson 2015). Furthermore, most L2/FL feedback research has been conducted in ESL and/or college contexts, with EFL school contexts clearly under-presented (Lee, 2014; see also Guénette & Lyster 2013; Üstünbaş & Çimen, 2016). Also, only a few L2/FL studies have investigated students' own views or experiences of feedback (Lee, 2005, 2008; Üstünbaş & Çimen, 2016). Most importantly, the bulk of FL/L2 feedback research has primarily been concerned with *corrective* feedback (CF), i.e. oral or written error correction only (Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja, 2015; Jang & Wagner, 2013). Yet, classroom research on students' experiences of and responses to feedback in general, and not just to CF, would be important in order to further develop foreign language assessment practices that facilitate and foster learning (see also Hyland, 2010).

I will first present the concept of feedback as defined in educational sciences. Feedback in FL/L2 research will also be discussed briefly, but as FL/L2 research has regarded feedback predominantly as corrective feedback, and this study does not, the main theoretical emphasis lies in education. Next, I will introduce the present study, its methodology and findings. Finally, the findings, limitations and practical implications of this study will be discussed.

2 Theoretical background

2.1 Feedback, its functions and features in education

Hattie's syntheses (2009, 2012) of more than 900 meta-analyses, with over 200 million students at different ages and in different subjects, indicate that feedback has a powerful impact on student learning. However, not all feedback is good feedback, and sometimes feedback can have negative effects on learning (e.g. Hattie & Timperley, 2007; Shute, 2008).

What *is* feedback, then? Because feedback is a term used in so many different fields, it is variously defined. Sometimes all actions or comments involving an element of assessment or evaluation, such as advice, praise, grades or even a nod from the teacher in the classroom, are considered feedback. According to many scholars, this should not be the case, though (see e.g. Askew & Lodge,

2000; Burke & Pieterick, 2010; Wiggins, 2012). Feedback should not only state or describe how things are at any given moment, but it should also aim at improving future performance (e.g. Black & Wiliam 1998; Hattie & Timperley, 2007; Wiggins, 2012). Actually, there may be a gap between what teachers see as feedback and what students would expect. According to Hattie (2012, pp. 19–20), teachers describe feedback as "constructive comments, criticisms, corrections, content, and elaboration," whereas students would like to get feedback that would help them to know "where they're supposed to go".

Like several other scholars (e.g. Brookhart, 2012; Burke & Pieterick, 2010; Shute, 2008), Wiggins (2012) opens up the two functions by providing a list of key factors of effective feedback. Firstly, effective feedback is *goal-referenced*, which "requires that a person has a goal, takes action to achieve the goal, and receives goal-related information about his or her actions" (Wiggins, 2012, p. 13). Feedback has to focus on the task at hand, not, for instance, on students' personalities or on comparing students with one another (Brookhart, 2012; Shute, 2008; Wiliam, 2012). In a school environment, the problem sometimes is that the students do not have a clear goal, or they do not know what the goal is. Yet, there cannot be effective feedback without a goal (Brookhart, 2012; Wiggins, 2012).

Secondly, feedback has to be *tangible*, *transparent* and *user-friendly* as well as *actionable*, i.e. so clear, concrete and specific that the learners can easily understand it, accept it and also act upon it in order to reach their goals (Wiggins, 2012). Feedback should not be too complicated, long or technical, nor should it be so short, cryptic or vague that students do not know what it really means, which, according to various studies, often seems to be the case (e.g. Burke & Pieterick, 2010; Cohen, 1987; Leki, 1990). Also, phatic feedback, for example a nod from the teacher, or a short evaluative comment such as *Good!* may encourage students but they do not help them any further (Murtagh, 2014). As Hattie (2012, p. 20) puts it, students need to know "where to put their effort and attention". Brookhart (2012) also adds *differentiated*, i.e. meeting each student's own learning needs, as a criterion for good feedback. Effective feedback should also be *consistent* and *ongoing* as well as *timely*. Sometimes students get feedback so late that they cannot act upon it anymore. However, students need the opportunity to use the feedback to further their learning, not only to receive and understand it (Brookhart, 2012).

Researchers also emphasize the importance of getting positive feedback in order to encourage further learning (Brookhart, 2012; Burke & Pieterick, 2010). However, Hattie (2012, p. 22), among others, warns against mixing too much praise "with other feedback because praise dilutes the power of that information" and may also turn the focus of the attention from the task to the individual. Similarly, feedback comments given in addition to a grade or score may go unnoticed as students shift their attention from the learning task to the grade, and also onto themselves when comparing grades with their peers (Black et al., 2003; Butler, 1987).

Often, and rather too often according to Hattie and Timperley (2007, p. 101), students "view feedback as the responsibility of someone else, usually teachers, whose job it is to provide feedback information by deciding for the students how well they are going, what the goals are, and what to do next". Thus, one aspect of effective feedback is that it enables and *empowers* learners to take charge of their own learning, that it promotes and fosters self-regulated learning, self-

assessment and student autonomy (e.g. Burke & Pieterick, 2010). Accordingly, Askew and Lodge (2000) criticise the traditional view of feedback as a *gift*, i.e. the notion that feedback is something that the teacher gives to the student. They do not subscribe to the constructivist view of feedback as *ping-pong*, going back and forth between the teacher and the student, either. They prefer feedback as *loops*, as reciprocal dialogue and information where "nothing is ever influenced in just one direction" and both the teacher and the student share the responsibility for learning (Askew & Lodge, 2000, p. 13).

Wiggins (2012), however, notes that feedback can exist without a teacher, too. Not only can students give feedback to one another, but students themselves can take note of the effects of their actions as related to the goal, and thus get feedback in the situation, without the feedback being explicitly given by anybody. For instance, students can note if their homework is correct, or whether other students understand what they are saying in an oral exercise in a foreign language class. If self-regulated, autonomous, life-long learning is the ultimate goal of education, then so is successful self-assessment and self-feedback (Earl, 2003, p. 101).

Nevertheless, even if feedback should meet all the requirements for effective feedback mentioned above, it still may not work. Wiliam (2012, p. 32) believes that we actually focus on the wrong thing when trying to determine effective feedback: "What matters is what response the feedback triggers in the recipient."

There are, according to Wiliam (2012), altogether eight alternative ways the recipient may respond to feedback. First of all, the feedback given to a student may either indicate that the student's performance has fallen short of the goal, or that the performance has reached or even exceeded the goal. In either case, the student can respond to feedback in four different ways: by changing behaviour (in terms of effort), by modifying the goal, by abandoning the goal or by rejecting the feedback. Out of these eight responses, only two are desirable. These are: increasing effort, i.e. changing behaviour when the goal has not been reached, and increasing aspiration, i.e. modifying the goal when the goal has already been reached. And, as Wiliam (2012, p. 33) concludes, the response does not necessarily depend on the feedback itself:

> Feedback given by a teacher to one student might motivate that student to strive harder to reach a goal, whereas exactly the same feedback given by the same teacher to another student might cause the student to give up.

2.2 Studies on feedback in foreign or second language education

Previous research on students' views or experiences of FL/L2 feedback has shown that students appreciate and trust teacher feedback – and more so than other forms of feedback, such as self-assessment or peer feedback (e.g. Hyland & Hyland, 2006; Lee, 2008; Leki, 1991; Tarnanen & Huhta, 2011; see also Jakobson, 2015). Most students also want teachers to treat all their errors (Amrhein & Nassaji, 2010; Leki, 1991; Lee, 2005; McMartin-Miller, 2014). And they do: recent studies on teacher feedback on L2/FL writing have found that teachers primarily correct all student errors but they – secondary school L2/FL teachers, in particular – give rather little any additional feedback (e.g. Furneaux, Paran, & Fairfax, 2007; Guénette & Lyster, 2013; Lee, 2004).

However, although students say they value teacher feedback, prior studies have also shown that a significant number of students do not actually pay much attention to teacher feedback. For instance, in a study by Cohen (1987), approximately 20% of the surveyed L1, L2 or FL students did not give much attention to teachers' comments or corrections, and those students who did mainly just made a mental note of the feedback. Is this because much of teacher feedback seems to focus on errors, and may thus be considered negative (e.g. Cohen & Cavalcanti, 1990; Lee, 2008), or because of the possible discrepancy between what kind of feedback teachers provide and what students would like to get (e.g. Black & Nanni, 2016; Cohen & Cavalcanti, 1990)? There also appears to be a gap between what feedback teachers report giving and what students report getting (Cohen & Cavalcanti, 1990). For instance, some recent studies such as Tarnanen and Huhta (2011), Hildén and Rautopuro (2014) and Härmälä, Huhtanen and Puukko (2014) found that Finnish FL teachers reported giving much more feedback than students (aged 15–16) reported receiving; Tarnanen and Huhta (2011) also noted that boys reported receiving individual feedback significantly more than girls.

Although there is some recent FL/L2 literature that examines feedback in a broader sense, such as *diagnostic feedback* focusing on both learners' strengths and weaknesses (e.g. Alderson et al., 2015; Jang & Wagner 2013), much of the FL/L2 literature appears to regard informing students "of the accuracy of their response" as the primary purpose of feedback (see e.g. Leontjev, 2016, p. 18). Accordingly, most FL/L2 feedback research focuses on *corrective* feedback, i.e. correcting language errors (Alderson et al, 2015; Jang & Wagner, 2013). There has been a lively debate about the efficacy of corrective feedback in L2 writing and acquisition literature over the past couple of decades (see e.g. Bitchener & Ferris, 2012; Ferris, 2012; Guénette, 2007). Despite numerous studies and analyses, no consensus on which corrective feedback method is the most effective – or even whether corrective feedback is beneficial for future writing and grammatical accuracy – has been found (e.g. Guénette, 2007; Hyland & Hyland, 2006; Lee, 2005, 2008, 2014; see also Bitchener & Storch, 2016).

Lee (2008) points out that not many studies among this wealth of CF research have asked the students themselves what kind of feedback they would like to have. Quite recently, however, there have been some such studies. For instance, the studies by Amrhein and Nassaji (2010) as well as Black and Nanni (2016) compared teachers' and students' perceptions and preferences over different methods of written CF. The results of both these studies indicated that students' and teachers' preferences as well as their justifications differed somewhat (Amrhein & Nassaji, 2010; Black & Nanni, 2016).

Yet, Sayyar and Zamanian (2015) did not find much difference between the teachers' and students' views. Nonetheless, these studies have concentrated on error correction and not on feedback in a broader sense. Furthermore, few of these studies take into consideration the fact that individual students may have different learning needs, wishes and strategies and thus may respond differently to different forms of corrective feedback (Sheen, 2007; see also Jang & Wagner, 2014). However, recent literature on dynamic assessment has discussed *adaptive* corrective feedback (e.g. Leontjev, 2014, 2016; Poehner, 2008; see also Bitchener & Storch, 2016). Although not necessarily based on students' different feedback preferences or responses, CF is adapted according to the learners' Zone of Proximal Development, i.e. the level where the learners are able to perform

when mediated by the tutor, or a computer, but not yet unassisted (e.g. Leontjev, 2014, 2016; Poehner, 2008; see also Vygotsky, 1978). Also, some studies have explored the connection between students' proficiency and educational context with their feedback preferences (e.g. Chen, Nassaji, & Liu, 2016). Yet, these studies focus on corrective feedback.

3 The present study

3.1 Aims

The present article is part of a larger study, the purpose of which was to discover what the students at our school think of assessment received during their upper secondary English studies. One topic area of the study was the feedback that they had received, which is the focus of this article.

This article has two broader research questions:

1) What are our students' experiences of feedback?
   - Do they feel they get enough feedback?
   - Does the feedback facilitate and guide their learning, i.e. does it serve its purpose as a tool for formative assessment/assessment for learning?
   - If students are not happy with the quality and/or quantity of the feedback, what kind of feedback would they like to have, and why?
2) As the efficiency of feedback is believed to depend on students' different reactions to it, what kinds of responses to feedback did the students have in this data?
   - Were there any differences in the responses to feedback in regard to background factors such as gender, previous grade or year?
   - Were there any other factors that might have a connection with the responses?

3.2 Educational setting

Practically all participating students had started studying English in Year 3 in primary school. Thus far, they had studied EFL for nearly nine or ten years, totalling around 700 or 800 lessons.

Finnish upper secondary school studies are divided into courses, each with approximately 35 lessons. At the time of this study, there were six compulsory and two advanced courses of Advanced English, and their general guidelines and syllabi were defined by the *National core curriculum for upper secondary schools 2003*. Each school could also offer additional school-based courses. Each course was assessed as an independent entity with a numerical grade (4–10, 10 being the best). According to the *Core curriculum 2003*, the primary purposes of assessment were to provide students with feedback on their progress and learning results as well as to guide and encourage them in their studies (p. 224). In addition to the grade, the student could also be given more detailed assessment and feedback either in writing or orally. (For further information, see *National core curriculum for upper secondary schools 2003*).

All course assessment is teacher-based assessment. The only national high-stake test in Finland is the Matriculation Examination, which the students sit towards the end of their upper secondary school studies.

3.3 Participants

The second- and third-year students of our upper secondary school were invited to participate in this study. Out of 199 students, 146 answered the questionnaire (response rate 73.4%), and 140 of them answered all the questions regarding feedback. Out of those 140 students, 76 were second-year students (54.3% of the respondents), who answered the questionnaire during one of their English lessons. Third-year, i.e. final-year, students answered in their own time (64 students, 45.7% of the respondents). Eighty-four respondents were female (60%), 56 male (40%). The average of the students' self-reported previous English grade was 8.6 (range 6–10). So far in upper secondary school, they had studied, on average, 6.7 English courses (range 4–11) and had 3.7 different English teachers (range 2–7). The respondents represent the total student population in our school at the time of the study well, regarding both gender and grades.

3.4 Methods

The data of this study was gathered through a comprehensive web-based questionnaire, specifically designed for the study, with altogether more than 100 statements and questions (see Pollari, forthcoming). They cover the following topic areas: students' goal orientation, the assessment methodology and criteria used in English courses, students' views on their usefulness, their personal experiences and views on the accuracy, fairness, guidance and agency of assessment, as well as feedback. The data explored in this article comes primarily from the feedback section of the questionnaire.

Principally, the data of this article was analysed quantitatively. There were 15 Likert-scale items dealing with feedback (see Table 1 in the Findings section). There was also one open-ended question whose answers offered additional, illuminative data in original student voices. Students' gender, year and previous English grade were used as independent variables. Furthermore, several sets of data from the other topic areas of the questionnaire were used as variables. Pearson correlation coefficients were calculated to analyse the correlations between variables. Independent samples T-tests were also conducted to test the statistical significance of the differences of means of gender and year. Varimax-rotated principal component analyses were also run to summarise the variables of different topic areas into sum variables.

4 Findings

As this study has two broader research questions, the results are also reported in two sections. The students' experiences of feedback in general are discussed first. Then, in order to see different responses to feedback, the four different response types extracted by the varimax-rotated principal component analysis are reported.

4.1 Students' experiences of EFL feedback

First, to show students' overall experiences of the feedback, their answers to the 15 Likert-scale statements are introduced in percentages. To give the students' personal experiences a voice, the percentages are illuminated with students' answers to the open-ended question "If you haven't received enough feedback, how and what kind of feedback would you like to get?" Each comment is first shown in its original wording in Finnish and then translated in English. The comments are identified by a student code indicating the student's year, gender and data number.

Consistent with earlier research (e.g. Lee, 2008; Leki, 1991), this study also found that students appreciated and craved teacher feedback. Nearly 70% of them wanted to have more feedback on their skills and even more, 75%, wanted to have more feedback on how to improve their studying (see Table 1).

Teacher feedback was also considered effective as roughly two-thirds of the students said that the feedback had both helped them to improve their language skills (68.5%) and also helped and guided their studying (64.2%). Furthermore, over half of the students felt assessment and feedback had motivated them. Over 50% of the students thought that the course grade they had received had guided their studies during the next English course. However, one in five, i.e. 20%, disagreed on both of these counts. Peer feedback, on the other hand, was not regarded quite as efficient as teacher feedback. Yet, over half of the students would welcome more peer feedback.

Table 1. Student answers (*n*=140) to feedback statements in percentages, with means and standard deviations (I strongly agree=5, I strongly disagree=1).

	I strongly agree	I agree	I do not know	I disagree	I strongly disagree	M	SD
I would like to have more teacher feedback on my skills.	17.7	51.1	20.6	9.2	1.4	3.74	.906
I would like to have more teacher feedback on how to develop my studying.	31.9	43.3	14,2	8.5	2.1	3.94	.998
Teacher feedback has helped me to improve my language skills	7.1	61.4	17.9	12.1	1.4	3.61	.846
The assessment and feedback I have got have helped and guided my studies	7.1	57.1	23.6	8.6	3.6	3.56	.884
Assessment and feedback I have got have motivated me	13.6	42.9	23.6	17.1	2.9	3.47	1.021
The course grade I receive guides my studies on the next course	3.6	50.7	20.7	19.3	5.7	3.27	1.002
Feedback I have got	6.4	36.9	30.5	19.9	6.4	3.17	1.028

from other students is useful.							
I get enough feedback from other students.	1.4	28.4	25.5	30.5	14.2	2.72	1.070
I do not know what my strengths and /or weaknesses in English are.	1.4	15.6	8.5	44.7	29.8	2.14	1.060
I assess my knowledge and skills myself when we check (homework) exercises in class.	8.6	55.7	14.3	17.9	3.6	3.48	1.000
I get enough information about my knowledge and skills through doing and checking exercises, for instance.	5.7	37.9	34.3	20.0	2.1	3.25	.915
My teacher writes enough feedback at the end of the essay, for instance.	12.8	49.6	7.1	28.4	2.1	3.43	1.097
I get enough feedback about my knowledge and skills during the course so that I can influence or adjust my studies during the given course.	7.1	48.9	13.5	27.0	3.5	3.29	1.052
The test grade interests me more than the teacher's comments or corrections.	6.4	31.4	17.9	35.0	9.3	2.91	1.137
I always check my mistakes and corrections carefully when I get my tests or essays back.	13.5	49.6	9.2	22.7	5.0	3.44	1.130

Do students feel able to assess their own skills? Nearly 75% felt that they know their strengths and weaknesses in English; yet 17% did not think so. Moreover, over 60% of the students said they assessed their skills when checking exercises or homework in class. About 30% did not think they had received enough feedback during the course so that they could have changed their studying during that particular course. Overall, approximately a third of the students would probably have hoped for additional feedback, specifically *during* the course, not only afterwards.

Nevertheless, even though many students seemed to want additional feedback, nearly 40% of the students said that the test score or grade interested them more than the teacher's comments or corrections on the test paper. Furthermore, nearly 30% admitted that they did not necessarily read the feedback or corrections that carefully.

The open-ended question "If you haven't received enough feedback, how and what kind of feedback would you like to get?" produced 65 answers (out of 140

respondents), of which 46 answers were written by female students. The guiding *feed-forward* dimension of feedback was mentioned in 15 answers:

> *Toivoisin, että opettaja voisi kertoa miten pitäisi kehittyä että oppisin paremmin. 2F62*
> I wish the teacher could tell me how to progress so that I would learn better.

> *No esim. kokeitten ja kirjoitelmien loppuun voisi ihan selkeästi laittaa, että mitkä asiat onnistuvat jo hyvin ja mitkä kaipaisivat lisäharjoitusta. 2F99*
> Well, teachers could clearly write the things I already master and those that need more work at the end of tests and essays, for example.

