JYVÄSKYLÄ STUDIES IN HUMANITIES

284

Dmitri Leontjev

ICAnDoiT: The Impact of Computerised Adaptive Corrective Feedback on L2 English Learners

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN HUMANITIES 284

Dmitri Leontjev

ICAnDoiT: The Impact of Computerised Adaptive Corrective Feedback on L2 English Learners

Esitetään Jyväskylän yliopiston humanistisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Historica-rakennuksen salissa H320
huhtikuun 23. päivänä 2016 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities of the University of Jyväskylä,
in building Historica, auditorium H320, on April 23, 2016 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2016

ICAnDoiT: The Impact of Computerised Adaptive Corrective Feedback on L2 English Learners

JYVÄSKYLÄ STUDIES IN HUMANITIES 284

Dmitri Leontjev

ICAnDoiT: The Impact of Computerised Adaptive Corrective Feedback on L2 English Learners

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2016

Editors
Ari Huhta
Centre for Applied Language Studies, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-6586-0
ISBN 978-951-39-6586-0 (PDF)
ISSN 1459-4331

ISBN 978-951-39-6585-3 (nid.)
ISSN 1459-4323

Jyväskylä University Printing House, Jyväskylä 2016

ABSTRACT

Leontjev, Dmitri
ICAnDoiT: The Impact of Computerised Adaptive Corrective Feedback on L2 English Learners
Jyväskylä: University of Jyväskylä, 2016, 152 p.
(Jyväskylä Studies in Humanities
ISSN 1459-4323; 284 (nid.) ISSN 1459-4331; 284 (PDF))
ISBN 978-951-39-6585-3 (nid.)
ISBN 978-951-39-6586-0 (PDF)

The present dissertation examines the impact of (computerised) adaptive corrective feedback, that is, feedback dynamically adapting to learners' abilities, in English as a second/foreign language (L2) and explores the ways to maximise this impact. The study was inspired by the sociocultural perspective on development, which had implications for the interpretation of the results, including those obtained through statistical data analyses.

The dissertation comprises five articles and a synthesis. In the first article, a positive effect of adaptive corrective feedback on the learners' ability to formulate L2 English wh-questions is established. The second article explores how learners' beliefs about corrective feedback guide their performance on dynamic assessment and reflection on it, and how reflection on experience with dynamic assessment mediated in social interaction helps to transform these same beliefs. The results suggest that due to their beliefs, some of the participants skipped the feedback they believed to be useless, but also illustrates how the learners, whose utterances were mediated by the interviewer, other learners in the interview, their teacher's voice and feedback practices, and experience of dynamic assessment, began appropriating beliefs about corrective feedback that were jointly constructed by the participants in the interaction. Studies reported on in articles three and four aim at addressing the issue of lack of research on L2 English word derivational knowledge (to an extent), the latter being the assessment target in the study reported on in article five. This final article reports on a case study that builds upon the previous studies conducted as a part of my doctoral research project and studies whether generalisations made based on the other two studies add to the usefulness of adaptive corrective feedback in dynamic assessment of L2 word derivational knowledge. The available evidence for the validity of the computerised system and the dynamic test of learners' ability to formulate wh-questions with auxiliaries is presented in a separate chapter of the synthesis.

The theoretical importance of the study lies above all in that it presents quantitative evidence for the beneficial role of corrective feedback provided within learners' Zone of Proximal Development. The findings also suggest that learners' performance during computerised dynamic assessment is mediated not only by the adaptive corrective feedback per se but also by their beliefs about corrective feedback and expectations of what a test should look like, their beliefs being, thus a potential threat to validity of computerised dynamic tests but also suggests a way these can be accounted for. As regards practical implications, the findings suggest that the assessment/tutoring system created in the course of the study, or a similar one, using a similar approach to mediating learners' performance, can be used in the classroom.

Keywords: dynamic assessment, corrective feedback, sociocultural theory, beliefs, English as a second/foreign language

Author's address

Dmitri Leontjev
Centre for Applied Language Studies
University of Jyväskylä
P.O. Box 35, 40014 University of Jyväskylä
dmitri.leontjev@jyu.fi

Supervisors

Ari Huhta
Centre for Applied Language Studies
University of Jyväskylä

Riikka Alanen
Department of Teacher Education
University of Jyväskylä

Reviewers

Prof. Claudia Harsch
Faculty 10
University of Bremen

Assoc. Prof. Matthew E. Poehner
Center for Language Acquisition
Pennsylvania State University

Opponents

Assoc. Prof. Matthew E. Poehner
Center for Language Acquisition
Pennsylvania State University

ACKNOWLEDGEMENTS

What a wonderful journey it was—long, often tough and tiring, but wonderful nonetheless! This work would not have been possible, though, without the support and encouragement from a number of people along the way that ultimately brought me here, writing these words.

My deepest gratitude goes to my supervisors, Ari Huhta and Riikka Alanen. It was a complete privilege to write this dissertation under their guidance. The combination of Riikka's and Ari's expertise was the best I could possibly think of for the kind of research that I conducted. They encouraged me when I felt desperate and gently directed me along the way. It is because of their guidance that I learned what mediation truly means. I cannot thank them enough. Thank you!

I am grateful to my pre-examiners, Professor Claudia Harsch and Associate Professor Matthew Poehner, whose constructive comments on the manuscript helped me to improve its quality. Specifically, I would like to thank Claudia Harsch for meticulously marking all the places where the manuscript could be improved and suggesting what could be done to improve it. I thank Matthew Poehner for his discussion, suggestions, and for helping me to better understand the results of my work. All the remaining shortcomings in the dissertation are, certainly, my own. I would like to add that Matthew Poehner is one of the people whose work inspired me throughout the course of my doctoral studies (and still does).

I would like to thank all the people (previously and currently) working at the Centre for Applied Language Studies (CALS/Solki)—they have all created a seemingly relaxed but exceptionally productive atmosphere where one can always share one's ideas and receive feedback in return. During my years in Solki, I always felt their support. I thank Helena Mackay, who proofread several of my manuscripts. I miss our discussions in Solki coffee room. I am grateful to Riikka Ullakonoja for our *chto-novogo* discussions, where we shared what we had been doing and exchanged feedback. I hope Saku is doing great, too. Thank you, Tatjana Rynkänen, for helping me to make sense of the mononuclear sentences in my interview data. Thank you, Marita Härmälä and Teija Kangasvieri, for translating the summary of my dissertation into Finnish. I could have ended up Google-translating it if it was not for your help. I am thankful to Tarja Nikula-Jäntti, the head of the Centre for Applied Language Studies, for her support, including providing the excellent working conditions and several short-term employment periods at CALS—these really helped me a lot. Lyhyesti sanottuna: kiitos teille, solkilaiset! This was not Google-translated, by the way.

Doctoral studies are not easy to pursue without financial support. I would like to thank the Centre for Applied Language Studies and the Faculty of Humanities of the University of Jyväskylä for funding my studies and the development of the ICAnDoiT system.

I thank WSOYpro Oy for granting me access to the online exercises designed to accompany the study materials they publish. This helped me to select the task types used in the ICAnDoiT system.

I am deeply grateful to the learners and the teachers who kindly agreed to participate in the studies. I am especially thankful to M, who worked through all those exercises I made him do and patiently sat through the long interviews with me.

I want to thank my parents who taught me to believe in myself and to try again if I don't succeed at first. Much of my faith in myself comes from my late grandmother, who never doubted a single moment that I can do it—I miss her every day.

Above all, however, I owe it to my wife, Anna, without whose love and support I would not have completed this work. And without her patience, too. I kind of think that she was more used to seeing the top of my head, as I was sitting there in front of my notebook, than to seeing my face. Not to mention that she had to put up with all those papers piling up on the bookshelves, in the drawers, and in the corners of the room. I really have to organise them one day.

Jyväskylä, March 22, 2016
Dmitri Leontjev

FIGURE

FIGURE 1 Ringbom's (1987) model of lexical knowledge. 47
FIGURE 2 Timeline of the studies. 54
FIGURE 3 Bachman and Palmer's (1996) perspective on test validity. 68
FIGURE 4 Means plots of learners' performance on the treatment tasks (from an analysis of covariance). 76
FIGURE 5 Sample instructions to test designers. 88

TABLE

TABLE 1 Studies comparing/contrasting recasts with other oral CF types. 22
TABLE 2 Written CF in different typologies. 23
TABLE 3 Studies comparing/contrasting direct and indirect written corrective feedback types. 23
TABLE 4 Stages in the development of L2 English questions (adapted from Pienemann, 2005; Spada & Lightbown, 1999). 45
TABLE 5 Teaching order of L2 English affixes (Bauer & Nation, 1993). 48
TABLE 6 Research questions in relation to the original publications. 53
TABLE 7 Validity facets (Messick 1989: 20) 67
TABLE 8 Aspects of usefulness of the LA2 English *Questions Test* in the *ICAnDoiT* system 72
TABLE 9 LA1's pretest and posttest performance on task 1 (E-mail writing using prompts). 83
TABLE 10 A sample from a learner profile (Pilot study). 85
TABLE 11 A Sample from a learner profile (Intervention study). 85

CONTENTS

ABSTRACT
ACKNOWLEDGEMENTS
FIGURES AND TABLES
CONTENTS
LIST OF ORIGINAL PUBLICATIONS

1 INTRODUCTION 13
1.1 Aims of the study 13
1.2 Methodological framework 16
1.3 Structure of the dissertation 17
1.4 Summary of the research questions 17

2 CORRECTIVE FEEDBACK IN SECOND LANGUAGE ACQUISITION RESEARCH 18
2.1 Early views on feedback 18
2.2 Defining corrective feedback 19
2.2.1 Corrective feedback typologies 19
2.3 Efficacy of different CF types 21
2.4 Does corrective feedback help at all? 24
2.5 Corrective feedback from a sociocultural perspective 25
2.6 Levels of Feedback 26
2.7 Sociocultural perspective and Hattie and Timperley's feedback levels 28
2.8 Learners' beliefs and corrective feedback 30
2.9 Drawing the threads together 32

3 DYNAMIC, DIAGNOSTIC, AND DYNAMIC DIAGNOSTIC ASSESSMENT OF L2 34
3.1 Assessment, testing, measurement, evaluation 34
3.2 Assessment of learning and pro-learning assessment 35
3.2.1 Approaches to dynamic assessment 36
3.3 Diagnostic and dynamic diagnostic assessment 39
3.3.1 What is diagnostic assessment? 39
3.3.2 Approaches to L2 diagnosis 41
3.4 Dynamic diagnostic assessment and the assessed construct 44
3.4.1 L2 English questions 45
3.4.2 L2 English word derivation 46
3.5 Dynamic diagnostic assessment and feedback 49

4 METHODOLOGY 51
4.1 Aims and Research questions 52
4.2 Research questions in relation to the original publications 53
4.3 Method 55

4.3.1 Summary of Article I ... 57
4.3.2 Summary of Article II ... 58
4.3.3 Summary of Article III ... 60
4.3.4 Summary of Article IV ... 61
4.3.5 Summary of Article V ... 62
4.4 Ethical considerations ... 64

5 VALIDATION OF THE *ICANDOIT* SYSTEM AND THE *QUESTIONS TEST* ... 66
5.1 Validity and test validation frameworks ... 66
5.1.1 Messick's validation framework ... 67
5.1.2 Bachman and Palmer's aspects of test usefulness ... 67
5.1.3 A priori and a posteriori validation ... 71
5.2 Aspects of usefulness of the *ICAnDoiT* system and the *Questions Test* ... 71
5.2.1 Reliability ... 73
5.2.2 Construct validity ... 76
5.2.3 Authenticity ... 78
5.2.4 Interactiveness ... 80
5.2.5 Impact ... 84
5.2.6 Practicality ... 86
5.2.7 Usability ... 87
5.3 Final considerations ... 88

6 RESULTS ... 90
6.1 Research question 1 ... 90
6.2 Research question 2 ... 92
6.2.1 Research question 2.1 ... 92
6.2.2 Research question 2.2 ... 93
6.3 Research question 3 ... 95
6.3.1 Research question 3.1 ... 95
6.3.2 Research question 3.2 ... 96
6.3.3 Research question 3.3 ... 98

7 DISCUSSION AND CONCLUSION ... 100
7.1 Discussion ... 100
7.2 Implications ... 108
7.3 Limitations and some future directions ... 110
7.4 Conclusion ... 113

YHTEENVETO ... 114

REFERENCES ... 116

APPENDIX 1: TEST SPECIFICATIONS ... 130

LIST OF ORIGINAL PUBLICATIONS

I Leontjev, D. (2014). The effect of automated adaptive corrective feedback: L2 English questions. *APPLES: Journal of applied language studies, 8*(2), 43–66. Retrieved from http://apples.jyu.fi/ArticleFile/download/459.

II Leontjev, D. (2016). Exploring and reshaping learners' beliefs about the usefulness of corrective feedback: A sociocultural perspective. *ITL International Journal of Applied Linguistics*, *167*(1), in press.

III Leontjev, D. (2016). L2 English derivational knowledge: Which affixes are learners more likely to recognise? *Studies in Second Language Learning and Teaching*, *6*(2), in press.

IV Leontjev, D., Huhta, A., & Mäntylä, K. (forthcoming). Word derivational knowledge and writing proficiency: How do they link? *System*. doi: 10.1016/j.system.2016.03.013

V Leontjev, D. (2016). Dynamic assessment of word derivational knowledge: Tracing the development of a learner. *Eesti Rakenduslingvistika Ühingu aastaraamat [Estonian Papers in Applied Linguistics]*, *12*, 141–160. doi: 10.5128/ERYa12.09

1 INTRODUCTION

Corrective feedback (CF) has a long history in education. This history is also controversial. So far, the research on the amount and detail of corrective feedback that is beneficial for the acquisition of a second or a foreign language (L2) has not arrived at a definitive conclusion. In fact, whether corrective feedback has any use in the L2 classroom has been challenged, too.

Instead of trying to contribute to the research on corrective feedback in instructed second language acquisition by going the same route that most of the previous research had gone, that is, studying whether and how different types of CF are more effective for learning an L2, I decided to consider the problem from a different perspective, getting inspiration from the sociocultural theory of learning (e.g., Vygotsky, 1978; **Chapter 2.5**). The premise of the present doctoral research is that it is not the type and amount of corrective feedback per se that is effective or otherwise useful, but CF with reference to each learner's reciprocity to it, the latter shaped by learners' abilities and beliefs about CF, but also the context in which feedback is provided.

The present synthesis summarises five articles that together with the synthesis itself comprise my doctoral dissertation. It should be added, that the synthesis also includes some data and analyses pertaining to the process of validation of the Web-based assessment/tutoring system and a computerised dynamic test that were not reported in any of the articles.

1.1 Aims of the study

The study was simultaneously a learning experience for me, as in the course of it, I designed and contributed to the implementation of the computerised assessment/tutoring system allowing for providing adaptive corrective feedback and designed two dynamic tests delivered via this system. The system received the name *ICAnDoiT* (Interactive Computer-Adaptive Diagnostic and Tutoring system). The name reflects the idea that considerably more learners can improve their language skills when the assistance they receive is tailored to their abilities. The system is currently hosted at https://solki4.cals.jyu.fi/icandoit.

The overall aim of the dissertation is to study the impact of adaptive corrective feedback, that is, feedback provided within learners' Zone of Proximal Development (see **Chapter 2.5**). In **Chapters 2** and **3** of the present synthesis, I will identify several research gaps that can weaken the argument for the benefits of adaptive corrective feedback, but will also briefly mention them in the following. To start with, being focused on the process of interaction rather than the product of it, research studying corrective feedback from this perspective has been largely qualitative/descriptive whereas research on static corrective feedback has been predominantly quantitative. I suggested that computerised dynamic assessment (see **Chapter 3**) can be a suitable research tool for accumulating quantitative (in addition to qualitative) findings regarding the role of adaptive CF, thus strengthening the claim for its beneficial role in promoting learners' L2 abilities and improving the comparability of the findings with those on static corrective feedback.

This formed the aim of Article I, in which I studied the effect of automated adaptive corrective feedback provided during a computerised dynamic test of learners' ability to form L2 English wh-questions with auxiliaries and compared this effect with that of knowledge of results feedback. I hypothesised that the effect of the former should be stronger than of the latter.

The second aim of the present doctoral research project rooted in the problem of not being able to identify learners' responsiveness to adaptive corrective feedback in computerised dynamic tests apart from being able to trace when learners skip feedback. Drawing on contextual approaches to studying beliefs (**Chapter 2.8**), I suggested that learners' beliefs about the usefulness of corrective feedback can mediate their decision to accept or reject the feedback they are provided with during dynamic assessment even when this feedback is provided within their ZPD. At the same time, I suggested, experience of dynamic assessment has a potential to transform these beliefs.

Finding out whether learners' beliefs (e.g., about corrective feedback) mediated their DA performance and whether DA experience, in its turn, guided the development of their beliefs was the aim of Article II, in which I reported on two studies. The first was a case study of one learner of English, M, reporting on his beliefs about corrective feedback before and after his experience of a human-mediated dynamic test of word derivation knowledge. What makes the study interesting is that the design of the feedback in the dynamic test of word derivation was in part informed by the design decisions made in the study reported on in Article I (e.g., the feedback types in the studies were similar). In the second study, the participants were selected among the participants in the study reported on in Article I. This allowed me to (a) select the participants of different abilities for the interviews, which were the major research tool used in the study, and (b) explain the findings of Article I with reference to the learners' beliefs identified in the study reported on in Article II.

Articles III and IV address the problem of operationalisation of assessment targets (constructs) in computerised dynamic assessment when the assessed constructs are little-researched. Specifically, the assessment/training target of the test in Article I was based on a rather well developed theory and empirical research, whereas the word derivational knowledge (the object of the two studies), while presenting a problem to L2 learners, is severely under-researched. Article III aimed at finding empirical evidence for the theoretically grounded teaching order of L2 English derivational affixes proposed by Bauer and Nation (1993). My contribution to this joint study included involvement in designing the study (including creating some of the tasks), piloting the tasks, collecting the data in Estonia, coding, and conducting some of the data analyses. Article IV aimed at providing a deeper understanding of the aspects that L2 English word derivational knowledge contains.

The validation of the system and the computerised dynamic test of learners' ability to form L2 English wh-questions with auxiliaries (hereinafter *Questions Test*; **Chapter 5**) was not reported in detail in a separate article, as it does not explicitly relate to answering any of the research questions in the study. At the same time, I found the process of validation as outlined in Article I insufficient, as the system and the test were the main research tool used in the study. Thus, I decided to outline this process in a separate chapter of the synthesis.

Finally, the study reported on in Article V builds on the findings of the rest of the studies I conducted as a part of my doctoral research project. These findings were synthesised to design and implement a computerised dynamic test of learners' word derivational knowledge. Established in the previous studies forming the present dissertation, in the study, some general principles of (automated) adaptive corrective feedback were operationalised and the ways the proposed DA design promoted one learner's unassisted performance on tasks eliciting his L2 English word derivational knowledge. The particular focus of the study was on the ways that adaptive corrective feedback promoted strategic learning, as I assumed that to promote learners' performance on such idiosyncratic feature as L2 English word derivation, training learners in the use of separate affixes would not be sufficient to improve their performance beyond the use of those single affixes.

To summarise, the present doctoral research project aims to strengthen the argument for the beneficial impact of adaptive corrective feedback by

a) collecting experimental evidence for its facilitative effect in computerised dynamic assessment;
b) studying how learners' beliefs mediate their DA performance and how experience of dynamic assessment can be used to mediate their beliefs about CF;
c) studying empirically whether particular features of adaptive corrective feedback, such as promoting the use of certain strategies, allow for the feedback to improve learners' performance on tasks requiring them to demonstrate their knowledge of L2 English word derivation, an under-researched and an idiosyncratic feature.

1.2 Methodological framework

The study was informed by a sociocultural theory of learning, which traditionally focuses on the process of learning rather than on its product. Despite that, due to the aims set in the doctoral research project, the mixed-methods approach was employed, that is, in Article I, for example, the data and data analyses were predominantly quantitative, but to add to the interpretation of the findings, a qualitative analysis of the participants' questionnaire responses was conducted. On the other hand, in the studies reported on in Article II the data were analysed only qualitatively and in those reported on in Articles III and IV, only quantitative research methods were used. However, since the findings of separate studies were interpreted with reference to other studies (e.g., the findings reported on in Article I were also interpreted with reference to the findings reported on in Article II), the present doctoral research, considered as a whole, adopts mixed-methods design.

It should be noted that adopting the sociocultural perspective had its implications for the interpretation of the results obtained in the quantitative analyses as well. Hence, the results in Article I were interpreted bearing in mind that it was not only that adaptive CF mediated the difficulty of the items to match the learners' abilities (which the knowledge of results did only for more able learners), but also the learners' beliefs about corrective feedback mediated their performance on the tasks. On the other hand, the idea of the two articles aiming to add to the research on the construct of L2 English word derivational knowledge was to discover learners' unassisted performance (before attempting to find ways of mediating it). What is more, studying learners' performance in these two studies qualitatively to obtain a deeper understanding of their performance (e.g., studying what mediated this performance) would be impractical due to the number of participants. Thus, their results were interpreted following the traditional experimental research paradigm, and the issue was instead studied in Article V.

Some elaboration should also be given as to the role of the researcher during the interviews. Vygotsky (e.g., 1978; 1987) suggested that higher forms of human mental activities are mediated by physical and symbolic tools. That is to say, we use these tools, language being one of them, to mediate the relationship between ourselves and the world. I, being the interviewer in most of the interviews, considered my utterances (and the utterances of the interviewer in the Pilot study; see **Chapter 5**) to be such means (similarly to the learners' experiences, authoritative voices, and other learners' utterances), mediating the learners' utterances. Thus, instead of trying to minimise the interviewer's intervention, which, within the perspective adopted in the studies, would still mean that that the interviewer mediated the learners' utterances, only less explicitly, I considered the interviewer to be a participant in the interaction and studied how the mediational means (including the interviewer's utterances) guided what the learners reported.

1.3 Structure of the dissertation

The present synthesis is organised as follows. In **Chapter 2**, I will present the theoretical background on corrective feedback. **Chapter 3** gives an overview of the previous research on dynamic (and diagnostic) assessment of L2. **Chapter 4** presents the specific research questions and gives an overview of the methodology. In **Chapter 5**, the validation of the *ICAnDoiT* system and the *Questions Test* will be detailed (Test Specifications listed in **Appendix 1**). In **Chapter 6**, the results will be presented. **Chapter 7** will discuss the findings and their implications, list conclusions made based on the findings, outline the limitations of the study, and sketch some directions for the future research. Finally, the original articles forming the dissertation are supplied.

1.4 Summary of the research questions

The general research questions of the present research project, which will be listed in full detail in **Chapter 4.1**, are the following:

1. Does automated adaptive corrective feedback facilitate the development of learners' ability to form L2 English questions?
2. How do learners' beliefs about corrective feedback and their performance on / experience of dynamic assessment mediate one another?
3. What are some ways of ensuring the usefulness of automated adaptive CF in a computerised dynamic assessment of an under-researched construct, such as L2 English word derivational knowledge?

As it has been mentioned above, the following two chapters will give an overview of the previous research that motivated these questions.

2 CORRECTIVE FEEDBACK IN SECOND LANGUAGE ACQUISITION RESEARCH

In the present chapter, I will outline the research on corrective feedback in second language acquisition (SLA), indicating issues that emerged from the CF research and suggesting an epistemology that has the potential to explain these problems. Acknowledging the vastness of the field, I will exclude peer feedback from the review. I will first briefly sketch the history of feedback, as it illuminates the current perspectives on corrective feedback in SLA. To avoid confusion in the use of the term **SLA**, in the present synthesis, it will be used to discuss **instructed second or foreign language acquisition** only.

2.1 Early views on feedback

According to some sources (e.g., Barbour, 2003), the term **feedback** originated from the field of cybernetics in the late 1940s. Others suggest it appeared already in 1860s to refer to loops of momentum or signals in mechanisms (e.g., Stone & Heen, 2014). As regards SLA, feedback is now considered to be one of the essential elements of instruction and is defined as any communication given to learners' performance, usually to inform them of the accuracy of their response (Mory, 2003: 745).

Perhaps, the earliest feedback studies were conducted by Thorndike (e.g., 1911: 244) in the field of animal psychology, resulting in the formulation of the Law of Effect, which states that:

> Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur.

It was due to Law of Effect that feedback started to be identified with reward, and error correction was considered detrimental to learning (e.g., Kulhavy & Wager, 1993). Kulhavy and Wager (1993: 5) added that it was only later that feedback started to be considered as (a) an incentive for improving performance, (b) a reinforcing process, and (c) "information which learners could use to validate or change a previous response." This triad of function influenced views on feedback, such as Hattie and Timperley's (2007), which will be discussed in **Chapter 2.6**.

With the spread of view of **feedback as information**, which emphasised error correction, research of corrective feedback began to develop although nowadays the utility of error correction is still challenged (**Chapter 2.4**). In the following section, I will define corrective feedback as will be used in the present synthesis.

2.2 Defining corrective feedback

Schachter (1991) wrote that **corrective feedback**, **negative evidence**, and **negative feedback** are often used interchangeably. Bruton (2000), on the other hand, saw correction as superordinate to **negative** and **positive evidence**. Long (1996: 143) defined the former as "information about what is ungrammatical" and the latter as "models of what is grammatical and acceptable."

In the present synthesis, siding with Bruton (2000), I will use Lightbown and Spada's (1999: 171) definition of corrective feedback, which is "any indication to the learners that their use of the target language is incorrect." This definition is rather broad, incorporating both negative and positive evidence, and allows for discussing different types of feedback as CF (**Chapter 2.2.1**). One problem still with this definition is whether to consider confirmation that the response is correct as corrective feedback. On the one hand, its primary function is not correction. On the other hand, it can, for example, be at the end of a corrective episode, signalling that the previous corrective feedback given to a learner was beneficial. Thus, for convenience, I will consider it together with different CF types, which will be detailed in the following section.

2.2.1 Corrective feedback typologies

Corrective feedback can be classified in different ways. In terms of directness of CF, Ellis (2009a), for example, differentiated between:

- direct corrective feedback, i.e., overt correction;
- metalinguistic feedback, i.e., metalinguistic clues, e.g., grammatical descriptions;
- and indirect corrective feedback, i.e., indicating that there is an error and/or locating it.

Somewhat similarly, CF can be considered within **implicit-explicit** distinction (e.g., Long, Inagaki, & Ortega, 1998; Ellis, Loewen, & Erlam, 2006). An example of implicit feedback can be a face expression, whereas a phrase *this is wrong; the correct response is…* is explicit. It has also been suggested that the dichotomous operationalisation of explicitness of corrective feedback can be problematic, and thus it is better to consider it as a dimension (e.g., Egi, 2007a).

Corrective feedback can also be studied based on its **complexity** (e.g., Dempsey, Driscoll, & Swindell, 1993; Whyte, Karolick, Neilsen, Elder, & Hawley, 1995; Schimmel, 1998), as presented below:

- confirmation feedback (also knowledge of results / knowledge of response (KOR)), i.e., informing the learner whether the answer was correct or incorrect;
- knowledge of correct response (KCR) feedback;
- explanatory feedback, i.e., why the response is incorrect.

Additionally, corrective feedback can be classified based on the **approach to correction**:

- reformulation in full or in part, without the error (e.g., recasts or explicit correction);
- clarification request—indication that the utterance has not been understood or was ill-formed (e.g., *Excuse me?*);
- metalinguistic feedback (clues)—providing metalinguistic information in the form of comments or questions (e.g., *Do we need Past Simple here?*);
- elicitation—eliciting the correct form (e.g., *How do we say that in English?*);
- repetition—repeating the erroneous utterance;
- models—examples of what is acceptable.

(Ellis, 2009b; Long, Inagaki, & Ortega 1998; Lyster, 2004; Lyster & Ranta, 1997; Yoshida, 2008; Zourou, 2008).

These CF classifications are not mutually exclusive. For example, *How do we say that in English?* is somewhat implicit, is an elicitation, and does not contain much detail about the error whereas *dogs* in response to learners' *cats* would be an implicit reformulation KCR feedback (i.e., a recast). Certainly, the same feedback message can include several CF types (e.g., repetition and metalinguistic feedback) as well.

Corrective feedback has also been considered from the point of its **immediacy**. Dempsey and Wager (1988), for example, noted that in different studies, feedback delivered after each item/response, after each section, and, immediately after a whole activity (e.g., a test), was all referred to as immediate, which makes it reasonable to consider the immediacy of feedback as a dimension.

Furthermore, corrective feedback can be studied from the point of view of its **modality.** Although CF can also take forms of, for example, imagery (e.g., Hew & Ohki, 2013), to keep the argument more focused I will only mention studies on written or oral CF. It has been noted (e.g., Ferris, 1999; 2010; Sheen, 2010) that while the research on written CF is above all embedded in writing composition theories, studying the way it improves the effectiveness of learners' writing, studies of oral CF are mostly based upon SLA theories, i.e., focuses on the acquisition of L2. Nevertheless, as Ellis (2010) noted, recently there has been a shift of focus to the theoretical question of whether written CF promotes learners' acquisition of L2.

Other differences between the two modalities include the inherently explicit and delayed nature of written corrective feedback which are not necessary properties of oral feedback (Ellis, 2010). As regards explicitness, it is understandable that learners should expect that most written comments on their writing are correction. Thus, referring to written CF in terms of being direct (i.e., directly correcting the error) and indirect (i.e., pushing learners to self-correct by not revealing the correct response) should be more defensible (Ellis, 2010).

It is also apparent that conventional written CF cannot be immediate. However, computerised written CF can. For example, learners' errors can be automatically marked while they are writing online, or written feedback is provided after each separate test item (e.g., Alderson & Huhta, 2005; Huhta, 2010). I would also argue that computerised (or web-based) CF can be considered within the implicit-explicit dimension, as learners receive different kinds of computerised written feedback doing, for example, online exercises, not all of it corrective (e.g., *next item* or *press ? for help*).