> *Opettaja voisi osoittaa tarkasti osa-alueet, joita kannattaisi kehittää, eikä yksittäisiä virheitä sieltä täältä. 3F30*
> The teacher could clearly indicate the areas that should be developed and not just odd mistakes here and there.

> *Palautetta on tullut määrällisesti riittävästi, mutta siinä pitäisi kertoa aina mahdollisimman tarkkaan, millä tavalla oppilas voisi parantaa taitojaan. Näin ei aina käy. 3F38*
> The amount of feedback has been quite adequate but teachers should always tell the students as precisely as possible how they could improve their skills. That doesn't always happen.

> *Toivoisin, että opettajat voisivat kertoa kurssin aikana esim. tehtävien yhteydessä asioista mitä pitäisi vielä harjoitella. 2F61*
> I wish teachers could tell us during the course, for example when checking or doing exercises, what things should be practised more.

All in all, students wanted feedback that is personalised (18 mentions), actionable and tangible (15), on-going and timely (5) as well as constructive and balanced (5).

The results above indicate that although feedback seems to guide and facilitate our students' learning quite adequately, we teachers should pay more attention to feedback in EFL teaching. However, as Hattie and Timperley (2007, p. 101) put it: "Simply providing more feedback is not the answer, because it is necessary to consider the nature of the feedback, the timing, and how a student "receives" this feedback". But how could a teacher know how individual students react to feedback? Is there a way to discover what factors might correlate with students' different needs and reactions?

4.2 Four different responses to feedback

To analyse the feedback data more closely, the varimax-rotated principal component analysis was conducted to summarise the covariance of the 15 variables into a few principal components, i.e. variable clusters, in order to discover the main components of the feedback response. The principal component analysis does not explicitly assume normal distribution (Chatfield & Collins, 1980, p. 58). However, as the components were used in a further statistical analysis, it is worth mentioning that most variables used in the PCA were slightly skewed to the right. The SPPS software was used for the statistical analyses.

The analysis extracted four components of intercorrelating variables (Eigenvalue >1, in total explaining 57.8% of variance). The amount of variance explained by the last two components was somewhat lower (16.2%) than by the

first two components (41.6%). Nonetheless, since the four component solution was pedagogically logical and relevant and the loadings were high enough (Metsämuuronen, 2008, p. 31), this solution was accepted.

Next, on the basis of these four components, four sum variables were formed by selecting the variables with the strongest loadings in each component. The original scale (1–5) of the variables was retained. Thus, the minimum value of each sum variable is 1 and maximum 5. The statements and their loadings in each sum variable are shown in Figure 1. Cronbach's alpha, which indicates the internal consistency of the sum variable, ranged from .72 to .49 (see Figure 1). Although two sum variables did not reach .60, which often is considered the adequate value for Cronbach's alpha (e.g. Jokivuori & Hietala, 2007, p. 104), they were included because they are pedagogically informative and easily interpretable. The most crucial reason for including or excluding some sum variable was the relevance of its content (e.g. Metsämuuronen, 2008).

The following four sum variables are thus clusters of the items that deal with students' wishes, views and experiences regarding feedback. Each student has a value (1-5) for each sum variable. Inspired by Wiliam (2012), I will call the resulting sum variables *responses to feedback*.

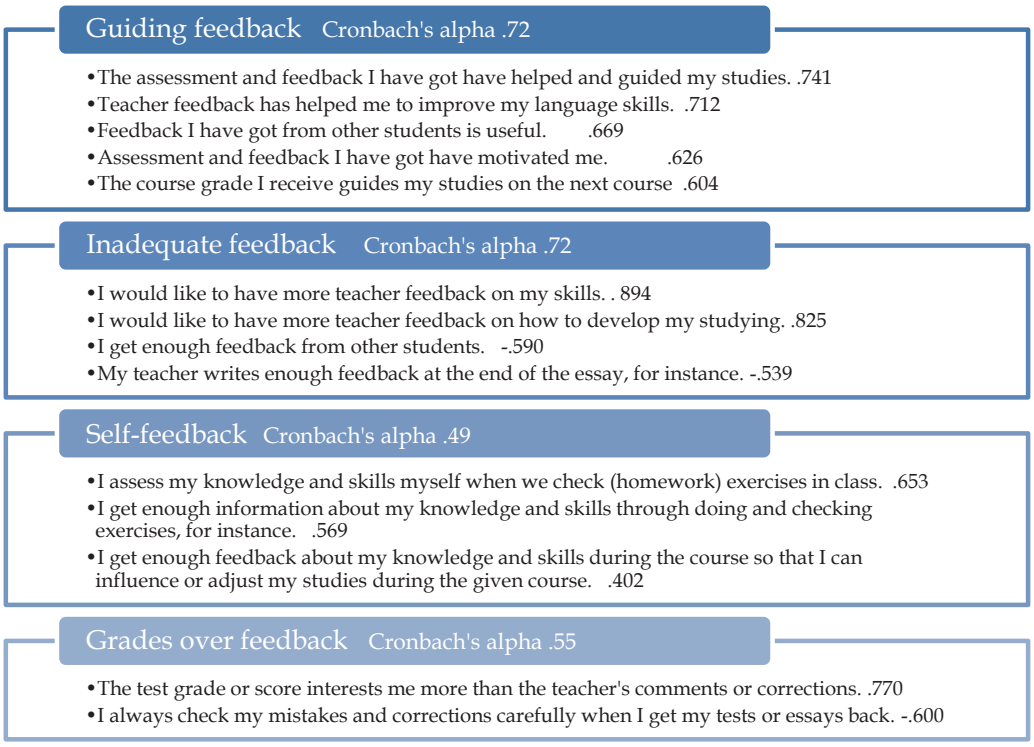

Figure 1. The four sum variables with the items and their loadings.

Similarly, with the same principles and methods as the sum variables in Figure 1, several sets of data from the other topic areas of the questionnaire were transformed into sum variables, one topic area at a time. The varimax-rotated principal component analyses resulted into altogether over 20 sum variables that

deal with different aspects of fairness, accuracy and versatility of assessment, as well as students' personal experiences of power, agency and anxiety related to assessment (see Appendix 1).

4.2.1 Guiding feedback

Undoubtedly, every teacher giving feedback hopes that the feedback is beneficial and that their students make good use of it. The first sum variable, or response type, epitomises that. For instance, the assessment and given feedback have helped and guided students' studies but also helped students to improve their language skills. Furthermore, the grades given may have guided their studies on the next course and both assessment and feedback are considered motivating. Also peer feedback is regarded as useful.

There is no statistically significant difference between genders in this sum variable (girls M = 3.42, SD = .648; boys M = 3.41, SD = .683), $t(138)$ = .021, p = n.s, d = .00. The third-year students had a slightly higher average (M = 3.57, SD = 559) than second-year students (M = 3.29, SD = .712). The difference is statistically significant, $t(138)$ = -2.57, p <.01, d = .43, and the effect size nearly medium. However, no significant correlation between previous grades and *Guiding feedback* was found (r =-.072, p = n.s).

Looking at the data from other parts of the questionnaire, *Guiding feedback* correlates with the experience that assessment and its methods have been varied and good (r = .364, p <.01). Furthermore, experienced empowerment over assessment, i.e. students' experience that they have had a chance to influence the assessment methods and criteria themselves as well, correlates significantly with this sum variable (r =.393, p <.01), as does the experience of not being defeated, disillusioned or disempowered by assessment (r =-.419, p <.01). Furthermore, self-assessment skills and the usefulness of received feedback seem to go hand in hand: there is quite a strong correlation between this sum variable and that of self-feedback (r =.456, p <.01).

When tracing this sum variable back to the individual students' answers in the questionnaire, there were 31 students (out of 140) whose value for this sum variable was 4 or more. Thus, they could be considered as having experienced feedback as beneficial and also as having made good use of it. Even though these students appeared to be quite happy with both feedback and assessment in general, they made some suggestions on how to improve feedback in their open answers:

> *Opettajat voisivat opastaa, millä keinoin voisin parantaa kielitaitoani ja antaa palautetta osaamisestani pitkin kurssia. 3F45*
> Teachers could show me in which ways I could improve my language skills and give feedback all along the course.

> *Just kirjoitelmissa suullinen ja kirjallinen palaute mikä meni pieleen mikä jo hyvää. Mitä voisi tehdä paremmin. Mikä unohtui? Myös yhteisesti on hyvä saada palautetta vaikka yleisistä virheistä ryhmän kesken. 3F122*
> In essays in particular both oral and written feedback, what went wrong, what is good already. What could be done better. What was forgotten? It's also good to get general feedback on common mistakes in class, for instance.

> *Toivoisin etenkin juuri, että kerrottaisiin, mitä pitäisi kehittää ja miten, eikä vain todeta, että tuo kohta meni väärin. 3M128*

> Especially I'd hope that teachers would tell us what to improve and how and not just state that that went wrong.

Out of those 31 students, 18 were female (58.1%) and 13 male (41.9%). As the female/male ratio in all the respondents was 60/40%, this also indicates that there is no link between the gender and *Guiding feedback*. However, the distribution of second-year and third-year students (13 and 18 students respectively, i.e. approximately 42/58%) shows an overrepresentation of third-year students when compared to all respondents (54.3/45.7%).

4.2.2 Inadequate feedback

In turn, the second sum variable focuses on feedback that does not work well, because, even though feedback is seen as important and valuable, there has not been enough of it. Thus, more feedback is called for on both language and studying skills. Furthermore, more peer feedback as well as more comments at the end of essays are needed.

Female students (M = 3.52, SD = .726) seemed to experience inadequacy of feedback a little more than male students (M = 3.18, SD = .750). The difference is statistically significant, $t(139) = 2.644$, $p < .01$, $d = .45$. The effect size is nearly medium. The previous grade may play a minor role as well: the lower the grade, the bigger the feedback inadequacy on average. However, the correlation is rather low ($r = .190$, $p < .05$). There is no statistically significant difference between second- and third-year students in this sum variable, (second-year students M = 3.40, SD = .751; third-year students M = 3.36, SD = .757), $t(139) = .338$, p = n.s, $d = .06$.

Again, when looking at the data from other parts of the questionnaire, *Inadequate feedback* correlates with insecurity of one's own skills ($r = .378$, $p < .01$) as well as the experience of being defeated, disillusioned or disempowered by assessment ($r = .355$, $p < .01$). The view that assessment methods have *not* been good and varied, as well as the wish to have more power to influence the assessment, correlate with *Inadequate feedback* ($r = -.323$, $p < .01$ and $r = .313$, $p < .01$ respectively). Also, there is a preference for softer, lower-stake assessment: a wish to have more formative assessment ($r = .320$, $p < .01$) and less weight on the course exam ($r = .370$, $p < .01$) correlate with this sum variable. Furthermore, the fear of the Matriculation Examination (=high-stake final examination) and the inadequacy of feedback correlate quite strongly ($r = .460$, $p < .01$).

In this data, there were 39 students whose value for this sum variable was 4 or more. Out of those 39 students, 27 were female (69%) and 12 male (31%) students. The female-male ratio in all respondents being 60/40%, this also demonstrates that there seems to be a link between gender and the need for more feedback. The ratio of second-year and third-year students (21 and 18 students respectively, i.e.54/46%) is the same as in all the respondents.

In the open answers of these 39 students, 16 students (out of 26 who volunteered comments) hoped for personal oral feedback from the teacher, usually in addition to written feedback. In a way, the students sound rather dependent on external feedback and do not view feedback as their own task.

> *Henkilökohtainen suullinen palaute. 3F120*
> Individual oral feedback.

Haluaisin saada opettajalta suoraan suullista palautetta, usein!! 2M111
I'd like to get oral feedback directly from the teacher, often!!

Kirjallisesti ja suullisesti mahdollisimman usein. 2F108
Orally and in writing, as often as possible.

4.2.3 Self-feedback

Whereas the previous sum variable demonstrated a need for ample teacher feedback, this sum variable is quite the opposite. *Self-feedback* refers to utilising different learning situations, such as checking homework or other exercises, for gauging one's learning and skills in a quite self-directed manner. In Wiggins' (2012, p. 13) words, "feedback is just there to be grasped", it does not need to be given to them by a teacher or a peer. The experience of having received enough feedback during the course in order to monitor their progress and possibly adjust studying strategies is part of this sum variable as well. As could be expected, there is a negative correlation between the sum variables of *Self-feedback* and *Inadequate feedback* (r = -.309, p <.01).

As mentioned above, the sum variables of *Self-feedback* and *Guiding feedback* share a strong correlation and, accordingly, are similar in many respects. For instance, there is a negative correlation between *Self-feedback* and the experience of being defeated or disempowered by assessment as well (r = -.307, p <.01). Furthermore, regarding assessment as good and many-sided (r = .357, p <.01) and having felt able to influence assessment and its methodology (r = .406, p <.01) correlate positively with this sum variable.

As was the case with *Guiding feedback*, there was no statistically significant difference between genders (girls M = 3.30, SD = .684; boys M = 3.34, SD = .718), t(138) = -.808, p = n.s, d = .14. The previous grade do not correlate with this sum variable (r = .055, p = n.s), but, once again, third-year students score higher here (M = 3.48, SD = .642) than second-year students (M =3.22, SD = .723), t(138) = -2.191, p <.05, d = 37, the effect size being between small and medium.

Even though Cronbach's alpha was not very high, and this group shares a rather strong correlation with *Guiding feedback*, *Self-feedback* has features of its own, and thus these two sum variables were kept separate. For instance, *Self-feedback* correlates more positively with the awareness of assessment criteria and goals (r = .337, p <.01) than any other of the four feedback sum variables. This is not surprising: in order to be able to assess their learning and skills, students need to know, and also understand, the goals and criteria of their learning tasks (e.g. Black & Wiliam, 1998; Earl, 2003; Sadler, 1998).

A total of 37 students had the value of 4 or more in this sum variable. The number of male students was 16 (43.2%) and female students 21 (56.8%), as it was with second-year and third-year students, i.e. 16 and 21 respectively. This again indicates an overrepresentation of third-year students when compared to all respondents. Even though these 37 students seem quite self-directed, some of them welcomed more detailed and personalised feedback.

Sanallista arviointia, ei pelkkiä erittäin hyvä, kiitettävä, hyvä jne. -asteikkoa.3F144
Verbal assessment, not just mere scales like excellent, very good, good, etc.

Olen saanut riittävästi, joskus toivoisin kuitenkin kirjoitelmien palautteiden olevan hieman pidempiä. 3F139
I've got enough but anyways, sometimes I'd like to get a bit longer feedback for essays.

Välipalautetta. 2F54
In-between feedback

4.2.4 Grades over feedback

In the final sum variable, the students' interest in their grades is bigger than in the teacher's feedback or corrections, which are not necessarily even checked so carefully.

Unlike in Cohen's (1987) study, where students who did not attend to teacher comments very carefully tended to rank themselves as poorer students, no correlation between different previous grades and this sum variable was found in this data (r = -.011, p = n.s). Furthermore, there was no statistically significant difference between genders (girls M = 2.70, SD = .945; boys M = 2.78, SD = .943), $t(138)$ = -.457, p = n.s, d = .08. Then again, second-year students seemed more grade-oriented (M = 2.94, SD = .945) than third-year students (M = 2.48, SD = .882), $t(138)$ = 2.935, p <.01, d = .48, the effect size nearly medium.

Surprisingly, grade orientation does not correlate with success orientation either (r = .069, p = n.s.). There is, yet again, a positive correlation between this sum variable and the feeling of defeat or disempowerment caused by assessment (r = .337, p <.01). Yet, and contrary to *Inadequate feedback*, *Grades over feedback* does not correlate with the wish to have more power to influence the assessment used (r = .074, p = n.s). Also, there are negative correlations between *Grades over feedback* and the wishes to have formative assessment or self-assessment (r = -.245, p <.01 and r = -.235, p <.01 respectively). Finally, quite expectedly, this sum variable correlates negatively with *Guiding feedback* (r = -.294, p <.01) and *Self-feedback* (r = -.299, p <.01).

There were 25 students whose value for this sum variable was 4 or more. Twelve of them were female (48%) and 13 male (52%); furthermore, 17 of them were second-year students (68%) and eight third-year students (32%). In other words, both male students and second-year students were overrepresented in this group. In their open comments some of these students hoped for more personal, guiding and also encouraging feedback, and some sounded slightly disappointed with their feedback:

> *Enemmän saisi kertoa sitä, mitä voi kehittää ja parannella. Liikaa keskitytään yksittäisiin, pienempiin virheisiin ja unohdetaan kokonaisuus sekä se, mistä virheiden tekeminen johtuu. 3F136*
> More could be said about what to develop and improve. Too much focus on separate, smaller mistakes and the whole is forgotten, as well as the reason why these mistakes are made.

> *Olen saanut tarpeeksi palautetta, niin hyvässä kuin pahassa. 2M4*
> I've got enough feedback, both in good and in bad.

4.2.5 Empowerment and the wish for external feedback

In general, background factors such as gender, year or previous grades did not appear to explain all the four different feedback responses effectively. However, as the analyses above reveal, there were other factors that seemed to be related to the feedback responses more clearly.

Both in *Guiding feedback* and in *Self-feedback*, assessment in general could be seen as *empowering*. First of all, assessment had been beneficial for the students' learning and studies. It had also allowed these students agency, i.e. the students felt that they had had a chance to influence the assessment methods and criteria themselves. Furthermore, assessment was regarded as fair, accurate and versatile. In other words, assessment was considered to serve students well.

In contrast, assessment was considered a *disempowering* element in *Inadequate feedback* and *Grades over feedback*. It had not facilitated students' learning adequately, nor had it been versatile enough to allow students to show their real skills. Students felt discouraged, even defeated, by assessment. In short, assessment had not worked well and instead of giving students more power or resources, it had impacted negatively on them. Therefore, on the basis of empowerment – or disempowerment – related to assessment, the four sum variables can be divided into two groups: the empowered and the disempowered (Figure 2).

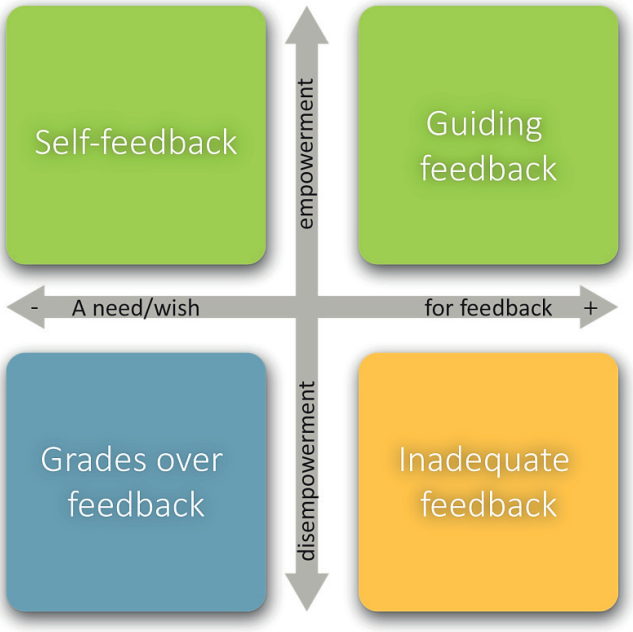

Figure 2. Different responses to feedback in relation to assessment empowerment or disempowerment and a wish or a need to receive feedback.

However, whereas students in *Inadequate feedback* wanted more feedback, in the category *Grades over feedback,* students tended to focus on grades and more or less rejected the teacher's comments or corrections. This resembles Wiliam's (2012) classification, where rejecting feedback is one of the unsuccessful responses to feedback. Therefore, a wish or a need for teacher feedback is another dimension on the basis of which these four sum variables can be divided into two groups: those needing or wanting external feedback, and those not. In the categories *Guiding feedback* and *Inadequate feedback*, students clearly welcomed feedback because they found it beneficial for their learning. *Self-*

feedback, on the other hand, did not seem to require feedback from teachers or peers since students could infer feedback from different learning situations themselves.

5 Discussion

Primarily, the aim of this teacher research was to evaluate and develop EFL assessment and feedback practices in our school. This study had therefore a very practical starting point: Do our students feel they get enough feedback? Does the feedback they receive facilitate and guide their learning, in other words, does it help to 'fill the gap' between their performance and the goal (Sadler, 1989)? If they are not happy with the quality and/or quantity of the existing feedback, what kind of feedback would they like to have, and why?

The findings proved that a vast majority of students wanted *more feedback* on their language skills and, moreover, on their learning and studying skills. At the same time, most of our students seemed content with the feedback they had received and found it helpful and motivating. There were, nonetheless, also a considerable number of students who were not completely happy with the existing feedback and gave several suggestions on how to improve feedback.

Several researchers, such as Black and Wiliam (1998), Sadler (1989, 1998), Taras (2005), Hattie (2009, 2012) and Wiggins (2012) to name but a few, have maintained that the quality of feedback is important. The students of this study agreed with them. Firstly, feedback should not only refer to the present state but feed forward: just as Hattie has suggested (2012), we teachers appeared to have concentrated more on the students' current performance while the students craved feedback that would improve their future performance and learning. Secondly, the students wanted feedback that is individual and personalised, and so clear, concrete and specific that they know what it means and what they should do (see e.g. Sadler, 1998; Wiggins, 2012). Furthermore, they wished to have more feedback during the course, not only at the end of it, so that they could act upon it. They also aspired to constructive and balanced feedback, not just error correction, and they wanted more varied methods of giving feedback. Hence, in the light of the results of this study, it seems that the traditional FL/L2 approach to feedback as corrective feedback does not satisfy the needs and wishes of all students. However, not one student mentioned *goals* in their comments. Yet, being goal-referenced is considered paramount not only in efficient feedback but in learning. Do we not make the goals of different exercises, assignments or learning in general clear enough to our students? However, as Sadler (1989, p. 119) phrases it, "for students to be able to improve, they must develop the capacity to monitor the quality of their own work during actual production" – for that end, they need to understand the goals as well as the criteria for good work (see also e.g. Black & Wiliam, 1998, 2012; Taras, 2005).

As pointed out in earlier research, (e.g. Hattie & Timperley, 2007; Wiliam, 2012), the efficiency of feedback does not seem to depend only on its quality or quantity, but also on students' different responses to it. Thus, another aim of this study was to discover what kinds of responses to feedback our students had. The principal component analysis extracted four sum variables, which showcased that students differed greatly in their responses to the feedback they had received. Feedback could be highly appreciated and work well, as was the

case in *Guiding feedback*. Or feedback could work well, but feedback given by teachers or peers was not necessary because of the students' good self-assessment skills, as seen in *Self-feedback*. In a way, this is the ultimate goal of feedback: external feedback has worked so well that it has made itself redundant. Feedback could, for one reason or another, also fail. *Inadequate feedback* did not meet all students' needs for external feedback, which they valued and craved for. Or, as was the case with *Grades over feedback*, feedback in the form of teacher comments or corrections was not much valued or welcomed.

In addition to differences in the appreciation of, or need for, teacher feedback, there were also clear differences in the experiences of empowerment and disempowerment related to assessment. With *Guiding feedback* and *Self-feedback*, assessment in general could be considered empowering. Assessment was seen as versatile, appropriate and just, and it seemed to serve students well. Therefore, assessment empowered students in their learning process: it gave them power and useful resources to conduct their studies. By contrast, with *Inadequate feedback* and *Grades over feedback* assessment was experienced as a disempowering factor that had not succeeded in motivating, guiding and helping students in their learning, nor had it given them a chance to show all their English skills.

Whereas the previous success in English studies did not correlate with any of these four feedback responses, gender may have an influence on *Inadequate feedback* and *Grades over feedback*, which both also correlated with experienced assessment disempowerment. Female students manifested a stronger tendency towards *Inadequate feedback*. One explanation for this may be test anxiety: earlier research has shown that female students experience more stress over testing, and in particular over high-stake tests (e.g. Hembree, 1988). To some extent, that seemed to be the case also in this data.

Male students, on the other hand, showed a stronger preference for *Grades over feedback* than female students, as also did second-year students. Do younger male students thus focus more on themselves, and on comparing their grades with their peers, than on the learning tasks (cf. Butler, 1987)? Third-year students, then again, seemed to be more capable or willing for *Self-feedback* and also experienced *Guiding feedback* more than second-year students. Do feedback, its importance and usefulness gain momentum as the stakes get higher with the nearing final examinations? Is this because students at that phase pay more attention to feedback and make better use of it, or do we teachers give more and better feedback for third-year students? Are their self-assessment and self-feedback skills also that much better at the point? Or is *learning* simply more important for them?

6 Conclusion

This study was limited to one Finnish upper secondary school only, and thus the findings cannot be generalised as such to other schools or contexts. Furthermore, the academic achievement of the student population in our school is above the national average. Thus, this data does not include many views or experiences of students who really struggle with their studies. With larger and more varied student samples, the feedback experiences and responses might look different. Furthermore, had there been more questions dealing with feedback, or different

questions, it might have changed the findings. Different data might have enabled the use of other data analysis methods as well. For instance, with more varied data, cluster analysis could have revealed different student types and their responses to feedback. There is plenty of room for further research on students' views on and experiences of feedback in foreign language education. Yet, to my knowledge, this study is the first attempt to analyse students' *responses to feedback in general*, and not only to corrective feedback, in FL education. And even if the descriptive statistics or the categories of feedback responses might not be similar in other schools or contexts, the pedagogical implications of this study could well be applicable to other FL education contexts as well.

What are the practical and pedagogical implications of this study? First, on the basis of this study, EFL feedback in our school works quite well in most respects. However, instead of feeling complacent, we should pay more attention to the quality of our feedback. Our feedback should aim at improving future performance, not just stating or describing how things are at that moment. Neither should we focus on error correction only. We should also strive to give more feedback during the learning process and not only after it. In short, more balanced and personalised feed-forward during the upper secondary courses is in order.

Building self-assessment skills needs to be addressed more, since self-feedback skills will be vital for our students' future studies and life-long learning (e.g. Hyland, 2010). Yet, a significant number of students do not engage in assessing their own skills or learning, in other words, they do not grasp feedback from the learning situations but depend on external feedback. One reason might be that they consider feedback "the responsibility of someone else" (Hattie & Timperley, 2007, p. 101). Another reason might be that they do not recognise their own strengths and weaknesses, or perhaps they do not know what the goals or criteria are. Hence, we should pay more attention to explaining the goals and criteria for good work to our students (e.g. Sadler, 1989; Black & Wiliam, 1998). More empowering assessment methodology and formative assessment – *assessment for learning* – is clearly required in our assessment practices. Further professional training for us teachers in how to give feedback which could foster future learning and not only focus on current errors would be welcomed.

In order to meet the different needs of our students better, feedback should be more differentiated. Dynamic assessment and adaptive (corrective) feedback may well be *one* tool towards this end. However, feedback should not be based only on students' skills, but also on their responses to feedback. This is a tall order since, at least according to this data, students' responses to feedback cannot be directly inferred from their gender, year or previous grades. Although the year and gender gave some clues in this data, there were many other factors that affected students' experiences of and responses to feedback more. For instance, feelings of empowerment or disempowerment linked with assessment turned out to play a significant role. This is an area that definitely calls for more research. I also urge for more FL/L2 research studying feedback in a broader sense and not only concentrating on correcting errors. Foreign language skills encompass much more than just correct language form and, accordingly, many students want, and deserve, more than error correction:

> *Kirjotelmissa voisi olla enemmän palautetta, sillä joskus pelkät punakynäkorjaukset eivät kauheasti motivoi:) 2F11*

> There could be more feedback on essays since sometimes the mere corrections with the red pen don't motivate you that much :)

I hope that future research and innovative classroom work will discover new ways to differentiate FL/L2 feedback so that it would be more beneficial for individual students – but not overburden the teachers at the same time. Then, feedback may truly achieve its real potential and feed learning forward.

References

Alderson, J. C., Haapakangas, E., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York: Routledge.

Amrhein, H. R., & Nassaji, H. (2010). Written corrective feedback: What do students and teachers think is right and why? *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquee, 13*(2), 95–127.

Askew, S., & Lodge, C. (2000). Gifts, ping-pong and loops – Linking feedback with learning. In S. Askew (Ed.), *Feedback for learning* (pp. 1–17). London: Routledge/Falmer.

Bitchener, J., & Ferris, D. (2012). *Written corrective feedback in second language acquisition and writing*. New York: Routledge.

Bitchener, J., & Storch, N. (2016). *Written Corrective Feedback for L2 Development*. Bristol: Multilingual Matters.

Black, D. A., & Nanni, A. (2016). Written corrective feedback: Preferences and justifications of teachers and students in a Thai context. *GEMA Online Journal of Language Studies, 16*(3), 99–114.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead: Open University Press.

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London: King's College London School of Education.

Black, P., & Wiliam, D. (2012). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and learning (2nd ed.)* (pp. 11–32). London: SAGE.

Brookhart, S. M. (2012). Preventing feedback fizzle. *Educational Leadership, 70*(1), 24–29.

Burke, D., & Pieterick, J. (2010). *Giving students effective written feedback*. Maidenhead: Open University Press.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*(4), 474–482.

Chatfield, C., & Collins, A. (1980). *Introduction to multivariate analysis*. London: Chapman and Hall.

Chen, S., Nassaji, H., & Liu, Q. (2016). EFL learners' perceptions and preferences of written corrective feedback: A case study of university students from Mainland China. *Asian-Pacific Journal of Second and Foreign Language Education, 1*(1), 5.

Cohen, A. (1987). Student processing of feedback on their compositions. In A. Wenden, & J. Rubin (Eds.), *Learner strategies in language learning. Language teaching methodology series* (pp. 57–83). Englewood Cliffs, NJ: Prentice-Hall.

Cohen, A., & Cavalcanti, M. (1990). Feedback on compositions: Teacher and student verbal reports. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 155–177). Cambridge: Cambridge University Press.

Earl, L. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.

Ferris, D. (2012). Written corrective feedback in second language acquisition and writing studies. *Language Teaching, 45*(4), 446–459.

Furneaux, C., Paran, A., & Fairfax, B. (2007). Teacher stance as reflected in feedback on student writing: An empirical study of secondary school teachers in five countries. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(1), 69–94.

Guénette, D. (2007). Is feedback pedagogically correct? Research design issues in studies of feedback on writing. *Journal of Second Language Writing, 16*(1), 40–53.

Guénette, D., & Lyster, R. (2013). Written corrective feedback and its challenges for pre-service ESL teachers. *Canadian Modern Language Review, 69*(2), 129–153.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hattie, J. (2012). Know thy impact. *Educational Leadership, 70*(1), 18–23.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research, *58*(1), 47–77.

Hildén, R., & Rautopuro, J. (2014). *Ruotsin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013.* [Learning outcomes for syllabus A in Swedish at the end of basic education in 2013]. Helsinki: Finnish Education Evaluation Centre/Finnish National Board of Education.

Hyland, F. (2010). Future directions in feedback on second language writing: Overview and research agenda. *International Journal of English Studies 10*(2), 171–182.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language teaching 39* (2), 83–101.

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

Jakobson, L. (2015). Holistic perspective on feedback for adult beginners in an online course of Swedish. *Apples – Journal of Applied Language Studies, 9*(2), 51–71.

Jang, E., & Wagner, M. (2013). Diagnostic feedback in the classroom. In A. Kunnan (Ed.), *The companion to language assessment* (p. II:6:42:693–711). Hoboken, NJ: John Wiley & Sons, Inc.

Jokivuori, P., & Hietala, R. (2007). *Määrällisiä tarinoita: Monimuuttujamenetelmien käyttö ja tulkinta [Quantitative stories: The use and interpretation of multivariate methods]*. Porvoo: WSOY.

Lee, I. (2004). Error correction in L2 secondary writing classrooms: The case of Hong Kong. *Journal of Second Language Writing, 13*(4), 285–312.

Lee, I. (2005). Error correction in the L2 writing classroom: What do students think? *TESL Canada Journal, 22*(2), 1–16.

Lee, I. (2008). Student reactions to teacher feedback in two Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*(3), 144–164.

Lee, I. (2014). Feedback in writing: Issues and challenges. *Assessing Writing, 19*, 1–5.

Leki, I. (1990). Coaching from the margins: Issues in written response. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 57–68). Cambridge: Cambridge University Press.

Leki, I. (1991). The preferences of ESL students for error correction in college-level writing classes. *Foreign Language Annals, 24*(3), 203–218.

Leontjev, D. (2014). The effect of automated adaptive corrective feedback: L2 English questions. *Apples – Journal of Applied Language Studies, 8*(2), 43–66.

Leontjev, D. (2016). *ICAnDoiT: The impact of computerised adaptive corrective feedback on L2 English learners.* Jyväskylä studies in humanities 284. Jyväskylä: University of Jyväskylä.

McMartin-Miller, C. (2014). How much feedback is enough? Instructor practices and student attitudes toward error treatment in second language writing. *Assessing Writing, 19*, 24–35.

Metsämuuronen, J. (2008). *Monimuuttujamenetelmien perusteet* [The fundamentals of multivariate methods]. Helsinki: International Methelp.

Murtagh, L. (2014). The motivational paradox of feedback: Teacher and student perceptions. *The Curriculum Journal, 25*(4), 516–541

National core curriculum for upper secondary schools 2003. (2004). Helsinki: Finnish National Board of Education.

Poehner, M. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. Berlin: Springer Verlag.

Pollari, P. (forthcoming). *(Dis)empowering assessment? Assessment as experienced by students in their upper secondary school EFL studies.* Jyväskylä: University of Jyväskylä.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*(1), 77–84.

Sayyar, S., & Zamanian, M. (2015). Iranian learners and teachers on written corrective feedback: How much and what kinds? *International Journal of Educational Investigations, 2*(2), 98–120.

Semke, H. D. (1984). Effects of the red pen. *Foreign Language Annals, 17*(3), 195–202.

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly, 41*(2), 255–283.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466–478.

Tarnanen, M., & Huhta, A. (2011). Foreign language assessment and feedback practices in Finland. In D. Tsagari, & I. Csépes (Eds.), *Classroom-based language assessment. Language testing and evaluation vol. 25* (pp. 129–146). Frankfurt am Main: Peter Lang Publishing House.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327–369.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*(4), 255–272.

Üstünbaş, Ü., & Çimen, S. (2016). EFL learners'preferences for feedback types for their written products. *The Online Journal of New Horizons in Education*, *6*(4), 68–74.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard Univeristy Press.

Wiggins, G. (2012). Seven keys to effective feedback. *Educational Leadership, 70*(1), 10–16.

Wiliam, D. (2012). Feedback: Part of a system. *Educational Leadership, 70*(1), 31–34.

Appendices

Appendix 1.

The 28 sum variables based on a varimax-rotated principal component analysis of each topic area of the questionnaire (each topic area is mentioned at the beginning of the name of the sum variable) as well as two additional variables (in italics) that were used in the analyses.

GOAL: empowerment as goal
Goal: self-expression as goal
Goal: success-oriented goals
Goal: education and knowledge as goal

EMPOWERMENT: experienced empowerment
Empowerment: given empowerment
Empowerment: self-grade empowerment
Empowerment: test empowerment

ASSESSMENT: badly assessed
Assessment: good and versatile assessment
Assessment: course test too weighted
Assessment: stressful and discouraging assessment
Assessment: personality affects assessment

USEFUL METHODS: oral
Useful: diagnostic and formative
Useful: no high-stakes tests at all
Useful: self-assessment
Useful: versatile and soft

VIEW: disempowerment
View: want more power
View: don't care
View: no to self-assessment
View: Assessment anxiety: "Assessment causes me anxiety and stress"

MATRICULATION EXAM: fear
Matriculation exam: English for life, not for the exam

FEEDBACK: guiding feedback
Feedback: inadequate feedback
Feedback: self-feedback
Feedback: grade over feedback
Feedback: "I don't know my strengths or weaknesses in English"

Received September 30, 2016
Revision received April 13, 2017
Accepted July 4, 2017

III

DAUNTING, RELIABLE, IMPORTANT OR "TRIVIAL NITPICKING"? UPPER SECONDARY STUDENTS' EXPECTATIONS AND EXPERIENCES OF THE ENGLISH TEST IN THE MATRICULATION EXAMINATION

by

Pollari, Pirjo (2016)

In A. Huhta & R. Hildén (eds.) *Kielitaidon arviointitutkimus 2000-luvun Suomessa. AFinLA-e. Soveltavan kielitieteen tutkimuksia 2016/n:o 9*, 184-211.

Huhta, A. & R. Hildén (toim.) 2016. Kielitaidon arviointitutkimus 2000-luvun Suomessa. AFinLA-e. Soveltavan kielitieteen tutkimuksia 2016 / n:o 9. 184–211.

Pirjo Pollari
University of Jyväskylä

Daunting, reliable, important or "trivial nitpicking?" Upper secondary students' expectations and experiences of the English test in the Matriculation Examination

The Matriculation Examination, the school-leaving exam taken towards the end of upper secondary education, is the only high-stakes examination in the Finnish school system. As the exam may have a strong impact on the students' further education opportunities, it evokes various feelings and thoughts in students. Yet, there is little research on these reactions. This article, based on a mixed-methods approach, sheds light on students' expectation and experiences of the English test in the Matriculation Examination. A total of 142 second- and third-year students from one upper secondary school shared their views on the possible washback effect and test anxiety caused by the exam. Also, the students expressed their ideas and experiences of the validity, reliability and fairness of the test. Although the test did not seem to cause excessive washback, it caused significant stress and anxiety. Furthermore, students seemed rather critical of its validity and reliability.

Keywords: Matriculation Examination, students' experiences, test anxiety, reliability, validity, washback

1 Introduction

In the Finnish school system, the Matriculation Examination is the only examination that can be considered a national, high-stakes examination (Atjonen 2015; Mehtäläinen & Välijärvi 2013). As school assessment is otherwise teacher-designed and quite low-stakes, the Matriculation Examination stands in marked contrast with it. Thus, it is no wonder the Matriculation Examination evokes various emotions, expectations and experiences in Finnish upper secondary school students. As students' experiences have rarely been studied their responses remain mainly anecdotal or based on hearsay.

This article aims to shed light on students' expectations for and experiences of the English test in the Matriculation Examination (ME) in one Finnish upper secondary school. The findings are based on a web-based questionnaire that 142 second- and third-year upper secondary students answered in March 2014. The study relies on mixed methods as both quantitative and qualitative data and methodology were used.

Firstly, the article will discuss high-stakes assessment and its characteristics as well as the Matriculation Examination. Then, the present study, its methodology and findings will be introduced. Finally, I will discuss the limitations and practical implications of this small study.

2 Theoretical background

2.1 High-stakes assessment

External, large-scale examinations, such as school-leaving examinations in various countries, are often labelled as high-stakes examinations. However, according to Heubert and Hauser (1999), for instance, what makes assessment high-stakes is not the assessment itself, nor its contents or form, but primarily the way its results are used and what their impacts are on the student, or on other stakeholders. Thus, in the educational setting, high-stakes tests normally refer to tests whose outcome has "high-stakes consequences for students – that is, when an individual student's score determines not just who needs help, but whether a student is allowed to take a certain program or class, or will be promoted to the next grade, or will graduate from high school" (Heubert & Hauser 1999: 14). High-stakes are therefore closely linked with pressure (Nichols, Glass & Berliner 2006).