Finally, corrective feedback can be classified based on whether it is **focused** or **unfocused**, that is, whether all errors are corrected or specific errors only. There have been several studies that contrasted the effect of focused with that of unfocused feedback. Ellis, Sheen, Murakami, and Takashima (2008) could not confirm that focused CF is superior to unfocused one, as both the focused CF and the unfocused CF groups were significantly better than the no CF group. On the other hand, Sheen, Wright, and Moldawa (2009) demonstrated that focused CF was more beneficial than unfocused one.

The conflicting results regarding the relative effectiveness of focused and unfocused CF are exemplary of corrective feedback research in SLA, where there is little consensus regarding the effectiveness of different CF types. In the section to follow, I will outline empirical research comparing and contrasting different CF types. I will concentrate on the effect of feedback on acquisition and make more emphasis on the studies comparing and contrasting explicit and implicit CF.

2.3 Efficacy of different CF types

As Ellis (2010) noticed, it is easy to manipulate CF in experimental settings. Thus, studies of CF have been predominantly conducted within the experimental research paradigm. Quite often these studies compare and contrast explicit and implicit feedback. Among oral implicit feedback types, recasts comprise the vast majority of research targets.

As it has been mentioned in **Chapter 2.2.1**, findings regarding the efficacy of different CF types vary, as will be exemplified with reference to studies comparing the effect of recasts with other oral CF types (**Table 1)**.

TABLE 1 Studies comparing/contrasting recasts with other oral CF types.

Studies	**Results**	**Comments**
Ellis, Loewen, & Erlam (2006); Lyster & Saito (2010); Varnosfadrani & Basturkmen (2009); Yang & Lyster (2010)	more explicit feedback is more beneficial than recasts (and no feedback)	• In Varnosfadrani and Basturkmen (2009), this was true for developmental-early features; for developmental-late features, the reverse was true. • In Yang and Lyster (2010), there was no difference between the prompts group and recasts group in the use of the English irregular past tense.
Ammar (2003); Ammar & Spada (2006);	On average, recasts were less effective than more explicit feedback.	• More able learners benefitted equally well or more from recasts.
Li (2010)	The effect of recasts was better maintained over time.	• The study was a meta-analysis.
Kang (2009); Loewen & Nabei (2007); Loewen & Philp (2006); Lyster & Izquierdo (2009)	Either there is no significant difference, or recasts are more effective.	• Loewen & Philp (2006) suggested that explicit recasts were more beneficial for learning.

Several interesting observations can be made based on the studies in **Table 1**. First of all, it seems that more able learners do benefit from recasts. It has also been found that sometimes, learners do not perceive recasts as corrective feedback or fail to see the discrepancy between the correct response and theirs (Egi, 2007b; 2010). Mackey & Philp (1998) suggested that less able learners have limited ability to notice recasts as corrective feedback, which, in light of Egi's findings, can explain the findings presented in **Table 1**. That is to say, (implicit) recasts appear to be more beneficial for more able learners, whereas less able learners might not perceive recasts as CF at all.

A similar lack of consensus exists as regards written CF. I will consider studies of written CF, above all, from the point of view of **direct** or **indirect** CF, the distinction prevalent in these studies. This distinction should not be regarded as necessarily corresponding to the implicit-explicit dimension, as direct feedback can be implicit (e.g., recasts) and indirect, rather explicit (e.g., underlining in red pen). I, therefore, looked at a group of studies considering written CF from the point of degree of its directness (**Table 3**) and also attempted to classify the CF in these studies based on its degree of explicitness and amount of detail (**Table 2**).

TABLE 2 Written CF in different typologies.

Feedback	Explicitness	Directness	Amount of detail
Error code	somewhat explicit	Indirect (more direct than underlining)	not detailed
Overt correction	Explicit	Direct	not detailed
Underlining	somewhat explicit (arguably, more explicit than error codes)	Indirect	not detailed
Comment	somewhat explicit	Indirect	Detailed

Getting ahead of the present discussion, neither of these distinctions can serve an explanation for the different findings in these studies (**Table 3**).

TABLE 3 Studies comparing/contrasting direct and indirect written corrective feedback types.

Study	Results	Comments
Lalande (1982)	Indirect CF (error codes) is more effective than direct CF (overt correction).	The study had no control group.
Eslami (2014)	The indirect feedback group outperformed the direct feedback group.	The author did not elaborate on what the indirect feedback looked like.
Ferris & Roberts (2001)	There was no significant difference between more direct feedback (error codes) and the more indirect feedback groups (underlining); both groups outperformed the no-feedback group.	
Bitchener (2008); Bitchener & Knoch (2009)	Regardless of the amount of detail of the direct CF, all groups improved their performance in the use of L2 English articles significantly and outperformed the no-feedback group.	

A similar picture emerges from studies of computerised corrective feedback. Rosa and Leow (2004) found that explicit prompts with explanations were more beneficial for the development of learners' ability to recognise and produce L2 Spanish conditional sentences than indirect/implicit feedback (KOR). On the other hand, Cabrera (2007), for example, found that groups receiving elicitations and metalinguistic feedback outperformed both the control group (no feedback) and the learners receiving error repetition and overt correction, the latter group also performing better than the control group, in their use of subjunctive and past tense in Spanish and English.

As the present section demonstrated, there is no clear answer as to which corrective feedback is more beneficial for learning an L2 (see also, e.g., Pica, 1994). In fact, not everyone agrees that CF should be used at all.

2.4 Does corrective feedback help at all?

Most studies discussed in the previous section found that CF was more beneficial than no CF. However, a different view exists, which, as Ellis (2010) noted, is inspired by Chomsky's (e.g., 2002) nativist perspective, considering acquisition to be predominantly promoted by positive evidence. In addition, perspectives stemming from Krashen's (e.g., 2009) Natural Order Hypothesis, predicting that learners acquire grammatical/lexical features in a fixed order, challenge the view that CF promotes acquisition.

Kepner (1991), for example, did not find any significant effect of both overt correction and reminding of rules on the performance of 60 learners of Spanish whose mother tongue (L1) was English. Similarly, Fazio (2001) found no significant difference in learners' accuracy in writing journal entries in L2 French following corrections, commentaries, or both. Polio, Fleck, and Leder (1998), who studied 65 ESL learners of English in an academic writing course, did not find a significant difference between the performance of the experimental group who had CF and the control group who did not.

Truscott (e.g., 1996; 1999a; 199b; 2007) is, perhaps, the most critical opponent of CF, claiming that it has either no effect or a small negative effect on L2 acquisition. One problem with error correction in the classroom emphasised by Truscott (e.g., 1996; 1999a) is that it can be notoriously difficult for teachers to provide useful CF, as it includes providing feedback that is understood by learners. This failure to understand the correction, he suggested, can be embedded in that learners are not developmentally ready for such feedback.

Truscott discussed the problem in terms of stages of development. However, an alternative perspective on development has the potential to explain the conflicting findings CF research has produced and address the problem delineated by Truscott (1996; 1999a).

2.5 Corrective feedback from a sociocultural perspective

In an attempt to address the conflicting findings regarding the role of CF in acquisition, several studies have considered it from a sociocultural perspective (e.g., Donato, 1994; Aljaafreh & Lantolf, 1994; Nassaji & Swain, 2000). At the heart of this perspective lies Vygotsky's concept of Zone of Proximal Development (ZPD), which is "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978: 86). The basic premise of this perspective is that knowledge is socially constructed and is a movement from other-regulation towards guiding one's own actions and behaviour (e.g., Vygotsky, 1978; 1986). It should be noted that in Piagetian (e.g., Ginsburg & Opper, 1979) perspective on development, the direction of the movement to becoming self-regulated is different. Specifically, whereas Piagetian perspective emphasises inner development directed at controlling the outer world, Vygotskian perspective on development maintains that cognitive development happens from the social to the inner mind (e.g., Alanen, 2003). What is more, the ZPD itself is not fixed but rather emerges and transforms in interaction, the latter providing learning opportunities which would be impossible otherwise (e.g., Wells, 1998).

Thus, in CF studies based on this perspective, there is a shift from trying to find evidence that one CF type is more useful/effective than another towards a view that assistance promoting development emerges in interaction. This guided assistance is known as **mediation** and can include, but is not limited to, corrective feedback, and it promotes development if provided within the individual's ZPD (e.g., Poehner, 2008; Vygotsky, 1987). That is to say, depending on the learner's ZPD, any CF can promote the learner's development.

In the present synthesis, I will refer to feedback as mediation as either **corrective feedback provided within learners' ZPD** or **adaptive corrective feedback** (to emphasise that it is adapted to learners' abilities). I took the latter term from Vasilyeva et al. (2007: 11), who refer to adaptive feedback as feedback dynamically adapting to users' characteristics and performance. I will refer to CF provided irrespective of learners' ZPD as **static CF**.

Before outlining the research of CF within a sociocultural perspective, it should be explained why Truscott's suggestion of the beneficial role of CF depending on learners' developmental stage would be difficult to confirm or disprove based on the findings of the present research. Dunn and Lantolf (1998), for example, discussed the incommensurability of the two perspectives, claiming that unlike theories based on Piagetian perspective on development, which perceive development as a movement from one development stage to another, the sociocultural theory maintains that development is directed by instruction, and there are no prescribed developmental stages (e.g., Leung, 2007).

The research on CF from the sociocultural perspective corroborated the prediction of the beneficial role of CF provided within learners' ZPD. Aljaafreh and Lantolf (1994), for example, collected qualitative data on three learners who received corrective feedback negotiated within their ZPD. The feedback provided to the participants addressed the errors/mistakes they made in their use of L2 English grammar. The analysis of the interaction revealed that every CF type was beneficial if negotiated between the learners and the teacher within the learners' ZPD. Having analysed the data, the authors designed a Regulatory Scale of 13 feedback messages arranged by the gradual increase in their degree of explicitness and level of detail.

Nassaji and Swain (2000) conducted a case study of two L1 Korean learners of English which aimed at discovering in what way adaptive CF is different from randomly provided static CF. Feedback on the use of articles was given to both learners, but one of them received CF provided within her ZPD, and the other, random CF. The results (both qualitative and quantitative) indicated that the learner who received adaptive CF improved more. The limitation of Nassaji and Swain's (2000) study arose from its small scale. That is, because of only two participants, no statistics beyond frequencies and percentages could be reported.

Antón (1999) studied the CF strategies of two university instructors, a teacher of L2 French and a teacher of L2 Italian, throughout one semester. While the French instructor invited learners to be active participants in the interaction, the Italian instructor adopted a teacher-centred approach. Furthermore, the Italian instructor mainly used overt correction whereas the French instructor used a variety of CF types emerging from the interaction. The author concluded that the learner-centred discourse provided ample opportunities for learning, as it allowed for negotiating both meaning and form. The teacher-centred discourse, on the other hand, provided rare opportunities for negotiation and fewer opportunities for learning than the former.

There is an apparent distinction between studies conducted within the conventional perspective on CF (**Chapter 2.3**) and the sociocultural perspective on it. Specifically, while studies conducted within the former framework emphasise learning as a product, and operationalise it within the quantitative research paradigm, in the latter, the emphasis is on the process, and thus these studies are predominantly qualitative. Nevertheless, it is not impossible to conduct research of adaptive CF within the quantitative research paradigm, as Nassaji and Swain (2000), who collected quantitative evidence for a greater development of the learner provided with adaptive CF than that provided with random CF, demonstrated.

2.6 Levels of Feedback

Another perspective on feedback was introduced by Hattie and Timperley (2007), who attempted to explain the effectiveness of feedback by adopting a broader view of it, involving products, processes, strategies, and personal characteristics of learners. They based their argument on Hattie (1999), who synthesised the findings of several meta-analyses on the effectiveness of CF. Considering that meta-analyses are, in turn, syntheses of the previous research on the same topic, this study produced strong evidence for Hattie and Timperley's argument.

Based on the accumulated data, the authors designed a feedback model attempting to explain what it is that makes feedback effective. They claimed that to be effective, feedback should answer three questions: (a) what is the goal? (feed up), (b) how am I doing in relation to the goal? (feed back), and (c) what should be done to improve? (feed forward). They suggested that the effectiveness of responses to these questions for learning/development depends on the focus of feedback, that is, on the level(s) at which feedback operates: the task level, the process level, the self-regulation level, and the self level.

The task level is about whether the response is correct or incorrect, but also what should be done in order to arrive at the correct response. The authors defined the process level as aiming at processing of information required for completing the task. The self-regulation level, according to them, can be directed at the increase of self-evaluation, encouragement to continue to engage in the task, and generally on learners' beliefs about learning. Finally, feedback directed at the self level, they suggested, is personal feedback, such as *well done*.

Based on their data, the authors argued that feedback about self is the least effective (also, Kluger & DiNisi, 1998) whereas feedback types directed at processes and self-regulation are the most effective. They added that the task-level feedback is effective when it also contributes to strategy processing and self-regulation (which rarely happens).

The most problematic level in their model, as was argued by Alderson et al. (2015), is the process level, as it is difficult to separate this level from feedback aimed at promoting learners' self-regulation, as both include transfer of knowledge to other contexts. Thus, it seems that both of these levels aim at promoting learners' self-regulation.

Alderson et al. (2015) suggested that the process and the self-regulation levels can be joined. They continued that since both of these levels include learning strategies, these could be joined into strategy level. In the following, for a better understanding of this feedback level, I will provide a brief overview of learning strategies based on Alderson et al. (2015), but also some other sources.

O'Malley and Chamot (1990: 1) defined learning strategies as "special thoughts or behaviours that individuals use to help them comprehend, learn, or retain new information." Dörnyei (2005) added that later, to acknowledge the problematic relationship between thoughts and behaviours, these were replaced with methods and techniques.

Alderson et al. (2015) suggested that strategy level of feedback could be divided into feedback on metacognitive, cognitive, and social/affective strategies (see also, e.g., Oxford, 1993; Purpura, 2014).

Metacognitive strategies involve thinking about the learning process and planning this process and include such strategies as monitoring, evaluating.

Feedback on these strategies can, for example, ask learners to stop at regular intervals when reading to check what they have understood (Alderson et al., 2015: 174).

Cognitive strategies involve manipulation of material to be learned. These include, but are not limited to, such strategies as repetition, note-taking, classifying, and inferencing. (e.g., O'Malley & Chamot, 1990). Feedback at this level can instruct learners to think about meanings of words in order to understand a sentence.

Social and affective strategies involve interaction with peers to solve problems and adjusting self-beliefs and feelings related to L2 learning. Alderson et al. (2015: 177) suggested that feedback on these strategies can involve asking learners to seek help from their peers.

It should be mentioned though that Dörnyei (2005) noted the 'fuzziness' of the definition of learning strategies and noted the problem of conflicting results due to adopting different methodologies as a result of this fuzziness of the construct. Instead, he suggested that the construct should be reconceptualised by shifting the focus from the product, that is, strategy, to the process of becoming self-regulated. He defined self-regulating capacity as an aggregate of commitment control (i.e., preserving the original goal), metacognitive control (i.e., monitoring and controlling concentration), satiation control (i.e., fighting boredom), and environmental control (i.e., using the environment to help you to achieve the goal) (Dörnyei, 2005: 113).

Rose (2012), however, noted that Dörnyei's (2005) approach is not incompatible with the learning strategies approach, the former emphasising the process of becoming self-regulated whereas the latter, the product of it. He added that development of self-regulation / learning strategies requires a more qualitative exploration as contrasted with earlier research on learning strategies (e.g., O'Malley & Chamot, 1990; Oxford, 1993), which used questionnaire as the principle data collection tool.

Alderson et al. (2015: 170) also classified some common CF types in relation to levels of feedback. This is a useful classification in that it demonstrates what level these feedback types have most of. However, it disregards predictions made within the sociocultural paradigm about the information that different feedback types contain for different learners, as I will illustrate in **Chapter 2.7.** For this, I will use some feedback types discussed by Hattie and Timperley (2007) and Alderson et al. (2015) as examples.

2.7 Sociocultural perspective and Hattie and Timperley's feedback levels

Hattie and Timperley's (2007) view of feedback bears some similarities with the sociocultural perspective on it. To start with, it considers both learners and teachers as active participants in the teaching/learning process. In addition, the authors suggested that feedback should match learners' ability to understand it.

However, synthesising these two perspectives can still be problematic, as Hattie and Timperley (2007) discussed different feedback types as having either one or several of the levels embedded in them whereas from the sociocultural point of view, these should depend on learners' development and their ZPD, as will be exemplified below.

Alderson et al. (2015: 179), siding with Hattie and Timperley (2007), suggested that self-feedback (e.g., *well done* or *you can do better*) is ineffective, as "it does not contain much information about the learners' performance on the task and also is not related to learning goals."

While in static CF research, such feedback might indeed be found not very beneficial for learning, the reason for that can different from the reasons that Alderson et al. (2015) suggested. The sociocultural perspective on CF predicts that learners who can benefit from self-feedback are already almost fully self-regulated and thus can understand what is wrong in their response or why it was correct with as little task-level information as such feedback contains. I would also suggest that this feedback, when it is provided within learners' ZPD, can promote learners' self-regulatory capacity, as it indicates whether the techniques learners used were beneficial for finding the correct response or not, and thus, helps them to adjust their approach to solving similar problems respectively. On the other hand, when self-feedback is outside learners' ZPD, it only has self level (i.e., either praise or an indication that the task is failed). Thus, an explanation of the ineffectiveness of self-feedback that the previous studies have found can be that usually, in experimental designs, samples are selected such that the participants are not skilled in the ability being assessed, that is, not many participants are close to being self-regulated in the abilities selected as targets of the instruction in these studies.

A feedback message classified by Alderson et al. (2015:174) as a combination of strategy feedback and task feedback can serve a clearer example of how feedback can include some or all the three levels. The feedback message *The answer is incorrect. What do you think words X and Y tell you about the writer's attitude?,* above all, has the task level. The second part of the feedback message includes the strategy level, as the learner is referred to a specific detail in the text. I would consider the first part of the feedback message as also having the self level, as learners receiving the feedback should realise that they failed the task. However, if the amount of assistance provided in the feedback message is not enough to find the correct response, what stays is the self level (learners failed the task). On the other hand, if a learner is able to complete the task with less assistance than this feedback provides, it will unlikely to result in the development of his/her ability (and increase in self-regulation) although will result in a self-correction. That is to say, every CF type, to a varying degree and depending on learners' ZPD, has the potential to target some or all Hattie and Timperley's (2007) levels, and it is when feedback is provided within learners' ZPD that the full potential of the feedback is revealed.

Judging by that, what Hattie and Timperley's (2007) classification allows for is maximising these three levels for learners of different abilities, thus increasing the CF usefulness. For example, more of task level can be added to the *well done* feedback by repeating the learner's response. The same, in fact, was also suggested by Hattie and Timperley (2007: 91). All in all, combining the two paradigms should allow for maximising the usefulness of feedback for learners having different ZPDs.

It should be noted at this point that feedback usefulness, especially **usefulness of adaptive CF**, in the context of the present doctoral research, is largely equivalent to the **validity** of this feedback (see **Chapter 5**). That is to say, if the feedback is not useful for the learner, it does not serve its role of promoting this learner's L2 abilities (i.e., it is not valid). Furthermore, if this same feedback is not *considered* useful by the learner, chances are that he or she will simply disregard it, as will be discussed in the chapter to follow. In other words, this usefulness can still be hindered if learners do not believe that CF they receive is useful.

2.8 Learners' beliefs and corrective feedback

While the role of educators/assessors in providing CF is hard to overestimate, it should not be forgotten that learners are also participants in the interaction. Thus their reciprocity to feedback is important to consider when discussing corrective feedback and mediation in general.

It has been argued that learners' **reciprocity**, that is, responsiveness to feedback/mediation (see **Chapter 3.2.1**) should be considered an indication of their development, as it reflects their ZPD and allows the mediator to adapt the amount of assistance to the learner's needs (e.g., Feuerstein, Rand, & Rynders, 1988; Poehner, 2005).

Poehner (2005), for example, studied responsiveness to mediation of 6 learners of French. Based on the recordings of the interaction, he concluded that the learners' reciprocity provided important information about their ZPD.

However, alternative explanations exist for what learners' reciprocity to feedback/mediation can be indicative of. It has been found that the utility of corrective feedback, to an extent, depends on how useful learners believe it is for them (e.g., Kern, 1995, Leki, 1991; Schulz, 1996; 2001). For example, Thouësny (2011) found that learners skipped CF in a computerised dynamic test (see **Chapter 3**) and suggested that one reason for that could have been because they believed they would not be able to correct their mistakes with the help of it. Therefore, it appears that learners' beliefs about corrective feedback can mediate the way they receive this same feedback.

There have been several studies that aimed at finding out learners' perceptions of and beliefs about CF. Hedgcock and Lefkowitz (1994), for example, using questionnaire responses of 247 learners of English, discovered that while, generally, their participants were positive towards CF, there was some variation in their self-reported beliefs, which led the authors to conclude that teachers' feedback practices influenced these beliefs. At the same time, they also found that learners' and teachers' beliefs about CF did not always match.

Hedgcock and Lefkowitz' finding was not an exception. Other studies have also demonstrated that learners and teachers can have different beliefs about corrective feedback (e.g., Brown, 2009; Diab, 2005; Saito, 1994). Specifically, learners appear to be more in favour of explicit CF rather than implicit (e.g., Amrhein & Nassaji, 2010; Ashwell, 2000; Lee, 2008; Leki, 1991), whereas teachers can be more in favour of less explicit feedback (e.g., Amrhein & Nassaji, 2010; Yoshida, 2010). Amrhein and Nassaji (2010) suggested that the reason for that could be learners' belief that it is teachers' responsibility to (overtly) correct their mistakes/errors.

This belief can hinder learning in that it can be at odds with teachers' beliefs and practices, resulting thus in lack of communication in the L2 classroom (e.g., Barcelos & Kalaja, 2013; Kern, 1995). More importantly, it can be an obstacle in the process of learners becoming self-regulated in their L2 use, as this hinders learners' autonomy (e.g., Amrhein & Nassaji, 2010).

A way to account for and counter a negative influence of learners' beliefs on their perception of CF can be found in the contextual, especially in sociocultural, perspectives on learners' beliefs, which maintain that beliefs are in a constant state of flux, being constantly socially constructed, influenced by social contexts, and in turn, mediating these contexts (e.g., Alanen, 2003; Aro, 2009; Barcelos, 2003; Barcelos & Kalaja, 2013; Dufva, 2003; Mercer, 2011).

An example of a study of learners' beliefs from a sociocultural perspective is Alanen's (2003) study, in which it was demonstrated how learners' beliefs were co-constructed (and appropriated) in the interaction, and how authoritative others, including those not directly present in the interaction (e.g., parents), mediated the construction of beliefs about learning an L2 of 16 learners before the formal L2 instruction began. Alanen (2003) suggested that the unit of analysis in studies of beliefs can be *mediated action*, a system in which the agents, the mediational means, and the context in which the interaction occurs are a part of the same system and co-influence each other. Some of the properties of mediated action as identified by Wertsch (1998) are the following:

- mediational means can not only enable the action but also impede it;
- new mediational means have the power to transform the action;
- the relationship of the agents towards the mediational means are often manifested in terms of appropriation;
- mediational means are often associated with power and authority.

Based on these properties of mediated action, Alanen (2003) suggested that beliefs start to be used as mediational tools when they are appropriated. She further proposed that the degree of appropriation of beliefs can be manifested in (a degree of) agency transpiring in learners' utterances. Finally, the acceptance of a certain belief by an agent is often associated with power and authority. This explains, for example, the relationship with teacher's practices and learners beliefs.

An example of mediated action can be the development of Eeva's, a participant in Alanen's (2003) study, belief. During the first interview, the interviewer explicitly suggested that English can be important for Eeva when she visited her grandmother in Singapore, which Eeva did not seem to acknowledge, simply confirming that she wanted to visit her grandmother. However, as it appeared 18 months later, the interviewer's mediation and Eeva's following confirmation were central to the development of her belief about the importance of the English language. As soon as another interviewer asked her if she would like to learn English, she immediately responded *I would! Because my godmother lives in Singapore!* (Alanen, 2003: 75). Thus, it transpired that Eeva appropriated the utterance of the first interviewer to report on the importance of English for her 18 months later. The action, which was the dialogue during the second interview, was mediated by (a) the authority of the first interviewer, (b) a similar context in which her belief was constructed, that is, a research interview, and (c) the fact that her grandmother indeed lived in Singapore. Notably, this would not have been possible had Eeva been more talkative and not left this opening for the interviewer to mediate her talk during the first interview. This example demonstrates the interdependence of the context, the mediational means, and the agents in interaction.

However, it should not be assumed that beliefs only transform over longer periods of time. Dufva (2003) suggested that learners' beliefs can also transform on a micro level, that is, during one specific situation. Alanen (2003: 78-79) illustrated how it can happen by presenting a short sample from the interaction between two learners and the interviewer, in which when the interviewer asked whether it would be easier for children to learn English than for adults one of the learners immediately repaired his initial utterance from *well yes it* to *well it wouldn't* when the other learner responded differently. Certainly, it is impossible to tell whether this belief was appropriated further later on and, due to situatedness of beliefs, whether in a different interview, the same belief would have emerged. However, this approach to the study of beliefs allows for determining how they are co-constructed by the participants in the interaction.

These studies suggest that a sociocultural approach to studying learners' beliefs can both produce important insights into the way these beliefs develop and is a way to transform these same beliefs.

2.9 Drawing the threads together

In the present chapter, I argued that considering corrective feedback from the perspective of the sociocultural theory of learning has the potential to explain the conflicting findings that the research on CF in SLA has produced.

Other perspectives on corrective feedback, such as that of Hattie and Timperley (2007), enrich the sociocultural perspective on corrective feedback, illuminating the way that corrective feedback promotes learners' self-regulation in Vygotskian sense by suggesting that the same feedback can function on different (and on different number of) levels, that is, the self level, the task level, and the strategy level.

However, CF research within this epistemology is largely descriptive. This is understandable, as its emphasis is on process rather than product. However, this limits the comparability of findings of more traditional (mostly experimental) studies of CF with those of studying CF from sociocultural perspectives, disallows the use of meta-analyses, and, overall, limits the strength of the claim of the usefulness of adaptive CF.

As I will try to argue in the following chapter, computerised dynamic assessment is one practical way of accumulating experimental data on the effectiveness of corrective feedback provided within learners' ZPD, and thus it can potentially address this research gap. That said, qualitative research should enrich of our understanding of how CF provided within learners' ZDP promotes their development as well as, for example, disclose how learners' beliefs influence their responsiveness to corrective feedback, how these beliefs change over time, and what mediates these changes.

3 DYNAMIC, DIAGNOSTIC, AND DYNAMIC DIAGNOSTIC ASSESSMENT OF L2

In the present chapter, I will provide an overview of dynamic assessment (DA) and list some benefits that research on diagnostic assessment can have for the development of and research on DA. Additionally, my intention will be to find out what insights into corrective feedback the paradigms underlying dynamic and diagnostic assessment can provide. First, I will introduce some important terms used in the assessment field and detail how these will be used in the present synthesis.

3.1 Assessment, testing, measurement, evaluation

Bachman and Palmer (2010) noted that in the field of Applied Linguistics, the terms **assessment**, **measurement**, and **testing** are used more or less synonymously to refer to collecting information about learners' L2 abilities. Following Bachman (1990), they defined **assessment** (measurement/testing) as collecting information using procedures that are clearly defined and based on accepted theory, methodology, and/or practice. They, however, made a distinction between **assessment** and **evaluation**, defining the latter as involving making judgements and decisions based on the information collected during the assessment. Lynch (2001), on the other hand, conceptualised **assessment** as a superordinate term including **measurement** and **testing**. He also perceived measurement as including testing, limiting testing to using quantifiable methods only.

I side with Lynch (2001) in that assessment should include, but is not limited to, testing, as it can also include, for example, the use of portfolios. However, I would not limit testing to procedures that quantify learners' performance. The purpose of a test can as well involve, for example, making diagnostic decisions, which should not require obtaining a numerical score (e.g., Alderson, 2005). Nevertheless, in the present synthesis, since I will not discuss assessment tools other than tests, I will mostly use the term **assessment** to refer to *testing*. I will use the term **tests** to refer to specific tools designed for the purpose of evaluation of learners' abilities. Likewise, I will use the terms **test-takers** and **test designers** to refer to particular stakeholders involved in creating/using these tools. I will, however, use terms like **proficiency testing** or **achievement testing** as these are conventionally used to refer to these types of assessment.

3.2 Assessment of learning and pro-learning assessment

Assessment has become an indispensable part of the teaching/learning process, its major aim being to gain insights into learners' abilities. At the same time, while it should complement learning, it is often dissociated from the goals of education or is even perceived to be in opposition with them, also resulting in that test preparation becomes the aim of instruction (e.g., Linn, 2000; Lynch, 2001; Poehner, 2008; Rea-Dickins, 2004; Shohamy, 2001).

Lynch (2001: 360), suggested that any assessment should consider:

- instruction and assessment as a unified process;
- learners as active participants in the development process of assessment;
- that a more detailed (qualitative) profile be given to test-takers rather than / in addition to a score.

He added that within the traditional assessment paradigm, these qualities are considered only if psychometric properties of the test, such as validity and reliability, are secured. What is more, as Tzuriel (2005) noted, traditional assessment can, for example, erroneously indicate the lack of learning strategies, motivation, and learning opportunities as a lack of intellectual abilities. The problems of the traditional assessment outlined above resulted in the appearance of alternative perspectives on assessment.

One of these is **dynamic assessment**, which both argues that assessment and instruction should be seen as a single process and challenges the way learners' development is traditionally perceived. Within the traditional assessment paradigm, development is perceived as a process of a learner moving through several predefined stages. This way of thinking is informed by Piaget's Theory of Cognitive Development (e.g., Ginsburg & Opper, 1979; **Chapter 2.5**). Lantolf and Poehner (2004), based on Valsiner (2001), term this approach as **past to present**, as, within this approach, learners' current performance is perceived as indicative of the stages they have moved through. In fact, as Leung (2007) notices, in past-to-present view, learners' future performance is also known, as it is the following developmental stage.