The proponents of high-stakes testing have argued that today's high-stakes tests are of state-of-art quality: they are, for instance, "highly reliable; free from bias; relevant and age-appropriate" (Cizek 2005: 41). Hence, because of their outstanding validity and reliability, they can have a positive washback effect: when teachers prepare

their students for testing, they will be "teaching to the standards", which leads to better learning (Cizek 2005: 42). Furthermore, the high stakes attached to the test outcomes are believed to motivate students to study harder in order to gain rewards (e.g. admission for further education) and to avoid punishing consequences such as retention or denial of graduation (see e.g. Heubert & Hauser 1999; Kornhaber & Orfield 2001; Nichols et al. 2006). Along the same lines, high-stakes test scores have increasingly been used for other accountability purposes, such as evaluating an individual teacher's effectiveness or a school's performance (Cizek 2005), which they were not necessarily designed for (e.g. Jones, Jones & Hargrove 2003; Stobart 2008). As the rewards or threats are closely linked with money, job security and other significant factors (Amrein & Berliner 2002), they are believed to act as highly effective incentives and thus improve educational quality and effectiveness (e.g. Cizek, 2005).

Those who are critical of high-stakes testing say that instead of improving teaching and learning, the washback effect leads to teaching to test (e.g. Cheng, Watanabe & Curtis 2004; Madaus & Clarke 2001; Stobart 2008). The pressure of accountability means that schools and teachers want to make sure their students do well in exams and start to prepare them for the exams: as teachers devote more time for test revision and practice tests, it narrows both the content and methodology of teaching and learning (Kornhaber & Orfield 2001; Stobart 2008; see also Alderson & Hamp-Lyons 1996). High-stakes tests are also believed to make learning shallower as often students' primary purpose is to pass the test, not to learn the topics or skills per se (Harlen 2012). Furthermore, most tests focus only on the assessment of the outcomes of learning rather than the process of learning. These factors affect the learning strategies chosen by students when studying. All this may contribute to superficial rote learning instead of real conceptual understanding (e.g. Harlen 2012; Volante 2004).

As high-stakes tests are often one single test with highly pressurised time and place constraints, they may also cause considerable stress and test anxiety (e.g. Aydın 2009). According to research, female students in particular seem to suffer from test anxiety, which may weaken their test performance (Cassady 2010; Hembree 1988). Underperforming in the test, in turn, can affect students' motivation, self-efficacy and self-esteem as learners (Harlen & Deakin Crick 2003).

A low-stakes test has no highly significant consequences for the student (Heubert & Hauser 1999). The defining factor not being the test itself but the use and perceived consequences of the test results, what may be a high-stakes test for one student may not necessarily be so for another.

2.2 The quality of assessment

The quality of any assessment or test is attributed to several characteristics, such as validity, reliability, fairness, impact and practicality or cost-efficiency of the assessment (see e.g. Bachman 1990; Race, Brown & Smith 2005). With high-stakes testing, these characteristics are all the more crucial; as Bachman puts it (1990, 56): "The more important the decision, the greater the cost of making an error".

Validity is traditionally "taken to mean how well what is assessed corresponds with the behaviours or learning outcomes that are intended to be assessed" (Harlen 2010: 36; see also Bachman 1990). Validity, however, is a broad concept and includes various types of validity. *Content validity* is about "the relevance of the test content to the content of a particular behavioural domain of interest and about the representativeness with which item or task content covers that domain" (Messick 1993: 17). *Consequential validity* (Messick 1993), then again, refers to the impact of the assessment (Harlen 2010) and *construct validity* to what is assessed (Harlen 2010; see also Messick 1993, 1996). Black and Wiliam (2012: 244) discuss the notions of construct under-presentation and construct-irrelevant variance (see also Messick 1996), defining them as follows:

> "Construct under-presentation therefore occurs when an assessment fails to assess things it should. The opposite threat to valid interpretation – when an assessment assesses things it should not – is called construct-irrelevant variance."

Often the variation in student scores that is caused by random factors is discussed under the heading of reliability, "reliability being the consistency or accuracy of the results" (Harlen 2010: 36). According to Black and Wiliam (2012), three main sources of construct-irrelevant variance are generally addressed when discussing threats to reliability. One of them is rater reliability, i.e. whether different raters give the same score to the same piece of student work, also known as inter-rater reliability. Intra-rater reliability, i.e. whether the same rater gives the same score to the same answer consistently, is also a significant issue when considering the consistency of scores (see e.g. Bachman 1990; Harlen 2010). The second source of construct-irrelevant variance is the variance in student performance from one day to another: in other words, the student may perform better or worse on different occasions and at different times. The third source is differences in student performance caused by the particular choice of questions or items in the test (see also Bachman 1990). In sum, Black (1998: 54) characterises reliability as follows: "Reliability depends on whether the results are reproducible with different markers, grading procedures, test occasions, and different sets of questions".

In addition to validity and reliability, several authors also include factors such as transparency and fairness. Tests or any forms of assessments should not have nasty

surprises and they should be in line with the intended learning outcomes. Moreover, "students should not be playing the game 'guess what's in our assessors' minds'" (Race et al. 2005: 2). Also, assessment practices should not discriminate or favour any individuals or groups of students. One way of ensuring fairness and equity is a balanced array of different types of exercises.

2.3 The Matriculation Examination

2.3.1 The Matriculation Examination and its history

The Chinese civil service examinations (c. 600–1905), which in some form date back to the times BCE, are credited as the first large-scale, high-stakes examination system in the world (e.g. Elman 2000). However, the standardised, high-stakes tests have dominated educational assessment mainly since the births of the IQ test and the multiple-choice test in the early 20th century (Hanson 1993; Nichols et al. 2006). The current proliferation of high-stakes testing in the USA and Britain, for instance, dates back to the 1980s (e.g. Amrein & Berliner 2002; Black 1998; Kornhaber & Orfield 2001).

The Finnish tradition of testing is quite different from those of the English-speaking countries. The only external, high-stakes examination that we have in the Finnish school context is the Matriculation Examination (Atjonen 2015; Mehtäläinen & Välijärvi 2013). As the word matriculation suggests, its roots lie in an oral entrance examination for Turku Academy. The first modern Matriculation Examination was arranged in 1852. Organised by the Matriculation Examination Board, the new examination was based on upper secondary school syllabi (Kaarninen & Kaarninen 2002; Lindström 1998). Thus far, the Matriculation Examination had still been a university entrance examination. However, in 1919 the Matriculation Examination became the final examination of the upper secondary school, and passing it ceased to mean automatic matriculation to the university. All parts of the examination, both written and oral, were to be organised at schools themselves at the very same time and under strict regulations (Kaarninen & Kaarninen 2002; Lindström 1998).

As with many other high-stakes examinations, the results of the Matriculation Examination were used for assessing the quality of the school until 1918 (Lindström 1998). So, the recent media interest to rank upper secondary schools on the basis of the Matriculation Examination results is not a new phenomenon in Finland. Neither is the washback effect of the Matriculation Examination: according to Lindström (1998) many teachers and principals criticised the Matriculation Examination for narrowing the curricula and teaching methodology into teaching to the test over a hundred years ago.

The Matriculation Examination has also undergone some changes more recently. For instance, a listening comprehension part was added to major foreign or second language tests in the 1970s. Since 1994, it has been possible to divide the examination over three consecutive exam periods, instead of taking the whole exam in just one term. Separate tests for each of the natural sciences and humanistic subjects, instead of an all-encompassing test including all subjects, were introduced in 2006. Currently, the Matriculation Examination is undergoing a process of digitalisation: all of its tests should be computerised by 2019. (For further information, see the Matriculation Examination Board.)

2.3.2 Earlier research on the Matriculation Examination

Research on student assessment in general is rather scarce is Finland, and so is research on the Matriculation Examination. There is some research that focuses on the comparability and reliability of the Matriculation Examination grades (Mehtäläinen & Välijärvi 2013), the history of the Matriculation Examination (Kaarninen & Kaarninen 2002; Lindström 1998) as well as its status (Vuorio-Lehti 2006, 2007). Furthermore, Anckar (2011) has investigated the processes and strategies that students used when answering multiple-choice questions in one French listening comprehension test of the Matriculation Examination. Her findings showed that items with flaws, such as too 'tricky' questions or options as well as items with excessive textual information load or difficulty, represented threats to the reliability of item scores.

Two or three studies have touched upon students' own expectations or experiences of the exam. First, Syrjälä (1989) studied students' and teachers' views and experiences on student assessment as part of studying and teaching. One question in the questionnaire that was part of this small-scale assessment experiment dealt with the Matriculation Examination: over 60% of the respondents, who all were third-year students, found the Matriculation Examination useful while 35% did not.

Some years later, Välijärvi and Tuomi (1995) investigated upper secondary school as a learning environment. Their sample totalled 2,850 first- and second-year students: 75% of them said that their teachers emphasised the importance of the Matriculation Examination either very often (43%) or fairly often (32%). Välijärvi and Tuomi (1995: 49) concluded that the "shadow of the Matriculation Examination is cast, according to students' observations, quite strongly on the everyday work of upper secondary school". Considering that the respondents were all first- or second-year students, this conclusion seems well warranted: at that time the Matriculation Examination was taken at one time only, which generally was during the spring term of their third year, so the respondents had a rather long time left before taking the exam.

In 2009, the evaluation of pedagogy in Finnish upper secondary education (Välijärvi, Huotari, Iivonen, Kulp, Lehtonen, Rönnholm, Knubb-Manninen, Mehtäläinen & Ohranen 2009) surveyed 8,500 third-year upper secondary students. One item in their questionnaire dealt with the Matriculation Examination: "The teachers teach only for the Matriculation Examination". Thirty-five percent of the respondents agreed with the statement while 45% disagreed with it.

3 The present study

3.1 Aims and research questions

The present article is part of a larger study, the purpose of which was to discover what the students at our school thought of assessment in their upper secondary English studies. One topic area of the study was the Matriculation Examination, which is the focus of this article.

The research questions of this article are:

1. Does the Matriculation Examination cause an excessive washback effect? In other words, do students feel that English teachers 'teach to the test' in the upper secondary school? Do students themselves feel that they study for the test alone?
2. Does the English test in the Matriculation Examination evoke test anxiety?
3. Do students consider the Matriculation Examination test a more valid and reliable way of showing their English skills than teacher-based assessment?

3.2 Educational setting

Practically all participating students had started studying English in Year 3 in primary school. Thus far, they had studied EFL for nearly nine or ten years, totalling around 700 or 800 lessons.

Finnish upper secondary school studies are divided into courses, each with 38 lessons. In 2014, there were six compulsory and two specialisation courses of English (Advanced syllabus) and their general guidelines and syllabi were defined by the *National core curriculum for upper secondary schools 2003*. In addition, each school could also offer school-based courses. Although each English course has a theme, they comprise all areas of both oral and written language skills. Hence, course assessment does not focus on any one area, such as grammar, writing or speaking only, but should include them all. Each course is assessed as an independent entity with a numerical

grade (4–10, 10 being the best). In addition to the grade, the student could also be given more detailed assessment and feedback either in writing or orally. (For further information, see *National core curriculum for upper secondary schools 2003*).

All course assessment is teacher-based assessment. The only high-stakes test is the Matriculation Examination, which the students take towards the end of their upper secondary studies. Although the English test is not a compulsory part of the Examination, nearly all students take it, both nationally (Mehtäläinen & Välijärvi 2013) and in this school.

3.3 Participants

The second- and third-year students of our upper secondary school were invited to participate in this study in March 2014. Out of 199 students, 146 answered the questionnaire (response rate 73.4%), and 142 of them answered all the questions regarding the Matriculation Examination. Out of those 142 students, 77 were second-year students, who answered the questionnaire during one of their English lessons. Third-year, i.e. final-year, students answered in their own time (65 respondents) as most of them, preparing to take several subtests of the Matriculation Examination (ME) that spring, did not have lessons at school any more.

Altogether 63 students (44.4% of all respondents) had already taken the English ME test, or part of it. They all were third-year students. Fifty-five of these students had passed the test, but seven of them were retaking the test that spring in the hope of improving their grade. In addition, eight students said that they were in the process of taking the test for the first time that spring. As the students answered the questionnaire in their own time sometime in March, some of these eight students had already completed the whole English test, some had only taken the listening comprehension part (in February). Nevertheless, all these eight students are included in the group of students who had taken the test.

Seventy-nine students (55.6%) had not yet taken any part of the English ME test: among them, there were two third-year students, but all the rest were second-year students. However, they, too, had probably had some personal experience of the exam format, in particular of the listening comprehension part, as sections of them had most likely been used in some of their most recent English courses.

Eighty-five respondents were female (59.9%), 57 male (40.1%). The average of the students' self-reported previous English grade was 8.58 (min. 6, max.10). So far in upper secondary school, they had studied, on average, 6.7 English courses (range: 4–11) and had 3.7 different English teachers (range 2–7). Although the results cannot be generalised, they give quite an accurate picture of the situation in our school at

the time of the study as the respondents represent the total student population well, regarding both gender and their grades.

3.4 Data collection and analysis methodology

The data was gathered through a web-based questionnaire with altogether more than 100 items and questions. The questionnaire covered several topic areas, for instance students' goal orientation, the assessment methods used in English courses and their usefulness, students' personal experiences of and views on the accuracy, fairness, guidance and agency of assessment, as well as the Matriculation Examination.

The data explored in this article come primarily from the Matriculation Examination section of the questionnaire and its Likert-scale items (see Appendix 1) as well as from the goal-orientation questions (see Appendix 2). Those/these data were analysed quantitatively using descriptive statistics. Students' gender, previous (self-reported) grade as well as the fact whether they had taken (part of) the English ME test or not were used as independent variables. Independent samples T-tests were conducted to test the statistical significance of the differences of means of gender and the test-taking. Pearson correlation coefficients were calculated to analyse the correlations between variables.

There were also two open-ended questions dealing with the Matriculation Examination in the questionnaire. Their answers offered qualitative data which were analysed through content analysis (e.g. Patton 2002). First, the content analysis started as inductive analysis "discovering patterns, themes, and categories in one's data" (Patton 2002: 453). However, as the emerging categories and themes, in particular with open-ended answers to Question 9, seemed to match the quality criteria for assessment presented in literature, the second round of content analysis turned into deductive content analysis (e.g. Patton 2002). In other words, at that stage the data were re-categorised according to already existing quality characteristics of validity, reliability and fairness.

4 Findings

My original hypothesis was that whether the students had already taken the test, or not, would somehow affect their answers. Therefore, the results show the descriptive statistics of all respondents first, but also those of the sub-categories of the students who had already taken the test and those who had not as well as female and male students.

4.1 Does the Matriculation Examination cause an excessive washback effect?

The critics of high-stakes testing have been concerned that high-stakes testing narrows teaching and learning. Therefore, the first research question of this study focused on a negative washback effect: Do students feel that English teachers 'teach to the test' in the upper secondary school? Do teachers teach to the test only? Do students themselves feel that they study for the Matriculation Examination alone?

In general, 70 percent of the respondents did not think that their teachers taught to the Matriculation Examination only (see Figure 1.). However, the number of students who said that teachers did indeed teach to the test only was greater among the students who had either already passed the exam or were in the middle of taking it.

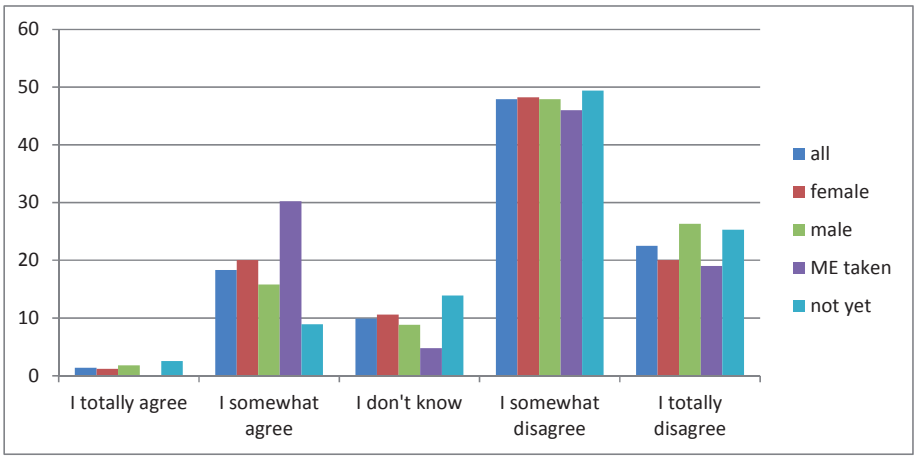

FIGURE 1. Item Teachers teach for the Matriculation Examination only and its responses.

Furthermore, 40% of the students who had not taken the test yet said that their teachers had guided and instructed them too little for the ME test (see Figure 2). Only a good 10 percent of the students who had taken the test shared the same view and almost 80% considered the guidance for the ME test adequate. The difference between those who had taken the test (m=1.98) and those who had not (m=2.91) was statistically very significant (p=.000; r=.391**). Female students (m=2.74) seemed to consider the guidance for the ME test somewhat less adequate than male students (m=2.26, t=2.453, df=140, p=.015; r=-.203*).

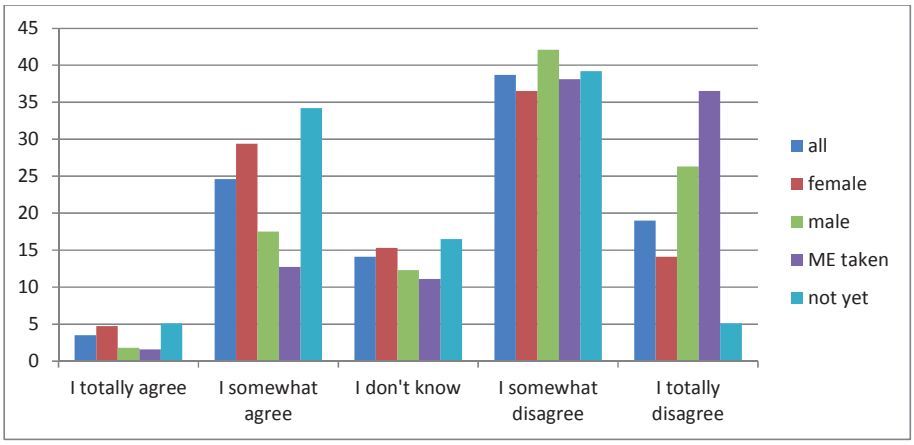

FIGURE 2. My teachers have instructed me too little for the Matriculation Examination.

At the beginning of the questionnaire, the students were asked how much some goals had influenced their studies in the upper secondary school (see Appendix 2). Over 85% of all the respondents said that a good success in the Matriculation Examination had been a goal that had affected their studies either very much or quite a lot. The Matriculation Examination thus seemed to be an important goal – and even more important than a good upper secondary school certificate (see Figure 3).