Dynamic assessment has a different epistemological (and ontological) basis. Theoretically and conceptually, it is based on Vygotsky's concept of ZPD (see **Chapter 2.5**) and thus maintains that no evaluation of learners' abilities can be complete without knowing how they perform under guidance (e.g., Leung, 2007; Poehner, 2008; Sternberg & Grigorenko, 2002; Vygotsky, 1998), that is, without being able to understand their **potential abilities** (i.e., performance with assistance) in addition to their **actual abilities** (i.e., unassisted performance). This is achieved through mediation of learners' performance provided based on their responsiveness to assistance (e.g., Poehner, 2008).

Furthermore, differently from traditional static assessment (SA), within this epistemology, learners' abilities are conceptualised as modifiable, rather than fixed, and thus the notion of developmental stages is rejected, as development, that is, future performance, is considered to be directed by instruction rather than following a predefined route (e.g., Feuerstein & Falik, 1999; Lantolf & Poehner, 2008; Poehner, 2008). Hence the approach is termed as **present to future** (Lantolf & Poehner, 2004).

Lantolf and Poehner (2004: 50) defined DA as follows:

> Dynamic assessment integrates assessment and instruction into a seamless, unified activity aimed at promoting learner development through appropriate forms of mediation that are sensitive to the individual's (or in some cases a group's) current abilities. In essence, DA is a procedure for simultaneously assessing and promoting development that takes account of the individual's (or group's) zone of proximal development.

As Leung (2007) terms it, DA is **assessment as teaching** and **pro-learning assessment** rather than **assessment of learning**, which is the aim of static assessment.

At the same time, this shift away from the traditional assessment paradigm has consequences, for example, for the way that test validity and reliability are defined and operationalised (e.g., Lantolf & Poehner, 2008; Poehner, 2005; 2008; 2011; Poehner & Lantolf, 2013; Chapter 5). It is, perhaps, not surprising that DA has been criticised by SA proponents (e.g., Glutting & McDermott, 1990) for the lack of psychometrical orientation and generalisability. However, as Poehner (2008) rightfully noted, this criticism is unsubstantiated, as DA has different aims from those of SA. Having said that, Poehner (2008) added that under the umbrella of dynamic assessment, one approach is more psychometrically oriented and thus can address such criticism. I will discuss the major approaches to DA in the section to follow.

3.2.1 Approaches to dynamic assessment

Lantolf and Poehner (2004) noted that mediation (including adaptive CF, see also **Chapter 2.5**) can range from support emerging in dialogic interaction to standardised hints. The two sides of the spectre of ways mediation can occur represent two general approaches to DA—interactionist and interventionist.

As the name suggests, in **interactionist** DA, assistance emerges in the interaction between the learner and the mediator (e.g., Lantolf & Poehner, 2004). An example of interactionist DA is Feuerstein's Mediated Learning Experience (MLE). Interestingly, Feuerstein, Rand, and Hoffman (1979) insisted that MLE was developed independently of Vygotsky's theories, and is instead based on the Structural Cognitive Modifiability Theory. According to this theory, humans' cognitive abilities are modifiable rather than fixed. Feuerstein et. al (1988) claimed that adequate development of cognitive functioning can only happen through MLE.

Feuerstein and Feuerstein (1999) listed twelve characteristics of MLE, of which they considered **intentionality** and **reciprocity**, **transcendence**, and **mediation of meaning** the most essential. **Intentionality** in MLE means that there should be an intent to mediate the learner's performance (but also to share this intention with the learner), **reciprocity** being the learner's response to the mediator's intentionality. **Transcendence** in MLE is the need for mediation to stretch beyond the present interaction (the task level in Hattie and Timperley's model) and promote the learner's area of knowledge being mediated. To make sure that learners' abilities are promoted, assessment based on the principles of MLE includes **transfer items**, that is, items of increasing difficulty assessing the already trained features but also going beyond what has been trained and assessed (e.g., Poehner & Lantolf, 2013). Transcendence in MLE specifically addresses the concern of teaching to the test (Poehner, 2008). Finally, **mediation of meaning** is the mediator's attempt at making the meaning relevant to the learner. Overall, this can be perceived as a feedback function not listed in Hattie and Timperley's model of feedback (**Chapter 2.6**), which answers the question of *why it is important*.

Several studies have been based on Feuerstein's concept of MLE. Antón (2009), for example, discussed an application of interactionist DA for diagnosis of university students in an advanced Spanish programme. She found that DA allowed for a richer diagnosis and more individualised approach to learners' needs.

Kozulin and Garb (2002) studied the applicability of DA to develop reading comprehension abilities of 23 L2 English academically at-risk students. The mediation stage aimed at helping learners to develop reading comprehension strategies. They found that the DA provided richer information than SA did. Interestingly, although Kozulin and Garb's (2002) study was based on the principles of MLE (which meant that the mediation emerged in the interaction), as it included a static pretest and a posttest, it can also be considered to include elements of **interventionist** DA, which will be discussed next.

While interactionist DA is particularly sensitive to learners' ZPDs, it also requires increased resources from the mediator. In contrast to interactionist DA, in **interventionist** DA, the mediator's freedom is limited by the list of standardised mediational moves arranged in a predefined fashion, usually from implicit and less detailed to explicit and detailed (and often in the form of CF), which the mediator has to follow (e.g., Lantolf & Poehner, 2004). Aljaafreh and Lantolf's (1994) Regulatory Scale (**Chapter 2.5**) serves an example of such arrangement. Although their study was rather a case of interactionist DA, the Regulatory Scale can serve a useful reference for designing mediation in interventionist DA (e.g., Lantolf & Poehner, 2011).

Interventionist DA departs from Vygotsky's thinking somewhat in that according to Vygotsky, development occurs and mediation should emerge in dialogic interaction (e.g., Poehner, 2008). On the other hand, interventionist approach to DA allows for establishing some of psychometric test properties in rather conventional ways (**Chapter 5**).

DA can also be classified into **sandwich** and **cake** formats, these two metaphors, introduced by Sternberg and Grigorenko (2002), aptly capturing the differences between the formats. In **sandwich** format dynamic tests, the dynamic part is conducted between the static pretest and the posttest, the former serving as a baseline of learners' unmediated performance and the latter indicating the progress made. As Poehner (2008) noticed, it was Budoff (e.g., Budoff & Friedman, 1967) who pioneered in the use of this format basing his *Learning Potential Measurement* on it. An important contribution that Budoff made to the field of DA was that he demonstrated that learners performing similarly during a pretest can perform differently on a posttest following DA.

In cake format DA, mediation is provided during the administration of the assessment whenever assistance is required, and there is neither a pretest nor a posttest. An example of interventionist cake format DA is described in Lantolf and Poehner (2011). The uniqueness of this study is in that it reports on the implementation of DA by a teacher following her own understanding of ZPD. Before the lesson, she designed a list of mediational moves, which she used during the lesson, noting the amount of assistance her learners required. The study indicated that the DA was beneficial for the learners' L2 Spanish abilities and that it was not only the length of the treatment that was important for the learners' development but also the quality of the treatment.

Interventionist DA also allows for computerised assessment, where mediation often takes the form of corrective feedback. Speaking of the latter, I will refer to corrective feedback during computerised DA as **automated adaptive corrective feedback** to account for the fact that it adapts to learners' abilities automatically. Computerised DA addresses the issue of practicality of human-mediated DA, which usually involves dyadic interaction between each test-taker and the mediator.

An example of computerised DA is the computerised version of Guthke's (1982) Leipzig Lerntest (LLT), an intelligence test used for diagnosing children's learning problems (Guthke & Beckman, 2000). An assumption underlying the test design was that there is no one ZPD, but rather a separate ZPD in each domain, such as mathematical calculations or L2 English (as opposed to L1) reading (also Garb, personal correspondence).

An important change to Budoff's approach introduced in LLT is that, in line with the principles of diagnostic assessment (**Chapter 3.3**), learners' performance on LLT served the basis for the subsequent teaching. Guthke and Beckman (2000) illustrated the benefit of the computerised LLT with examples of separate learners' performance but did not report on any experimental research findings, which computerised modality allows for.

Tzuriel and Shamir (2002) developed an interesting DA procedure for assessing seriational thinking ability of pre-school children, in which both interventionist and interactionist approaches were combined. Specifically in their study, in addition to the standardised computerised mediation, the assessor could also interact with the learners in an unstructured way. The authors concluded, perhaps not surprisingly, that the learners provided with both forms of mediation benefitted the most.

In the field of second language acquisition, Teo (2012) reported on a computerised sandwich format dynamic test of L2 English inferential reading abilities. Based on the quantitative results of the study, the author argued for the beneficial effect the test on the learners' abilities. She corroborated her findings with the qualitative analysis of learners' written reflections on their experience of the test, which demonstrated that with the help of mediation, the learners were able to use a number of strategies appropriately to read between the lines.

Poehner and Lantolf (2013) reported on an implementation of a cake format computerised DA of learners' L2 listening and reading comprehension. They used transfer items to operationalise the participants' development within one DA session (see also e.g., Aljaafreh & Lantolf, 1994; Lantolf, 2000; Lantolf & Poehner, 2011). The test calculated three sets of scores, the unassisted performance score, the mediated score, and learning potential score (LPS) calculated based on Kozulin and Garb's (2002) formula adapted for the cake format DA and suggesting the amount of instruction that would potentially be required by the learners for developing their abilities. Poehner and Lantolf (2013) demonstrated that learners having the same unassisted performance score could have different mediated performance scores, LPS, and performance on transfer items, thus confirming that computerised DA can provide richer information than SA can.

As the discussion above suggests, DA has a definite potential for diagnosing learners' strengths and weaknesses. In fact, Poehner (2008) claimed that the diagnostic value of DA is in that it establishes abilities that are fully developed, abilities in the process of development, problems that learners have, and ways to address these problems in instruction. In the following section, I will discuss diagnostic assessment in some detail and argue that research on diagnostic assessment can contribute to studies of DA.

3.3 Diagnostic and dynamic diagnostic assessment

3.3.1 What is diagnostic assessment?

A common definition of diagnostic assessment is that it is assessment that identifies strengths and weaknesses of test-takers (e.g., Hughes, 1989; Alderson, Clapham, & Wall, 1995; Bachman and Palmer, 1996). Judging by this definition, any test can be diagnostic to an extent, which suggests that notwithstanding the long history of diagnostic assessment (see e.g., Stobart, 2008), research of L2 diagnosis is still lacking and confusion exists as regards what diagnostic assessment is.

Alderson (2005) noted that in L2 assessment, the emphasis had been on standardisation and high-stakes assessment to the extent that such areas as diagnosis had been under-researched and confusion of what diagnostic assessment is appeared. He then proposed an explicit difference between diagnostic assessment and other assessment types, in that diagnostic tests should be *primarily* designed to establish strengths and weaknesses in learners' abilities and inform teachers of these with the intention of remediation of classroom instruction if required. This definition has been accepted in many following studies (e.g., Alderson & Huhta, 2011; Alderson et al., 2015; Huhta, 2008; Lee, 2015).

Having synthesised the previous research findings, Alderson (2005: 11-12) then listed 18 hypothetical features of diagnostic tests of L2. The features that I find particularly relevant for interventionist DA are presented below:

- diagnostic tests are designed to identify learners' strengths and weaknesses;
- a greater emphasis in diagnostic assessment is on weaknesses;
- diagnostic tests give detailed feedback which can be acted upon by learners and enable remediation of classroom instruction;
- this feedback is often presented in form of detailed profiles of learners abilities;
- diagnostic tests provide immediate results;
- the content of diagnostic tests has either been addressed in previous instruction or which will be covered shortly;
- diagnostic tests are based on a detailed theory of language development and SLA research;
- diagnostic tests are more likely to be discrete-point than integrative;
- tests of detailed grammatical knowledge are difficult to construct as the range of contexts that need to be covered either hinder reliability or practicality;
- diagnostic testing is likely to be enhanced by being computer-based.

Later, Alderson et al. (2015: 169) elaborated on features of diagnostic feedback which:

- is not limited to learners' errors;
- is based on the reasons underlying these errors;
- informs learners and/or educators of what can be done to improve the ability/skill in question.

The latter point in the second list is what dynamic assessment can do, too, and judging by the research outlined in **Chapter 3.2**, DA both provides information on learners' actual and potential performance and suggests assistance that can help learners to address their problems, thus, allowing for a more fine-grained diagnosis and self-diagnosis. The latter is especially important, since, as it has been noted (e.g., Lee, 2015) the specificity of feedback required for (self-)diagnosis and remediation is difficult to establish in static diagnostic tests. However, as the first list also suggests, both for an adequate diagnosis and an adequate mediation, the assessed construct needs to be carefully defined. This, as Lee (2015) and Lee and Sawaki (2009) suggested, includes components forming the construct in question and relations between these components. In addition research into the development of the construct should help to identify reasons behind learners' mistakes, and thus, suggest how they should be addressed. This should allow for maximising the three functions of useful feedback as defined by Hattie and Timperley (2007; see **Chapter 2.6**), that is, informing learners (and teachers) where they are heading, how they are doing in relation to this goal, and what they should do to improve their performance.

In the following, I will outline current approaches to diagnosis, which seem to be rather different, but in essence, aim to address the three functions of feedback as discussed by Hattie and Timperley (2007).

3.3.2 Approaches to L2 diagnosis

Alderson et al. (2015) discussed three general approaches to L2 diagnosis: retrofitting existing proficiency tests for making diagnostic inferences, static tests that have been specifically designed to be diagnostic, and dynamic assessment.

In the first approach, as its name suggests, instruments are not initially designed to be diagnostic. Instead, existing proficiency tests, such as TOEFL IBT are retrofitted for diagnosis of learners' weaknesses and strengths through the use of statistical procedures known as Cognitive Diagnostic Models (Alderson et al., 2015). These procedures enable estimating attributes of items in proficiency tests that tap into learners' cognitive abilities, which allows for discovering where improvement is required (e.g., Jang, 2009; Lee & Sawaki, 2009). As regards the perceived usefulness of feedback that such tests provide, Jang's (2009) study, for example, demonstrated that the majority of learners considered the feedback useful for understanding what their problems are. However, especially low-achieving learners reported that the feedback was not enough for them to learn (i.e., in their opinion, it did not have the feed forward function).

Alderson's et al. (2015) major criticism of this approach is that tests *designed* for the purpose of being diagnostic achieve the same without the disadvantages that retrofitting proficiency tests entails, such as underrepresentation, or lack, of certain attributes in test items. Instead, the authors suggested that designing new assessment tools with the primary purpose of diagnosing learners' abilities should be a better alternative.

Perhaps the most well-known of such tests is DIALANG—a free computerised diagnostic assessment system of listening, reading, writing, vocabulary, and grammar available in 14 languages (e.g., Alderson, 2005; Alderson, 2007; Alderson & Huhta, 2005; Huhta, 2010; available at http://dialangweb.lancaster.ac.uk/). The first part of the procedure is an optional vocabulary-size placement test followed by a self-assessment based on the Common European Framework of Reference (CEFR) can-do statements. The aim of the placement part is to roughly estimate learners' abilities and select a test of appropriate difficulty based on it. Then the main part of the assessment starts, with the option of receiving immediate item-by-item KCR feedback. Finally, a profile of test-taskers' performance is presented including:

- the evaluation of learners' level on the CEFR scale with a description of their level and the levels below and above and pieces of advice of how to improve the performance;
- the results of self-assessment, including a comparison of the self-assessment with the actual level;
- the results of the vocabulary-size placement test;
- and the correct and incorrect responses grouped by subskills (Alderson, 2005; Alderson et al., 2015).

Feedback is thus a unique feature of DIALANG. In fact, it stretches beyond the task level, as it also focuses on strategies, for example, the advice for improving reading skills to level B2 includes suggesting adjust their reading style based on the reason for reading.

What is especially important in the case of DIALANG feedback is that some of DIALANG test-takers' experiences with it have been studied (e.g., Floropoulou, 2002; Huhta, 2010; Yang, 2003). Floropoulou (2002), for example, found that some learners reconsidered the way they evaluated their proficiency as they were able to identify their weaknesses and strengths with the help of DIALANG. Yang (2003) found that those learners who indicated that they wanted to improve their English accepted DIALANG advice more readily. Yang (2003) also found that learners' experiences with proficiency tests seemed to have shaped their beliefs of what tests should look like, which prevented them from recognising the difference between DIALANG and proficiency tests. Finally, Huhta (2007; 2008; 2010) reported on a survey of 557 learners' experiences with feedback in DIALANG, which demonstrated, for example, that the participants found the feedback on the mismatch between their self-assessment and the actual level the least useful whereas the overall test result and classification of the items/responses by subskills, as useful. Huhta (2007) suggested that the reason for the low perceived usefulness of the self-assessment feedback could be due to the participants' unfamiliarity with such feedback. However, this can also be explained with reference to Yang's (2003) findings. This adds to the argument that learners' beliefs about feedback/assessment should be studied in order to get a deeper insight into learners' performance.

Another interesting diagnostic test is DELTA (Diagnostic English Language Tracking Assessment). It provides detailed profiles of test-takers' performance on its listening, reading, vocabulary, and grammar sections. In the profiles, subskills, vocabulary bands, or grammatical elements (depending on the test) where learners have problems are listed (similarly to DIALANG). In addition, those subskills, grammatical elements, and vocabulary frequency bands that are below the average performance of test-takers are highlighted. Further details on diagnostic profiles in DELTA can be found at http://gslpa.polyu.edu.hk/eng/delta_web/doc/Sample_Report.pdf. DELTA exemplifies the advantages of computerised modality to assess learners' abilities, which, in its case, for example, includes establishing the learners' performance that is below the average performance of all the test-takers on the fly.

So far, the discussed diagnostic instruments were static tests. However, as Ableeva (2012) noted, dynamic assessment has become one of the major approaches to diagnosis. Importantly, DA is based on Vygotsky's and Feuerstein's developmental theories (see **Chapter 3.2**), which provides a theoretical ground to diagnosis, allowing, for example, for predicting that a learner who demonstrates certain performance with mediation will be able to demonstrate the same performance independently in future. Dynamic assessment also addresses the problem of identification of the degree of specificity of diagnostic feedback that promotes the development of abilities being assessed (cf. Lee, 2015).

The dynamic assessment paradigm has certain implications for diagnosis. Naturally, in human-mediated DA, the mediator is not a neutral being but is rather an active participant in the assessment. This means that diagnosis in DA is ongoing during the whole assessment process rather than (only) presented as a post-hoc profile of learners' performance. Connected to the previous, all dynamic tests can be considered as enabling self-diagnosis, as when the difficulty of the assessment target is mediated within learners' ZPD, learners should realise what their problems are, why they occur, and what the expected performance should be. It also appears that there is little sense in assessing constructs that are outside learners' ZPD, as this will likely only result in learners' frustration rather than in learners' development (**Chapter 2.7**; Ableeva 2010; Haywood & Lidz, 2007).

One DA study that speaks in favour of using DA for diagnostic purposes is Ableeva's (2010) doctoral research project. The aims of the study were to explore the diagnostic capacity of DA and to establish how DA can promote the development of L2 English learners' listening proficiency. The study adopted the interactionist DA format. The procedure included three pretest sessions (an SA session, a DA session, and a transfer session), an enrichment programme, and five posttest sessions (one static, one dynamic, and three transfer sessions) (Ableeva, 2010: 171). The transfer sessions were mediated as well. The aim of the transfer sessions was to establish whether the learners' listening skills developed. The enrichment programme aimed at addressing the learners' problems that were identified during the pretest sessions.

Following the enrichment, there was a clear improvement in the learners' performance on the original texts, but their performance on the transfer sessions varied. Ableeva (2010) argued that this improvement was mainly due to the mediation during the DA and the enrichment programme. Importantly, through a qualitative analysis of the dyadic interactions between the learners and the mediator, the author demonstrated that the DA allowed for a more refined diagnosis of the learners' abilities than SA would have been able to do. Specifically, she found that the explicitness of mediation depended on the severity of the learners' problems and their overall listening proficiency. Overall, the author found that the learners' unassisted listening comprehension ability was rather limited, but with appropriate mediation, different for different learners and for different problems, the learners demonstrated a far better understanding of the texts.

What unites Ableeva's (2010) study with the rest discussed in the present subchapter is that in all of them, the assessed construct was general language proficiency. In the following section, basing my discussion on two particular linguistic/grammatical features in L2 English, and bearing in mind the previous discussion, I will try to suggest how a more focused construct than L2 proficiency can be defined with the intention of using it as an assessment target in dynamic diagnostic assessment.

3.4 Dynamic diagnostic assessment and the assessed construct

As the research on diagnostic assessment suggests, without carefully defining the construct, the aspects forming this construct, and the way they are interrelated, it can be difficult to decide what the reasons for the specific learners' problems are and how to approach these problems.

Ableeva (2010) addressed this issue by discovering the specific problems learners had during the pretest procedures. The interactionist approach adopted in her study allowed for noting the learners' reciprocity during the pretest and for adjusting the mediational strategies at later stages respectively. However, this is not the case with interventionist DA, and especially the computerised modality, as the decisions of how to approach learners' mistakes should be made a priori.

Thus, using SLA research findings to define and operationalise assessment targets in computerised DA seems to be a more viable option. In the two following subchapters, I will present two research areas to exemplify problems that can arise and decisions that could be made when defining and operationalising assessment targets in (computerised) DA.

These two areas, that is, L2 English questions and L2 English word derivation, were selected because the development of L2 English questions has been substantially studied (though not as a target of DA) whereas L2 English word derivation is a heavily under-researched area. I found the contrasts interesting to study from the point of view of operationalising them in DA. Specifically, I suggested that designing a DA of L2 English questions would illuminate some of the decisions regarding the operationalisation of L2 English word derivational knowledge and adaptive CF in a DA of word derivational knowledge. What is more, particularly as regards L2 English word derivational knowledge, I found it interesting to explore whether and how DA promoted it, which, I suggested, had implications for teaching L2 English derivational affixes.

A further reason for selecting the two areas was practical. That is to say, both the development of L2 English questions and L2 English word derivational knowledge were studied in the CEFLING (http://www.jyu.fi/cefling) and the TOPLING (www.jyu.fi/topling) research projects. Therefore, considering that both of my supervisors were in the project research groups, I had a better idea of some instruments and know-how in these areas.

3.4.1 L2 English questions

Perhaps the most influential work in the field of L2 that inspired further research on question development (e.g., Dyson, 2008; Spada and Lightbown, 1999) was Pienemann, Johnson, and Brindley's (1988) study, which led to the formulation of what has become known as Processability Theory. The theory was different from previous studies (see Carroll, 1998) in that it was based on a solid theoretical basis, that is, Levelt's model of speech production (e.g., De Bot, 1992) and Lexical-Functional Grammar (e.g., Horn, 2011). One of the claims Pienemann and his colleagues made was that the order of acquisition of structures in SLA depends on their processing complexity, and this hierarchy is universal in SLA. The stages in L2 English question development are presented in **Table 4**.

TABLE 4 Stages in the development of L2 English questions (adapted from Pienemann, 2005; Spada & Lightbown, 1999).

Stage 1	Single words and phrases: *How are you?*
Stage 2	SVO: *The tea is hot?*
Stage 3	Do fronting: **Do he work? Does he work?* Wh- fronting: **Where the station is?* Other fronting: **Is the boy is beside the bus?*
Stage 4	Yes/No inversion: *Has he seen you? *Have he seen it?* Pseudo Inversion: *Where is John?*
Stage 5	Do/Aux 2nd: *Why did he sell that car?*
Stage 6	Cancel Inversion: *I wonder where he has gone?*

Several studies have confirmed the availability of similar stages in syntax of learners with different L1 backgrounds and studying different languages (e.g., Di Biase & Kawaguchi, 2002; Glahn et al., 2001; Dyson, 2008), adding to the validity of the stages.

What makes the Processability Theory perspective on the development of questions a viable basis for diagnosis is that it allows for making inferences about which questions learners can and/or cannot form. It is also possible to inform teachers what question types should be taught next.

On the other hand, there is an apparent problem with adopting the concept of developmental stages to dynamic assessment. According to Pienemann's (e.g., 2005: 255-256) Teachability Hypothesis, developmental stages cannot be skipped even with help of feedback although it can have a positive effect on the rate of acquisition. However, this contradicts the assumption of the sociocultural perspective that learners' abilities are modifiable.

It is worth mentioning, though, that the stages in question development were designed for oral production, not writing. Alanen and Kalaja (2010) found the same stages in their participants' written performance. However, they also found that it was rather the frequency of questions at higher stages of the developmental scale and accuracy in their production that increased at higher levels of L2 proficiency. In addition, with regard to written performance, Spada and Lightbown (1999) found that learners who were initially at stage 2 and 3 of question development showed knowledge of stage 5 questions, but tended to do so only if subjects in the sentences were pronouns. These results speak in favour of modifiability of learners' ability to produce questions in writing.

This means that whereas stages in question development and DA rest upon different epistemologies, in practical terms, the former can be used as an assessment target in DA, the more so as it has been found that instruction can direct the development of questions (e.g., McDonough, 2005; Spada & Lightbown, 1993). The increase in the frequency of use of these structures (and the frequency of use of accurate structures) in learners' unassisted performance can be an indication of this development. On the other hand, the relevance of question development stages for making diagnostic inferences regarding learners' future performance (e.g., which question stage to teach next) based on the results of a dynamic assessment is questionable. Perhaps, the best course of action would be to focus the assessment on one particular type of questions (or a particular stage), find out what problems learners can have with these questions, and decide how adaptive CF can address these problems.

That the development of questions has been thoroughly researched has distinct advantages for diagnosis. However, it can be even more important to select an area in which learners usually have problems. One such area will be discussed in the following section.

3.4.2 L2 English word derivation

L2 English word derivation (WD) presents a problem to learners (e.g., Friedline, 2011; Schmitt & Meara, 1997; Schmitt & Zimmermann, 2002; Silva & Clahsen, 2008). At the same time, unlike the development of questions, little is known about how WD knowledge develops, what aspects the construct of L2 English WD knowledge consists of, and how its separate aspects are interrelated, which makes it problematic to design dynamic diagnostic tests of word derivational knowledge (cf. Alderson, 2005; Lee, 2015).

This said, some insights into learners' L2 English word derivational knowledge have been produced. For example, some studies have found a link between learners' L2 English WD knowledge and their L2 proficiency although the results here are mixed. Mäntylä and Huhta (2013), in a cross-sectional study, found moderate positive correlations between learners' writing proficiency and their performance on affix elicitation tasks. Friedline (2011) had mixed results as regards the relationship between learners' WD knowledge and their proficiency, finding it only for a word relatedness task, in which the learners were asked to determine whether the words in the word pairs (e.g., *decorative–decoration*) were related. Schmitt and Meara (1997) did not find significant correlations between learners' proficiency operationalised as their TOEFL scores and their performance on a productive measure of WD, in which the participants were asked to attach all the allowable suffixes to the given base forms, and receptive measure of it, in which the learners were asked to select all the allowable suffixes to the given base forms.

One explanation for the conflicting results can be that similarly to vocabulary knowledge (e.g., Ringbom, 1987; Nation, 2001), word derivational knowledge is a multidimensional construct. Thus, the measures that were used in these studies tapped into different aspects of word derivational knowledge and different combinations of them, such as, syntactic knowledge (e.g., Schmitt & Meara, 1997), semantic knowledge (e.g., Mäntylä & Huhta) and morpho-orthography (e.g., Friedline, 2011), and these all had different relationship with the learners' proficiency.

As regards the construct of WD knowledge, there is still no clear understanding what aspects/dimensions the construct of WD knowledge consists of. However, Ringbom's (1987; 1990) model of lexical knowledge (**Figure 1**) can serve a starting point in defining this construct, as it is rather detailed and is both multidimensional and developmental.

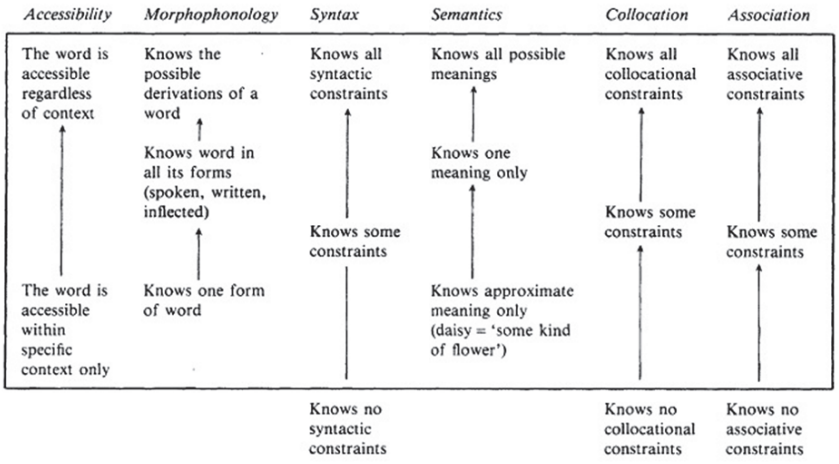

Accessibility
Morphophonology
Syntax
Semantics
Collocation
Association
The word is accessible regardless of context
The word is accessible within specific context only
Knows the possible derivations of a word
Knows word in all its forms (spoken, written, inflected)
Knows one form of word
Knows all syntactic constraints
Knows some constraints
Knows no syntactic constraints
Knows all possible meanings
Knows one meaning only
Knows approximate meaning only (daisy = 'some kind of flower')
Knows all collocational constraints
Knows some constraints
Knows no collocational constraints
Knows all associative constraints
Knows some constraints
Knows no associative constraints

FIGURE 1 Ringbom's (1987) model of lexical knowledge[1].

Still, it is difficult to say how to operationalise this model in terms of mediation in computerised DA. What is more, as Ableeva (2010: 279) noted, "inappropriate mediation can undermine learners' opportunities to develop abilities that may be ripening." Therefore, more research into the different aspects of WD knowledge (e.g., their relationship with L2 proficiency) would be required before attempting to design adaptive CF in computerised DA of L2 English WD knowledge.

In addition, the conflicting results of the studies of the relationship between the learners' L2 proficiency and their WD knowledge can also be explained by such factors as semantic transparency and frequency of the items used in the measures in these studies, but also by the difficulty of the deriva-

[1] The figure is reproduced with kind permission from Multilingual Matters.

tional affixes used in the studies. Marslen-Wilson (2007), for example, found that semantic transparency (i.e., the ease of understanding the meaning of morphologically complex words from their parts) affects the processing of morphologically complex words, that is, the more semantically transparent the word is, the easier it is analysed. Hayashi and Murphy (2010) confirmed that in their study, where all of the participants successfully segmented such words as *rewrite*, *enable*, or *disorder* in a word segmentation task, all of these words being, arguably, semantically transparent.

There is also research (e.g., Clahsen & Neubauer, 2010) suggesting that the word frequency influences the way learners process morphologically complex words, processing frequent words as wholes and attempting to analyse less frequent words. Therefore, the frequency of the items could have interacted with the learners' proficiency in these studies (especially as regards word segmentation tasks), in that the more proficient the learners were the more words were stored as wholes in their mental lexicons, making it more difficult to complete these tasks. I discuss frequency and semantic transparency in detail in Article III, suggesting a way of controlling for these two factors in research on word derivation.

As regards the difficulty of derivational affixes, it is still not clear whether and why some derivational affixes are more difficult to learn than others. A theoretical order in which L2 English affixes can be taught, based, among other factors, on frequency and semantic transparency of these affixes, proposed by Bauer and Nation (1993) can be a starting point in this research (**Table 5**).

TABLE 5 Teaching order of L2 English affixes (Bauer & Nation, 1993).

Level 1	A different form is a different word.
Level 2	Inflectional affixes, e.g., -ed, -s, etc.
Level 3	The most frequent and regular derivational affixes: -able, -er, -ish, -less, -ly, -ness, -th (fourth), -y, non-, un- (unusual)*.
Level 4	Frequent and regular affixes, e.g., -ation, -ful, -ism, -ist, -ise (-ize), -ment, in-, etc. *
Level 5	Infrequent but regular affixes, e.g., -ance, -ant, -ship, en-, mis-, un- (untie), etc.
Level 6	Frequent but irregular affixes, e.g., -ee, -ic, -ify, -ion-, re-, etc.
Level 7	Classical roots and affixes, e.g., -ate, -ure, etc.

*All with restricted uses; see **Appendix 1** in Bauer and Nation (1993) for details.

To my knowledge, only a small number of studies have tried to challenge the order (Chuenjundaeng, 2006; Mochizuki & Aizawa, 2000), and based on the results of these studies, it is hard to say whether the order holds or not. Chuenjundaeng (2006), for example, mostly used affixes at level 4 of Bauer and Nation's order. In Mochizuki and Aizawa (2000), as the authors themselves suggested, English loan words in Japanese could have influenced the learners' performance on some affixes (e.g., sub-), making them easier to recognise.

Bauer and Nation's (1993) order of affixes could be used as a reference for affix difficulty in DA of word derivational knowledge. However, basing a test on too many difficult affixes which are outside a particular learners' ZPD can result in this learner's frustration, affecting his/her performance (cf. Haywood & Lidz, 2007; Lee, 2015). Thus, in my opinion, empirical support for the order should be provided before the order can be used to operationalise affix difficulty in a dynamic assessment of learners' L2 English WD knowledge.

Furthermore, little is still known about how to instruct learners in the use of L2 English affixes, especially considering the idiosyncratic nature of L2 English word derivation. In this respect, mediating learners' use of strategies during a dynamic test (e.g., Kozulin & Garb, 2002; 2004; Teo, 2010) can be an interesting alternative (or addition) to, for example, instructing learners in what certain affixes mean.

Many questions (e.g., *What is included in word derivational knowledge?* or *How does it develop?*) should be answered before an adequate diagnosis of learners' word derivational knowledge can be provided and ways to mediate learners' performance to promote it can be established. At the same time, a DA of word derivational knowledge can, too, contribute to operationalisation of this construct. Thus, an exploration of what dynamic assessment of L2 English WD knowledge can look like can be interesting.

3.5 Dynamic diagnostic assessment and feedback

In the present chapter, I outlined research on dynamic, diagnostic, and dynamic diagnostic assessment. The purpose of the overview was (a) to get an idea what a dynamic test of L2 can look like, including possible assessment targets/constructs of such tests, (b) to find which features of useful feedback emerge from the paradigms underlying dynamic and diagnostic assessment and (c) to find out what it is important to consider in order not to hinder feedback usefulness in DA. I summarise the ways of maximising the usefulness of CF based on the research I outlined in **Chapter 2** and in the present chapter below:

- for learners to accept feedback more readily, not only teaching goals should be revealed to learners, but it should also be explained to them why they are important;
- learners' held beliefs can hinder usefulness of feedback, but feedback, being a mediational means, can help transform these same beliefs; DA where learners receive CF adapted to their abilities, can be a suitable tool for mediating these beliefs;

- getting feedback from learners in order to get a deeper insight into the ways learners use CF in computerised DA; this could be done by way of conducting surveys (cf. Huhta, 2010) or more interactively, through discussions;
- in addition to strengthening validity of DA (see **Chapter 5**), discussions with learners (but also, perhaps, questionnaires) can help transform learners' beliefs about CF (cf. Alanen, 2003; Floropoulou, 2002);
- it is important to carefully define assessment targets, including the aspects forming the constructs, and the way these aspects are interrelated (e.g., Lee, 2015);
- to do so, computerised DA of L2, where decisions about mediation of learners' performance are made a priori, should be informed by SLA research and theory;
- a multidimensional view of a construct presupposes that some aspects of the construct can be more developed than others; thus feedback should both list the problems that learners have and what they do right;
- assessment target(s) should be somewhat, but not excessively, beyond learners' unassisted performance (Ableeva, 2010; Haywood & Lidz, 2007; cf. Hattie & Timperley, 2007).

The research summarised in **Chapters 1** and **2** allowed for singling out features of feedback that should promote learners' abilities. In addition, several research gaps were identified, which motivated the present doctoral research.

Specifically, it appears that:

a) experimental research on the effect of adaptive corrective feedback is lacking;
b) not much is known about the way learners' beliefs about corrective feedback and testing/assessment mediate their DA performance and how DA, in its turn, mediates learners' beliefs;
c) not much research has been conducted specifically targeting the operationalisation of assessed constructs in computerised DA although some aspects to consider can be generalised from the available research on DA and diagnostic assessment;
d) it is not clear how to design a computerised dynamic assessment of an under-researched construct, such as L2 English word derivational knowledge, when it is not even known how this construct develops.

In the following chapter, I will describe the studies conducted to address these gaps, listing the aims and the research questions of the present doctoral research, as well as the participants, the data, and the analyses of the separate studies.

4 METHODOLOGY

In the overview of the previous research on corrective feedback and dynamic assessment, I suggested that while the sociocultural perspective on corrective feedback can account for the mixed results that the research on static CF has produced, the research on corrective feedback within the sociocultural paradigm has been predominantly qualitative, above all, as it is focused on the process rather than the product of interaction. I then suggested that interventionist DA (due to being more psychometrically oriented than interactionist DA), especially its computerised modality (due to its increased practicality), can be a suitable tool for providing quantitative evidence for the superiority of adaptive over static CF in promoting learning of L2 English, thus producing a stronger argument for this approach to giving feedback.

However, as the previous research has demonstrated, dynamic assessment of L2 has been under-researched and although several approaches to DA have been developed over time, little is known, for example, about how to define and operationalise the assessed constructs in DA (especially in interventionist DA), how learners perceive dynamic assessment, especially computerised DA (e.g., as a conventional test, as teaching), and how their beliefs about corrective feedback direct their DA performance (e.g., promoting or hindering it). The gaps identified through the study of the previous research (**Chapters 2** and **3**) inspired the aims of the dissertation, as outlined in **Chapter 1.1**.

As suggested in **Chapter 3**, the research on diagnostic assessment provides some directions for defining the assessed construct and operationalising mediation in DA. One feature of diagnostic assessment that I found particularly important for computerised dynamic assessment is that it should be based on SLA theory and research. Thus, the computerised dynamic diagnostic tests in the present study, which focused on specific grammatical/linguistic features, were designed based on SLA research and theory. Specifically, the operationalisation of the assessed constructs are based on (a) the previous research outlined in **Chapter 3.4,** (b) the findings of the *CEFLING* and the *TOPLING* research projects, and (c) the findings of the two studies conducted as a part of the present doctoral research project (**Chapters 4.3.3** and **4.3.4**).

These tests were designed following the interventionist sandwich DA and were evaluated in order to address the aims the study, that is, to establish the impact of computerised DA (including its effect on learning and facilitation of transformation of learners' beliefs) and to study empirically whether and how some general principles of computerised DA (including the importance and peculiarities of defining and operationalising the assessed constructs and promotion of strategic learning) (see **Chapter 1.1**) resulted in the increase of the positive impact of the tests on the abilities being assessed and on learning in general. These aims were expanded and formulated as several detailed research questions, which I will present in the following.

4.1 Aims and Research questions

To reiterate, the overall aim of the present research was to study the impact of adaptive corrective feedback. Specifically, I aimed at:

a) collecting experimental evidence for the effect of adaptive CF;
b) studying the way learners' beliefs guide their DA performance and the experience of DA, in turn, mediates their beliefs; and
c) studying empirically whether (and how) the features of adaptive CF synthesised from the previous research and the present doctoral research project allow the feedback to promote learners' performance on tasks requiring them to demonstrate their L2 English WD knowledge.

As it has already been mentioned, these aims did not take shape simultaneously, but rather emerged in the course of the research process. In order to realise the aims, the following research questions were posed:

1. Does automated adaptive corrective feedback provided during a computerised diagnostic test facilitate the development of learners' ability to form L2 English questions?
2. Based on learners' experiences with the dynamic diagnostic tests and, specifically, with the feedback designed in this study, how do learners' beliefs about corrective feedback and their performance on / experience of dynamic assessment mediate one another?
   2.1 How do learners perceive the usefulness of the corrective feedback in the study, what mediates this perception, and how does this perceived usefulness of the feedback mediate their performance on dynamic test?
   2.2 How does experience of dynamic assessment help to transform learners' beliefs about the usefulness of corrective feedback?
3. Based on the findings pertaining to research questions 1 and 2 (but also the previous research), what are some ways of ensuring the usefulness of automated adaptive CF in a computerised dynamic assessment of an under-researched construct, such as L2 English word derivational knowledge?
   3.1 What general principles for designing and administering computerised dynamic tests emerge from the studies of L2 English questions and of learners' beliefs about corrective feedback?

3.2 What is characteristic of the construct of L2 English word derivation?
   3.2.1 Are some L2 English derivational affixes more difficult than others?
   3.2.2 What is the relationship between L2 English word derivational knowledge, and its different aspects, and learners' L2 proficiency?
3.3 How, if at all, does automated adaptive corrective feedback guide the development of L2 English word derivational knowledge?

It should be noted though that these research questions are not those asked in the separate articles forming the present dissertation although, largely, the research questions posed in the articles can be perceived as sub-questions to the research questions above (see separate articles for the corresponding research questions).

This allowed me, on the one hand, to largely address each of the research questions of the present research in a separate article and, on the other hand, to support the findings with the results reported on in other articles, thus considering the same questions from different perspectives and using different data and methods, which I will detail in **Chapters 4.3** and **6**. Before presenting the methods used is separate articles, however, to make it easier for the reader to follow the research process more clearly, I will present the research questions posed in the current research in relation to the articles and will also position the studies on the timeline.

4.2 Research questions in relation to the original publications

The present doctoral research consisted of several separate (though related) studies which all ultimately added to the major aim of the study, that is, increasing the understanding of what adaptive corrective feedback (as a kind of mediation) in computerised dynamic assessment can look like and what its impact is. Thus, the compilation dissertation format was found to be the most appropriate, as it allowed for concentrating on one particular aspect in each article. This is not to say, though, that the findings from only one article will be used to answer each of the research questions (see **Table 6**).

TABLE 6 Research questions in relation to the original publications.

Research questions		**Article**
RQ 1		*Article I*; Article II; Chapter 5
RQ 2	RQ 2.1	*Article II*; Article I; Chapter 5
	RQ 2.2	*Article II*; Article I
RQ 3.1		*Article I; Article II; Chapter 5*
RQ 3.2	RQ 3.2.1	*Article III*
	RQ 3.2.2	
		Article IV
RQ 3.3		*Article V*

The article number in italics opposite each of the research questions was the primary source for finding answers to these questions. In addition, findings reported on in the papers listed following the main article were used to support the answers. Chapter 5 mentioned in **Table 6** refers to the chapter of this synthesis that covers the process of the validation of *Questions Test* and the *ICAnDoiT* system. It was decided not to publish the process of their validation in a separate article, as the validation did not explicitly relate to the aims of the study. At the same time, I felt that the extent to which the creation and validation of the test were covered in the articles forming the present dissertation was insufficient.

This does not mean though that the studies conducted in order to provide answers to the research questions followed in the order the articles are referred to. Rather, the research process can be visualised as illustrated in **Figure 2**.

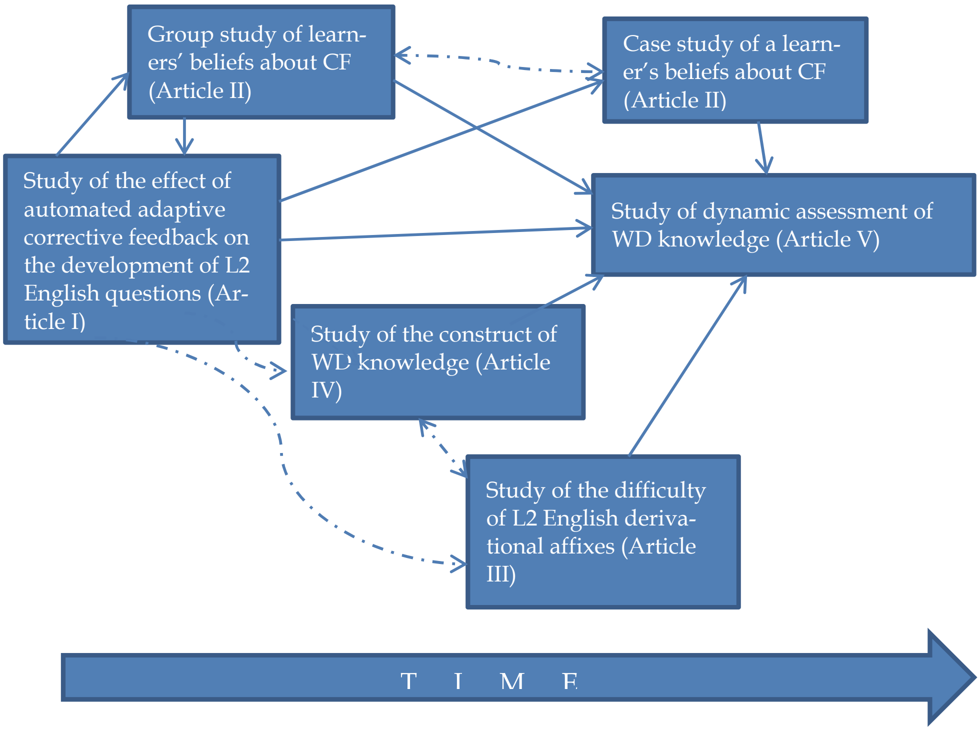

Group study of learners' beliefs about CF (Article II)
Case study of a learner's beliefs about CF (Article II)
Study of the effect of automated adaptive corrective feedback on the development of L2 English questions (Article I)
Study of dynamic assessment of WD knowledge (Article V)
Study of the construct of WD knowledge (Article IV)
Study of the difficulty of L2 English derivational affixes (Article III)
T I M E

FIGURE 2 Timeline of the studies.

The solid arrows represent stronger links between the studies, where findings of the studies were used in / inspired other studies or helped to interpret the findings in these studies. The dotted arrows indicate weaker links, that is, a similar method and analysis used in the Group and the Case study of learners' beliefs about CF and the results of the study of the effect of adaptive CF on the development of L2 English questions corroborating my decision to study the construct of L2 English word derivational knowledge.

4.3 Method

Generally speaking, the present doctoral research project adopted a mixed-methods approach. In essence, mixed-methods research is a combination of qualitative and quantitative research philosophies, methods, and/or concepts in a single research project (Dörnyei, 2007; Sullivan, 2009; Teddlie & Tashakkori, 2010). Dörnyei (2007: 42) noted that this type of research has been referred to under various names, such as multitrait-multimethod research, interrelating qualitative and quantitative data, mixed-methods research, and even methodological triangulation. As Dörnyei (2007) noted, this approach has evolved into the third major research paradigm, integrating the qualitative and quantitative paradigms.

Among the advantages of mixed-methods research is a better understanding of phenomena being researched. Importantly, as Teddlie and Tashakkori (2010) noted, it is not just for confirmation of results that mixed methods are used for but also for finding the dissimilarities in the results obtained by different methods.

It should not also be assumed that only an equal contribution of qualitative and quantitative methods should be considered a mixed-methods research. Teddlie and Tashakkori (2010) suggested that this perception of the paradigm should rather be replaced with a continuum view, where purely qualitative and quantitative research represents the ends of the spectrum of research methods.

Several typologies of mixed-methods research according to the relative contribution of qualitative and quantitative methods were proposed. Johnson and Christensen (2008), for example, proposed a typology of mixed methods depending on the assigned priority of the qualitative and quantitative methods in the research design and whether the data collection using the methods from these two paradigms is sequential, that is, in different stages of the research process, different methods are used (e.g., a survey an a follow-up interview) or parallel, that is, different research strands are conducted simultaneously, and are at some point (usually during the interpretation of the results) converge to enrich the overall picture.

As regards what can be considered mixed-methods research, Dörnyei (2007), for example, discussed research including *quantifying* qualitative data or *qualitising* quantitative data as a variant of mixed-methods research. The former, he suggested, refers to producing numerical representations of certain aspects of the otherwise qualitative data (e.g., frequency of themes). The latter is less common and refers to studying quantitative data qualitatively (e.g., using a background questionnaire to inform the interpretation of the qualitative data).

In the present doctoral research project, the choice of the research paradigm for each of the studies reported on in the articles forming the present doctoral dissertation was informed by their aims and the research questions. Thus, the study reported in Article I (see **Chapter 4.3.1**), which aimed at finding experimental evidence for the positive effect of adaptive corrective feedback adopted the experimental research paradigm and most of the data collected (i.e., learners' performance on the pretest and the posttest and questionnaire responses) and the analyses of the data were quantitative.

On the other hand, the two studies reported on in Article II aimed at understanding how learners' beliefs about corrective feedback guided their DA performance and how their DA experience, in its turn, mediated their beliefs. Therefore, the interviews were the primary data collection tool in the two studies. The qualitative data analyses, including those reported on in Article II, were informed by the sociocultural paradigm.

In other words, what the learners reported in the two studies reported on in Article II (and other studies) was considered to be mediated in one way or another. Specifically, in the analysis, I considered the learners' utterances as a type of mediated action (Alanen, 2003; Wertsch, 1991; 1998; **Chapter 2.8**).

However, this outlook also had implications for interpreting the results obtained in the study reported on in Article I, despite the latter designed following the experimental research paradigm. That is, I assumed that the learners' performance during the computerised DA would be mediated both by the adaptive feedback and by their beliefs.

Certainly, it can also be assumed that the performance of the participants in the studies aiming at adding to the understanding of the construct of L2 English WD knowledge reported on in Articles III (**Chapter 4.3.3**) and IV (**Chapter 4.3.4**) were also mediated by something other than their WD knowledge and the tasks per se. However, for practical reasons and because this was studied in detail in Article V (**Chapter 4.3.5**), the two studies reported on in Articles III and IV adopted exploratory quantitative design.

Getting back to the interpretation of the results reported on in Article I, the findings reported on in Article II helped to confirm the interpretation of the results reported on in Article I. That is to say, it was towards the stage of the interpretation of the results of the present doctoral research project that I decided that interpreting the results of one study without referring to the findings of the rest of the studies would be underusing the opportunities what a combination of qualitative and quantitative research methods allowed for. Thus, I also interpreted the findings of the two studies with reference to one another. Similarly some findings of Article V were interpreted with reference to the findings of the Case study in Article II. In fact, the Case study reported on in Article II and the study reported on in Article V can be considered as one strand of research, where the impact of DA on the learner's beliefs and strategic learning was studied. In other words, combining the findings of the separate articles allowed for a better interpretation of the results of the present doctoral research project.

That said, it is not just the research process taken as a whole that can be considered to have adopted the mixed-methods research paradigm. Some of the separate studies also had elements of both the qualitative and the quantitative research paradigms, as I will elaborate on in the following.

4.3.1 Summary of Article I

Leontjev, D. (2014). The Effect of Automated Adaptive Corrective Feedback: L2 English questions. *APPLES: Journal of applied language studies, 8*(2), 43-66. Retrieved from http://apples.jyu.fi/ArticleFile/download/459.

This article was designed to address the lack of experimental evidence for the effect of corrective feedback provided within learners' ZPD. The aim was to compare the effect of automated adaptive corrective feedback on learners' ability to form L2 English wh-questions with auxiliaries (stage 5 in the order in question development; see **Chapter 3.4.1**) with that of static knowledge of results feedback. The study also aimed at finding out whether learners generally considered adaptive feedback as more useful than knowledge of results feedback.

The participants were a total of 47 learners of English at grade 8 studying in four different groups taught by two teachers. Eight-graders were one of the target populations of the substudy of questions in the CEFLING project (e.g., Alanen & Kalaja, 2010). Thus, it was decided to recruit the participants at this grade, so that they had not mastered stage 5 questions by the time of the study but the structure had been in their ZPD (also **Chapter 5.2.2**). The learners were randomly assigned to two treatment conditions: dynamic assessment (experimental group, n = 26) and static assessment (control group, n = 21). The learners' performance on two exercises, an E-mail writing according to the prompt (k = 8) and a gap-filling exercise (k = 9), was measured before and after the treatment. The treatment exercises included two ordering and three multiple-choice exercises. The validation of the exercises will be discussed in **Chapter 5**. Both groups completed the same pretest and the posttest. The treatment exercises were also the same for both groups. The difference between the two conditions was the feedback: the experimental group received automated adaptive CF, whereas the control group received knowledge of results feedback. The adaptive feedback messages were arranged from more implicit and less detailed to more explicit and more detailed, an order similar to Aljaafreh and Lantolf's (1994) Regulatory Scale. The tasks were administered in the *ICAnDoiT* system. The performance of the two groups on the pretest/posttest tasks was then compared statistically using t-tests to compare the increase in performance (the difference between the pretest and the posttest performance) of the two groups and Wilcoxon signed-rank tests to analyse the change in the learners' performance within the two groups. In addition, a questionnaire was administered to the learners aiming to find out their perceived usefulness of feedback. The two groups' questionnaire results were studied and compared both qualitatively and quantitatively.

The quantitative data analysis aimed at finding out whether the two groups rated the usefulness of the feedback they received during the treatment significantly differently (using nonparametric Mann-Whitney U and Chi-square tests). The open-ended questionnaire items (e.g., *How did the hints help you?*)

were studied for recurrent patterns in the learners' responses to find out if and in what way these were different in the two groups. The qualitative element in the otherwise quantitative study was introduced to illuminate possible reasons for the differences in the perceived usefulness of the CF between the two groups.

4.3.2 Summary of Article II

Leontjev, D. (2016). Exploring and reshaping learners' beliefs about the usefulness of corrective feedback: A sociocultural perspective. *ITL International Journal of Applied Linguistics*, *167*(1), in press.

The aim of Article II was to understand how learners' beliefs about the usefulness of corrective feedback emerge and start transforming in social interaction following learners' experience of dynamic assessment (both human-mediated and computerised). To fulfil this aim, two small-scale studies were conducted. The Group study came first and was informed by the question that arose in the study reported on in Article I, that is, what it was that made some learners in the experimental group consider some (or all) of the feedback in the study useless. Considering the research outlined in **Chapters 2.8** (e.g., Amrhein & Nassaji, 2010; Ashwell, 2000; Leki, 1991) and **3.3** (e.g., Yang, 2007), I suggested that learners' beliefs about CF mediated their DA performance. Thus, I decided to first study the issue cross-sectionally (considering that I had the participants in the study reported on in Article I to select from) and then follow it up with a longitudinal Case study. Thus, although in Article II, the Group study followed the Case study, in the following, I will present the studies in chronological order.

In the Group study, the aim was to understand how drawing on the experience with DA in social interaction (constrained by the interview activity) can change the way learners formulate their utterances about the usefulness of CF over a short period of time—during one research interview. The participants in the Group study were six learners selected among the participants in the study reported on in Article I based on their unassisted performance on the pretest tasks (see **Chapter 4.3.1**) and on their teacher's evaluation of their abilities. Specifically, two high-achieving (HA1 and HA2), two middle-achieving (MA1 and MA2), and two low-achieving (LA1 and LA2) learners were selected and divided into two groups, a group of two high achievers and one middle achiever (MA1) and a group of two low achievers and one middle achiever (MA2).

Research interview was the main data collection tool in the study. The two groups of learners were interviewed on the following day after the DA (see **Chapter 4.3.1**).

The aim of the Case study was similar to that of the Group study. However, the changes in the learner's beliefs were traced longitudinally. For the study, it was decided to recruit a learner studying English at grade 10 in a school in Estonia. This decision was informed by the assumption that by this grade, learners should have reached level B1 of their L2 English proficiency on the CEFR scale, when L2 word derivation has been advised to be taught. Nation (2001), for example, suggested that derivational affixes can be taught to learners at the lower-intermediate level, which corresponds to B1 on the CEFR scale (Council of Europe, 2001). Furthermore, the Case study was conducted after the data in the study reported on in Article IV were collected and an initial data analysis was conducted. This analysis also suggested that a learner whose proficiency on the CEFR scale is at level B1 would be a suitable candidate for the study (see **Chapter 6.3**). As regards the particular grade at which the learner was selected, it is indicated in the Estonian State Curriculum that learners' proficiency in English as the first foreign language should generally be at level B1 on the CEFR scale.

The participant in the Case study, M, an L1 Russian learner of English whose utterances about the usefulness of corrective feedback were collected in three interviews, one before, one in a week after, and one six months after three weekly human-mediated DA sessions (see also **Chapter 4.3.5**), their target being L2 English word derivation. At the onset of the study, M studied at grade 10 of a school in Estonia.

In both the Group study and the Case study, the emphasis was on finding out the learners' beliefs about corrective feedback emerging in the interaction with the interviewer and other learners and the way these transformed, or started transforming, in the course of this interaction. The unit of analysis in both studies was mediated action (**Chapter 2.8**), specifically the learners' utterances.

When transcribing the interviews, I also noted the learners' intonation, pauses, and the degree of agency transparent in their utterances. Following the contextual approaches to the study of learners' beliefs (Chapter 2.8), in the analysis I noted how the participants in the interaction, including the interviewer, co-constructed the context in which the learners revealed and challenged their beliefs about corrective feedback, including whether and how they used the utterances from earlier in the interviews and what other agents they brought into their utterances. The secondary data (i.e., the learners' performance on the dynamic assessment and, in the Group study, learner questionnaire and teacher interview) were used to better interpret what the learners reported and overall, to produce a richer picture.

If the study reported on in Article I and the Group study in Article II are considered together, it is not straightforward whether to perceive them as two strands of research conducted concurrently or sequentially. On the one hand, the selection of the participants was, in part, based on the unassisted performance of the group they were sampled from (those participating in the study reported on in Article I). In fact, the results of the piloting of the DA procedure used in the study in Article I (see **Chapter 5**), and this piloting can be perceived as a preparation for the study of the effect of adaptive corrective feedback on learning, inspired the studies reported on in Article II. In addition, having confirmed the participants in the Group study with their teacher, I looked through their DA performance logs and questionnaire responses—which were a part of the data in Article I. So in this sense, these two studies can be perceived as conducted sequentially, the results of the study of the effect of adaptive corrective feedback informing the procedure in the Group study.

On the other hand, strictly speaking, the Group study started in the middle of the data collection in the study reported on in Article I (i.e., after the treatment but before the posttest), and to an extent, the data analysis in the two studies as well as the interpretation of the results happened in parallel, the results obtained in the two strands of research informing one another. In this sense, these two research strands can also be jointly considered as a parallel mixed-methods design. However, what I think is important, regardless of the interpretation of the overall design, is that the synthesis of the two studies allowed me to obtain a deeper understanding of the results and interpret them more fully, as I will detail in **Chapter 6**.

As regards other mixed-methods research elements in the Group study, I used both the qualitative and the quantitative data to create the learners' profiles (see Appendices **C** and **D** in Article II). The latter data included the learners' responses to the Likert-type and dichotomous questionnaire items and the learners' DA performance (i.e., the level and the number of times they received the feedback and the time they spent reading the feedback). This allowed me to create a better overview of the learners' beliefs about CF prior to the interview and also to interpret their utterances during the interview with reference to their questionnaire responses.

4.3.3 Summary of Article III

Leontjev, D. (2016). L2 English Derivational Knowledge: Which Affixes Are Learners More Likely to Recognise? *Studies in Second Language Learning and Teaching*, *6*(2), in press.

The aim of Article III was to find empirical evidence for (or against) Bauer and Nation's (1993) teaching order of derivational affixes. A motivation for the study was that a confirmation of the order would (a) make the diagnostic feedback provided on the basis of learners' performance more meaningful, as one could suggest instructing learners in the use of easier affixes first (cf. Bauer & Nation, 1993) and (b) to manipulate the difficulty of the tasks/items in the test, designing the test so that more difficult items appear later (see Article I and **Chapter 5.2.1** for a similar decision made for the *Questions Test*). That is to say, the results of the strand of the present doctoral research regarding the role of adaptive corrective feedback in the development of L2 English questions inspired the two studies of the construct of word derivational knowledge (i.e., the study reported on in Articles III and IV).