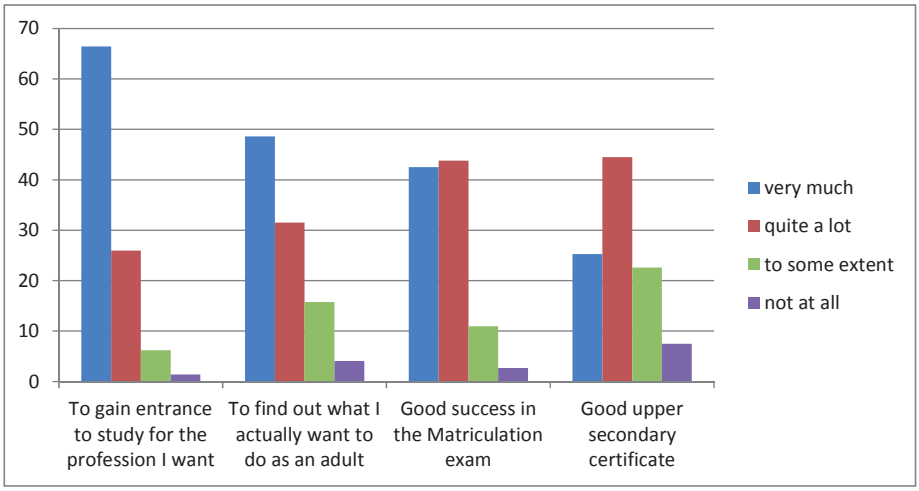

FIGURE 3. To what extent have the following goals guided your upper secondary studies?

Yet again, when asked why they studied English, the results changed. Only about 30% of the respondents said that a good grade in the Matriculation Examination was the most important goal of their upper secondary English studies whereas approximately 55% of the respondents disagreed (see Figure 4). Quite unanimously, the respondents agreed that they studied English primarily for their own future and not for the Matriculation Examination (see Figure 5).

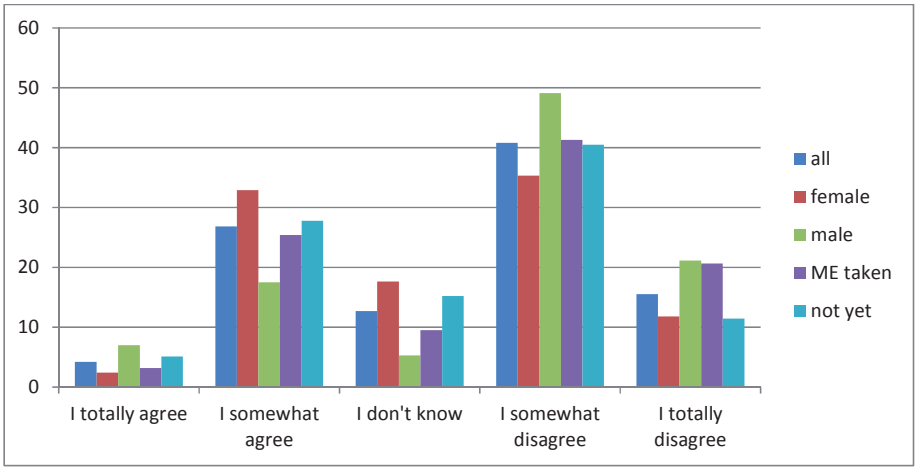

FIGURE 4. The most important goal for me in my English studies is a good grade in the Matriculation Examination.

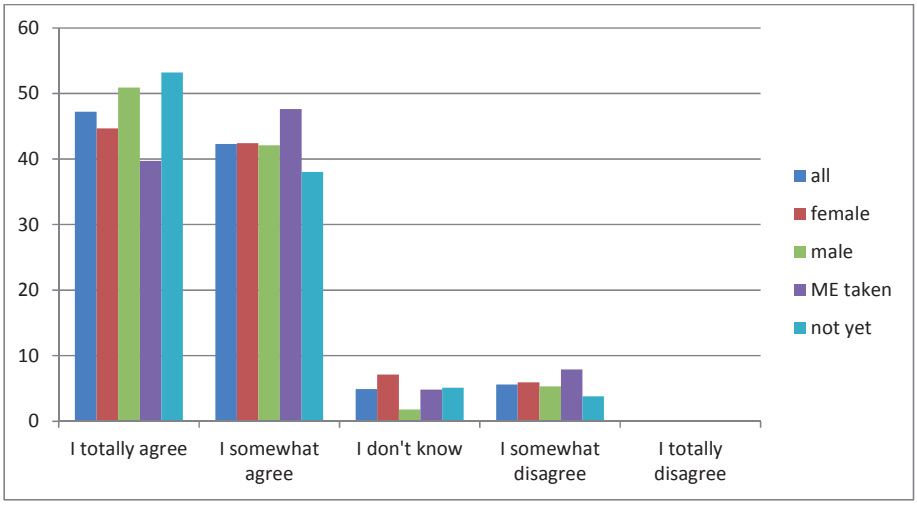

FIGURE 5. I study English for life and for my future, not for the Matriculation Examination.

So, the Matriculation Examination was an important goal that strongly influenced upper secondary studies. However, the primary goal for students' English studies was their own future, not the Matriculation Examination. According to the students in this study, although teachers did not seem to teach to the test at least during the first two years of the upper secondary studies, teaching to the test seemed to increase when the exam approached. Students' earlier success in the English studies, i.e. their previous English grade, did not correlate with any of these items discussed above.

4.2 Does the Matriculation Examination evoke test anxiety?

The second research question dealt with test anxiety, another concern that the critics of high-stakes testing have raised. The results of this study seem clear: the Matriculation Examination evoked some fear or test anxiety in about 60% of the respondents. However, the fear or anxiety seemed to grow a bit milder with the passing of the test, as can be seen in Figure 6. Female students were clearly more susceptible to ME anxiety, and among them, the anxiety was significantly higher (m=3.84) than among male students (m=2.70, t=5.108, df=101.933, p=.000; r=-.411**). In fact, half of the male students did not seem to suffer from any Matriculation Examination anxiety. Students' previous grades did not correlate with anxiety (r= -.125).

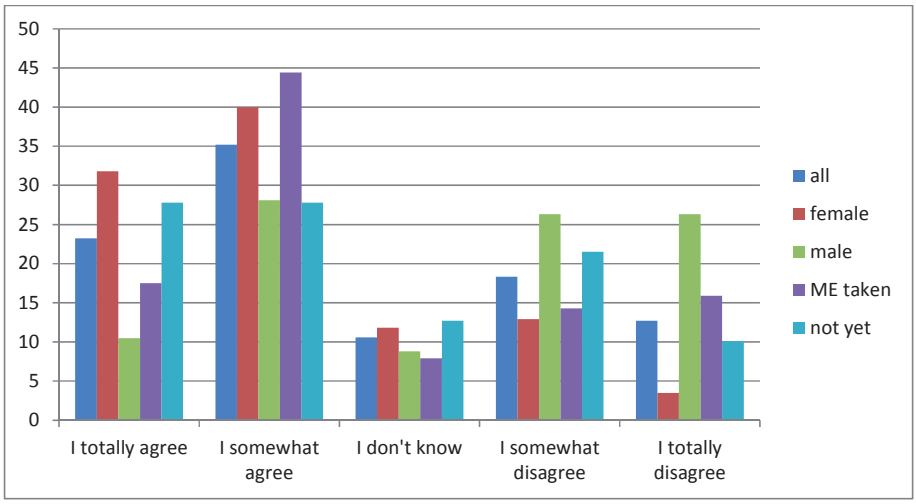

FIGURE 6. The Matriculation Examination scares me.

Approximately one student in four also mentioned either stress or anxiety in their open-ended answers to Question 9: *"What do you think of the Matriculation Examination in Advanced English? What kinds of thoughts/emotions does the examination evoke?"*

Eight of the students who mentioned stress or anxiety had already taken the test. Their stress or anxiety was mostly linked with the test-taking situation or with the high-stakes of the exam:

> The test situation itself is unnerving and exhausting, so the test performance does not always correspond with the real performance.
>
> In listening comprehensions, in particular, stress decreases scores and thus doesn't give a totally reliable picture of the skills.
>
> I'm scared of the ME. test because it affects further studies so much.

Perhaps surprisingly, the students who had not taken the test yet mentioned anxiety or apprehension more often: twenty students (out of 61) mentioned that they were anxious because of the test. Their anxiety or fear ranged from slightly anxious excitement to strong fear that had affected their study plans:

> Haven't done it yet. Mostly anxiety and fear, because I'm scared that I will totally fail in the test even though my English skills are quite good in my opinion.
>
> I fear that ME test the most in the upper secondary school. I chose Advanced Maths so that I won't have to take the Advanced English exam. I'm beside myself with fear because I don't believe I'll pass it with dignity.

On the other hand, ten students were confident of their skills and not worried or anxious about the test:

> I'll pass it even if I had my eyes shut and hands tied behind my back.
>
> It's quite normal, doesn't evoke any feelings, really.

In sum, expectations, perhaps based on other students' anecdotes, seemed somehow stronger, either more anxiety-ridden or more relaxed and confident, than the actual experiences. As with the quantitative answers, attitude and also anxiety seemed to grow more realistic and perhaps milder when lived through.

> As expected, it was quite difficult for me. I had quite a lot of pressure in the test, but I managed well, considering my skills.

Yet, in sum, approximately 60% of all the respondents, and over 70% of female respondents, said that the English test of the Matriculation Examination frightened them at least to some extent.

4.3 Students' views on the validity, reliability and fairness of the ME English test

The proponents of high-stakes tests say that high-stakes testing, designed by assessment experts, is more valid and reliable as an assessment tool than, say, teacher-based assessment (e.g. Cizek 2005). The last research question of this article was to see if the students agreed with this notion.

The questionnaire had two items that addressed the validity and reliability of the Matriculation Examination English test. One of them read as follows: *In the Matriculation Examination, I can reliably show how good my English skills are.* Nearly 60% of the students who had already taken the test did not agree with the claim (see Figure 7). However, approximately a third agreed with the statement. Male students (m=3.18) in general seemed to trust the reliability of the Matriculation Examination more than female students (m=2.39, t=-3.943, df=140, p=.000, r=.316**). There were no statistically significant differences between the students who had already taken the test and those who had not. Students' previous English grade did not correlate with this item (r=.098).

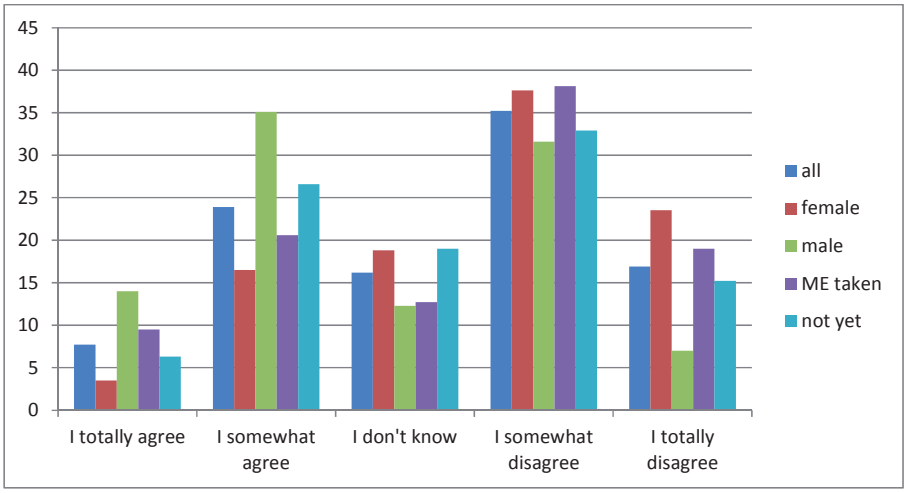

FIGURE 7. In the Matriculation Examination, I can reliably show how good my English skills are.

The second item compared the accuracy of teacher assessment with that of the Matriculation Examination: *The assessment given by the teacher gives a more accurate picture of my skills than the ME test.* Once again, nearly 60% of those who had taken the test thought that the course assessments gave a more accurate assessment of their skills than the Matriculation Examination test (see Figure 8). Only approximately 13% of them disagreed with that claim, leaving 30% undecided. Somewhat surprisingly,

although female students seemed to consider the Matriculation Examination a clearly less reliable format to demonstrate their skills than male students, there were no statistically significant differences between male (m=3.47) and female students (m=3.58) in this item; nor did the previous grade correlate with this item (r=.121).

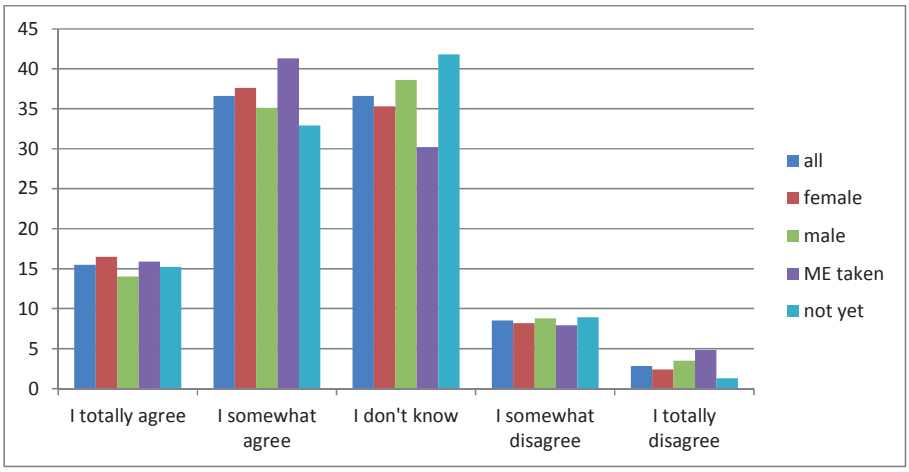

FIGURE 8. The assessment given by the teachers gives a more accurate picture of my skills than the ME test.

In addition, students' open-ended answers illuminated the students' experiences of and expectations of the reliability and validity of the Matriculation Examination test. The students readily volunteered answers: 58 out of the 63 students who had already taken (part of) the test answered the following question: (Q9) *What do you think of the Matriculation Examination in Advanced English? What kinds of thoughts/emotions does the examination evoke?* In the following account, I will concentrate on their answers, because of their first-hand experience. However, I will also briefly mention the expectations of those students who had not taken the test yet. All these answers are categorised according to the main quality requirements of assessment, i.e. validity, reliability and fairness.

4.3.1 Validity: does the test measure what it is supposed to measure?

Out of those 58 students who had taken the test and volunteered open-ended answers, three complimented the test whole-heartedly:

> Good, versatile test. Seems that they know their business in the Matriculation Board.

In addition, eight students regarded the test as good, but also offered some criticism or suggested some improvements:

> Listening comprehensions are quite difficult but otherwise it is suitable. Essays have sometimes rather bad [topic] options as you should have specific knowledge or experience on things.

> In my opinion, the test was good but to my mind an oral test should be part of the package because oral communication is important.

Nonetheless, 35 students (out of 58) questioned the validity of the Matriculation test in one way or another. First of all, the test did not assess students' oral skills in any way, which was criticised in 11 answers:

> The test is deficient in the sense that it doesn't measure the student's ability to communicate orally in English.

> Oral component is missing. Yet, it's one of the main elements of language skills.

The students also commented on the difficulty or 'excessive difficulty' of the test (21 mentions), which surpassed the difficulty level of the English courses (4 mentions):

> Quite challenging, but some structures are really challenging and the teaching during the courses doesn't match their difficulty.

> You can't do well by just attending the English courses offered at school - - The vocabulary and reading comprehensions are more difficult than in the English courses.

Students mostly criticised the test for testing too detailed grammatical knowledge (12 mentions) or vocabulary (8) which were "not important or relevant for good language skills" or real life:

> We learn languages so that we could encounter new people and get to know different cultures. - - This is something the Matriculation Board doesn't seem to understand when they include excessively difficult lottery exercises that test the grammatical knowledge of the exceptions to the exceptions.

> Vocabulary was impossible for an average student wishing for a good grade.

> The difficulty level is rising all the time and the vocabulary needn't be quite so scientific.

However, four students understood the difficulty of the test:

> The English ME test is frighteningly difficult but I guess that separates the best from the rest.

Four students mentioned consequential validity – i.e. the impact – of the Matriculation Examination test on their further studies in their answers:

> Despite the unfairness of the test, in the eyes of further studies institutions, the applicant's English skills are directly comparable with the letter that stands in the Matriculation Examination certificate.

Overall, the students who had taken the test seemed quite critical of its validity. How about the students who had not taken the test yet? Altogether 61 out of 79 students volunteered answers to Question 9. Their answers were not as detailed as the answers of those who had personal experience, and quite a few students also expressed that their answers were based on expectations and other students' stories, not their own experiences. Nonetheless, similar validity issues emerged:

> It's a bit too hard for an ordinary Finn because even native speakers have problems with it at times.

> It contains too much of all sorts of nitpicking that isn't really that much relevant in the development of practical English language skills.

Altogether, eight students criticised exercises for focusing on too detailed knowledge and four students for the lack of an oral part in the test. The difficulty of the test was mentioned ten times and its irrelevance for real life languages skills an additional five times.

All in all, many students seemed quite critical of the validity of the Matriculation test. Its content validity was not regarded as particularly good because speaking was not tested. Furthermore, too detailed knowledge of grammatical exceptions or rare vocabulary was considered irrelevant for real-life language skills. The difficulty level was also seen as too demanding when compared to the syllabi of Advanced English courses.

4.3.2 Reliability: is the ME test a reliable and accurate way to show one's skills?

Once again, many students who had already taken the test (25 out of 63) mentioned various threats to reliability, i.e. many sources of construct-irrelevant variance (Black & Wiliam, 2012). The greatest threats, according to them, were trick questions and red herrings (17 students). Many students compared answering these questions to the draw of the lottery numbers.

> The questions and answers lead you astray, to answer wrong…even if you understand the text/what you hear, the options in the answers trick you to answer wrong, and that is not right.

> The listening comprehension test includes too many so called trick questions and thus doesn't really measure your language skills

> Those bloody multiple choice trolls irk me every time, but of course similar situation may come up in real life too.

In addition to those 17 students, three students also mentioned luck as a possible factor affecting results.

> Because there's only one exam, the result depends very much on test exercises, and doesn't necessarily give the right overall picture of the student's skills.

Another threat to reliability was the test-situation itself with its time constraints and pressure. Four of these seven students specified listening comprehension tests.

> The stressful situation affects your results too and all your skills won't necessarily come out as well as possible.

> I think the listening comprehensions are unfair because they try to bluff the student deliberately and the pauses are so short that you don't have time to read the questions then. So, the results don't give the right picture of your skills then.

Nevertheless, and perhaps slightly surprisingly, none of the 58 students mentioned any concerns about reliability in the sense of inter- or intra-rater reliability.

Out of the students who had not yet taken the test, ten mentioned trick questions. Chance or luck with the topics of the test or with the test's difficulty was mentioned in three answers.

> Exercises made weird and tricky on purpose and not a test that is made on the basis of the real language skill needs.

> Scared, because the difficulty level varies so much between years.

Furthermore, five students also concluded that the test did not necessarily measure or capture one's real English skills.

> Although I feel that I'm pretty good at English, I'm scared that the test will go badly and everybody will get the wrong image of my skills.

4.3.3 Fairness: Is the scoring and grading of the Matriculation Examination fair?

The other open-ended question dealt with the Matriculation Examination grade and its accuracy and fairness: *If you have already taken the Matriculation Exam in English, did you get the grade you deserved in your opinion? Why/why not?(Q5)*

Within the Finnish Matriculation Examination system, if a student has passed any ME sub-test, he or she can retake it once in an attempt to improve the grade. The better grade of these two attempts will be the official grade. Out of those 48 students who had already completed the English ME test and were not going to re-sit it, 45 students answered the question. Twenty-five of them said that the grade had been what they had deserved; for one it was more than she had expected.