Due to the complex nature of the construct, as I discussed in **Chapter 3.4.2**, I limited the aim of the article to finding evidence for Bauer and Nation's order as a/the difficulty order of recognising these affixes. Limiting the study to recognition only reduced the generalisability of the results. On the other hand, Bauer and Nation (1993) stated that the order of affixes they proposed should reflect the ease/difficulty of recognising the affixes while reading. What is more, considering the lack of studies confirming or disproving the order empirically, any findings for or against the order could help to operationalise the assessed construct in dynamic assessment of learners' word derivational knowledge. It should be stressed at this point that the intention was not to use the findings of the study later in the present research project as a way of establishing an order of acquisition of derivational affixes (which would have little use from the point of view of a sociocultural perspective on development). Rather the findings were planned to be used to make sure that earlier in the DA procedure, those affixes appeared that learners would be more likely to recognise as affixes (see also **Chapter 4.3.5**).

I studied the learners' unassisted performance on a word segmentation task. The task included a list of words containing 12 derivational affixes at each of Bauer and Nation's affix levels 3 to 6. In addition 6 distractors were included, that is, words that did not contain any derivational affixes. The task consisted of a total of 50 words, of which 44 were formed with the help of a total of 48 affixes, among them 10 prefixes. I reduced the possibility that the words were known to the participants by selecting lower frequency words as the items. The learners were also asked to write definitions or translations of any of the words in the task that they knew.

Initially, the participants in the study were 76 learners for English, which average proficiency level was B1 on the Common European Framework of Reference scale, operationalised as the median across the learners' self-evaluation of their writing and reading ability (using CEFR level descriptors) and their teachers' evaluation of their writing and reading ability. However, 14 learners supplied more or less accurate translations or definitions of one or several items (or their bases). Therefore, to account for frequency effect (Clahsen & Neubauer, 2010) and the effect of semantic transparency (Marslen-Wilson, 2007), their performance was not considered.

In the analysis, I grouped the affixes in the task by Bauer and Nation's (1993) levels, the number of affixes at each level forming a separate variable. I, then, conducted a repeated measures ANOVA to establish whether there were significant differences between the numbers of affixes the learners were able to recognise at different Bauer and Nation's levels.

4.3.4 Summary of Article IV

Leontjev, D., Huhta, A., & Mäntylä, K. (forthcoming). Word derivational knowledge and writing proficiency: How do they link? *System*. doi: 10.1016/j.system.2016.03.013

The aim of this cross-sectional exploratory study was to establish whether and in what way learners' ability to derive words in L2 English is related to their English proficiency. The hypothetical aspects of the construct of L2 English word derivational knowledge were informed by Ringbom's (1987; 1990) model of lexical knowledge. This study allowed for determining which of these aspects related to learners' proficiency. Thus, it aimed at promoting the understanding of the construct of L2 English word derivational knowledge, which had relevance for designing adaptive corrective feedback for a dynamic test of learners' word derivational knowledge. That is to say, if, for example, learners' syntactic knowledge of derivational affixes grows as their proficiency grows, it can be assumed that mediation eliciting syntactic role of affixes should promote the rate of this development.

The participants in the study were a total of 117 L1 Finnish, Estonian, and Russian learners of English in their tenth year of school (upper-secondary education) in Finland and Estonia. To measure the learners' word derivational knowledge, a battery of tasks was used: three measures designed earlier (Mäntylä & Huhta, 2013) and six measures designed specifically for the study (see **Appendix A** in Article IV). The measures were designed or adapted for the computerised delivery in the *ICAnDoiT* system.

To estimate the participants' proficiency, two writing performance samples were collected from each learner, rated by two raters independently on the CEFR scale, and analysed with Facets software. The fair average figures across the two samples and two raters' evaluation served as a measure of their proficiency. For practical purposes, in some of the analyses, the figures were rounded back to proficiency levels on the CEFR scale.

To establish the relationship between the measures and the learners' proficiency, correlational analyses were conducted. Following that, to determine whether there was a more rapid increase in the learners' performance on the measures of their WD knowledge at a particular proficiency level, a series of one-way ANOVAs were run. Finally, a linear regression analysis was conducted to find out which of the measures predicted the learners' proficiency.

4.3.5 Summary of Article V

Leontjev, D. (2016). Dynamic assessment of word derivational knowledge: Tracing the development of a learner. *Eesti Rakenduslingvistika Ühingu aastaraamat [Estonian Papers in Applied Linguistics]*, *12*, 141–160. doi: 10.5128/ERYa12.09

This article reported on a case study proposing a procedure for computerised dynamic assessment of L2 English word derivational knowledge and exploring the way DA, both human-mediated and computerised, promoted one learner's word derivational knowledge. The general aim of the study was to find out whether and what DA features informed both by the previous research and the present doctoral research resulted in a DA procedure that promoted learners' ability to derive words in L2 English. The study, thus, built on all of the previous studies conducted as a part of the present doctoral research project. For example, the findings reported in Article III allowed for manipulating the difficulty of the derivational affixes in the computerised dynamic test. That is to say, in earlier DA sessions, easier affixes were used (see **Chapter 6.2.1** for the results pertaining to the decision regarding the difficulty of the affixes). In addition, the same affixes used in different items were used at later DA sessions as transfer items (e.g., Poehner & Lantolf, 2013). The findings reported on in Article IV al-

lowed for deciding which aspects of word derivational knowledge should be elicited in the adaptive corrective feedback provided to the learner.
However, and more importantly, considering the lack of a clear understanding of how L2 English word derivational knowledge develops (see **Chapter 3.4.2**) and the idiosyncratic nature of L2 English word derivation, the study aimed at finding out how DA promoted strategic learning. This, according to Hattie and Timperley (2007) and judging by such studies as Kozulin and Garb (2002), could enable the improvement beyond the mediated performance (see also **Chapter 3.4**). The participant in the study was M (see **Chapter 4.3.2**).

The corrective feedback in the study was based on the feedback design used in Article I. In both DA modalities, depending on the mistakes, the adaptive CF was designed to elicit both the syntactic roles and semantics of the derivational affixes in the items (see **Chapter 6.3.2**). However, in the human-mediated DA, to discover whether the designed order of feedback could be improved, some variation in level of detail was present (e.g., *which part of speech do we need here*? versus *the suffix you added forms nouns but think what the suffix we need means and what part of speech it forms*). While the mediation did not explicitly instruct the learner to analyse the words morphologically, that is, not saying, for example, *you should find the affix and the base*, it still presupposed that the learner does so in response to mediation.

In addition, as discussed in **Chapter 4.3.2**, informed by the findings of the Group study (Article II), the participant's beliefs about corrective feedback were discovered and mediated by the interviewer, who elicited the participant's experience with the human-mediated DA prior to administering the computerised dynamic test. This was done in order to reduce the possibility that M's beliefs hindered his performance on the computerised DA.

The major data collection tools in the study were think-aloud protocols (during the static assessment) and research interviews (immediately following it). During the data collection, the interviewer elicited the learner's use of strategies. M's performance on both the static and the dynamic assessment sessions was also analysed. It should be noted that M's performance on the human-mediated DA was a part of the data in both the Case study in Article II and the present study.

The overall procedure was the following:

- a static assessment session;
- three weekly human-mediated DA sessions;
- a static assessment session;
- a year and a half gap;
- a static assessment session;
- three weekly computerised DA sessions;
- a static assessment session.

The static assessment tasks were taken from the battery of tasks used to collect the data in Article IV. The dynamic assessment tasks were designed for the study. The data analyses were predominantly qualitative. Quantitative data (M's scores on the static assessment tasks) were collected and studied to confirm that M's WD knowledge developed due to the DA. Since this was the study of only one learner, no inferential statistics were calculated, that is, only the raw numbers of affixes across the SA sessions were compared. In the study, a decision had to be made whether to concentrate on strategies, that is, the product, in line with studies like Nassaji (2003) or more generally or the self-regulatory processes using the framework of, for example, Tseng et al. (2006). As Nassaji (2003) provided a classification that could be easily adapted for word derivation and since, in my opinion, a change in the learner's use of strategies (i.e., products) would allow for making a stronger case for the influence of DA, I decided to, above all, concentrate on M's strategies and knowledge sources. However, I both studied the strategies that were more beneficial for M's performance and noted more qualitative changes, for example, in the way M used these strategies when working on the static assessment tasks and also interpreted the data with reference to Tseng's et al. (2006) framework.

While conducting the initial data analysis, I noticed that simply classifying M's strategies and knowledge sources might not be revealing as regards the changes in M's strategic learning, as the types of strategies and knowledge sources used by M at the onset of the study were the same as those he used after the DA. Thus, I noted the contexts in which M used these strategies / knowledge sources and which combinations of them he used and also counted the frequencies of the separate strategies / knowledge sources. The latter was a case of quantifying the data (cf. Dörnyei, 2007), which, together with the M's scores during the static assessment, formed the quantitative element of the study and, as I will demonstrate in **Chapter 6.3**, allowed for tracing the changes in M's strategic learning due to the DA more straightforwardly and clearly.

4.4 Ethical considerations

Before closing the methodology chapter of the present synthesis, I will mention the way ethical issues were addressed in the present doctoral research project.

Before data collection, consent was obtained from the learners, their teachers, and school administration to use the data obtained from the learners, and, in some cases, the teachers for the purposes of research and reporting. The parties were informed about the aims of the research (to the extent it did not hinder the validity of the procedures; see, e.g., Chapter 5.2.1), the procedures, and the ways the data will be used.

Regarding the latter point, it was explained that for the most part, the learners' group performance would be analysed; thus, no performance that could be associated with individual learners would be reported in these cases. The online tools used for data gathering, that is, the online questionnaire and the *ICAnDoiT* system used secure encrypted connections. What is more, even in the cases where some individual learners supplied their names rather than the codes they were assigned, these were replaced by the codes in the databases before the analysis started. In the cases when the individual learners' performance was analysed and reported on (as in Article II), it was emphasised that the learners' (and the teachers') names would not be revealed and no sensitive personally identifiable information (except for gender) would be reported. The performance of those learners who did not give their permission to use their data for research purposes was excluded from the analyses.

In this chapter, I outlined the methods used in the studies forming the present doctoral dissertation and mentioned the ethical considerations of the present doctoral research project. In the following, I will summarise the process of the validation of the *ICAnDoiT* system and the *Questions Test*. Above all, the validation of the test and the system served as a part of method validation in the studies where the *Questions Test* and the system were used as data collection tools. However, some of the findings used for the validation will also be used to corroborate the findings reported in the articles forming the present doctoral dissertation.

5 VALIDATION OF THE *ICAnDOIT* SYSTEM AND THE *QUESTIONS TEST*

In the present chapter, the process of the validation of the *ICAnDoiT* system and the *Questions Test* (Articles I and II) will be discussed. The *ICAnDoiT* system was designed as a web-based tutoring/assessment system in which dynamic tests could be compiled (**Chapter 4**). The pre-/posttest tasks of the *Questions Test* were E-mail writing according to the prompts and a gap-filling task (**Appendix 1** of Article I). The treatment task types were ordering tasks and multiple choice tasks (**Appendix 2** of Article I).

In this chapter and elsewhere, I will use the terms **usefulness** and **validity** interchangeably. It should also be noted that discussing the usefulness of the adaptive CF in the study, I, in effect, discuss its validity, above all its impact (which, within the validation framework discussed in the present chapter, is closely linked to other aspects of usefulness; see **Chapter 5.1.2**). I decided to utilise the validation framework similar to the one used by Huhta (2010), which is mainly based on Bachman and Palmer's (2006) framework but also incorporates some ideas of Messick (1989) and Weir (e.g., 1993). I will mostly use Bachman & Palmer's (2006) framework, as I found the latter the most clearly structured and practically oriented. The fact that Bachman and Palmer (2006) provided detailed examples of utilising the framework added to my decision. One major difference of the framework I use from the framework used by Huhta (2010) is that I studied the **usability** aspect (Fulcher, 2003) separately.

In the following subchapter, I will present the theoretical framework I used for the validation. Following that, I will discuss the validation of the system and the test proper.

5.1 Validity and test validation frameworks

Messick (1989: 13) defined **validity** as an "evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores or other modes of assessment" (emphasis in the original). That is to say, it is not test performance (e.g., scores) that are to be validated but rather the inferences made based on this performance.

Among the existing validation frameworks, Messick's (e.g., 1989) and Bachman and Palmer's (1996) are the most well-known although argument-based validation (e.g., Kane, 2006) has also been used quite often, including the validation of TOEFL (e.g., Chapelle, Enright, & Jamieson, 2008). In what follows, I will give a short overview of Messick's framework and discuss Bachman and Palmer's framework in some detail but also mention Weir's (e.g., 1993) **a priori** and **a posteriori** validation, as I used this distinction for organising the evidence for the validity of the *ICAnDoiT* system and the test.

5.1.1 Messick's validation framework

Similarly to Cronbach and Meehl (1955), Messick considered validity to be an integrated, though multifaceted concept. His framework consists of two major facets: the basis of justification consisting of **evidence** and **consequences** and function or outcome consisting of test **interpretation** and **use** (**Table 7**).

TABLE 7 Validity facets (Messick 1989: 20)

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity + Relevance/utility
CONSEQUENTAL BASIS	Value implications	Social consequences

Messick (1989) specified that the **evidential basis** of **both test interpretation** and **test use** is **construct validity** proper, but in the latter case, it is supported by evidence for the relevance of the instrument for the intended purpose. Messick (1998) stressed that the **consequential basis** of test interpretation appertains to unintended consequences of otherwise valid test use and interpretation. These include personal and/or social values evoked by the interest of the test-designer in the construct and labels attached to that construct as well as the values imposed by the theory underlying the construct and more globally, by social ideologies that influenced the development of this theory (Messick, 1989). The **consequential basis** of **test use** involves **consequences** the test has for society (and separate individuals). It should be mentioned here that Messick (1998) considered consequences of the misuse of the test to be irrelevant for the validation process.

5.1.2 Bachman and Palmer's aspects of test usefulness

Bachman and Palmer's (1996) validation framework is somewhat similar to Messick's framework and is inspired by it. Similarly to Messick, Bachman and Palmer (1996; 2010) connect validity to the test **purpose** (or use). In line with that, I will discuss the validity of the *Questions Test* in the context of diagnosis and the development of learners' abilities being assessed as informed by the sociocultural theory. That is to say, the conventional definitions of aspects of validity, such as reliability and construct validity, should be expanded or reconceptualised, and so should validation procedures, to account for the differences between the paradigms underlying the conventional and dynamic assessment, as I will detail later in the present chapter.

At the core of Bachman and Palmer's validation approach lies the correspondence between language test performance and **target language use** (TLU), that is, the use of language in the domain of generalisation of learners' test performance. In other words, if learners' performance is not generalisable beyond their test performance, such test lacks validity.

Bachman and Palmer (1996) differentiated between six **aspects** (or **qualities**) of test usefulness: reliability, construct validity, authenticity, interactiveness, impact, and practicality. However, similarly to Cronbach and Meehl (1955), they perceived validity as a nomological network (i.e., a network of interrelated aspects, all contributing to a whole). Below, I present the way I visualise Bachman and Palmer's (1996) perspective on test validity (**Figure 3**), following which, I will define separate aspects of test usefulness.

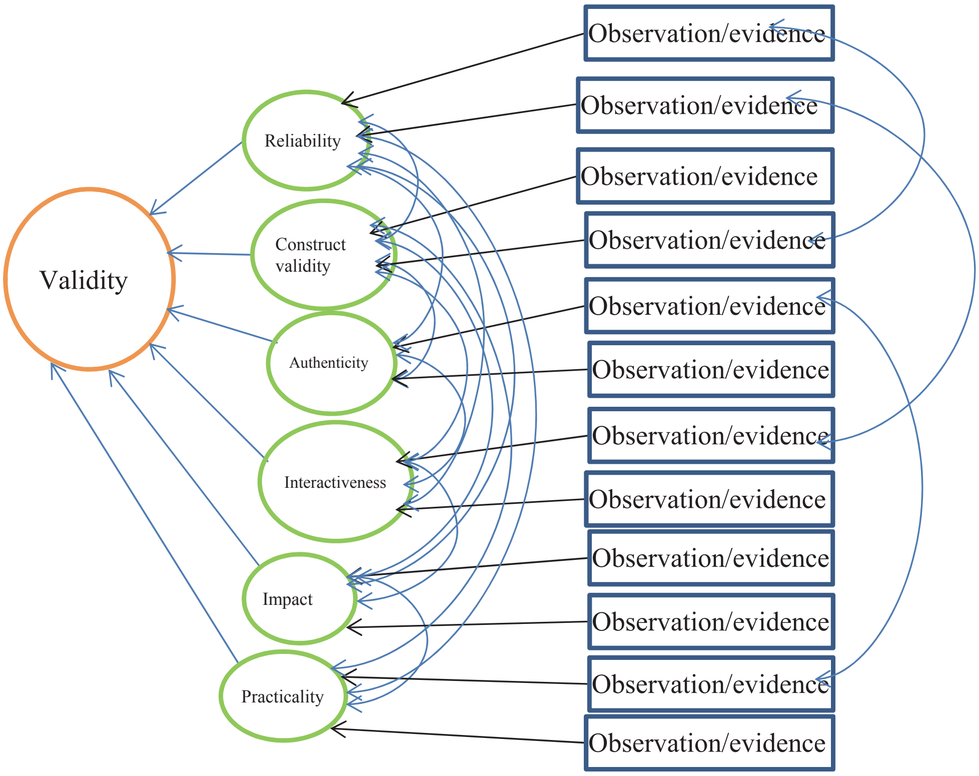


FIGURE 3 Bachman and Palmer's (1996) perspective on test validity.

As regards static assessment, a test possesses **reliability** if it consistently measures what it purports to measure. The situation is different in DA. Poehner (2008) noticed that the main caveat in defining reliability in DA lies in that whereas in SA, the learning effect has to be minimised, DA is all about promoting development, and thus a consistency across test administrations indicates that the test is not valid. Poehner (2008) suggested that reliability and validity of dynamic assessment should be reconceptualised taking into consideration the epistemological basis of DA. Drawing on discussions of DA reliability (e.g., Ableeva, 2010; Haywood & Lidz 2007; Lantolf & Poehner, 2004; Poehner, 2008; Poehner & Lantolf, 2005), it can be suggested that a dynamic test that consistently results in the development of the ability(ies) being assessed is reliable.

Construct validity refers to meaningfulness, appropriateness, and justifiability of inferences made on the basis of test-takers' performance. Thus, establishing construct validity should naturally start with defining the assessed construct. Defining the construct includes defining development, which in DA, is fully identifiable only when mediation is taken into account (e.g., Lantolf & Poehner, 2008).

In relation to the above discussion, Poehner (2011) discussed two interrelated kinds of validation processes involved in dynamic tests—validation of mediation on **micro** and **macro** levels. The **micro** level validation pertains to the evaluation of separate mediational moves, i.e., evaluation of the interpretations of learners' performance and evidence to support these interpretations. Validation on **macro** level involves studying the test as a whole, aiming to find out whether, and to what extent, mediation reveals and promotes the abilities being assessed.

Bachman and Palmer (1996) argued that construct validity also refers to the generalizability of the interpretations to the TLU domain. The latter closely connects construct validity with **authenticity**, which is defined as the correspondence of characteristics of the test tasks with those of the TLU tasks. That is to say, a test possesses authenticity if the inferences made on the basis of its scores are generalisable beyond the performance on the test to the 'real life' language use. Bachman and Palmer (1996) added that the importance of this quality of usefulness is the more so high, as it can potentially influence the way test-takers perceive the test and, consequently, their performance on the test.

As regards dynamic assessment, it seems that authenticity should include **authenticity of mediation**. In practice, if the mediation/feedback that test-takers receive during a dynamic test is not generalizable beyond the test context, that is, in Hattie and Timperley's (2007) words, it only has task level, the test lacks both construct validity, authenticity, and impact, as the mediation test-takers receive during the test would not be useful in other contexts, such as classroom instruction.

Bachman and Palmer (1996: 25) defined **interactiveness**, as "the extent and type of involvement of the test taker's individual characteristics in accomplishing the test task." The authors added that these characteristics include language ability, topical knowledge, and affective schemata (i.e., values, beliefs, and experiences).

I would suggest that interactiveness should also be evaluated as the interactivity between the personal characteristics and the whole test, as it is not just the task, but the whole path a learner moves through the test that interacts with his/her personal characteristics. This is especially evident in DA, where the whole test and the paths learners work through it should be considered when making inferences based on their performance. What is more, in human-mediated DA, the evaluation of interactiveness should include studying the interactivity between the personal characteristics of the test-taker and those of the mediator. Here, such characteristics of Feuerstein's MLE as intentionality, reciprocity, and mediation of meaning (Chapter 3.2.1) can be used as a starting point for such evaluation. The sociocultural paradigm also predicts that in DA, the relationship between the learner's personal characteristics and those of the test is reciprocal.

This implied influence of test (tasks) characteristics on test-takers' characteristics relates **interactiveness** to test **impact**, which is the influence that the test has on society and educational systems (macro level) and individuals, such as teachers and learners (micro level). Bachman and Palmer (1996; 2010) maintain that taking a test implies the interaction of values and goals embedded in the test with those of test users. Huhta, Kalaja, and Pitkänen-Huhta (2006) and Huang (2009), for example, extended this interaction to teachers' and learners' beliefs. That is to say, a lack of (or not intended) interactiveness can result in a lack of impact (or a **negative impact**) of the test. Regarding the latter, Alderson et al. (2013: 239) suggested that it might be incorrect to talk about negative impact of diagnostic assessment, as it does not result in abandonment of goals of instruction (i.e., teaching to the test) and aims at improving the quality of teaching and learning. That is to say, no impact of diagnostic assessment can be considered as purely negative. The same should be true for dynamic assessment.

Practicality differs from the rest of aspects of usefulness in that it does not refer to uses of and inferences made on the basis of performance on a test but rather means that the resources (e.g., time and money) that a test requires are efficiently allocated. Ensuring practicality, generally speaking, means ensuring that the test will be used at all. As regards dynamic assessment, computerised DA is more practical than human-mediated DA, as it addresses the impracticality of face-to-face interactions between the mediator and the learner in that several learners can take the test simultaneously. On the other hand, it should not be forgotten that human-mediated interactionist DA makes it possible, for example, to fine-tune mediation to learners' needs more precisely and quicker than in computerised DA, which can be considered a practical aspect of human-mediated DA.

Usability refers to the extent the interface of a computerised test, but also instructions, item types, and scoring rubrics, are easy and unambiguous to use (Fulcher, 2003). Among the considerations vital for test usability are usable navigation and controls, terminology, text size and colour(s), icons and graphics (especially used as metaphors), available help facilities, and the selected task types. Usability is, above all, connected with construct validity in that lack of usability can result in construct-irrelevant variance, thus hindering construct validity. However, if, for example, learners are not familiar with task types, the test will also lack interactiveness, as the interactivity between the learner and task characteristics will decrease. Finally, there is also a connection between usability and practicality, as a test that lacks usability is clearly not practical.

The above discussion illustrates that all aspects of usefulness as discussed by Bachman and Palmer (1996) are interconnected (to a certain degree) and are best to be visualised as a nomological network. There is, however a distinction which Bachman and Palmer (1996) did not explicitly make, but which, following Huhta (2010), I would like to mention in the following subsection.

5.1.3 A priori and a posteriori validation

Despite being rather comprehensive, Bachman and Palmer's validation framework lacks an explicit distinction between theoretically-based test design decisions and changes introduced based on empirical validation, a distinction made by Weir (1993; 2005), who differentiated between **a priori** and **a posteriori** validation. As the terms suggest, the former refers to theoretically based decisions made before piloting the test. As Huhta (2010: 9) noticed, Weir's argument reflects the theoretical rationale for validity as discussed by Messick (1989). **A posteriori** validation refers to empirical evidence for validity collected from test pilotings and other test administrations to target groups. This evidence should also include evidence against test validity, as is argued in more recent approaches to validation (e.g., Bachman & Palmer, 2010).

I will use this distinction when presenting the currently available evidence for the validity of the *ICAnDoiT* system and the *Questions Test*. My rationale for doing so is that such distinction makes the presentation more structured and easier to follow.

5.2 Aspects of usefulness of the *ICAnDoiT* system and the *Questions Test*

In the present section, I will discuss the evidence for (and against) the validity of the *ICAnDoiT* system and the *Questions Test*. Before I discuss the separate qualities of usefulness in some detail, I would like to summarise the sources the validity evidence came from (**Table 8**).

TABLE 8 Aspects of usefulness of the LA2 English *Questions Test* in the *ICAnDoiT* system

Aspects of usefulness	**Theoretical validation**	**Empirical validation**
Reliability	- the detailed test specifications (**Appendix 1**);	-evidence for the reliability of the unmediated exercises, e.g., reliability coefficients (Article I); - evidence for the validity of the adaptive corrective feedback within the mediated exercises (Article I; this chapter).
Construct validity	- operationalising the knowledge of L2 English questions in terms of stages in question development; - operationalising the adaptivity of the feedback based on the Regulatory Scale.	- evidence for construct validity of the unmediated exercises (e.g., learners' and teachers' reports; Article I, Article II, this chapter); - the correspondence between the mistakes the learners made in the pretest and those that the DA addressed (Article I; this chapter); - evidence for validity of the feedback within the mediated exercises (Article I; Article II).
Authenticity	- task types contextualised in a possible real-life situation; - tasks types similar to the ones used in English textbooks at grade 8; - basing the feedback the learners received on Aljaafreh and Lantolf's (1994) study, where the authors determined an implicational scale of the feedback messages based on those emerging in interaction.	- learners' and teachers' reports regarding their perceptions of correspondence of the test tasks and the feedback and the TLU tasks (Article I; Article II; this chapter).
Interactiveness	- designing the tasks and the feedback to activate learners' knowledge of L2 English questions as well as cognitive and metacognitive strategies (e.g., using context clues or evaluating).	- learners' and teachers' reports on their experience with the exercises and the feedback (Article I; Article II); - the lack of relationship between the amount of feedback and the increase in the learners' performance (this chapter).

Impact	- designing the test to promote learners' ability to form and use L2 English questions and construct/remediate their beliefs about the usefulness of corrective feedback; - based on the previous research (**Chapters 2** and **3**), selecting and arranging the feedback messages based on their explicitness and level of detail.	- change in the usefulness of CF as perceived by the test-takers following their experience of the dynamic test (Article II); - learners' comments on their experience of taking the test (Article I; Article II); - teachers' reported experiences with the test and their observations of their learners working through the tasks (this chapter).
Practicality	- design decisions made for the *ICAnDoiT* system, including its accessability due to the Web-based modality, its flexibility, and the inexpensiveness of its design (Appendices **A** and **B**).	- teachers' and learners' reports on the practicality of the system / the test (this chapter).
Usability	- basing the design of the system and the test on the previous research (Fulcher, 2003; **Appendix 1**).	- learners' and teachers' reports regarding the usability of the interface (Article I; Article IV; this chapter).

Since the aspects of usefulness in Bachman and Palmer's framework are interrelated, I will discuss certain data in connection with the aspect they have more to do with (in my opinion). I will also present some evidence collected in the studies but not reported in the original publications. The quoted interview transcript excerpts will be from two teachers (in English) and learners both from the Pilot study (in Finnish) and the studies reported on in Articles I and II (in Russian). I will refer to the latter two collectively as Intervention study. To avoid confusion, I will refer to the intervention part of the *Questions Test* (both dynamic and static) as to *treatment*. Unless otherwise stated, the Intervention study findings obtained from the adaptive CF group will be discussed. The English translations will be given when required. The transcription symbols used in the quoted interview excerpts are supplied in **Appendix A** in Article II.

5.2.1 Reliability

Decisions made during the process of designing the system and the test allowed for establishing the reliability of the *Questions Test* and the *ICAnDoiT* system. Test and system specifications (**Appendix 1**) being products of those decisions, were the documents that served the basis for the design.

The decisions made to ensure reliability included designing the scoring rubric for the pretest, divided into separate aspects of the construct, that is, the word order of correctly formed wh-questions with auxiliaries (i.e., stage 5 questions), and learners' problems with questions with auxiliaries *do*, *does*, and *did* identified in analysing of the CEFLING project data (**Chapter 5.2.2**). However, the previous stage in question development was also taken into account (see **Chapter 3.4.1**).

Another part of the a priori evaluation of reliability included ensuring the adequacy of the translation (of the instructions, system messages, etc.) into the languages of the test-takers. This allowed for enabling a uniform test taking experience for different learners (cf. Huhta, 2010).

To minimise the possibility of learners' beliefs negatively influencing their performance (**Chapter 2.8**), in the instructions to learners, the procedure was referred to as exercises aiming to help them to find out about their problems with question formation. The same was stressed in the instructions to teachers/proctors regarding responding to the test-takers' queries.

A part of the a posteriori process of ensuring/evaluating the reliability of the procedure was piloting of the pretest and the treatment tasks. The Pilot study was conducted among 19 L1 Finnish learners of English and their teacher. The Pilot study data mostly come from the participants' performance on the tasks but also learner questionnaire, its major aim being to study the usability of the ICAnDoiT system (**Chapter 5.2.7**), the teacher's think-aloud protocol, and semi-structured interviews with the teacher and six of the learners (three high-achieving and three low-achieving, as evaluated by their teacher), the aim of the interviews being to evaluate several aspect of usefulness of the Questions Test and the ICAnDoiT system. The frequencies of the learners' responses to dichotomous, multiple-choice and Likert-type questionnaire items were calculated. The patterns of responses on the open-ended questionnaire items were studied. As regards the analysis pertaining to the reliability of the *Questions test*, a classical item analysis was conducted based on the learners' performance on the pretest tasks. To summarise the Pilot study results pertaining to the reliability, the study demonstrated that overall, the test was reliable as to its consistency. For example, the Cronbach's alpha for the first task of the pretest was .77. I nevertheless, modified several items, for example item 3 in the first task of the pretest, which was added to assess the learners' ability to form stage 5 questions, but two learners out of three whose responses to this item were correct produced stage 4 questions (e.g., what pets are there). Therefore, the prompt *what pets the shop has*, which in Finnish sounded like *millaisia lemmikkejä kaupassa on* (and could be translated back to English as *what pets there are in the shop*) was changed to *what pets they sell*. I also added several more items to account for Spada and Lightbown's (1999) finding that learners are more likely to form an accurate L2 English question when the subjects in the sentence were pronouns than when they were nouns (**Chapter 3.4.1**). In addition, I added a line to each feedback message explicitly informing the learners that the following items in the exercises will be similar.

As regards the Intervention study, the results demonstrated that the pretest tasks had sufficient reliability (Article I). In the following, I will corroborate the finding with the results of the classical item analysis performed on the whole sample (*n* = 58), that is, not limited to those learners who took all the three parts (i.e., the pretest, the treatment, and the posttest) and including the learners from both the adaptive and the static CF group. I will, however, exclude those learners who were caught cheating (*n* = 5).

The analysis conducted with jMetrik software (Meyer, 2013) demonstrated that apart from item 8 (a *does*-item, its prompt being "chto oznachayet nazvanie magazina" [what the name of the shop means]), the difficulty values and the item-total (point-biserial) correlations were acceptable, the former ranging from slightly above .20 to .51 and the latter, from .36 to .68. The difficulty value of item 8 was .15, which was undesirably low, and its item-total correlation, .24 (which could be called acceptable). However, after considering the consequence of removing the item (i.e., under-representation of *does*-items), I decided to retain the item.

The internal consistency of the pretest items (k = 17) was rather high, Cronbach's α = .87. On the other hand, the Rasch reliability of the pretest tasks was somewhat low if the reliability of the scale containing only stage 5 questions was estimated (Article I). However, that I added stage 4 questions to the pre-/posttest exercises (**Chapter 3.4.1**) allowed for creating a variable that included both stage 5 and stage 4 questions, which had a higher reliability than that including stage 5 questions only. What is more, the pretest performance of the Pilot study (L1 Finnish) participants was not significantly different from the Intervention study learners' performance (Article I), which added to the reliability of the pretest.

However, establishing the reliability of the treatment (i.e., the mediated part) required different procedures (**Chapter 3.2**; **Chapter 5.1.2**). To ascertain that the CF provided to the learners during the treatment guided the way they performed across the tasks, I traced the learners' performance on the treatment tasks separately for the adaptive and the static CF groups controlling for their pretest performance (**Figure 4**).

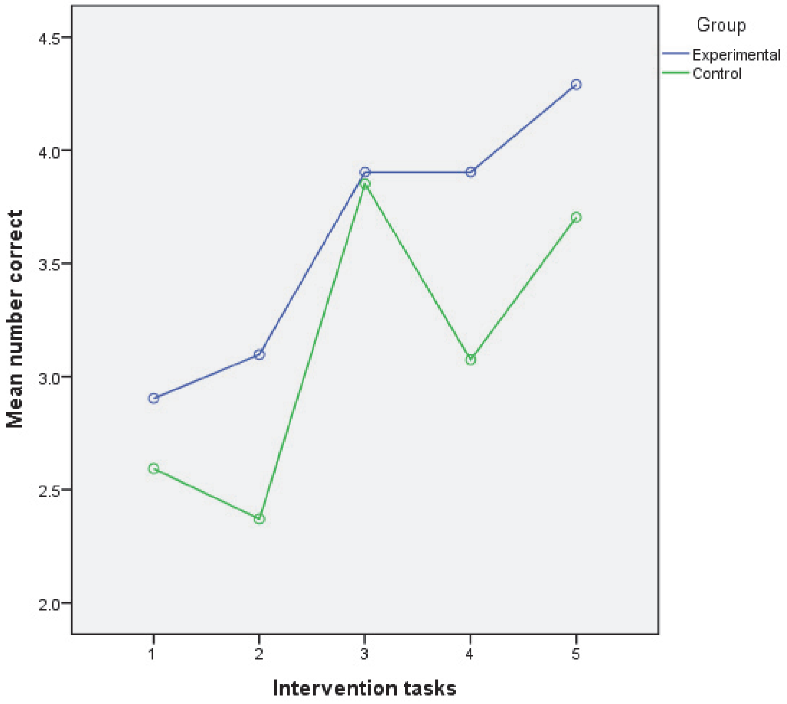


FIGURE 4 Means plots of learners' performance on the treatment tasks (from an analysis of covariance).

Figure 4 demonstrates that while the performance was somewhat similar in both groups, there was a gradual increase in the performance of the adaptive CF group, whereas the change in performance of the static feedback group across the treatment tasks was much less linear. When the learners' pretest performance was not controlled for, the trend was similar though there was a somewhat bigger drop in the performance of the experimental group on task 4. It can thus be argued that the adaptive CF resulted in a gradual improvement of the learners' performance during the (about 40 minutes long) treatment.

5.2.2 Construct validity

The a priori evaluation of the construct validity of the test included, above all, making sure that the construct and its development were clearly defined. As I have mentioned in **Chapter 5.2.1**, the ability to form L2 English questions was defined in terms of stages in question development (see **Chapter 3.4.1**), and stage 5 questions were selected as the assessment target.

This is, however, where the clash between Piagetian and Vygotskian perspectives on development had to be addressed. While I was unable to find a common theoretical basis for the two perspectives (**Chapter 2.5**), I decided to explore what questions the learners' would be able to form after the treatment, the more so, as I mentioned in **Chapter 3.4.1**, in written performance, it is the frequency of use of questions of different stages and the accuracy of their use that differentiates between learners of different proficiency rather than the emergence of certain stages in question development.

The validity of the automated adaptive CF was ensured by basing it on Aljaafreh and Lantolf's (1994) Regulatory Scale. That is, the learners' abilities were defined in terms of the amount of assistance they required to formulate stage 5 questions and their development in terms of the number of accurately formed stage 5 questions after the treatment. It should be elaborated at this point how the accuracy of stage 5 questions was operationalised. I considered cases where wrong auxiliaries or wrong forms of the main verbs were used (e.g., *What does the shop name means?*) to be inaccurate stage 5 questions whereas, for example, spelling mistakes, provided that the auxiliary and the main verb were used correctly (e.g., *What does the name of the store meen?*), as accurate.

The pre-/posttest and the treatment tasks elicited the use of stage 5 questions. The exceptions were two items in the first pre-/posttest task where both stage 4 and stage 5 questions were possible and one item in the second pre-/posttest task that elicited the use of stage 4 questions. The CEFLING project data (see Alanen & Kalaja, 2010) was studied in order to single out the common mistakes that learners' made when constructing stage 5 questions. These data were used for the creation of distractors in the multiple-choice tasks of the treatment part of the *Questions Test*.

To reduce the construct irrelevant variance, it was checked that the task types and the vocabulary used in the test were known to the participants. For this purpose, a number of websites offering online grammar/vocabulary exercises for practicing English and several textbooks used in Finland for teaching/learning English at grade 8 were studied (**Section 5B** of **Appendix 1**).

Regarding targeting of the exercises to grade 8 learners, this was done above all to ensure that stage 5 questions were within the learners' ZPD, but, at the same time, the learners were not self-regulated in their use (**Chapter 3.3**). Both in Finland and in Estonia, learners are expected to be at level B1 of the CEFR at the end of the lower secondary school, i.e., at the end of grade 9 (Põhikooli riiklik õppekava õigusakt: Lisa 1, 2010; Finnish National Board of Education, 2004). According to Alanen and Kalaja (2010), it is at B1 level that the number of accurately formed stage 5 questions starts increasing.

The a posteriori validation included making sure that the treatment tasks addressed the mistakes that the learners made in formulating stage 5 questions. The analysis of the participants' (both the Pilot study and the Intervention study) performance demonstrated that the learners used all the distractors in the multiple-choice tasks (Article I), and the sentences they produced during the pretest were similar to the distractors in the multiple-choice tasks (e.g., **How is you magazine get my e-mail?* or **how they got my email adress?*).

The learner and teacher interview results supported the theoretical evidence for the adequacy of construct definition (i.e., both the ability to form L2 English questions and development of this ability). The Pilot study teacher, for example, reported: "*I think they* [the exercises] *are appropriate. I think they measure* *the things that you wanted them to measure*." Later she added, "*I think they* [the feedback messages] *were quite clever because they got somehow more and more intense and gave me more information. When the programme recognised that I had made lots of mistakes, and very stupid mistakes, they were more detailed*."

Many learners also perceived the feedback as useful, as illustrated by a report of one Pilot study learner: "*no siis just jos oli jotain tommosii virheitä mitä luuli, et ne menee silleen niin sit siin ku tuli ne ohjeet, et miten se niinku pitäs oikeesti tehä tai siis just noi [shows], niin sit siitä niinku oppi silleen, et se ei meekään nii*" [*well, when there were mistakes that I thought were not mistakes, these instructions appeared, like how to really do it, or these ones [shows]; with help of these you learn that it doesn't go like this*]. Generally, there were significantly more learners in the adaptive CF group who considered CF in the test as useful for learning than in the static CF group, and their questionnaire responses demonstrated that they benefitted from different feedback types (Article I). The latter adds to the validity of adaptive feedback on micro level (**Chapter 5.2.1**).

Furthermore, the participants reported that they were familiar with the task types used in the *Questions Test*, for example, "*koulussa me ollaan tehty mutta en mää kauheesti kotona oo tehny*" [*we have done [those] at school, but I have not really done [those] at home*].

More evidence for the construct validity can be observed in the learners' performance on the two ordering tasks (**Figure 4**). While in the adaptive CF group, the performance on the second ordering task (where the subjects were nouns) is higher than that on the first ordering task (where the subjects were pronouns), the reverse is observed in the static CF group. This difference can be interpreted with reference to Spada and Lightbown (1999), who found that formulating questions was easier when the subjects were pronouns, which is observed in the performance of the static CF group. In the adaptive CF group, however, the mediation of the learners' performance resulted in a higher performance on the second ordering task.

Most importantly, the adaptive CF group learners improved their ability to form stage 5 questions significantly after the treatment, and their improvement was significantly higher than that of in the static CF group (**Chapter 6.1**), which indicates that the *Questions Test* served its purpose and that the mediation was valid (on macro level; see **Chapter 5.1.2**). These findings add to the construct validity (and reliability) of both the dynamic treatment, where the learning effect was expected, and the static version of it, where the learning effect had to be minimal.

All in all, the a posteriori construct validation produced rather positive outcomes. However, problems identified with interactiveness of the procedure could hinder construct validity, as will be discussed in **Chapter 5.2.4**.

5.2.3 Authenticity

The a priori authenticity was, above all, established by means of contextualising the pre-/posttest and the treatment tasks within a hypothetical situation which is likely to happen in real life, that is, inquiring for additional information in an

E-mail. Specifically, the learners were asked to imagine that they moved to London one day, got an advertisement from a pet shop, and decided to inquire for more information from the shop. The authenticity was reinforced by including additional reading tasks to the pretest—the advertisement and the reply from the pet shop—which the test-takers were asked to read but their reading performance on which was not evaluated.

In this regard, task 1 of the pre-/posttest is the most authentic, as it asks test-takers to write an E-mail. Task 2 of the pre-/posttest and the treatment tasks are less authentic as regards the generalisability of TLU outside the classroom. However, the second context into which the inferences were planned to be generalised was L2 English use in the classroom (**Section 2** of **Appendix 1**).

When collecting the empirical evidence for the authenticity, I was interested in whether learners and teachers acknowledged the link between the exercises and the TLU domains. I also made sure that the participants, especially the teachers, perceived the procedure as enabling diagnosis of learners' abilities (i.e., in the authenticity of the adaptive feedback). The excerpts quoted to support these are from both the Pilot study and the Intervention study. Hereinafter, I will refer to the Pilot study teacher as T1 and the Intervention study teacher as T2.

Regarding the E-mail writing task, T1 reported that the overall outline, the buttons, the test-takers' names included in the letters, the pictures, etc. made it *[l]ike a real thing*. She then added that the E-mail task was a good way to encourage pupils to try to convey their message and formulate the questions. This suggests that this were not just the tasks per se, but also the elements of the interface of the *ICAnDoiT* system that added to the authenticity of the procedure.

The learners' questionnaire reports, for example, *"opin miten esitetään kysymyksiä kohteliaasti" [I learned to ask questions politely]* or *"ainakin nyt osaa kysyä lemmikeistä kaikenlaista" [now at least I am able to ask about all kinds of pets]* (from the Pilot study) suggest that at least some of them acquired skills that they could potentially use outside the classroom. In addition, this suggests that the feedback was able to explain the goals of the procedure to the learners (cf. Hattie & Timperley, 2007). Moreover, the recurring theme in the learners' questionnaire responses and interviews was the correspondence between the test tasks and the tasks the learners had been doing in the classroom, for example, *"No siis se mis piti laittaa niinku, ku siit paino siitä jutusta ja siit tulee silleen alas niitä, et mitä vaihtoehtoja niin semmosia. Ja sitte, just se, et mis pitää niinku täyttää niit juttuja." [well like the one where you had to click and different options dropped down an' all. Well, and where you had to fill in these things]*, which added both to the reliability and the authenticity of the test.

In addition, as demonstrated in Article II, one learner even suggested that the feedback he received during the DA can also be used in the classroom: *"mozhno snachala nameknut' na to, chto u tebja oshibka (.) uchenik poprobuet sam ugadat' eyo" [one may at first hint that there is a mistake* (.) *the learner will try to guess it himself]*. Both teachers also thought that the adaptation of the feedback as operationalised in the *Questions Test* is suitable for the classroom use (e.g., Excerpt 1).

Excerpt 1

> T2: I must know why or where or at what level he doesn't understand. So if I ask him simply "correct your mistake". For example if I say "correct your mistake" or "you are wrong". Maybe he will not understand me because he has some patterns (.) Or maybe he will remember or he can't remember (.) something.

All in all, while the a posteriori evidence for authenticity was somewhat scarce, there was no evidence against it.

5.2.4 Interactiveness

The interactiveness was mostly evaluated with reference to the adaptive CF group. Perhaps, the smallest interactiveness was expected to be between learners' topical knowledge and the tasks characteristics, as learners were not expected to know anything about pets and pet shops. Thus, the prompts in the pre-/posttest tasks were designed so that questions could be formed without any knowledge other than the knowledge of L2 English wh-questions with auxiliaries. Moreover, special care was taken so that other learners' abilities, such as vocabulary knowledge, did not influence their performance on the tasks (**Chapter 5.2.2**).

The interactiveness between learners' abilities and the treatment tasks was achieved by means of the CF their received (**Chapter 5.2.1**). Thus, the degree of interactiveness was different in the adaptive and in the static CF groups. On the other hand, the interactiveness between learners' L2 abilities of and the characteristics of the pre/post-test tasks was expected to be low. Thus, an easier task (gap filling) was presented after a more difficult task (E-mail writing) so that learners' performance on easier items would not mediate their performance on more difficult ones.

High interactiveness was expected between the characteristics of the tasks and learners' strategies. At the same time, the possibility that learners start using conventional test-taking strategies, such as guessing, was minimised (**Chapters 5.2.1** and **5.2.2**). Instead, the adaptive CF was designed to facilitate the use of other strategies, such as identifying discrepancies in a sentence by looking at a model sentence (e.g., *Look at the following examples. How are they different from your sentence?*), using context clues (*How often* ***do*** *you'****re*** *clean the shop? … do we need the verb* ***are*** *here?*), evaluating, etc. As the previous research (e.g., Kozulin & Garb, 2002) has demonstrated, dynamic assessment benefits learners in that they are able to apply the acquired strategies in other similar contexts (Alderson's et al, 2015, strategy feedback level). Thus, I hypothesised that the strategies encouraged during the dynamic treatment would manifest themselves during the posttest.

The empirical evidence for interactiveness was mostly based on the qualitative data regarding the extent to which the learners and the teachers considered language abilities, topical knowledge, and affective schemata to interact with the tasks. In addition, I studied whether the learners skipped any feedback (through the analysis of both the participants' reports and performance logs, where the time learners spent before closing the feedback window was recorded).

Both teachers found that the major advantage of the dynamic part of the *Questions Test* was that it interacted with their learners' ability to formulate L2 English wh-questions. For example, the Pilot study teacher reported the following (**Excerpt 2**).

Excerpt 2

> *T1:* Because it gives (.) when I think of not of the most low-achievers, but my ordinary pupils... average pupils. It gives them, you know, ideas, and then they kind of solve themselves. ↑Aha, oh yeah (.) what's this, mm.↓ It makes them hopefully think and analyse the structure of the sentence.

The Intervention study teacher was of the similar opinion saying that all of feedback levels "are useful and on a certain level, they need this one, for example. If they are advanced students, they need the other one, and so on."

Some quantitative analyses of the data allowed for finding evidence for the interactiveness of the treatment tasks and the learners' abilities. Specifically, the median feedback level the adaptive CF group learners received in the Intervention study did not correlate with the variable representing the increase in their use of stage 5 questions (the difference between their pretest and posttest performance), r_s= -.162, p = .430. That is to say, regardless of the amount of assistance, the experimental group learners improved more or less the same, which shows that regardless of its explicitness and amount of detail, the feedback interacted with their abilities.

Although the interactiveness between the characteristics of the tasks and learners' topical knowledge was specified as low, the test included some training in the use of questions in semi-formal E-mails. That is to say, while in informal communication, such stage 3 questions as *Where I can find more information about dogs?* are possible, they are not used in formal and semi-formal correspondence. This was reflected in the CF displayed when the learners formulated such questions. As a result of that, several learners reported that they remembered that convention of (semi-)formal correspondence, for example, "*opin miten esitetään kysymyksiä kohteliaasti" [I learned how to ask questions politely]* or "*Kak bolee official'no stroit' vopros" [How to form questions more officially]*. In addition to interactiveness, this suggests that the mediation of meaning (see **Chapter 3.2.1**) that this feedback aimed at was achieved at least for some learners.

As regards the interactiveness between the characteristics of the tasks (and the whole test) and learners' affective schemata (i.e., beliefs, strategies, etc.), the results of both the qualitative and quantitative analyses were varied. On the one hand, the adaptive CF group learners reported to have used different strategies when working on the tasks, the choice of which seemed to be influenced by different feedback types, among them identifying discrepancies by looking at model sentences (e.g., *"Oni mne pomogli primerami kak pravil'no kuda stavit"* [They helped my with examples how to put it correctly to the right place]) or looking for a variety of clues elicited by different CF types (e.g., *"Oni podskazyvali, chto v predlozhenii postavleno slovo ne pravil'no ili cho-to eshtsho."* [They hinted that in the sentence, a word is at the wrong place or something else.] or *"Kakoe slovo pravilnoe, kakoe net – eti"* [Which word is right, which not—these ones]). With reference to Hattie and Timperley's (2007) discussion of useful feedback (**Chapter 2.6**), this demonstrates that different feedback types instructed the learners both in how they performed in relation to the goal (i.e., formulating questions in English) and what should be done to improve their performance. In contrast, the static CF group learners did not generally report on the use of strategies, except for single cases, such as "ya nauchilsya byt' bolee vnimatelnym" [*I learned to be more attentive]*, which may indicate that the feedback actually helped this learner to stay concentrated on the task (see Article I for more examples).

That said, in general, the experimental group learners still tended to rate more explicit feedback higher than implicit feedback. In the questionnaire, they were asked to rate the feedback of different levels on a scale from 1 (the most useless) to 5 (the most useful). A Friedman's ANOVA demonstrated that the adaptive CF group learners rated level 1 (mean rank 1.94), level 2 (mean rank 2.54), level 3 (mean rank 2.99), level 4 (mean rank 3.57), and level 5 feedback (mean rank 3.98) significantly differently, $X^2(n = 27, df = 4) = 31.13$, $p < .001$. This trend seemed to be irrespective of the learners' abilities, as both the questionnaire and the interview demonstrated. For example, in the questionnaire, HA2 (a learner in the high-achievers group in the Group study) reported that explicit explanation and correction was the most useful for him although he had not even received this feedback during the treatment. What is more, a similar situation arose during the piloting, where two of the high-achieving interviewees had not seen the feedback during the dynamic test but still reported that it was the most useful, for example, *[a]inakin noi kaks alimmaista, niin aika hyödylliset silleen, et niinku tajuaa et se ei mikään niinku mitä on luullu [At least those two at the bottom are the most useful].*

The preference for more explicit feedback is in line with the previous research (Amrhein & Nassaji, 2010; Ashwell, 2000; Lee, 2008; Leki, 1991). However, as I have argued in **Chapter 2.8**, this can result in that learners skip the feedback they *believe* to be useless and not because it is outside their ZPD.

As regards, the Intervention study, there appeared to be only three learners who did not skip any of the feedback, spending at least four seconds on each of the feedback messages they received following their incorrect responses. This suggested that it was not always the case that the learners skipped the feedback because they were unable to benefit from it.

Perhaps the most extreme case was LA1 (Article II). It appears that although being told that it was not a test, she still considered it to be one. The fact that the teacher told her to work on her own added to her perception of the procedure. Intertwined with her belief in the superiority of good marks over knowledge and not being able to understand her mistakes with help of implicit CF (which led to her frustration; see also Haywood & Lidz, 2007), this resulted in that she skipped the feedback in most of the tasks, which hindered the interactiveness between her abilities and the test.

However, during the interview following the treatment, it seemed that LA1's belief that getting correct answers without understanding is always useful weakened. In fact, it was not only LA1 who benefitted from the discussion (see Article II). Thus, discussions with learners following their experience with DA have a potential to remedy interactiveness issues of computerised DA arising due to learners' beliefs (cf. Thouësny, 2011).

What is important to note in case of LA1 is that she improved her ability to produce questions in the first task of the posttest (i.e., E-mail writing according to the prompts) but only in the use of wh-questions with modal auxiliaries (**Table 9**), which was the target of the ordering tasks where she reported to have to read the feedback (Article II).

TABLE 9 LA1's pretest and posttest performance on task 1 (E-mail writing using prompts).

Pretest	**Post-test**
Exercise 1	
1)Where your shop is located?	1) Where are you stay?
2)When you are open?	2)When shop is open?
3)What pets you selling?	3)Which pets you sell?
4)How much cost pets ?	4)How much coast pets?
5)Where I can search more info about pets,for example fotocards?	5)Where can I get photo about pets?
6)What info i can get about your pets?	6)Which information can i get else?
7)Where you was searching mu e-mail?	7)Where did you find my e-mail?
8) What does maen name your shop?	8)What does it mean name of shop?

It also seems that LA1 improved her performance on the items with *did*, but judging by her interview response and performance log, it appears she read at least some feedback in the first task of the pretest where wh-questions with auxiliary *did* were trained. That is to say, when LA1 read the feedback, the *Questions Test* served its purpose. Interestingly, her performance also speaks against the implicational order of stages in question development, as after the treatment, she was able to formulate several stage 5 questions while being unable to form any of the stage 4 questions.

Overall, most of the evidence for the interactiveness of the *Questions Test* has been favourable, which also has relevance for the impact of the procedure.

5.2.5 Impact

For the most part, the evidence for the impact of the *Questions Test* has already been discussed in the previous sections. Moreover, as the present doctoral research project is, above all, **impact-driven**, the articles forming the present dissertation are, in essence, about studying the impact of DA. Thus, I will give only a brief overview of the impact evidence.

The theoretical evaluation of the impact of the *Questions Test* on learners included making sure that the participants are informed about the procedures and designing the test so as to develop the ability being measured (**Chapter 5.2.2**) and result in changes in learners beliefs about corrective feedback and strategies they use to complete the tasks successfully (**Chapter 5.2.4**). As regards teachers, the major impact of the procedure was planned to be the remediation of classroom instruction. For this purpose, detailed learner profiles, including their mistakes and the feedback they required to self-correct these mistakes were designed.

Much of the a posteriori impact evidence has already been discussed (e.g., the development of the learners' abilities in **Chapter 5.2.1**). In the present section, I will briefly reiterate these pieces of evidence and present some additional evidence for the impact on learners and teachers.

Some indirect evidence for the impact of the test on the learners' beliefs about corrective feedback can be traced in the available data set. To start with, there was a clear difference between the experimental and the control group learners' perception of the usefulness of the feedback (**Chapter 5.2.2**). However, judging by the fact that the learners rated explicit and detailed feedback significantly higher than implicit, this could have been because the feedback in the adaptive CF group was generally more explicit and detailed than the KOR (knowledge of results) feedback in the static CF group.

As not all the adaptive CF group learners considered the feedback useful (**Chapter 5.2.4**), it appears the experience of the DA alone might not be enough to lead to changes in learners' beliefs about corrective feedback, as, perhaps, it was not long enough (see Article II for the changes in M's beliefs as a result of a longer DA experience). On the other hand, discussions drawing on the learners' recent experience with the adaptive CF resulted in changes in the learners' utterances, which can be interpreted as the beginning of a transformation of these beliefs.

As has been mentioned earlier, learner profiles based on learners' performance on the *Questions Test* (both the pre-/posttest and the treatment) were designed to serve the major feedback for teachers aiming to lead to remediation of their classroom instruction. These profiles (**Table 10**) were tried out during the Pilot study.

TABLE 10 A sample from a learner profile (Pilot study).

Is rather consistent in not inverting/using the auxiliary in wh-questions. Showing him the place of the error and hinting what is wrong in his sentence might help him to self-correct his questions with both the modal aux. and do, but not with does and did. May occasionally invert both the main verb and the auxiliary in his wh-questions.
In wh-questions with *does*, very often adds the *-s* ending to neither the aux. nor the main verb. Feedback did not seem to help him.

To check the validity of the profiles, these were sent to the teacher, who was asked to rate each of the statements as either correct (if these corresponded with her own knowledge of her learners' abilities), incorrect (if these were different from her knowledge of her learners), or as new information (if this information was new to her). The Pilot study teacher, who was an adherent of scaffolding (and thus had an idea how much assistance could have helped her learners), marked forty-eight out of sixty-four statements (75%) as correct and eight as incorrect (12.5%). The further eight messages were marked as new information. This suggests that the profiles were rather valid for her, but she also learned something new about her learners. It should be noted that by the time of the Intervention study, the outline of learner profiles was modified so that the profiles could be read more easily (**Table 11**).

TABLE 11 A Sample from a learner profile (Intervention study).

Common mistake(s)	**Feedback**
Often, puts the modal auxiliary after the subject.	Probably, an example with the correct word order accompanied the feedback that she needs to pay attention to the word order will be enough for her to correct her mistake.

It should be mentioned that at the onset of the Pilot study, T1 thought the DA would not be useful for her less able learners. This, however, changed, when she observed one of them (**Excerpt 3**).

Excerpt 3

> *T1:* I noticed one of the low-achieving pupils doing it. She kind of got the idea with the help of these. And finally she noticed that. I didn't say anything. I was just watching behind. And she was reading it. And I noticed how she went to the correct alternative. And clicked it there.

Thus, it can be suggested that T1 realised she had been wrong about the usefulness of the test for the low-achieving learners. It might also be suggested that the Intervention study teacher, too, saw the usefulness of the test for her own feedback practices. Namely, she reported that she was not sure whether her feedback practices were useful for learners, adding that teachers needs need *some special references to* (.) *how to do it*. She also found the feedback useful (e.g., **Excerpt 1**), adding *if he makes a mistake, I must know why or where or at what level he doesn't understand.*

The teachers' reports cannot serve as direct evidence for the impact of the test on their feedback practices, as no data were collected regarding the change in their feedback practices. However, I think that they still add to the evaluation of the test impact.

5.2.6 Practicality

As regards the practicality of the system, most of the design work has been done by a coding specialist and me (which greatly reduced the monetary expenses) although at certain stages of the development, other people contributed, for example, translating the interface, the instructions, etc. into the languages of the system, i.e., Finnish, English, Russian, and Estonian.

The practicality of administering the tests via the *ICAnDoiT* system is the more so high as the system does not require any additional software to be installed and can be accessed from all major Internet browsers (**Chapter 5.2.7**). Moreover, the advantage of the computerised modality includes the possibility of assessing a number of learners simultaneously, automatically scoring their DA performance. Finally, the data that the computerised modality allows for recording are not limited to learners' performance and the feedback that they receive, including, for example, the amount of time learners spend on each item and on reading the feedback. This is a practical way of tracing the time aspect of the test-taking process and, specifically, in the present study, allowed for making decisions regarding the interactiveness (**Chapter 5.2.4**) and, more generally, the impact (**Chapter 5.2.5**) of the *Questions Test*.

The Intervention study Teacher, while mentioning the latter benefit of the system, reported on its other practical advantage (**Excerpt 4**).

Excerpt 4

> *T2:* If you have a group of students (.) you don't know them well (.) When they are working separately and they have their own speed, they have their own abilities, and you can watch them. And you can see how they work—what way they work. Who is the quickest, who is the slowest, and what they stop at […]. I think that while they are working you can watch them, but, for example, if they are sitting in the classroom, you cannot see (.) you cannot observe their reaction.

That is to say, T2 suggested that the advantage of the computerised modality is that teachers can observe their learners working on the tasks, noting things that the system does not record.

For learners, the practicality of the *Questions Test* lies in that it does not take long to complete. The dynamic part, which was arguably the longest, took all the learners about one academic period (45 minutes) although the time was still variable, as high-achieving learners completed the test faster than the low-achieving learners. What is more, considering that DA is an activity that contributes to teaching/learning, this time should be considered as the time that took the learners to practice wh-questions with auxiliaries, and for many of them, to improve their ability to produce these questions.

The variable time that learners spend on the test might be found impractical by teachers although neither T1 nor T2 reported on that. What the Pilot study learners found impractical was the absence of the progress bar, which cannot be implemented, as depending on learners' responses a variable number of items (from 5 to 7) is displayed (e.g., **Section 6A** of **Appendix 1**).

A potential practicality problem of the procedure rises from the issue encountered with the interactiveness of the *Questions Test*. That is to say, to increase the interactiveness and, consequently, the construct validity of the test, discussions with learners might be necessary, where the usefulness of feedback not revealing correct responses is elicited.