> Yes, the grade corresponds with my skills and is in line with my course grades.
>
> Yes, I went there to get a certain grade, and I got it in the end.
>
> I got a far higher grade than I thought so I was happy with the result.

An additional five students said that even if they were not quite satisfied with their grades, they thought they had deserved it for one reason or another:

> I would have wished for a better grade, but in my opinion I deserved that grade because I just could not do better then.

Eleven students, however, did not consider the grade to be what they would have merited. In their opinion, the grade was not in line with their course grades or with their real skills. Furthermore, some students criticised both the excessive difficulty as well as the focus and format of the ME test – in other words, the same issues discussed earlier with validity or reliability:

> The vocabulary in the test was really challenging, and it went badly. In my opinion, I can use English much better than what the grade suggests.
>
> Multiple choices have often questions and options that are somehow bad: several right answers, no completely correct option or a question that can be interpreted in several ways. In open-ended questions you can't guess/deduce what sorts of things they want in the answer. I understand everything but can't always get my answer 'right'.

All seven students who were going to retake the exam answered the question. Five of them said they had not got the grade they felt they deserved:

> In my opinion, no, because I got better results from the tests we did as prep tests than from the real one and that bugs me.
>
> No, I didn't. In my opinion, the grade doesn't reflect my skills because the listening test was a very unnerving experience for me and therefore, anxiety probably ruined my performance. After that when the written part came I was as if I had lost all my hope since I knew I couldn't reach the grade I wanted by any means.

> No because it doesn't match my course grades.

Two of them, however, regarded the grade as deserved:

> Yes, I put too little effort into it.

To summarise, most students seemed to think that they had been quite fairly scored and graded in the ME English test, and that the grade they got was mostly deserved for that particular test. However, they did not seem to think that the Matriculation Examination test itself was a most valid or reliable way to show their skills.

5 Discussion and conclusions

Several opponents of high-stakes tests blame them for a negative washback effect that narrows the curriculum into teaching to the test. Therefore, one of the aims of this article was to find out whether the Matriculation Examination caused a washback effect in this school, and whether students considered the potential washback effect excessive. In the 1995 study by Välijärvi and Tuomi that seemed to be the case. In this study, the results were quite the contrary: 40% of the second-year students felt that their teachers had instructed them too little for the future Matriculation Examination.

In another study by Välijärvi et al. (2009), a good third of the respondents, who all were third-year students, said that their teachers taught only for the Matriculation Examination; nearly half of the respondents, however, disagreed. Although two thirds of all the respondents in this study did not think that their teachers taught only for the Matriculation Examination, 30% of the third-year students thought that they actually did. In that respect, the result is somewhat in line with that of Välijärvi et al. (2009). In sum, the Matriculation Examination seems to have quite a strong washback effect during the final upper secondary courses, but not earlier. Thus, the washback effect cannot perhaps be considered excessive. However, although the Matriculation Examination should be based on the upper secondary school curriculum and its syllabi, the examination is not part of the upper secondary curriculum per se.

Yet, the Matriculation Examination is still a highly important goal for the students, and, therefore, probably also for their teachers. First and foremost, however, the students regarded English as a life skill; almost all the respondents said that they studied English for their future, not for the Matriculation Examination.

Although the Matriculation Examination does not influence the upper secondary studies as much as it may have done in the past, it still 'casts a shadow' on students' daily work in the form of apprehension, stress and anxiety, with nearly two-thirds of the

respondents saying that the exam scared or frightened them. Female students seemed to be more anxious than male students. The reasons the students mentioned for anxiety were, for instance, the pressurised test-taking situation as well as the consequences of the exam for their further studies. Hence, this study corroborates the findings of earlier studies that high-stakes testing causes test anxiety and that female students are more vulnerable to it (e.g. Cassady 2010; Hembree 1988).

Is anxiety a necessarily evil, in other words, is the examination so important and excellent that it is worth the anxiety it seems to cause? The third and final research question of this article focused on students' experiences of and expectations for the validity, reliability and fairness of the Matriculation Examination test as a test of their English skills. Out of those students who had already taken the test, over half thought that the Matriculation Examination test was not a reliable way to show their skills and considered teacher-based assessments a more accurate assessment of their skills. Not everybody agreed with them, though, and students did not seem totally convinced that teacher-based assessment would necessarily be much better.

Yet, many students seemed quite critical of the validity of the Matriculation test in their open-ended answers. Its content validity, or content relevance and coverage, was not regarded as particularly good because speaking was not tested. Furthermore, too detailed knowledge related to grammatical exceptions or rare vocabulary was considered irrelevant for real-life communication skills. The difficulty level was also seen as too demanding when compared to the goals and syllabi of Advanced English courses.

The reliability of the test was not considered very high, either. Students who had already taken the test mentioned various sources of construct-irrelevant variance (Black & Wiliam 2012; Messick 1996), in other words, several threats to reliability. Deliberately tricky questions "that lead you astray, to answer wrong" were considered the greatest threat to reliability (see also Anckar 2011). Quite a few students compared answering tricky multiple-choice questions to pure guessing (see also Anckar 2011). The pressurised test-taking situation and luck were also regarded as threats to the reliability of the test. Hence, the students did not seem convinced that they could show their English skills very reliably in the Matriculation Examination test. Yet, although not necessarily happy with the test and its format, the students seemed to consider the scoring and grading of their test papers quite fair.

This study has many limitations. First of all, it was limited to one school only, and thus the findings cannot be generalised. Furthermore, the academic achievement of the student population in this school is well above national average, also in the Matriculation Examination. Thus, these data do not include many views or experiences of students who struggle with their upper secondary school studies or who risk failing the ME English test. Although the previous English grade did not correlate with any of

the findings in this study, the experiences of and expectations for the ME English test might look different in larger and more varied student populations. Furthermore, had there been more questions dealing with the Matriculation Examination test, or different questions, the findings might have changed. Also, the data analysis methods employed in this article were quite basic. Thus, other data collection and analysis methods would most probably have yielded additional, or different, information. However, this small study sheds some light on students' experiences of the English ME test and also brings forth many interesting questions that still remain unanswered. Thus, further research on students' experiences of the Matriculation Examination with more varied student groups as well as data collection and analysis methodology is clearly needed.

Even though public discussion on the possible abolition of the Matriculation Examination sometimes surfaces, the exam enjoys a high status in Finland (Vuorio-Lehti 2007). Students were not asked directly whether they considered the exam needed or not in this study but my guess, a pure hunch, is that most of the students who have passed the exam would not like to abolish the examination. It seems to be a rite of passage that is part of the school-leaving tradition (Vuorio-Lehti, 2006).

Nevertheless, the students in this study voiced several concerns over the English test which are worth careful attention. Firstly, assessing speaking should somehow be part of the examination. The Matriculation Examination Board has announced that oral production will, sometime in the future, be included in the test. Secondly, the test format should perhaps be reconsidered. As cost-efficient and seemingly reliable (at least in the sense of rater reliability) as the multiple-choice questions are, is there over-reliance on them in the foreign/second language tests? At the moment, approximately half of the total test score, and most of the reading and listening comprehension score, is based on multiple-choice questions. Furthermore, although the difficulty and trickiness of the items may create variance in test results conveniently, is this variance necessarily fair? Moreover, is that variance not too much based on construct-irrelevant variance (Black & Wiliam, 2012; Messick, 1996)? Also, because of the pressurised test-taking situation (as it is the case with the listening comprehension part, in particular) is the test equally fair for all students – including those who suffer from test anxiety?

What should be done? The idea of using the Matriculation Examination results even more extensively for the admission to further education, as suggested, would raise the stakes of the examination considerably. That would also increase the pressure. That, in turn, might have detrimental effects on teaching and learning, as several studies have shown elsewhere. The shadow of the Matriculation Examination would certainly grow longer, and probably darker, again. How would all that accord with the new *National core curriculum for general upper secondary schools 2015* that emphasises versatile assessment methodology, assessment *for* learning, promoting and encouraging students' learning, as well as self-assessment, for instance? No matter how excellent a

test, one single test should never have too much power over a student's future. And as the students in this study have pointed out, there is room for much improvement in the present Matriculation Examination and its English test.

References

Alderson, J. C. & L. Hamp-Lyons 1996. TOEFL preparation courses: a study of washback. *Language Testing, 13* (3), 280–297.

Amrein, A. & D. Berliner 2002. High-stakes testing & student learning. *Education Policy Analysis Archives, 10* (18). DOI: http://dx.doi.org/10.14507/epaa.v10n18.2002

Anckar, J. 2011. *Assessing foreign language listening comprehension by means of the multiple-choice format: processes and products*. Jyväskylä: University of Jyväskylä.

Atjonen, P. 2015. *Kehittävä arviointi kasvatusalalla*. Joensuu: Kirjokansi.

Aydın, S. 2009. Test anxiety among foreign language learners: a review of literature. *Journal of language and linguistic studies, 5* (1), 127–137.

Bachman, L. F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Black, P. J. 1998. *Testing, friend or foe? The theory and practice of assessment and testing*. London: Palmer Press.

Black, P. & D. Wiliam 2012. The reliability of assessments. In J. Gardner (ed.) *Assessment and Learning.* (2nd ed.) London: Sage, 243–263.

Cassady, J. C. 2010. Test anxiety: contemporary theories and implications for learning. In J. C. Cassady (ed.) *Anxiety in schools: the causes, consequences, and solutions for academic anxieties.* New York: Peter Lang, 5–26.

Cheng, L., Y. J. Watanabe & A. Curtis 2004. *Washback in language testing: research contexts and methods.* Mahwah N.J: Lawrence Erlbaum.

Cizek, G. J. 2005. High-stakes testing: contexts, characteristics, critiques, and consequences. In R. Phelps (ed.) *Defending Standardized Testing.* Mahwah, NJ: Lawrence Erlbaum, 23–54.

Elman, B. A. 2000. *A cultural history of civil examinations in late imperial China.* Berkeley and Los Angeles: University of California Press.

Hanson, F. A. 1993. *Testing testing: social consequences of the examined life*. Berkeley: University of California Press.

Harlen, W. 2010. What is quality teacher assessment? In J. Gardner, W. Harlen, L. Hayward & G. Stobart (eds.) *Developing teacher assessment.* Maidenhead: Open University Press, McGraw-Hill Education, 29–52.

Harlen, W. 2012. The role of assessment in developing motivation for learning. In J. Gardner (ed.) *Assessment and learning.* (2nd ed.) London: Sage, 171–183.

Harlen, W. & R. Deakin Crick 2003. Testing and motivation for learning. *Assessment in Education, 10* (2), 169–208.

Hembree, R. 1988. Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58* (1), 47–77.

Heubert, J. P. & R. M. Hauser 1999. *High stakes: testing for tracking, promotion, and graduation.* Washington D.C.: National Academy Press.

Jones, G., B. Jones & T. Hargrove 2003. *The unintended consequences of high-stakes testing.* Lanham, USA: Rowman & Littlefield Publishers.

Kaarninen, M. & P. Kaarninen 2002. *Sivistyksen portti: Ylioppilastutkinnon historia*. Helsinki: Otava.

Kornhaber, M. L. & G. Orfield 2001. High-stakes testing policies: examining their assumptions and consequences. In G. Orfield & M. L. Kornhaber (eds) *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press, 1–18.

Lindström, A. 1998. *Ylioppilastutkinnon muotoutuminen autonomian aikana.* Jyväskylä: Jyväskylän yliopisto, Koulutuksen tutkimuslaitos.

Madaus, G. & M. Clarke 2001. The adverse impact of high-stakes testing on minority students: evidence from one hundred years of test data. In G. Orfield & M. L. Kornhaber (eds.) *Raising standards or raising barriers? Inequality and high-stakes testing in public education.* New York: The Century Foundation Press, 85–106.

Matriculation Examination Board. https://www.ylioppilastutkinto.fi/fi/. See also https://www.ylioppilastutkinto.fi/images/sivuston_tiedostot/Ajankohtaista/HS_28052016.pdf [retrieved 20.9.2016].

Mehtäläinen, J. & J. Välijärvi 2013. *Ylioppilaskokeiden arvosanojen vertailtavuus eri aineissa vuosina 2007–2011.* Jyväskylä: Jyväskylän yliopisto, Koulutuksen tutkimuslaitos.

Messick, S. 1993. Validity. In R. L. Linn (ed.) *Educational measurement.* (3rd ed.) Phoenix (AZ): Oryx, 13–103.

Messick, S. 1996. Validity and washback in language testing. *ETS Research Report Serie*s (1), i–18.

National core curriculum for upper secondary schools 2003. (English translation printed in 2004.) Helsinki: Finnish National Board of Education.

National core curriculum for general upper secondary schools 2015. (English translation printed in 2016). Helsinki: Finnish National Board of Education.

Nichols, S., G. Glass & D. Berliner 2006. High-stakes testing and student achievement: does accountability pressure increase student learning? *Education Policy Analysis Archives, 14* (1). DOI: http://dx.doi.org/10.14507/epaa.v14n1.2006.

Patton, M. Q. 2002. *Qualitative research & evaluation methods*. (3rd ed.) Thousand Oaks, CA: Sage.

Race, P., S. Brown & B. Smith 2005. *500 tips on assessment.* (2 ed.) London: RoutledgeFalmer.

Stobart, G. 2008. Testing times: the uses and abuses of assessment. New York, N.Y.: Routledge.

Syrjälä, L. 1989. *Oppilasarviointi osana lukio-opiskelua ja opetusta: oppilaiden ja opettajien näkemyksiä ja kokemuksia Alppilan lukiossa.* Oulu: Oulun yliopisto.

Volante, L. 2004. Teaching to the test: what every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy*, (35). Available at http://files.eric.ed.gov/fulltext/EJ848235.pdf.

Vuorio-Lehti, M. 2006. *Valkolakin viesti: ylioppilastutkintokeskustelu Suomessa toisen maailmansodan jälkeen*. Turku: Turun yliopisto.

Vuorio-Lehti, M. 2007. Valkolakin hohde: keskustelua ylioppilastutkinnon merkityksestä Suomessa toisen maailmansodan jälkeen. *Kasvatus & Aika, 1* (1), 19–33.

Välijärvi, J., N. Huotari, P. Iivonen, M. Kulp, T. Lehtonen, H. Rönnholm, G. Knubb-Manninen, J. Mehtäläinen & S. Ohranen 2009. *Lukiopedagogiikka*. Jyväskylä: Koulutuksen arviointineuvosto. Koulutuksen arviointineuvoston julkaisuja, 40.

Välijärvi, J. & P. Tuomi 1995. *Lukio nuorten valintojen ja oppimisen ympäristönä.* Jyväskylä: Kasvatustieteiden tutkimuslaitos.

APPENDIX 1.

The nine Likert-scale items dealing with the Matriculation Exam with their percentages, means and standard deviations.

	Täysin samaa mieltä ***I strongly agree***	**Jokseenkin samaa mieltä** ***I agree***	**En osaa sanoa** ***I don't know***	**Jokseenkin eri mieltä** ***I disagree***	**Täysin eri mieltä** ***I strongly disagree***	**m**	**sd**
Opiskelen englantia elämää ja tulevaisuuttani enkä yo-kirjoituksia varten. *I study English for life and my future, not for the Matriculation Examination.*	47.2	42.3	4.9	5.6		4.31	.809
Yo-kirjoitukset pelottavat minua. *The Matriculation Examination scares me.*	23.2	35.2	10.6	18.3	12.7	3.38	1.357
Opettajien antama arviointi antaa oikeamman kuvan osaamisestani kuin yo-koe. *The assessment given by the teachers gives a more accurate picture of my skills than the ME test.*	15.5	36.6	36.6	8.5	2.8	3.54	.950
Yo-kirjoitusten arviointi ei vastaa opettajien arviointikäytänteitä. *The assessment and grading of the Matriculation Examination doesn't correspond with those of the teachers.*	7.7	36.6	38.0	16.9	0.7	3.34	.874
Voin yo-kirjoituksissa luotettavasti osoittaa, kuinka hyvin englantia osaan. *In the Matriculation Examination, I can reliably show how good my English skills are.*	7.7	23.9	16.2	35.2	16.9	2.70	1.225

Lukiossa pitäisi käyttää vain samoja arviointitapoja kuin yo-kirjoituksissakin. *In upper secondary school, only the same assessment methods that are used in the Matriculation Examination should be used.*	4.9	26.8	10.6	41.5	16.2	2.63	1.183
Englannin opintojeni tärkein tavoite minulle on hyvä arvosana yo-kirjoituksissa. *The most important goal for me in my English studies is a good grade in the Matriculation Examination.*	4.2	26.8	12.7	40.8	15.5	2.63	1.158
Opettajani ovat opastaneet minua liian vähän yo-kirjoituksia varten. *My teachers have instructed me too little for the Matriculation Examination.*	3.5	24.6	14.1	38.7	19.0	2.55	1.158
Opettajat opettavat vain ylioppilaskirjoituksia varten. *Teachers teach for the Matriculation Examination only.*	1.4	18.3	9.9	47.9	22.5	2.28	1.054

APPENDIX 2.

The goal-orientation items:

Missä määrin seuraavat tavoitteet ovat ohjanneet lukio-opiskeluasi? (%)
To what extent have the following goals guided your upper secondary studies? (in percentages)

	Erittäin paljon ***Very much***	**Melko paljon** ***Quite a lot***	**Jonkin verran** ***To some extent***	**Ei lainkaan** ***Not at all***
Päästä lukion jälkeen opiskelemaan tavoittelemaani ammattiin. *To gain entrance to study for the profession I want.*	66.4	26.0	6.2	1.4
Selvittää itselleni, mitä isona oikeastaan haluan tehdä. *To find out what I actually want to do as an adult.*	48.6	31.5	15.8	4.1
Hyvä menestyminen ylioppilaskirjoituksissa. *Good success in the Matriculation Examination.*	42.5	43.8	11.0	2.7
Hyvä päättötodistus. *Good upper secondary certificate.*	25.3	44.5	22.6	7.5
Oppia suunnittelemaan opintojani ja tulevaisuuttani. *To learn to plan my studies and future.*	24.7	47.9	24.7	2.7
Saada hyvä yleissivistys. *To get a good all-round education.*	24.0	50.7	23.3	2.0
Opiskella mahdollisimman paljon kiinnostavia kursseja. *To study as many interesting courses as possible.*	21.9	45.2	27.4	5.5
Oppia tuntemaan itseni, vahvuuteni ja heikkouteni. *To learn to know myself, my strengths and weaknesses.*	18.5	43.2	30.1	8.2
Opetella itse ottamaan vastuuta asioista. *To learn to take responsibility.*	15.8	45.2	32.9	6.2
Oppia tekemään päätöksiä ja valintoja. *To learn to make decisions and choices.*	11.0	44.5	37.0	7.5
Oppia ilmaisemaan itseäni. *To learn to express myself.*	8.9	43.2	40.4	7.5
Oppia tulemaan toimeen erilaisissa ryhmissä ja erilaisten ihmisten kanssa. *To learn to get along in different groups and with different people*.	8.9	34.9	49.3	6.8
Mennä samoille kursseille kuin kaverinikin. *To go to the same courses as my friends.*	2.7	15.1	43.8	38.4

IV

CAN A CHEAT SHEET IN AN EFL TEST ENGAGE AND EMPOWER STUDENTS?

by

Pollari, Pirjo (2015)

AFinLAn vuosikirja–AFinLA Yearbook, (73), 208-225.

Jakonen, T., J. Jalkanen, T. Paakkinen & M. Suni (toim.) 2015. Kielen oppimisen virtauksia. Flows of language learning. AFinLAn vuosikirja 2015. Suomen soveltavan kielitieteen yhdistyksen julkaisuja n:o 73. Jyväskylä. s. 208–225.

Pirjo Pollari
University of Jyväskylä

Can a cheat sheet in an EFL test engage and empower students?

Although occasionally used in language classrooms, cheat-sheet tests have not been explored in foreign or second language education research. This study experimented with cheat-sheet tests in the teaching of EFL in a Finnish upper secondary school. The participants, 101 students, could make a cheat sheet for the grammar part of their English test. A total of 92 students prepared the cheat sheet, nine did not. Students' cheat sheets, test results and comments constituted the data for this study, analysed both qualitatively and quantitatively. The existence of the cheat sheet and its quality (thorough, good or limited) correlated with the grammar test results: students with a thorough cheat sheet scored slightly higher points on average than other groups. Even though the cheat sheet did not markedly improve their test results, the majority of students felt that it had improved their learning and studying. Some students also reported reduced test anxiety.

Keywords: student assessment, testing aids, engagement, empowerment
Asiasanat: oppilasarviointi, arviointimenetelmät, kokeet, voimaantuminen

1 Introduction

Over the past decades, schools, curricula, teaching methods as well as theories of learning have undergone great changes. There is very little research on student assessment in Finland, but student assessment in foreign (FL) or second language (L2) education still appears to be somewhat test-based and limited in scope (Hildén & Härmälä 2015; Tarnanen & Huhta 2011). For instance, despite the requirements of the Finnish national curricula, self- and peer-assessment do not seem to play a significant role in FL/L2 assessment (Tarnanen & Huhta 2011). Yet, teaching and learning cannot really be reformed if assessment methods do not change as well. New avenues should therefore be explored in FL student assessment, both in research and in practice.

This study has a dual aim: it is both a teaching experiment exploring cheat sheets in an English test and a contribution to research on foreign language assessment. After a brief look at the theoretical background, I will introduce the experiment, its participants and the methods used. Then I will present the findings, based primarily on qualitative data. In addition, I will examine quantitatively if the cheat sheets had any measurable impact on students' test results. Finally, I will discuss the findings, their limitations and possible implications.

2 Theoretical background

2.1 The power of assessment

Research in FL/L2 education shows that testing has a significant yet quite complex washback effect (Cheng, Watanabe & Curtis 2004; Hughes 1989; Rea-Dickins & Scott 2007). Although the washback effect is not negative per se, evidence about how tests – high-stakes tests in particular – narrow the curricula into 'teaching to the test' abounds (Rea-Dickins & Scott 2007; Volante 2004). Also students, wishing to succeed, want to study for the test itself, which in turn influences their learning strategies. Many tests still focus on memory and accurate knowledge retention instead of high-order learning and thinking skills such as problem-solving or critical thinking (e.g. Atjonen 2007; Pickford & Brown 2006). So, students often try to memorise the information they think will be tested. This easily leads to superficial rote learning and real conceptual understanding, deep learning, takes a back seat (e.g. Harlen 2012; Volante 2004). Ultimately, passing the exam becomes far more important than learning itself (Harlen 2012).

Furthermore, test anxiety, which is rather common among female students, can weaken memory and knowledge retention and, thus, many students cannot show all that they actually know in test situations (e.g. Hembree 1988). Underperforming in the test can affect their motivation, self-efficacy and self-esteem as learners (Harlen & Deakin Crick 2003). Accordingly, several studies have shown that test anxiety, which is also closely related to foreign language anxiety (e.g. Horwitz 2001, 2010), may affect not only students' test results but also their FL/L2 learning processes, proficiency and motivation (e.g. Aydin 2009; Cheng, Klinger, Fox, Doe, Jin & Wu 2014; Liu & Huang 2011).

Finland has only one national high-stakes examination, the Matriculation Examination. However, the effects of testing on students and their learning are not only limited to high-stakes exams (Harlen & Deakin Crick 2003). Students may feel anxious and powerless in the face of any of the assessment situations that take place dozens of times throughout their school year (Atjonen 2007). Determining students' grades, they, too, have high stakes for students. Furthermore, even though socio-constructivist learning theories – the basis of the Finnish national curricula – emphasise the learner's active role and agency in the learning process (e.g. Tynjälä 1999; von Wright 1993), the test-taker has remained far more often than not an object of assessment, rather than an active agent.

2.2 Cheat sheets in a test?

During the past couple of decades, both teachers and researchers have developed alternative assessment methods that are better aligned with current learning theories. For instance, a cheat-sheet exam, also known as a crib-notes exam, refers to an exam or a test where students can bring into the exam notes they have written themselves for that particular testing situation. Sometimes the notes may be restricted, for instance with regard to their content or size. Some teachers have also insisted on using hand-written notes only.

Although cheat-sheet tests have not been really examined in FL/L2 education research so far, there are some published studies on cheat sheets in other contexts, mainly in psychology and mathematical subjects at the tertiary level. Most of the studies so far have advocated cheat sheets for a variety of reasons. For instance, they have concluded that the engagement in creating a personal cheat sheet – and not only using one in the test – improves studying and learning and thus also performance in the test (Block 2012; de Raadt 2012; Erbe 2007; Larwin 2012; Whitworth 1990). This is attributed to a coding process: when students review, select, organise and rewrite information on their cribs, they process the information more actively and more profoundly than when just trying to memorise it (e.g. Larwin 2012; Whitworth 1990). The improvement

in test results may be rather small (Gharib, Phillips & Mathew 2012) but there are other benefits, for instance decreased test anxiety (Block 2012; Butler & Crouch 2011; Erbe 2007; Whitworth 1990) or simply the fact that students find cheat-sheet exams useful and prefer them over closed-book exams (Block 2012; Erbe 2007; Gharib et al. 2012).

However, some studies have concluded that cheat sheets are not beneficial for learning (Dickson & Bauer 2008; Dickson & Miller 2005; Funk & Dickson 2011) even if they have improved test results and students have found them both helpful and stress-reducing (Dickson & Bauer 2008). Dickson and her colleagues argue that instead of really engaging in studying and learning, the students become dependent on their cribs. To test their dependency hypothesis, Dickson and Bauer (2008) organised a dual test on a course examination of developmental psychology at an American university. First, students had to take an unexpected pre-test without their crib notes and, immediately afterwards, they took the real exam, now with their crib notes. The questions were mostly identical multiple-choice questions. Dickson and Bauer argued that if the reason for an improved test performance lay in the engagement and improved learning, then students who had made the crib notes for the exam should perform just as well with or without the cribs in the actual test situation. As this was not the case (students performed better in the real test with their cribs than in the pre-test), Dickson and Bauer (2008: 117) concluded that "constructing crib sheets did not enhance learning, but use enhanced performance" because students depended on their notes in the exam. In fact, Dickson and Bauer (2008: 117) warned that crib sheets, or "crutches", actually cripple learning as "students do not learn the course material as well when they expect to use a crib sheet" as they would for a closed-book exam, and advised against using cheat sheets.

To measure the efficacy of learning in another way, Gharib, Phillips and Mathew (2012) gave students surprise post-tests two weeks after the exams. They could not find any significant difference in the retention quiz performance between the students who had taken open-book, closed-book or cheat-sheet exams. Furthermore, they found out that "scores among exam types are positively correlated – students who do well on one exam type tend to do well on the others" (Gharib et al. 2012: 476). They concluded that "all three types of exams are equally effective as teaching tools", but because of other beneficial factors, they deemed cheat-sheet and open-book exams more learner-friendly than closed-book exams (Gharib et al. 2012: 477).

Although prior research seems to be somewhat conflicting on the benefits of cheat sheets, in their recent meta-analysis Larwin, Gorman and Larwin (2013: 439) found out that "the use of either student-prepared testing aids or open-textbook exams can have a moderate impact on student performance on exams". Furthermore, on the basis of higher effect sizes for student-prepared testing aids, they concluded as follows:

> This outcome suggests some possible additional benefit to students who are required to prepare their own testing aids, thus requiring them to review, organize, and clarify the information on which they are being tested, as has been suggested by earlier research (...). This also suggests a potential benefit in the form of greater student engagement with the course material and information that, as other research has found, can ultimately help students to develop better study strategies and skills that they will incorporate into their other coursework. (Larwin et al. 2013: 439)

Therefore, Larwin et al. (2013) inclined towards favouring cheat-sheet exams.

3 The aim and setting of the study

Some years ago I introduced cheat sheets to some of my EFL courses. In order to examine the cheat-sheet test and its effects more thoroughly, I collected systematic data in 2013 and 2014. My main research interests were to find out how students react to cheat-sheet tests and what kinds of cheat sheets they construct and why. I also wanted to see if cheat sheets affect students' learning, test results as well as their learning and studying experiences. Above all, I wanted to explore if cheat-sheet tests could empower and engage students more in their assessment.

Altogether 101 students (61 females and 40 males, aged 17–18) took part in this study in 2013 and 2014 (47 and 54, respectively). They were on the penultimate compulsory English course (ENA5, the culture course) before the final Matriculation Examination. The cheat sheet was made for the written test, which comprised both grammar and reading comprehension exercises. The grammar exercise was a traditional multiple-choice exercise with 40 items and a maximum of 40 points. Although rather behaviouristic, some of the items required processing two grammatical constructs at once (e.g. articles and capital letters) and, admittedly, the exercise was quite detailed and challenging. The maximum score of the reading comprehension (RC) part was also 40 points.

The contents of the cheat sheet were limited to grammar (articles with proper nouns, punctuation, capital letters, sequence of tenses, conjunctions and linking words, some phrasal verbs). The size of the sheet was restricted (A4 on one side) so that students would have to process and summarise the information they selected. Even though some studies have suggested a link between hand-written notes and better test results (Larwin 2012), the students could prepare their cheat sheets outside lessons as they wanted. My only requirement was that each student made their own cheat sheet, i.e. they were not to copy somebody else's sheet. The cheat-sheet test was not the only assessment method in the course but one among many forms, such as two longer written pieces, an oral presentation, and smaller vocabulary and listening comprehension tests.

After the test, I collected the students' cheat sheets for analysis. Furthermore, using an open-ended questionnaire in Finnish (see Appendix 1), I collected their comments both before and after the test. The final student comments were collected after I had handed students their tests and cheat sheets back. The findings presented in this article are primarily based on the students' cheat sheets as well as their comments.

Inductive qualitative content analysis was used for analysing the students' cheat sheets and comments: after several readings, the cheat sheets as well as the comments were placed into categories emerging from the data (e.g. Elo, Kääriäinen, Kanste, Pölkki, Utriainen & Kyngäs 2014; Tuomi & Sarajärvi 2009). However, not all students answered every questionnaire item, and some answers were also rather vague. When a comment proved difficult to categorise, I consulted a second reader, an experienced educational researcher. Some student comments are used to illustrate the categories in the following text. This way, the reader can evaluate the trustworthiness of the analysis (Elo et al. 2014). The comments are identified by a student number, gender (F/M) and the quality of the cheat sheet (T=thorough, G=good, L=limited and N=no cheat sheet). Originally written in Finnish, I translated not only their meaning but also tried to retain the students' style, grammar and occasional ambiguities. The students' test results and previous grades are also used for additional quantitative analyses.

4 Findings

4.1 The initial reaction, construction and quality of the cheat sheet

When I introduced the idea of a cheat-sheet test to my students, it was a new idea to nearly half of them. In contrast, 15 students said they had used cheat sheets in two or more tests in various subjects, for instance foreign languages and mathematics. At this point, a great majority of the students said they liked the idea:

> *I was excited! The cheat sheet is familiar to me from the past and I like it a lot. Then you study properly for the test and you also feel more secure when you go to the test. If you get a black-out, you don't need to panic as you can check it from your cheat sheet. (15F, T)*

> *Nice to have some change. A new thing for me. The test didn't stress me so much. (42F, T)*

> *A good thing, we've had one sometime last year as well. Otherwise, it'd be a terrible task to remember all the exceptions. (57M, G)*

About 20% of the students had neutral or slightly mixed feelings. Mixed feelings were caused by concern about either the difficulty of the test or the quality of learning:

I liked it, I've had one once before. I was a bit worried if the test would be really hard - on the other hand, the cheat sheet relieves anxiety. (01F, T)

Good idea as such because nowadays similar things are used in the working life, for instance. On the other hand, you should know grammar in particular by heart so a cheat sheet for a grammar test may weaken your learning. This test form is new to me. (71M, L)

Wasn't a new thing, and for me, it doesn't really matter what kind of test it'd be. (24M, G)

Initially, five students felt that the cheat-sheet test would be a bad idea:

Cheat-sheet tests belong to the junior high. They are of no use when you are preparing for the Matriculation Exam. (03F, N)

Not a new thing but the cheat-sheet test will be harder, I don't like it myself. (94M, G)

After the test I collected and analysed the cheat sheets. A total of 92 students had made a cheat sheet, nine had not. The cheat sheets were divided into three groups: there were 42 thorough cheat sheets, 44 good ones and six limited cheat sheets that seemed very hastily constructed or consisted of only some short notes. The difference between a good and a thorough cheat sheet was basically in the quality, not in the quantity of information: thorough cheat sheets seemed more processed and organised with colour codes, pictures, the student's own rules or examples, for instance.

Altogether, over half of the female students (55.7%) prepared a thorough cheat sheet compared to 20% of the male students. In other words, 34 thorough cheat sheets were made by girls, eight by boys. Making a thorough cheat sheet therefore seemed to appeal more to the girls than to the boys. Another gender difference appeared among those who prepared the cheat sheet: with boys, the higher the previous grades, the more thorough the cheat sheet, but vice versa with girls (see Table 1).

TABLE 1. Previous English course grades (scale 4–10, 10 being the highest), the cheat sheet and the gender (n=101).

Cheat sheet n all (female / male)	Limited n=6 (2/4)	Good n=44 (23/21)	Thorough n=42 (34/8)	No cheat sheet n=9 (2/7)	All n=101 (60/41)
Mean of all prior course grades: all (female / male)	8.63 (9.25 / 8.31)	8.57 (8.64 / 8.50)	8.65 (8.63 / 8.72)	9.32 (9.68 / 9.21)	8.68 (8.69 / 8.65)
Previous grade: all (female / male)	8.50 (9.00 / 8.25)	8.50 (8.65/ 8.33)	8.43 (8.38 / 8.63)	9.44 (10.00 / 9.29)	8.55 (8.56 / 8.55)

However, nine students, seven boys and two girls, did not prepare a cheat sheet for the test. With one exception, the group was quite homogenous on the basis of their prior grades: the mean of their previous grades was clearly higher than that of those who made the cheat sheet. These students seemingly trusted their skills and preferred to take the test without a cheat sheet:

> *I don't find it useful in English because my language skills are so good anyway, but it must be helpful when trying to remember small, trivial things. (34F, N)*
>
> *I wanted to see how well I can do without. (70M, N)*

As mentioned earlier, one of the pedagogical premises of the cheat-sheet test is enhanced learning through an engagement in constructing the cheat sheet (e.g. Erbe 2007; Larwin 2012; Whitworth 1990). Also, the limited size of the cheat sheet makes students select material, which necessitates both the self-assessment of their own skills and an evaluation of the relevance of the information (e.g. Whitworth 1990). These ideas of engagement and self-assessment come across clearly in the majority of students' comments when they describe what they wrote on their cheat sheets and why. Some also mentioned their learning styles as a basis for selection:

> *I wrote nearly all grammar things on the cheat sheet because I revised things that way even if I had known some of the things beforehand. (09F, T)*
>
> *Mainly I chose things that were difficult for me but also the most important. (08M, G)*
>
> *For the cheat sheet, I selected things that I thought I might forget or that I didn't master yet so that they would stay better in my memory also for the future. (16F, T)*
>
> *Some rules, but more examples. I understand things better through examples. (66F, T)*

Furthermore, a few students wanted to make sure they would be able to perform well despite their possible test anxiety:

> *I wrote all I could squeeze in because I was afraid that I might have a 'blackout' in the test situation. (39F, T)*

There were, however, some students who did not base their selection on self-assessment but rather more on a presumption of what might be asked in the test. Some students also wanted to take full advantage of the cheat sheet:

> *I listed all the things that were on the list of the test topics because they will probably be asked. (33M, T)*

> *Almost all the information I could find. If you are allowed to use a cheat sheet, make the most of it then. (60F, T)*

In sum, most students liked the idea of a cheat sheet. They had also engaged in preparing their cheat sheets, which, in general, were of good quality: 42 of them were considered thorough, 44 good and six limited.

4.2 The effects of the cheat sheet on learning experiences and results

Immediately after the test, the students were asked if the cheat sheet had been useful or beneficial in the test. Although feeling almost unanimously that the cheat sheet had been helpful in one way or another, the students did not believe its impact on their actual test results would be strong, perhaps a couple of points on average. Overall, the students felt that the cheat sheet had rather helped them to learn better than offered them the right answers in the test, as the following comment shows:

> *I believe the cheat sheet improved my test results a bit, but not significantly. Or, actually, maybe the cheat sheet improved the result quite a lot. I noticed in the test that I knew the things I had written on my cheat sheet. So, making the cheat sheet had taught me. (82F, T)*

Over a third of the students said they had used the cheat sheet mainly for checking some of their answers. Eleven students mentioned that they had used their cheat sheets little or not at all in the test situation even though they had them:

> *I didn't actually need it more than in a couple of cases where I was wondering if there should be an article or not. (50F, G)*

> *I believe that making the cheat sheet helped my language skills. If I had realised to take it out of my rucksack, I could have checked exercise 1 from it. (06M, G)*

Conclusively demonstrating the efficacy of learning and the influence of the cheat sheet on the test results is unfeasible in a real classroom context, since the students cannot take the same test both with and without the cheat sheet. In Dickson and Bauer's (2008) study, college students first had an unexpected pre-test without their crib notes, and then the actual test with their cribs. Using the same research design would not have been possible in this experiment because of the practical time constraints of school life. Furthermore, Dickson and Bauer's (2008) assumption that the slightly better results in the real test with the crib meant dependency on the notes and thus inferior learning is, in my opinion, somewhat fallacious. There is ample evidence that students use different learning strategies when studying for different types of assessment situations (Atjonen

2007; Pickford & Brown 2006). For instance, students have been documented to use more deep learning strategies and skills that lead to better conceptual understanding as well as more self-directed study skills when studying for an open-book exam than for a closed-book exam, the latter evoking more rote learning and memorisation (Block 2012; Boniface 1985; Theophilides & Koutselini 2000). Thus, along the same lines, students who are told that they can use the cheat sheet in the test probably select things differently for the cheat sheet than they would if they knew they cannot use their notes in the actual test situation. For example, if there is something that is difficult to memorise by heart, such as a long list of exceptions, they write them down – as most students did in the present study. Also, if the students recognise their own weaknesses, they write down things they do not master – as, again, many students did in this study. Thus, using the cheat sheet very sensibly for their own needs, they can concentrate on learning and understanding more important concepts. If no cheat sheets are then allowed in the test situation, it handicaps students who have prepared well and rationally – but for a different kind of purpose and test situation.

However, in order to investigate if the cheat sheet had had any measurable effect on test results, I first compared the quality of the cheat sheet and the grammar results (see Table 2). I used the comparison of means and Pearson's correlation coefficient to analyse the test results; t-test, one-way analysis of variance as well as analysis of covariance were used to analyse the statistical significance of the differences of the means (see e.g. Jokivuori & Hietala 2007; Metsämuuronen 2009). On average, the students with thorough cheat sheets scored the highest in the grammar exercise and the six students with limited cheat sheets scored the lowest ($p<.01$, one-way ANOVA).