5.2.7 Usability

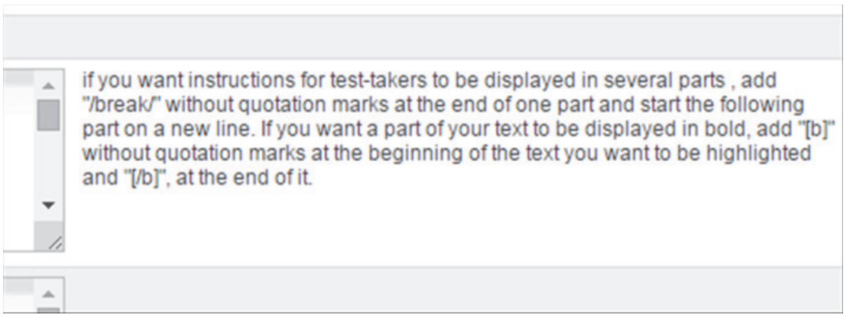
As I have mentioned in **Chapter 5.1**, usability can be assessed within construct validity. Alternatively, it can be assessed within practicality, as, for example, Huhta (2010) did. Nevertheless, I decided to discuss this aspect separately for the following reasons:

1. usability is an important quality of computerised tests;
2. usability is more a quality of the computerised system than a quality of the test per se;
3. usability, despite its importance, has rarely been discussed in the literature on validation of computerised tests.

To ensure the usability of the *ICAnDoiT* system and the *Questions Test*, the design of the system followed the framework presented by Fulcher (2003) (see **Section 5F** of **Appendix 1**).

Before the work on designing the system started, a usability checklist was compiled in co-operation with the coding specialist, which included making sure that the test and the system functioned the same under all the major Internet browsers, the errors resulting from the unintended actions were prevented, the elements of the interface were visible, the design was minimalistic, and sufficient help was provided to test-takers, raters, teachers, and test-designers, i.e., the potential major users of the system.

After the pilot version of the system was operational, the usability evaluation continued. As regards the help documentation, for example, initially a separate user manual for test designers was compiled. In the present version, instructions for test designers are built into the system (**Figure 5**), which increases its usability.



if you want instructions for test-takers to be displayed in several parts , add "/break/" without quotation marks at the end of one part and start the following part on a new line. If you want a part of your text to be displayed in bold, add "[b]" without quotation marks at the beginning of the text you want to be highlighted and "[/b]", at the end of it.

FIGURE 5 Sample instructions to test designers.

The usability of the system was evaluated in a multistage process, during which participants' experiences with the system were collected using interviews, think-aloud protocols, and questionnaire responses as data. These data were collected during an iterative process, in which after each stage, ranging from semi-formal triallings with colleagues to systematic pilotings, changes in the system interface were introduced.

All in all, the Pilot study, being the first major stage of testing the usability of the *ICAnDoiT* system, confirmed that most of the usability problems, e.g., the lack of the practice items, and font size, were addressed already during the less formal evaluation of the system. Nevertheless, for example, T1 noticed that the amount text in the instructions was too excessive to process it easily and suggested that instructions be divided into several parts. T1's suggestion was implemented (**Figure 5**), and during the next stage of the usability evaluation, which was conducted among 23 university learners of English and their lecturer (Article IV), none of the participants experienced any problems with the instructions. Other interface elements that were found problematic during the Pilot study and successfully corrected were the size of the *help* and the *log out* buttons.

All in all, due to the multistage evaluation of the usability, the usability of the *ICAnDoiT* system should be rather high. Nevertheless, a posteriori evidence for the usability of the interface for test designers and researchers is missing.

5.3 Final considerations

The evidence that has been collected for the validity of the *ICAnDoiT* system and the *Questions Test* is mostly favourable, indicating that the *Questions Test* serves its purpose of enabling learners' self-diagnosis and the development of learners' ability to form L2 English wh-questions with auxiliaries.

The process of the validation of the system was also a way to study whether the validation framework introduced by Bachman in Palmer (1996) can be used for validation of a computerised dynamic test. Judging by the current state of the validation of the *Questions Test*, the framework served its purpose, as it allowed for a fine-grained analysis of both qualitative and quantitative data and for observing the way these data were interconnected, contributing to different aspects of the test usefulness, which in their turn, were interconnected, too.

This is not to say that a different test validation framework would have been less useful. If Kane's (2006) argument-based approach had been used, for example, different aspects of test validity might have been highlighted and different data might have been used. This, however, does not make the evidence collected in the process discussed in the present chapter void. It rather suggests that these results should not be considered final. As Messick (1989) noticed, the validation process hardly ever comes to an end. As the system and the test continue to be used, more validity evidence (and/or counterevidence) will be collected.

6 RESULTS

In the present section, I will present the results of the studies separately for each research (sub-)question. I will then, at the end of each subchapter, summarise these results to present an answer to each of the posed question.

6.1 Research question 1

The answer to the research question asking **whether the adaptive feedback facilitated the development of the learners' ability to formulate L2 English wh-questions** included finding out whether this feedback enabled the learners' self-diagnosis. For the most part, the question of whether the DA enabled learners' self-diagnosis and development of their ability to form L2 English stage 5 questions was answered in Article I. The development of the ability being assessed was operationalised as the increase in the learners' unassisted performance, which made it straightforward to establish using the data in Article I. However, to be able to claim that the learners self-diagnosed their own problems as CF adjusted the difficulty of the items within the learners' ZPD, in addition to the learners' questionnaire responses discussed in Article I, findings from Article II (also **Chapter 6.2.1)** and **Chapter 5.2** will also be used to strengthen the claim. I will start by presenting the evidence for the positive effect of the automated adaptive CF.

The results reported in Article I confirmed that the learners who received automated adaptive CF significantly improved their ability to form L2 English wh-questions with auxiliaries, as demonstrated by Wilcoxon signed-rank test conducted on the variables representing the performance of the experimental group on the pretest and the posttest tasks. There was a moderate effect that this feedback had on the learners' ability. What is more, the finding that the experimental group learners improved their performance on the posttest regardless of the feedback they received (**Chapter 5.2.4**) suggests that depending on the learners' abilities any feedback was able to improve the learners' performance.

To support these results, I made sure that it was the adaptive CF that resulted in the changes in the learners' performance and not the treatment tasks per se by including a control (i.e., static CF) group, who completed the same exercises both during the pre-/posttest and the treatment as the experimental (i.e., adaptive CF) group learners did. The two groups' pretest performance was also not significantly different. The only difference between the two conditions was the feedback provided to the learners during the treatment.

The results demonstrated that the improvement of the experimental group's performance was significantly higher than in the control group. The results also demonstrated that there was a moderate effect of the treatment. It can thus be concluded that the significant increase in the experimental group learners' pre-/posttest performance and the difference in performance of the two groups were due to the effect of the automated adaptive CF.

The learners' questionnaire responses and interview utterances corroborated these findings. To start with, as transpired in Article I, the learners who were given adaptive CF considered it significantly more useful for learning than the learners provided with static KOR feedback and generally reported that they realised what the reasons for their mistakes were, for example, "*I did not remember the rule and the feedback helped me to* or I *understood my mistake*." It should also be noted that the learner whose response is used in the example above remembered the rule without it explicitly being formulated in the feedback messages he received during the DA.

What is more, in the interviews conducted in the Group study reported on in Article II, the learners also elaborated that the feedback helped them to learn because they were able to understand what was wrong. This is evident, for example, in **Excerpt 13** (Article II), where LA2 revealed that he was able to learn questions in the Past Simple tense because feedback made it "a little bit more understandable".

However, perhaps the most revealing example that feedback enabled learners' self-diagnosis was the performance of LA1, another participant in the low-achieving group in Article II. It transpired in her utterances that she only tried to read feedback in the ordering tasks because she could not use her other test-taking strategy (i.e., memorising the structure of her responses when these were correct by chance), or, perhaps, these can be interpreted such that she switched to her strategy because most of the CF in the ordering tasks was beyond her ZPD. Importantly, it resulted in that during the posttest, she improved her performance on the questions with modal auxiliaries, which were the target of the ordering tasks during the DA (see **Table 9** in **Chapter 5.2.4**). At the same time, her performance on the rest of the questions types remained the same during the posttest. The only other questions that she can be considered to have improved her performance on were wh-questions with auxiliary *did*, but as I reported in **Chapter 5.2.4**, there is a reason to suggest that LA1 resumed reading feedback at the end of the treatment. That is to say, the results of the

Group study (**Chapter 6.2**) helped to explain the apparent discrepancy between LA1's unassisted and DA performance, but also why LA1 failed the first two ordering tasks during the DA but did very well on the rest of the DA tasks. It appeared that this LA1's strategy was, in part, guided by her belief in the supremacy of good marks over knowledge. In fact, judging by the performance logs of several other learners and the discrepancy between their pretest/posttest and DA performance, other learners might have also used a similar strategy.

6.2 Research question 2

The findings reported on in Article I inspired my interest in studying **how learners' beliefs, especially those regarding the usefulness of corrective feedback, mediated their DA performance** (research question 2.1) and **the experience of DA, in turn, helped to shape their beliefs** (research question 2.2). I the following, I will present the results separately for each of the two subquestions.

6.2.1 Research question 2.1

While the experimental group learners thought that the feedback in the study was more useful than the control group learners did, not all the experimental group learners perceived all of the feedback they received during the *Questions Test* as useful, and some of them even considered all of it useless.

This is evident in the quantitative data analysis revealing that regardless of their abilities, learners tended to rate more explicit and detailed CF as more useful than implicit and less detailed, and the more explicit the feedback was, the higher it was rated (**Chapter 5.2.4**). A similar picture emerged from the qualitative analysis of the Group study data discussed in Article II, that is, as regards the beginning of the interviews. What is more, particularly the learners in the low-achieving group appeared to skip some (or most of the feedback). In fact, as reported in **Chapter 5.2.4**, it seems that only three learners read all the feedback that was displayed to them.

The Group study learners' (i.e., the six learners selected from the sample who took the Questions test; see **Chapter 4.3.2**) utterances during the interviews suggested that the learners skipped the feedback because they considered it useless. Interestingly, judging by the learners' performance logs (see **Chapter 5.2.4** and **Appendix C** in Article II), in some cases when the learners read the feedback types that they otherwise skipped, they were able to self-correct their mistakes (e.g., MA2) and/or improve their unassisted performance (e.g., LA1).

The Group study learners also tended to pay attention to more explicit feedback because, as they generally reported, it showed them what their mistakes were and explained what was wrong (cf. Amrhein & Nassaji, 2010). Interestingly, especially for low-achieving learners, their teacher's feedback practices could have been involved in the construction of the learners' beliefs about corrective feedback. Specifically, their teacher reported that she believed that her low-achieving learners expected only overt correction from her and that her usual corrective feedback involved explanation followed with correction (i.e., the most explicit feedback in the DA).

As regards the Case study, at its onset, the teacher's feedback and M's experience with it also appeared to be involved in the construction of M's beliefs about corrective feedback. However, M probably had negative experiences with the teacher indicating the location of the mistake, as in the beginning, he considered this type of corrective feedback useless.

To summarise, it appears that learners' beliefs about the usefulness of corrective feedback, at least in part guided by their teachers' practices, mediate their performance on DA, and can result in that learners skip feedback/mediation they receive during computerised dynamic tests even tough that same feedback is within their ZPD.

6.2.2 Research question 2.2

While the learners' beliefs about corrective feedback appeared to mediate their DA performance, the results of both the Group study and the Case study in Article II suggest that dynamic assessment has the potential for transforming learners' beliefs about correct feedback. In fact, the hypothesis that DA can transform learners' beliefs about corrective feedback was inspired by the findings reported on in Article I. Specifically, I suggested that the significant differences between perceived usefulness of the feedback of the control and the experimental group learners in the study (see **Chapter 6.1**) could in part be explained by the changes in the learners' beliefs about the usefulness of CF triggered by their DA experience. Certainly, though, the change was probably also due to the fact that with the exception of more able learners, the adaptive CF that the experimental group learners received was more detailed than that in the control group.

As regards the studies in Article II, above all, the changes due to the DA experience were evident in the Case Study, where M changed his initial opinion that feedback locating mistakes was the most useless to thinking that this feedback, but also feedback about the nature of mistakes, was among the most useful for him. He still reported that examples of the correct structures would be useful, but only for self-correction, not learning. It was also notable that these emerging beliefs about corrective feedback developed further.

In half a year after the second interview, M appeared to have considerably appropriated (judging by the lack of active mediation from the interviewer) the belief that both feedback locating mistakes and indirect feedback about the nature of mistakes would be the most useful, as these feedback types make him think. However, he reconsidered his opinion about the usefulness of examples of correct structures, thinking that such feedback would not help him to understand what his mistakes were.

M's performance during the dynamic assessment illuminated the reasons for these changes, as especially during the later human-mediated DA sessions, M was often able to self-correct with help of the CF types he reported to be useful after his experience of DA. He also often benefitted from the examples of correct structures, but seemed not to understand why these were correct. However, it was also the interviewer's mediation, for example alternative questions from the interviewer, but also the mere presence of the interviewer (who was also the mediator during the DA), that directed the way M reported on his beliefs.

Collectively the results of the Case study and the Group study revealed that it was a combination of the learners' experience with the DA, CF from their teacher, the teacher's voice, and the mediation from the participants in the interaction, including, but not limited to, the interviewer, that brought about these changes. Below, I will illustrate this by summarising the changes in LA2's (a participant in the low-achieving group) utterances about the usefulness of CF and what facilitated these changes.

In the questionnaire, LA2 considered explicit explanation and overt correction to be the most useful during the DA. During the interview, this belief started to change. The change was triggered by LA1, who, discussing the feedback from the teacher, reported that she believed it was useful for learners (including herself, as she used the pronoun *we*), as it made them think and make connections. While uttering it, she pointed at the printed out sample feedback messages from the *Questions Test* and noted the similarity of the feedback from the teacher and the feedback in the study.

This resulted in that soon after the episode with LA1, when discussing the most useful feedback in the study, LA2 initially pointed at the examples of the correct structure (i.e., level 4 feedback) adding that it made them think (i.e., used the same word that LA1 had used) but used a slightly rising intonation, which indicated hesitation. Only after some hesitation, he added that the overt correction was helpful, too. The interviewer then mediated this emerging belief by asking whether he had to think less when provided with the correct response, which LA2 confirmed.

Towards the end of the interview, when the interviewer asked the learners whether it would be more useful for them if they had been provided with the explicit explanation and the overt correction whenever they made mistakes during the DA, LA2 responded that it would not, as such feedback did not make them think. He, thus, verbalised the idea that feedback that did not explicitly explain their mistakes was useful because it make them think, this time without any hesitation.

There were other examples in Article II demonstrating how the learners' beliefs were co-constructed together with the other participants in the interaction. For example, the interviewer asked HA2 whether he had received implicit indications that there were mistakes during the DA. Following that, HA2 suggested that teachers should first of all hint that there is a mistake, so that learners could try to find it themselves. Before this episode, however, HA2 considered this feedback type to be useless. The best way to trace these changes would be to refer to the original publication, as the given examples do not allow for fully visualising the way that the learners' beliefs emerged and began transforming.

The results suggest that the participants' DA experience, the interviewer's mediation, other participants' mediation, and teacher's voice / feedback practices helped to co-construct and transform the learners' beliefs about corrective feedback.

To summarise the response to research question 2, learners' beliefs about corrective feedback appeared to mediate their DA performance and the way they reported on their DA experience. However, the experience of DA itself and the way it was mediated in social interaction resulted in that different beliefs emerged. What is more, the learners started appropriating these new beliefs already during the short amount of time the research interviews lasted, and in M's case, some of these beliefs were further appropriated after a six-month period and developed further.

6.3 Research question 3

Finding out **how to ensure the usefulness of adaptive CF in a computerised DA of L2 English word derivational knowledge** required a synthesis of the findings of all of the studies conducted as a part of the present doctoral research project. The generalisations made based on the results of these studies were evaluated empirically in the study reported on in Article V.

6.3.1 Research question 3.1

The results reported on in Articles I and II as well as the results discussed in Chapter 5 of the present chapter allow for making several **generalisations regarding some principles for designing (computerised) dynamic tests and, especially, adaptive CF in these tests**. In the following, I will summarise these with reference to assessed construct, mediation, and learners' beliefs.

Defining the development of questions in terms of question development stages had the advantage of interpretation of learners' unassisted performance in terms of these stages (i.e., whether the learners were able to produce stage 4 and stage 5 questions before, during, and after the treatment). In addition, basing it on the available research on the development of questions, allowed me, for example, to select the participants in the study so that the assessed structure was within their ZPD (see **Chapter 5.2.2**).

Furthermore, basing the items on the results of the *CEFLING* project (allowed for finding out the common mistakes learners' made and address them properly in the multiple-choice tasks and in the feedback in response to these mistakes. Additionally, following Lee's (2015) suggestion, the adaptive CF both hinted what was wrong in the learners' responses and elicited the parts which were correct. The results of Article I suggested that these decisions were correct, as the adaptive CF promoted the abilities being assessed whereas static corrective feedback did not. These results also suggested that it is essential to define / operationalise the development of the construct being assessed, that is, what it is that demonstrates that the development occurs.

To sum up, in order to provide adequate mediation, the assessed construct, its aspects, and the development of these aspects should be studied (see also **Chapter 3**). Care should also be taken so that the assessed features are within learners' ZPD. Finally, especially for computerised DA, learners' common mistakes should be studied so that both the tasks and the mediation address these adequately during the assessment.

Importantly, the results of the qualitative analysis of the learners' questionnaire reported in **Chapter 5.2.4** revealed that the adaptive CF in the Questions test directed the learners to use various strategies, including those that were not elicited by the feedback. Specifically, the qualitative analysis of the learners' open-ended questionnaire responses indicated that while the experimental group learners reported on using various strategies (not limited to the ones I hypothesised that the feedback should train), the control group learners did not. Combined with the fact that significantly more experimental group learners benefitted from dynamic assessment than control group learners, this suggested that the adaptive CF promoted the use of strategies which, probably, contributed to the improvement of the learners' ability to form wh-questions with auxiliaries beyond the context of the treatment exercises.

That said, the results of Articles I and II demonstrated that not all experimental group learners benefitted from the *Questions Test*, and, at least for some of them, this was due to their beliefs about corrective feedback. A remedy for the problem, as found in Article II, were discussions with learners in which they recall and reflect upon their DA experience.

To sum up:

- defining and operationalising the assessed construct, its aspects, and their development as well as informing assessment by and basing the mediation on findings of second language acquisition research is important in (computerised) DA;
- mediation should also promote the use of learning strategies (see also **Chapters 2.6** and **2.7**);
- learners' beliefs about corrective feedback should, at least, be studied prior to them taking a computerised dynamic test, and, if required, mediated so that these beliefs do not hinder learners' performance on the test (or, at least, minimise this possibility).

The following two sections will present the results of the two studies that aimed to address the first point of the list above.

6.3.2 Research question 3.2

The response to the question of **what is characteristic of the construct of L2 English word derivational knowledge** is based on the findings of Articles III and IV. I will supply the results informing the responses to questions 3.2.1 to 3.2.2 in the present section.

As regards **the difficulty of derivational affixes**, the results of the study reported on in Article III largely supported the order of derivational affixes proposed by Bauer and Nation (1993) as a (the) difficulty order. Specifically, the results of the repeated measures ANOVA indicated that there was a significant difference in the learners' ability to recognise derivational affixes at different Bauer and Nation's levels. The levels alone accounted for 66% of the variance in the learners' performance. The pairwise comparisons that followed indicated that with the exception of no significant difference between level 5 and level 6 affixes, the higher the level was, the significantly fewer affixes were recognised by the learners. The results of the non-parametric data analysis corroborated these findings. Thus, the results provide empirical evidence for Bauer and Nation's levels.

As regards **the relationship between learners' L2 English proficiency and word derivational knowledge**, the results of the study reported on in Article IV demonstrated that with the exception of the word segmentation task, the higher the learners' proficiency was, the better they performed on the tasks, which added to the previous research on the relationship between learners' word derivational knowledge and proficiency. In addition, as the results of the ANOVAs demonstrated, the increase in the performance across the CEFR levels was not stable in some of the measures. In most measures there was either a more or less similar increase in learners' performance across the CEFR levels or a bigger difference in performance of learners between levels B1 and B2 rather than between levels A2 and B1. It is also interesting to note that while the learners' performance all of the WD tasks (save the word segmentation task) correlated with their writing proficiency, only the learners' performance on three tasks predicted their proficiency, accounting for about 57% of the variance in it. These were a grammar recognition (where the learners were asked to select the missing word in the sentence among the three different parts of speech), a prefix elicitation (where the learners were asked to add a prefix to a word in the sentence, selecting it among provided), and a meaning recognition (a task similar to the grammar recognition, but the options were the same part of speech) tasks.

To sum up, the results of the two studies suggested that:

- the order of derivational affixes proposed by Bauer and Nation (1993) can be used to account for the difficulty learners have with these affixes;
- syntactic and semantic knowledge of derivational affixes are especially relevant for the development of their WD knowledge, increasing as their L2 English proficiency grows, and predicting this proficiency;
- there is a particularly strong improvement in learners' L2 English word derivational knowledge after learners reach level B1 of their L2 English proficiency on the CEFR scale.

In the following section, I will report on the results of the study where these findings were used to design and try out a human-mediated and a computerised dynamic test of learners' L2 English word derivational knowledge.

6.3.3 Research question 3.3

The aim of Article V was to find out **how, if at all, dynamic assessment promoted one learner's L2 English word derivational knowledge**. The design of the study was informed by the findings of the other studies forming the present doctoral dissertation (see **Chapter 4.3.5**).

The results indicated that with the help of both the human-mediated and the computerised DA, M' performance increased across most of the measures. It is notable that the computerised DA resulted in a similar increase in M's performance as the year and a half time period between the human-mediated and the computerised DA did. That is to say, the improvement of M's unassisted performance during the last SA session (i.e., after the computerised DA) as compared to the static assessment session 3 (i.e., before the computerised DA but a year and a half after the second SA session) was similar to that between SA sessions 2 (i.e., after the human-mediated DA) and 3.

It should also be noted that with only a few exceptions, M required either the same or, more often, less assistance with transfer items (i.e., items with the same affixes appearing at later DA sessions) during both the human-mediated and computerised DA (e.g., with suffixes *-ess* and *-ness* in human-mediated DA; see Articles II & V).

In part, M's progress can be attributed to the fact that he learned the meanings and syntactic roles of some affixes, as the results demonstrated. For example, in a year and a half after the human-mediated DA, he recalled the meanings of some affixes he had been taught during the human-mediated DA, using similar words to define them as the mediator had used when guiding his performance. However, after the DA, M also improved his performance on some affixes that were not trained during the DA sessions. The analysis of the think aloud and the interview data illuminated the reason for that. Before the DA, as the frequencies of the use of different strategies demonstrated, M rarely analysed words morphologically and excessively relied on syntactic knowledge, failing to think about the semantics of the words/affixes even in obligatory contexts (e.g., the meaning recognition task; see Article IV). Following the DA, M analysed the words more often, paying close attention to affixes and their meanings.

It is worth noting that M's performance reflected the way mediation was provided to him during both the human-mediated and computerised DA. During the human-mediated DA, M made recurrent errors in the parts of speech, which resulted in that the mediator elicited both the syntactic function and the meanings of the affixes. Following the human-mediated DA, this was the most usual approach M used when working on the static assessment tasks, that is, he first mentioned the parts of speech of the words/affixes and then thought about their meanings. He never made mistakes in parts of speech during the computerised DA, so these were never mentioned in the automated adaptive CF provided to him. Following the computerised DA, M mentioned the syntactic function and the meanings of the affixes together considerably less, preferring to think only in terms of the meanings of the affixes and even referring to his semantic knowledge in the tasks where syntactic knowledge would be helpful, too.

That certain strategies were trained during the DA does not mean that owing to the adaptive CF, M learned new strategies to utilise when working on the tasks requiring him to demonstrate his word derivational knowledge. On the contrary, M used all of the strategies identified in the analysis of the think-aloud protocol and the interview data already during the first static assessment session. It can even be argued that it was not the frequency of use of certain strategies that DA resulted in but the appropriate use of these strategies. For example, M learned not to rely only on syntactic knowledge in order to find the correct answer but started using several strategies (e.g., L2 analogy, mother tongue analogy, checking the response against the wider context, analysing the words morphologically, etc.) to evaluate and corroborate/disprove his initial assumption when not sure in his response or use just one (e.g., analysing the word and thinking of the meaning of the affix) when his certainty in the correctness of the response was high. In other words, the DA resulted in that M increased his self-regulatory capacity for solving the tasks requiring demonstration of L2 English word derivational knowledge.

Interestingly, M's performance logs recorded during the computerised DA indicated that he skipped the feedback only thrice during all three computerised DA sessions (spending three seconds or less on it). What is more, in two of these three cases can be considered the manifestation of his abilities, as his following response was correct (see Poehner, 2005, for a discussion of changes in learners' reciprocity to mediation as an indication of their development). Certainly, as there were no data demonstrating the way M read/skipped the CF during the DA prior to the interviews, it cannot be claimed that the interviews conducted with him in the Case study mediated his performance on the computerised DA. However, there is evidence that M's human-mediated DA experience and the interaction with the interviewer mediated M's beliefs about corrective feedback (see **Chapter 6.2**), which could have influenced his performance on the computerised DA.

The results of the study, thus, demonstrated that basing the automated adaptive CF on the findings of research of the assessed construct and designing the feedback so that it elicited the use of strategies allowed for promoting the learner's performance on the tasks measuring the knowledge of such idiosyncratic feature as L2 English word derivational knowledge.

In the following **Chapter 7**, the results reported on in the present chapter will be discussed with reference to the aims of the present doctoral research project. I will also list the limitations of the present doctoral research, including the consequence of studying only one participant in this latter study.

7 DISCUSSION AND CONCLUSION

7.1 Discussion

The present doctoral research project aimed at providing evidence for a positive impact of corrective feedback provided within learners' ZPD and at studying the ways the usefulness, specifically, the impact of such feedback can be increased. Collectively, the articles fulfil these aims by approaching them from different angles, that is, studying the effect of the feedback on / the role it has for learning, the changes it brings about in learners' beliefs and in the way they work through the tasks, and how to use theoretical and practical research findings to conceptualise and operationalise adaptive corrective feedback.

Not all of these aims, however, were clearly defined at the onset of the research, but rather they gradually developed over the course of the doctoral research project. That is, as answers to the posed questions were found, new questions arose, for example, *why is there no significant difference in the way learners perceive the KOR and the adaptive CF as useful for self-correction? Why did some learners skip the adaptive CF and report that it was useless for them? How can the development of such idiosyncratic ability as the ability to derive words in English be operationalised? Can learners' success of in recognising/producing only the affixes being trained be considered as a manifestation of this development?* Needless to say, the two major aims of the doctoral research project, that is, studying the impact of adaptive feedback and ways to increase the positive impact of it, were tightly interrelated, the findings regarding one aim also promoting the understanding of the other.

Grounding the study within the sociocultural theory of learning guided the whole research process, including the methods, the data analyses, and my own role in the studies, that is, a researcher on the one hand and a participant in the interaction on the other hand. Adopting this dual role had its advantages in that, for example, during the study of the learners' beliefs, I was able to analyse the way the interaction unfolded in real time, helping the learners to reconstruct (and co-construct) their experience with the dynamic assessment. On the other hand, when coding, analysing, and interpreting the data, it was also required to step aside and try to perceive what was going on as if not involved in the interaction. Asking a second person to assist me with coding and analysing the transcript and using methodological triangulation to corroborate the interpretations helped me to do it. In the following, I will discuss the results of the studies with reference to the adopted framework and demonstrate how they contributed to the two major aims of the study.

The study of the effect of the adaptive corrective feedback might be perceived as a departure from the sociocultural tradition. Instead of studying the interactivity between the learners and the feedback, the data were analysed statistically. However, despite the predominantly quantitative analysis of the data, which demonstrated that the adaptive CF in the study was indeed effective in promoting the learners' abilities, the learners' performance was also interpreted with reference to their questionnaire responses regarding the usefulness of the feedback. That is to say, their performance was assumed to be mediated not only by the feedback during the treatment but also by their prior experience with assessment and their beliefs about what corrective feedback should be like, which guided the way they perceived the usefulness of the feedback in the study and the whole assessment procedure. The instructions that were supplied to them, that is, not referring to the procedure as a test but rather instructing them that these were exercises which aim was to help the learners realise how well they could form questions in English (see **Chapter 5.2.1**), were also thought to direct the way the learners perceived the procedure. What is more, it was not just their ability to form wh-questions with auxiliaries that was constructed dynamically during the procedure but also their beliefs about corrective feedback. In fact, it was assumed that while taking the test, but also when reporting on their experience with it in the questionnaire, and especially during the interviews that followed for some of the learners, the learners challenged, reconstructed and co-constructed their beliefs about the usefulness of corrective feedback. Thus, the experimental group learners' questionnaire responses regarding the usefulness of corrective feedback collected immediately following the DA served both to demonstrate that they were indeed able to find their mistakes, realise the reasons for them, and self-correct them (see, e.g., Alanen, 2013, for a discussion of awareness with understanding) and to find out whether the learners (both in the experimental and the control group) believed it was useful for them, which was then used to corroborate the findings.

The experimental group learners found the feedback significantly more useful for learning than the control group learners did and generally reported that they realised what the reasons for their mistakes were, while the control group learners, for the most part, did not. I interpreted these findings and the findings regarding the effects of the two feedback conditions with reference to the learners' ZPD (see also Nassaji & Swain, 2000). The adaptive feedback adjusted the difficulty of the items to match the learners' abilities, which resulted in that more learners than in the control group realised what their mistakes were, what the correct responses were, and ultimately, improved their unassisted ability to formulate wh-questions with auxiliaries. One particular piece of evidence for this interpretation serves the performance of LA1, whose performance improved but only on the particular question types that were trained when she read the feedback during the DA (see **Chapter 5.2.4**). I would interpret it so that before the DA, LA1 was other-regulated in her use of stage 5 questions, which also transpires in her performance on the ordering tasks, as she could not find any correct responses and had to be given an explicit explanation of her errors and be provided with correct responses. However, this same feedback resulted in that she was able to improve her unassisted performance on wh-questions with modal auxiliaries. At the same time, since she skipped feedback in the rest of the tasks, she could not produce other stage 5 questions during the posttest, as her strategy for taking the rest of the tasks appeared not to be beneficial for promoting her development (although, true, helped her to find the correct responses). In contrast, in the control group, presumably, only the learners for whom KOR feedback was within their ZPD improved their unassisted performance.