TABLE 2. The means of the test results, earlier English grades and the quality of the cheat sheet (n=101).

	Limited (n=6)	**Good (n=44)**	**Thorough (n=42)**	**No cheat sheet (n=9)**	**All (n=101)**
Grammar score A (max. 40p)	29.33	31.39	33.93	33.67	32.52
Reading comp. score B (max. 40p)	31.92	32.50	32.62	36.89	32.91
A-B: the difference	-2.58	-1.11	+1.31	-3.22	-0.38
Total score (max. 80p)	61.25	63.89	66.55	70.56	65.43
Test grade	7.71	8.03	8.36	8.89	8.22
Previous course grade	8.50	8.50	8.43	9.44	8.55
Mean of previous course grades	8.63	8.57	8.65	9.32	8.68

Next, I compared the grammar results with the reading comprehension scores. The underlying assumption was that the reading comprehension results would give an idea of the student's overall language skills and could thus be used as a baseline for the cheat-sheet part of the test. As hypothesised, the students' reading comprehension scores turned out to be in line with the average of their earlier English grades ($r = .83$, $p<.001$). As could be expected on the basis of their previous grades, the students without a cheat sheet scored the highest of all in the RC part (36.89/40p.), but on average 3.22 points less in the grammar section. The students with thorough cheat sheets were the only group that on average had a higher score in grammar than in reading comprehension.

The mean of the previous grades also correlates strongly with the grammar score ($r=.71$, $p<.001$) but the correlation is smaller than with the reading comprehension score. To investigate the effect of the cheat sheet (both its existence and quality) on the grammar score, an analysis of co-variance was run (see e.g. Jokivuori & Hietala 2007). When the effect of the mean of the previous grades was removed, the difference in grammar scores was still statistically significant ($p<.001$, ANCOVA).

With most girls having prepared a thorough cheat sheet, I wanted to see if there were any statistical gender differences (t-test) in the results. As the scatter graph illustrates at the individual student level (see Figure 1), female students tended to score slightly higher in grammar ($m=33.57$) than in RC ($m=32.61$). Their grammar scores were also higher than those of the male students ($m=30.93$, $p<.01$), who, in turn, scored a little higher in RC ($m=33.35$, $p=ns.$).

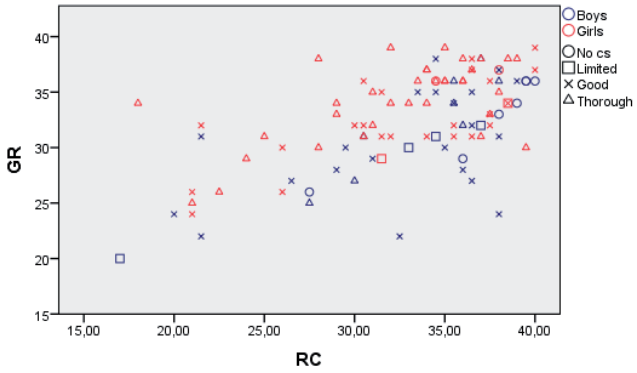

FIGURE 1. Scatter chart of grammar (Y axis) and reading comprehension scores (X axis) of female and male students, displaying the quality of the cheat sheet.

Female students' higher grammar scores further corroborate the effect of the quality of the cheat sheet on grammar results since females also prepared more thorough cheat

sheets than males. All in all, female students did slightly better in this test even though the means of the previous grades of both female and male students were almost identical (see Table).

4.3 The students' final verdict: was it worth it?

When I handed the tests and cheat sheets back, I asked the students for final feedback on the cheat-sheet test: "Now that you have seen your marked and graded test, what do you think of the cheat-sheet test, the cheat sheet itself and its effects? Was it worth it or not? Why? Would you do something differently now?" Out of the 92 students who had prepared a cheat sheet, six said the impact of the cheat sheet had been non-existent or negative, mainly because of the difficulty of the test, lack of time or lack of preparation:

> *Well, now it seems that it wasn't that useful. Of course, I could have made a bit better cheat sheet but I forgot and made it during the previous break/class so it didn't give me an awful lot of benefit. (40M, L)*

> *The test was really difficult and there was too little time. I felt that the cheat sheet didn't help although I was well prepared otherwise, too. If the test was more difficult because of the cheat sheet, it would have been better to have the test without it. (21F, T)*

Two of those six students also considered the cheat sheet detrimental for learning:

> *I would rather have done an ordinary test even though the grade would probably have been pretty similar. The cheat sheet was helpful in some things because I didn't cram at all, actually. A nice experiment but without a cheat sheet you would learn better. (79F, T)*

Still, a great majority of students felt that the cheat sheet had been beneficial because it had enhanced their studying and learning as well as recollection, for instance. Some also said that it had made them prepare for the test better than usually. Fourteen students, all girls, mentioned feeling less insecure or stressed because of the cheat sheet.

> *A very good thing. I noticed last week that I already knew by heart some pretty difficult phrasal verbs that I had written on my cheat sheet. In the test itself I didn't need the cheat sheet that much but still it was useful in general, e.g. phrasals got into my mind. (22M, G)*

> *The cheat sheet was quite handy because it forces you, in a way, to read and study properly grammar rules that you wouldn't normally bother to study so carefully. (Which probably was the whole point.) Things got into my head. I wouldn't do anything differently. (77F, T)*

> *Yes it was useful. Things went into several boxes in the brain so to speak when you read and wrote at the same time. (47M, T)*

It was worth it. Precisely because this test wasn't so stressful as well. (25F, G)

As can be seen from the comments above, preparing the cheat sheet had generally engaged students in the studying and learning process. Nevertheless, would students do something differently now that they had seen the results of their tests? Those 31 students who answered this question mentioned a few changes, for instance investing more effort in the construction of the cheat sheet or making the cheat sheet more condense. Five students mentioned changes that suggest some prior dependency on the cheat sheet: they would either study more or trust themselves more instead of trusting the cheat sheet alone

I wouldn't do anything differently except I would trust my gut feeling more than the cheat sheet. (38F, T)

If I did something differently, then I'd study more and not just trust the cheat sheet. (31F, G)

A few students were also disappointed when they noticed that they had made mistakes in the grammar section despite having had the correct information on their cheat sheets. Had they perhaps not had time to check, or had they not found the information on the cheat sheet – or had they had a false sense of knowing it by heart?

I'd concentrate better in the test because some mistakes were stupid, careless errors. (87F, G)

How about the students who had not made a cheat sheet at all? Would they construct one if given a second chance? Some perhaps would, some would not:

I don't regret my decision. My mistakes were in such small things that I wouldn't have written them on my cheat sheet anyway. (34F, N)

In some small things (like in the article exercise) I would have liked to have had it. (03F, N)

All in all, most students considered the cheat sheet helpful for the learning and studying process as well as for the test situation.

5 Discussion

The aims of the reported experiment were to find out how students react to cheat-sheet tests, what kinds of cheat sheets they construct and why. I also wanted to see if cheat sheets affect students' learning experiences and results. Above all, I wanted to

explore if cheat-sheet tests could give students a more engaged or empowered role in assessment. The qualitative findings showed that a significant majority of the students liked the idea from start to finish. Accordingly, 92 out of 101 students prepared cheat sheets, which in general were of good quality – 44 of them were regarded as good, 42 as thorough – so students clearly invested thought and effort in preparing them. Furthermore, quantitative analysis indicated that a thorough cheat sheet improved their test results a little.

This experiment has some limitations, though. First of all, this was primarily a teaching experiment in order to develop assessment methodology in my own teaching rather than a pure research experiment. The design was therefore not as rigorous as it could have been. Also, the number of students is rather limited for statistical analyses, and as they were not a random sample, the results cannot be generalised. Furthermore, my role as both the teacher and the researcher may have affected some of my decisions, both in teaching and in conducting the study. Some external factors such as poor handwriting may have in some subliminal way influenced the categorisation of the cheat sheets as well.

In addition, we can naturally argue whether the grammar results and those of the rest of the test are comparable and thus, whether they are feasible indicators of either improved test results or improved learning. We can also say that the differences were so small that they are not really significant. Furthermore, we can argue whether the small test result improvements were because of the students' dependency on the cheat sheet in the test situation or because of their engagement in learning while constructing the cheat sheet. However, the evidence clearly shows that female students, who seemed to have invested more in making their cheat sheets, performed better in the grammar exercise.

The students' personal experiences are perhaps the most pertinent issue here. First of all, the majority of the students experienced the cheat sheet as helpful for their learning process. Many students said that making the cheat sheet had improved their learning. Some also said that preparing the cheat sheet made them study better, in a more engaged way. Secondly, most students felt that although the cheat sheet did not increase their test scores much, it helped them in the actual test situation in one way or another: a few mentioned that the cheat sheet decreased their test anxiety and stress, and some said they could check their answers with the help of the cheat sheet. Finally, most of the students liked the cheat sheet, and quite a few would like to use it more often. Although there were a couple of students who may have depended on the cheat sheet, I agree with several researchers who claim that the use of cheat sheets can enhance both performance and learning through increased engagement (e.g. Block 2012; Erbe 2007; Larwin et al. 2013; Whitworth 1990). Even if test results had not improved, I would still

recommend cheat sheets because most students found them helpful. Unlike Dickson and her colleagues (Dickson & Bauer 2008; Dickson & Miller 2005; Funk & Dickson 2011), I believe that learning may manifest itself in many guises, not only in improved test results, and that students' own experiences and reactions are paramount. Positive attitudes, motivation as well as reduced anxiety are key components in learning. They simply cannot be ignored.

Furthermore, the cheat-sheet test empowered students in a very concrete way. First, the students could each decide whether or not to prepare the cheat sheet. Then, they could decide what to include and how. Constructing the cheat sheet developed their self-assessment as well as their learning-to-learn skills, and it introduced a new study method for those who had never written revision sheets. By reducing some students' test anxiety, the cheat sheet allowed them to focus on both studying and taking the test without excessive, disruptive stress. Finally, the students could decide when to use the cheat sheet in the test – a few decided not to consult it at all. So, the students had several opportunities to act as active agents – none of which a closed-book exam allows.

As shown in previous studies, the effect of the cheat sheet on the actual grade in general remained quite small: as Gharib et al. (2012) concluded, students who usually do well tend to do well regardless of the exam type. Thus, students and teachers who worry that cheat sheets might result in everybody getting (too) good grades, regardless of whether they really deserve them or not on the basis of their skills, need not worry. And the two or three disappointed students who had hoped that the cheat sheet would give them an easy escape route in lieu of studying for the test may have learnt more learner responsibility through their disappointment.

In a nutshell, I would argue that, in spite of its clear benefits, a cheat-sheet test is not a panacea, and it should not be used as the only assessment method. Assessment must be versatile and diverse enough to tap into the diverse skills of all students. Thus, further research in FL student assessment, whether dealing with cheat sheets or not, is needed. Yet, the findings of this limited study suggest that a cheat-sheet test is *one* learner-friendly assessment method that most students find beneficial for learning. It also engages and empowers them far more than traditional closed-book tests do. Quite justly, then, I will give the final word to a student whose comment summarises the findings of this study very well:

> It was nice to try this. I wouldn't like to have a cheat sheet test every time but it's good every now and then. There was so much stuff that's difficult to learn by heart in a minute. Making the cheat sheet made me study the test area better than I would have studied for an ordinary test. (13F, T)

References

Atjonen, P. 2007. *Hyvä, paha arviointi*. Helsinki: Tammi.

Aydın, S. 2009. Test anxiety among foreign language learners: a review of literature. *Journal of Language and Linguistic Studies, 5* (1), 127–137.

Block, R. M. 2012. A discussion of the effect of open-book and closed-book exams on student achievement in an introductory statistics course. *PRIMUS, 22* (3), 228–238.

Boniface, D. 1985. Candidates' use of notes and textbooks during an open-book examination. *Educational Research, 27* (3), 201–209.

Butler, D. & N. Crouch 2011. Student experiences of making and using cheat sheets in mathematical exams. In J. Clark, B. Kissane, J. Mousley, T. Spencer & S. Thornton (eds) *Mathematics: traditions and [new] practices: proceedings of the 34th annual AAMT–MERGA conference.* Adelaide: AAMT and MERGA, 134–141.

Cheng, L., D. Klinger, J. Fox, C. Doe, Y. Jin & J. Wu 2014. Motivation and test anxiety in test performance across three testing contexts: the CAEL, CET, and GEPT. *TESOL Quarterly, 48* (2), 300–330.

Cheng, L., Y. J. Watanabe & A. Curtis 2004. *Washback in language testing: research contexts and methods.* Mahwah, NJ: Lawrence Erlbaum.

de Raadt, M. 2012. Student created cheat-sheets in examinations: impact on student outcomes. In *Proceedings of the Fourteenth Australasian Computing Education Conference (ACE2012)*, Melbourne: Australian Computer Society Inc., 71–76.

Dickson, K. L. & J. J. Bauer 2008. Do students learn course material during crib sheet construction? *Teaching of Psychology, 35* (2), 117–120.

Dickson, K. L. & M. D. Miller 2005. Authorized crib cards do not improve exam performance. *Teaching of Psychology, 32* (4), 230–233.

Elo, S., M. Kääriäinen, O. Kanste, T. Pölkki, K. Utriainen & H. Kyngäs 2014. Qualitative content analysis: a focus on trustworthiness. *SAGE Open, 4* (1). Available at DOI: 10.1177/2158244014522633.

Erbe, B. 2007. Reducing test anxiety while increasing learning: the cheat sheet. *College Teaching, 55* (3), 96–98.

Funk, S. C. & K. L. Dickson 2011. Crib card use during tests: helpful or a crutch? *Teaching of Psychology, 38* (2), 114–117.

Gharib, A., W. Phillips & N. Mathew 2012. Cheat sheet or open-book? A comparison of the effects of exam types on performance, retention, and anxiety. *Psychology Research, 2* (8), 469–478.

Harlen, W. 2012. The role of assessment in developing motivation for learning. In J. Gardner (ed.) *Assessment and learning*. London: Sage, 171–183.

Harlen, W. & R. Deakin Crick 2003. Testing and motivation for learning. *Assessment in Education, 10* (2), 169–208.

Hembree, R. 1988. Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58* (1), 47–77.

Hildén, R. & M. Härmälä 2015. *Hyvästä paremmaksi: kehittämisideoita kielten oppimistulosten arviointien osoittamiin haasteisiin*. Koulutuksen seurantaraportit 6. Helsinki: Kansallinen koulutuksen arviointikeskus. Available at http://karvi.fi/app/uploads/2015/03/KARVI_0615.pdf.

Horwitz, E. K. 2001. Language anxiety and achievement. *Annual Report of Applied Linguistics, 21*, 112–126.

Horwitz, E. K. 2010. Foreign and second language anxiety. *Language Teaching, 43* (2), 154–167.

Hughes, A. 1989. *Testing for language teachers*. Cambridge: Cambridge University Press.

Jokivuori, P. & R. Hietala 2007. *Määrällisiä tarinoita: monimuuttujamenetelmien käyttö ja tulkinta.* Porvoo: WSOY.

Larwin, K. 2012. Student prepared testing aids: a low-tech method of encouraging student engagement. *Journal of Instructional Psychology, 39* (2), 105–111.

Larwin, K. H., J. Gorman & D. A. Larwin, 2013. Assessing the impact of testing aids on post-secondary student performance: a meta-analytic investigation. *Educational Psychology Review, 25* (3), 429–443.

Liu, M. & W. Huang, 2011. An exploration of foreign language anxiety and English learning motivation. *Education Research International*. Available at DOI: 10.1155/2011/493167

Metsämuuronen, J. 2009. *Tutkimuksen tekemisen perusteet ihmistieteissä: tutkijalaitos.* 4. laitos. Helsinki: International Methelp.

Pickford, R. & S. Brown 2006. *Assessing skills and practice*. Abingdon: Routledge.

Rea-Dickins, P. & C. Scott, 2007. Washback from language tests on teaching, learning and policy: evidence from diverse settings. *Assessment in Education: Principles, Policy & Practice, 14* (1), 1–7.

Tarnanen, M. & A. Huhta 2011. Foreign language assessment and feedback practices in Finland. In D. Tsagari & I. Csépes (eds) *Classroom-based language assessment. Language testing and evaluation, Vol. 25*. Frankfurt am Main: Peter Lang, 129–146.

Theophilides, C. & M. Koutselini 2000. Study behavior in the closed-book and the open-book examination: a comparative analysis. *Educational Research and Evaluation, 6* (4), 379–393.

Tuomi, J. & A. Sarajärvi 2009. *Laadullinen tutkimus ja sisällönanalyysi*. 6., uud. laitos. Helsinki: Tammi.

Tynjälä, P. 1999. *Oppiminen tiedon rakentamisena: konstruktivistisen oppimiskäsityksen perusteita*. Helsinki: Kirjayhtymä

Volante, L. 2004. Teaching to the test: what every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy, 35*. Available at http://www.umanitoba.ca/publications/cjeap/articles/volante.html.

Whitworth, R. 1990. Using crib notes as a learning device. The Clearing House: *A Journal of Educational Strategies, Issues and Ideas, 64* (1), 23–24.

Wright, J. von. 1993. *Oppimiskäsitysten historiaa ja pedagogisia seurauksia*. Helsinki: Opetushallitus.

APPENDIX.

Nimi/Name:

Kun opettaja ehdotti lunttilappukoetta, mitä mieltä olit siitä ajatuksena? Miksi? Oliko se sinulle uusi asia?
When the teacher suggested a cheat-sheet test to your group, what did you think of the idea? Why?
Was it a new idea to you?

(This question was answered after the initial discussion on a cheat-sheet test in class.)

****

Mitä asioita kirjoitit lunttilappuusi? Miksi valitsit juuri ne asiat? *Jos ET tehnyt lunttia, kerro mikset.*
What did you write on your cheat sheet? Why did you choose these things? *If you did NOT make a cheat sheet, tell me why.*

(This question as well as the following two questions were answered immediately after the test.)

Oliko luntista hyötyä kokeessa? Millaisissa asioissa/tilanteissa? Miksi (ei)? *(Millaisissa tilanteissa olisit kaivannut lunttia?)*
Was the cheat sheet useful in the test? In what kinds of things/situations? Why (not)? *(In what kinds of situations would you have liked to have had a cheat sheet?)*

Miten uskot luntin vaikuttaneen koetulokseesi? *(Tai sen, ettei sinulla ollut lunttia)*
How do you think the cheat sheet affected your test result? (*Or, how did you not having a cheat sheet affect your test result?)*

****

Nyt kun olet nähnyt kokeesi korjattuna, mitä mieltä olet lunttilappukokeesta, luntista ja sen annista? Kannattiko lunttilappukoe vai ei? Miksi? Tekisitkö nyt jotain toisin?
Now that you have seen your marked and graded test, what do you think of the cheat-sheet test, the cheat sheet itself and its effects? Was it worth it or not? Why? Would you do something differently now?

(This question was answered after the marked tests and the cheat sheets were returned to the students.)

V

HOW TO MAKE CORRECTIVE FEEDBACK MORE LEARNER-CENTRED? A FEEDBACK EXPERIMENT IN UPPER SECONDARY EFL STUDIES IN FINLAND

by

Pollari, Pirjo (submitted for review to *Innovation in Language Learning and Teaching*)

VI

MONOGRAPH

"THIS IS MY PORTFOLIO" PORTFOLIOS IN UPPER SECONDARY SCHOOL ENGLISH STUDIES

by

Pollari, Pirjo (2000)
Jyväskylä: Institute for Educational Research

The monograph is available at

http://urn.fi/URN:ISBN:978-951-39-6898-4