That said, it would not be entirely correct to speak about the learners' development in terms of their unassisted performance only. As the dynamic assessment of M's word derivational knowledge suggested (Case study in Article II; Article V), it was not just M's unassisted performance that developed due to the DA, but also his mediated performance did, as emerging in the change in the quality of mediation, especially that on M's performance on transfer items, provided to M across the different DA sessions (cf., e.g., Aljaafreh & Lantolf, 1994). Thus, should there have been several DA sessions in the study reported on in Article I, more evidence for the development of the learners due to the DA could have been collected.

In fact, as transpired in Figure 4, the mediation in the treatment part of the Questions test also resulted in that at later tasks in the treatment, on average, more items were solved correctly in the experimental group than in the control group and thus, on average, less mediation was provided in later tasks. Therefore, it can be suggested that some of the learners whose unassisted performance did not increase due to the DA, still benefitted from the procedure.

These findings add to the previous research on the effect of adaptive CF (and other mediation) provided within learners' ZPD (e.g., Teo, 2012) in that similarly to Nassaji and Swain (2000), I contrasted adaptive and static corrective feedback, but unlike Nassaji and Swain, studied the effect of adaptive CF through inferential statistical analysis.

It is, however, important not to forget that considering the usefulness of DA only in terms of whether and how it promotes learners' development would be underusing the possibilities that DA allows for. The diagnostic value of DA is in that it yields insights into learners' emerging abilities. Studying the diagnostic value of computerised DA beyond its diagnostic value for learners was not among the aims of the present doctoral research project. However, the findings used as evidence for the validity of the Questions test (**Chapter 5**), particularly those pertaining to the learners' performance logs can be discussed in this respect. For example, that the pilot study teacher considered some of the information in the learner profiles (built based on the learners' performance logs) new, despite positioning herself as an adherent of scaffolding and thus, probably being aware of how much assistance her learners needed, illustrates how learners' DA performance reveals more than static assessment can do, including the assistance that learners need at the time of DA with certain features and, importantly, what the learners will be able to achieve in future without any assistance.

Another source of diagnostic information about learners' ZPD in computerised DA, as the results of the present research suggested, can be the time that learners' spend attending to mediation/feedback. That is to say, depending on the following responses, that learners skip mediation/feedback in computerised DA can be an indication of either that they do not need the mediation because of being more self-regulated (i.e., the amount of assistance is needlessly too much for them) or that, in their opinion, the provided assistance is not enough to improve their answer. However, in addition to the learners' ZPD, their reciprocity is also indicative of their beliefs, as I will discuss next.

The fact that more learners in the experimental group considered the feedback more useful for learning than the control group learners, I suggest, should have also contributed to its beneficial effect in the former group. At the same time, there appeared to be no difference in the learners' perception of the usefulness of the feedback for finding correct responses. This motivated a deeper study of the way the learners perceived the usefulness of the feedback in the study, the way their beliefs about corrective feedback mediated their DA performance, and whether the experience of DA could indeed result in a transformation of such beliefs.

The contextual, specifically, the sociocultural approach adopted for the study of learners' beliefs allowed for producing interesting insights into the way the learners (co-)constructed their beliefs about corrective feedback with reference to their DA experience. As the results of Article II demonstrated (see also **Chapter 6.2**), at least some of the experimental group learners skipped the feedback because they believed that it was useless. At least for LA1, this seemed in part to be the consequence of her considering the procedure to be a common test. It can be, thus, suggested that depending on whether learners consider DA as merely an activity where they demonstrate their abilities (i.e., what they have learned so far) or an opportunity for learning, their reciprocity to mediation can be different. In fact, this can explain why the learners (both in the studies reported on in Article I, Article II, and in the Pilot study) were, initially, generally in favour of more explicit feedback. That is, if the learners considered the procedure as an assessment, they would consider that feedback the most useful which directed them to the correct answers quickly and with less effort on their part. Perhaps, referring to the Questions test as an activity helping the learners to learn how to formulate questions instead of exercises showing them where they had problems could have reduced the number of learners who assumed that their task was to demonstrate their abilities .

On the other hand, it should not be forgotten that a similar picture emerged in contexts other than DA (e.g., Amrhein & Nassaji, 2010; Ashwell, 2000). Therefore, Amrhein and Nassaji's (2010) explanation that learners learn to believe that it is teachers' responsibility to correct their mistakes is also plausible.

Related to the above discussion is the finding of Article II that less-able learners, at least those participating in the Group study, still appeared to use the feedback less than high-achieving learners (even when CF seemed to be aligned with their ZPD). From the sociocultural perspective, it can thus be suggested that learners' abilities should be developed enough for them to recognise the areas where they have difficulties and realise that particular forms of mediation can be helpful for them (see also Poehner, 2012). Therefore, the findings of the present research regarding the involvement of learners' beliefs in the way they interact (or choose not to interact) with mediation in computerised DA should be considered as adding to the findings of, for example, Poehner (2005) regarding learners' reciprocity as an indication of their ZPD (rather than seeing these in opposition to each other).

It seems, thus, that learners' beliefs can lead to that they underuse the learning opportunities that computerised DA offers. At the same time, it appears that within social interaction during which the learners' recollections of their DA experience was mediated by the interviewer's and their peers' utterances, their beliefs about corrective feedback started transforming. Particularly the Case study reported on in Article II demonstrated how DA helped to transform the learner's beliefs about CF and how these were appropriated by the learner and further transformed with time. The Group study, on the other hand, showed more clearly how the participants in the interaction co-constructed their beliefs about corrective feedback, using their DA experience, their experience with the teacher's feedback, and each other's utterances to mediate the way they reported on the usefulness of corrective feedback.

It should be highlighted that not only the learners entered the interviews with certain beliefs about corrective feedback in general and the corrective feedback in the study in particular, but also I, being the interviewer, did. However, as it also transpires in the transcript (see Article II), I consciously did not explicitly impose my beliefs on the learners, instead mediating their recollections of their DA experience by asking questions, helping them to remember something from their experience with the CF in the study that they otherwise did not notice (or, at least, did not report on it). Importantly, it was not just my own mediation that helped them to construct their utterances. Other learners who participated in the interview also guided the way that the interaction unfolded. One particular example was when LA1 reported on the usefulness of the feedback from the teacher, adding that it made them think about their mistakes, and then made a connection between the teacher's feedback and the feedback in the study. This episode resulted in that other learners also started referring to the feedback in the study as useful because it made them think.

These results are in line with the other studies of learners beliefs conducted within the contextual approaches (e.g., Alanen, 2003; Aro, 2009; Barcelos, 2003; Barcelos & Kalaja, 2013; Dufva, 2003; Kalaja & Barcelos, 2013). However, the originality of the study lies in that the epistemology (and methodology) applied to the study of beliefs about corrective feedback, and experience of dynamic assessment was used to mediate the way learners (co-)constructed their beliefs in social interaction. Thus, while the learners' beliefs about corrective feedback emerged as a potential threat to the validity of the dynamic assessment, the dynamic assessment proved to be useful for transforming these same beliefs.

One difference between the Case study and the Group study reported in Article II was that M required less mediation from the interviewer than the learners in the Group study did before his beliefs about corrective feedback changed. One explanation for that can be a longer treatment in M's case, that is, three DA sessions experienced by M as compared to one taken by the Group study participants. Alternatively, the modality of the DA, that is, human-mediated as compared to computerised, and that the interviewer was also the mediator during the DA could have played its role.

As implied in the contextual approaches to the study of beliefs, and as Article II demonstrated, contexts in which beliefs emerge are unique in each particular situation, and are co-constructed within the interaction which unfolds within these contexts. Therefore, especially considering the small scale and the designs of the two studies, I would advise against using the findings as a way of selecting specific types of mediation learners might require with other grammatical/lexical features or mediation learners might require in other contexts to start transforming their previously held beliefs. That said, the findings illustrate that discussions during which learners co-construct their beliefs about CF should facilitate transformation of at least some learners' beliefs. The findings also give an idea of how computerised mediation guides the development of learners' L2. Thus, they can provide a starting point for designing mediation in other computerised dynamic tests.

The results of the two studies allowed for making several generalisations which helped to design the automated adaptive CF in a dynamic test of L2 English word derivational knowledge (see Chapter 6.3.1). One of them was taking learners' beliefs into consideration or/and transforming these beliefs. This, as I have already argued, should be particularly important in computerised DA, where learners' reciprocity to mediation (except for studying the time they spend on mediation, e.g., as in the case of the present research, on reading the CF) is hard to trace, and thus, it is difficult to make certain that learners attend to all the mediation they are provided with and if they don't read it, to establish their reasons for skipping it.

Another suggestion was based on the previous research regarding diagnostic and dynamic diagnostic assessment and the findings of Articles I and II. It concerned the assessed construct, specifically, defining the development of the ability being assessed. Defining the development of questions in English in terms of question development stages had its disadvantages in that it is difficult, if not impossible, to combine the epistemological basis of the question development stages with the sociocultural paradigm underlying dynamic assessment (e.g., Dunn & Lantolf, 1998). On the other hand, it helped me to focus the assessment on one particular question type and interpret the learners' unassisted performance in terms of the stages, that is, what kind of questions (and how many of them) the learners were able to produce before, during, and after the treatment). In this respect, it is worth mentioning that all experimental group learners produced several correct stage 5 questions during the treatment (although, during the pretest, some of them failed to formulate any), which suggests that this construct was within their ZPD.

There was an apparent problem with defining the development of learners' L2 English word derivational knowledge although it was different from the one I faced when defining the development of questions in L2 English. Namely, unlike the latter, L2 English word derivational knowledge is a severely under-research area. I could have relied on the scarce (and sometimes conflicting) findings of the previous research (see **Chapter 3.4.2**), but this might have resulted in inaccurate mediation, learners' incorrect self-diagnosis, and, as a result, decreased positive impact of the DA. Adopting an interactionist approach to DA, as Ableeva (2010) did, would help to determine mistakes that particular learners make when working on the tasks. However, the problem would have still persisted how to operationalise WD knowledge and its development as well as to generalise these mistakes and ways to go about them to other contexts (and learners), which is an aim of interventionist DA, mediation in the latter being the object of the present study. This is not to say that I failed to appreciate the qualitatively different findings that could have been discovered based on M's performance on an interactionist DA, that is, the mistakes that are typical for him and the assistance which is tightly attuned to his abilities. In fact, one of the aims of the human-mediated DA of WD knowledge was to discover whether the usefulness of the designed order of adaptive feedback in promoting the learner's WD knowledge can be improved, and thus, some departures from the otherwise standardised mediation were allowed in the human-mediated DA (see **Chapter 4.3.5**). That said, relying on M's performance on an interactionist DA to operationalise the mediation in an interventionist computerised DA would have been unwise in terms of potential generalisability and future research into computerised DA of WD knowledge.

Thus, I decided to go a different route and instead to find (more) empirical evidence for the theoretical assumptions made regarding the construct of WD knowledge. This research strand had two separate directions—(a) attempt at finding empirical evidence for the difficulty order of derivational affixes proposed by Bauer and Nation (1993), and (b) an exploration of what aspects the construct can consist of and the relationship between these aspects and learners' proficiency.

The results of the two studies provided useful insights into designing and implementing the computerised DA reported on in Article V. Specifically, it allowed for

- deciding on the difficulty of the derivational affixes used in the tasks, more difficult affixes appearing in later DA sessions;
- eliciting first syntactic roles of affixes and then their meanings in the feedback;
- deciding on the proficiency level of the potential participant in the study.

The adaptive CF on the learners' performance on formulating L2 English questions appeared to result in that the learners generally approached the tasks in the *Questions Test* more strategically whereas the static feedback group did not. Thus, also drawing upon the discussion in **Chapters 2.6** and **2.7**, I assumed that for such idiosyncratic feature as L2 English word derivation, promoting learners' strategic learning should enable their development beyond the task at hand, that is, improve their performance on affixes other than those being trained during DA. Thus, the adaptive CF in Article V was designed to promote learners' ability to analyse words morphologically and, based on the results reported on in Article IV, to elicit both syntactic roles and semantics of derivational affixes.

Perhaps the most interesting finding of the final study was that M also improved his use of strategies not directly elicited by the feedback he received during the DA. That is to say, in line with the assumption made in **Chapter 2.7**, the adaptive corrective feedback, being delivered in M's ZPD, not only guided M in the use of the strategies that were hypothesised it should. It also increased M's overall self-regulatory capacity, as, presumably, he found out which strategies and combinations of them worked for *him* during the DA and was able to successfully apply these during the static assessment. It should be stressed that M did not learn any new strategies but rather learned to self-regulate his learning by finding out his way to solve the tasks requiring him to demonstrate his ability to derive words, using the strategies he already had in his repertoire more appropriately. This latter finding can be interpreted with reference to transcendence (e.g., Feuerstein & Feuerstein, 1999). That is, that M was able to successfully use his prior knowledge in new context (and with new derivational affixes) is a strong indication of his development.

As regards the smaller increase in M's performance after the computerised DA, although I assumed that there should not have been a novelty effect due to the human-mediated DA, as M reported he had been instructed in L2 English word derivation, there still could have been due to the qualitatively different instruction that M received in the classroom and during the study (see the Case study in Article II). Furthermore, during the year and a half gap between the human-mediated and the computerised DA sessions, M also increased his vocabulary knowledge (see **Chapter 6.3.3**), which could have been the reason for the increase of his unassisted performance between the second and the third static assessment session. Even so, the increase in M's performance after the computerised DA was comparable to, or higher than, that which happened in a year and a half time. Finally, during the last static assessment session, M performed at ceiling on several tasks, so it can be suggested that he would have had a higher score should there have been more items in the tasks.

That M referred to his vocabulary knowledge when working on the tasks designed to elicit his WD knowledge suggests that at least in the study reported on in Article IV, the learners, too, could sometimes refer to their vocabulary knowledge instead of attempting to analyse the words. With reference to some psycholinguistic research I outlined in the present synthesis (e.g., Clahsen & Neubauer, 2007) and in Article III (e.g., Ullman, 2004), In article III, I discussed this possibility with reference to the frequency and the semantic transparency of the words (also **Chapter 4.3.3**), but the same could also have happened due to M's and other learners' beliefs, for example, their belief of how vocabulary is learned the best (e.g., by memorising separate words). The latter is more of a speculation, as I did not study the reasons for M using his vocabulary knowledge. This finding, however, has implications for future research of word derivational knowledge, as I will discuss in the following chapter.

All in all, the results of the final study strengthened the decisions made based on the findings of the rest of the studies forming the present doctoral research project, as the results suggest that M developed his ability to derive words. That said, since there was only one participants in the study, its results lack generalisability, as I will discuss in **Chapter 7.3**. Before that, however, I will discuss some of the implications that the results of the present doctoral research have.

7.2 Implications

In addition to being informed by theory and research in SLA, the present research project was impact-driven (in the sense the term *impact* is used in test validation studies; see **Chapter 5.1.2**). In other words, throughout the research process, I studied how the experience with DA changed the learners' performance, the way they perceived the usefulness of corrective feedback, and the way they positioned themselves in the learning process.

The findings of the study have several practical implications. Above all, the results demonstrated that at least as regards L2 English questions, learners' performance can be improved with help of an interventionist computerised DA. Despite the lack of generalisability, the same can be suggested about a DA of learners' word derivational knowledge. That is, with help of the *ICAnDoiT* system or a similar one, teachers' can simultaneously assess their learners' abilities and promote them. Furthermore, the practicality of the computerised delivery and the interventionist approach to DA lies in that several learners can be assessed simultaneously, and teachers, thus, do not have to find out how to address each of their learners' problems, the issue mentioned by Truscott (1996). The computerised DA (provided it records sufficient details of learners' performance) should allow for making inferences about the amount and type of assistance their learners need with specific mistakes.

Another, and perhaps, more important, implication arises from the findings regarding learners' beliefs about corrective feedback and the role that DA can play in the development of these beliefs. Despite the lack of generalisability of these findings, they suggest that experience of DA reflected upon in discussions with learners can facilitate transformation of learners' beliefs about the usefulness of corrective feedback that does not give away correct responses, which should decrease the possibility that learners disregard such feedback from their teachers or during computerised DA. In fact, judging by the participants' responses, it can also be suggested that such discussions might result in that learners become more responsible for their own learning, thinking about the reasons for their mistakes instead of considering it to be their teachers' responsibility to correct their errors.

One suggestion that was made in Article II to account for the possible frustration of learners for whom implicit feedback is beyond their ZPD (see also e.g., Lee, 2015), as argued in Article II, could be adjusting the starting level of the feedback displayed to these learners in computerised dynamic tests so that these learners do not receive the feedback that is clearly beyond their ZPD (as can, for example, be determined by their teachers). After their ability develops, perhaps, next time they take the test, the adaptive feedback settings could be put back to the default, that is, all feedback levels are displayed.

Finally, the studies regarding the construct of L2 English word derivational knowledge have practical implications as well. As I argued in Article III, the results suggest that teachers can use Bauer and Nation's (1993) affix levels as a reference for the properties of the affixes that present difficulty for learners, for example, instructing less able learners in the use of easier affixes, leaving teaching of more difficult affixes for later instruction.

As regards the theoretical implications, a major motivation for the present doctoral research was finding evidence for the effect of adaptive CF, an aim which the present research achieved. Thus, it adds to the previous research on corrective feedback in general and corrective feedback provided within learners' ZPD, presenting a stronger claim in favour of using the latter.

In addition to the quantitative evidence for the effect of adaptive CF, the results suggest that the feedback in both the *Questions Test* and the computerised dynamic test of learners' word derivational knowledge promoted the use of learning strategies, including those that were not explicitly mentioned in the feedback. This supports my suggestion that feedback provided within learners' ZPD should have most or all the levels identified in Hattie and Timperley's (2007) model (see **Chapter 2.7**).

The exploration into the way learners' beliefs mediate the way they work through DA, skipping or accepting the feedback provided to them, should add to DA research and research on corrective feedback. Specifically, it can serve an explanation for findings of studies like Thouësny (2011) and other studies establishing that learners underuse the possibilities provided to them in computerised assessment and other computer-assisted language learning systems. It should also add to research on validation of (computerised) interventionist DA, suggesting that learners' beliefs about corrective feedback should be accounted for when validating such tests. Perhaps, mediating learners' beliefs by means of a tutorial DA session and a discussion following it before conducting the computerised dynamic test proper could be considered as a part of computerised DA procedures. In addition, adding the functionality to record time that learners spend on reading feedback in computerised DA (as in the *ICAnDoiT* system) should produce valuable insights into learners' responsiveness to mediation and can thus be considered when designing computerised dynamic tests.

The exploration of the construct of L2 English word derivational knowledge provided interesting insights into its operationalisation in studies on word derivation in English as a second/foreign language. For example, it provided empirical evidence for Bauer and Nation's order of derivational affixes as a difficulty order, which produced a stronger case for using it for manipulating the difficulty of derivational affixes. Importantly, it also provided insights into the ways of defining assessed constructs (cf. Alderson et al., 2015) and the importance of this for computerised DA.

The findings of the final study, suggested that learners can turn to their vocabulary knowledge when working on the tasks eliciting word derivational knowledge. I would suggest that unless the learners' vocabulary knowledge is controlled for (e.g., as in the study reported on in Article III), the possibility that it affected the results in some way should be considered when interpreting results of research of WD knowledge. That said, the study reported on in Article V had its limitations, as I will report in the following section. Moreover, it is also a question whether WD knowledge should be perceived as separate from vocabulary knowledge in the first place.

7.3 Limitations and some future directions

The results of the present doctoral research projects should not be interpreted such that this research produced comprehensive responses to the posed research questions. On the contrary, as regards the study reported on in Article I, for example, no delayed posttest was conducted. What is more, in light of Nassaji and Swain's (2000) finding that in case of random help, the learner benefitted from more explicit feedback, it would also be interesting to check if adaptive CF would be more beneficial for the development of learners' L2 abilities than explicit static feedback (e.g., overt correction). Finally, the study was rather modest in that the number of participants in it was not large. However, a virtue of quantitative research is that it enables meta-analyses (see Hattie & Timperley, 2007), so further studies of the effect of adaptive CF, accounting for the limitations delineated in the present section, can increase the reliability of these results.

Speaking of the sampling, it should be mentioned that the samples in the rest of the studies (except, perhaps, for the study reported on in Article IV) were rather modest as well. Therefore, caution should be exercised with regard to the generalisability of these findings, too. Further research could confirm or disprove the findings of the studies reported on in Articles III, and IV, as well.

As regards the studies reported on in Articles II and V, one advantage of qualitative research over quantitative is in providing a deeper understanding of certain phenomena (and not in its generalisability), that is, in the case of the present doctoral research, the reciprocity between learners' beliefs and their DA performance, the way they co-construct their beliefs about CF in interaction (Article II), and the way DA mediates learners' self-regulatory capacity / strategy use. That said, the results of the study reported on in Article V, for example, do not allow making generalisations regarding the effect of the DA of word derivational knowledge on other learners.

As regards other limitations of the studies, as far as the exploration of the construct of WD knowledge (Article IV) is concerned, we did not account for the effect that learners' general vocabulary knowledge had on their performance on the tasks we designed. Judging by M's use of vocabulary knowledge (which increased with time) when working through the static assessment tasks, controlling for it in future studies (e.g., by utilising structural equation modelling) should improve upon the design of the study reported on in Article IV and produce findings that can be interpreted with reference to learners' L2 English word derivational knowledge more validly.

It should also be noted that while the present research was impact-driven, it was only the impact of DA on micro level, that is, its impact on individual learners and small groups, which was studied rather than its impact on macro level. Implementing DA in several schools and studying the changes occurring due to it on institutional level would be an interesting development of the present research. Studying teachers' experiences with DA would be an interesting first step in moving from micro towards macro-level impact. Besides little data collected from two teachers regarding their own experiences with the *Questions Test* and observations of their learners working on the tasks (**Chapter 5.2**), no such data were collected during the present research project. This could be an interesting undertaking for the future.

With reference to the previous point, it should not be forgotten that teachers, too, have their own beliefs about what a test should look like. Therefore, an exploration of whether and how DA experience facilitates changes in teachers' beliefs about assessment and instruction should enrich our understanding of the impact of DA and adaptive CF. That the Intervention study teacher asked LA1 to work on her own improved the reliability of the finding regarding the effect of the adaptive CF on the performance of the experimental group. At the same time, it decreased the usefulness of that same feedback for LA1 (and for making diagnostic inferences based on her DA performance). An interesting possibility for further research would be studying the outcome of allowing learners' teachers and peers to act as mediators (instead of minimising this possibility) in addition to using automatic mediation provided by computerised assessment systems (cf. Tzuriel & Shamir, 2002).

In the context of Finland, further research of DA can be inspired by the changes in the educational system in this country. Specifically, the Finnish Matriculation Examination is planned to become fully computerised by 2019 (https://www.ylioppilastutkinto.fi/fi/ylioppilastutkinto/digabi; http://tucs.fi/news/article/ville.php). Introducing dynamic assessment that is computerised and, perhaps, has format that is similar to the English Matriculation Examination can, on the one hand, increase learners' familiarity with the procedure, and on other hand, reduce the possibility of teaching for the test, as DA improves the performance beyond the particular tasks, and instead of teaching some test-taking strategies, judging by the results of the present study, it helps learners to device their own ways of completing the tasks.

Technological advancement allowed for delivering DA via computer. However, so far, only some computerised dynamic tests of L2 have been designed, or, at least, reported on (e.g., Lantolf & Poehner 2013, Teo, 2012). All of them, including those in the present doctoral dissertation, were designed following the interventionist approach to DA. Designing a computerised dynamic test following the interactionist approach could be an interesting undertaking. Major questions to address here would be how to trace learners' responsiveness to feedback/mediation and how to use this information about learners' reciprocity to provide assistance which is sensitive to their ZPD.

Judging by the findings of the present doctoral research project and the previous research, this can be studied with reference to the time that learners spend reading feedback. If a learner skips the feedback and is able to correct his/her mistake following the feedback, this can be interpreted that s/he did not need to read the whole of the feedback message, perhaps, requiring registering that there was something wrong with his/her response to be able to self-correct. In other words, this would mean that less help could be provided to this learner. If a learner skips the feedback but is not able to self-correct, one or several questions could be asked to find out his/her reasons for doing so and subsequently, depending on the learner's responses, the same assistance could be displayed to the learner again to see whether s/he can benefit from it. If a learner does not skip the feedback and cannot self-correct, this can be interpreted so that the feedback is beyond the learners' ZPD and more assistance is required.

In addition, principles of computerised adaptive testing (CAT) based on Item Response Theory (IRT), that is, a procedure that tailors the test to the specific test-taker during test administration based on his/her previous answers, selecting from a collection of items only those that match this test-taker's estimated ability level (e.g., Chapelle, 2008; Jamieson, 2005; Sereci, 2003), can be used to establish the difficulty of each item with each possible assistance. In other words, the difficulty of the items can be established using the algorithms developed for conventional CAT, but when defining the difficulty, both the items and the assistance provided for the preceding item are considered. Clearly, the proposed approach should work only for designs similar to that in Article I, where a new item was displayed after the learners' every unsuccessful attempt. However, such approach could potentially allow both for selecting the items which are within learners' ZPD and for discovering learners' development from DA session to DA session both in terms of the amount of assistance they require to self-correct and the increase/decrease of their performance on an IRT scale.

7.4 Conclusion

The present research aimed at contributing to the under-researched area of dynamic assessment. This lack of research meant that it had to be carefully considered how to define and operationalise the assessed constructs and what to account for when designing the computerised adaptive CF. On the other hand, the situation in which there are more questions than answers introduced certain flexibility into what can be researched and how. So, as I have already mentioned, the approach that I took was that each following phase of research was, in part, inspired by the questions arising in the preceding phases.

This also means that the present research project was exploratory in nature, and especially considering the qualitative design of roughly a half of the studies forming the present doctoral dissertation, the conclusions should not be considered definite. The results of the research project suggest that adaptive CF has a positive effect on the development of L2 English wh-questions, can mediate/transform learners' beliefs about corrective feedback, and promotes learners' self-regulated, that is, strategic learning.

Future research conducted with different samples (e.g., learners of other L1s) and using different assessment targets (provided it arrives at similar conclusions) should increase the generalisability of the findings that the present doctoral research project produced. I cannot but feel that I barely scratched the surface of what impact adaptive corrective feedback can have, but I hope that more research on dynamic assessment and corrective feedback from the perspective of sociocultural theory will be conducted, which will encourage the reconceptualisation of classroom instruction and educational assessment and will help teachers in their everyday work, illuminating their learners' abilities and, at the same time, promoting these abilities.

YHTEENVETO

ICAnDoiT: Tietokoneella annetun adaptiivisen korjaavan palautteen vaikutus L2 englannin oppijoihin

Opetukseen liittyvää korjaavaa palautetta on tutkittu paljon ja erityisesti toisen kielen kouluoppimisesta tutkimusta on runsaasti. Usein niissä tarkastellaan ja vertaillaan erilaisia korjaavan palautteen tyyppejä kvasikokeellisissa tutkimusasetelmissa. Ei ole kuitenkaan yhteistä näkemystä siitä, minkä tyyppinen korjaava palaute on hyödyllistä toisen tai vieraan kielen oppimisessa – vai onko mikään (ks. esim. Nassaji & Swain, 2000).

Tässä väitöstutkimuksessa, samoin kuin Donaton (1994), Aljaafrehin & Lantolfin (1994) ja Nassaji & Swainin (2000) tutkimuksessa, valittiin erilainen lähestymistapa. Korjaavaa palautetta tutkittiin tarkastelemalla sen hyödyllisyyttä Vygotskyn sosiokulttuurisen teorian näkökulmasta. Sosiokulttuurisen näkökulman pääoletus on, että tieto on sosiaalisesti rakentunutta ja että oppijan kehittyminen nähdään 'siirtymisenä toisten tuella tapahtuvasta toiminnasta itsenäiseen työskentelyyn (Vygotsky, 1978). Tutkimuksissa, joissa korjaavaa palautetta tarkastellaan tästä näkökulmasta, todetaan, että millainen tahansa palaute on hyödyllistä, kunhan se on linjassa oppijan taitojen kanssa.

Tässä väitöstutkimuksessa analysoidaan dynaamisen arvioinnin aikana oppilaille annetun adaptiivisen korjaavan palautteen vaikutuksia. Tutkin erityisesti tietokoneella annetun adaptiivisen korjaavan palautteen vaikutusta oppijoiden L2 englannin taitojen kehittymiseen, palautteen roolia oppijoiden käsitysten muovautumisessa ja siitä, miten palaute edistää strategista oppimista. Tietokoneella annettava dynaaminen palaute soveltui hyvin sellaisen kvantitatiivisen aineiston keräämiseen, jossa tarkastellaan adaptiivisen korjaavan palautteen vaikutusta oppimiseen, sillä se mahdollistaa suuren määrän arviointeja samanaikaisesti. Tilastollisten analyysien ja laadullisten osioiden tuloksia tulkittiin sosiokulttuurisesta näkökulmasta. Oletuksena oli esimerkiksi, että oppijoiden suoritusta tietokonevälitteisen dynaamisen arvioinnin aikana ohjasi ei vain heidän saamansa palaute, vaan myös heidän uskomuksensa palautteen hyödyllisyydestä ja kielitaidon arvioinnista ylipäätään.

Väitöskirja koostuu viidestä artikkelista ja tästä yhteenvedosta. Artikkelissa I verrattiin adaptiivisen korjaavan palautteen ja oikein/väärin-palautteen vaikutusta oppijoiden taitoihin muodostaa L2 englannin wh-kysymyksiä. Sitä varten verrattiin kahden ryhmän suoritusta tehtävissä, joista toinen sai adaptiivista palautetta ja toinen vain palautetta siitä, oliko suoritus onnistunut vai ei. Tulokset osoittivat, että adaptiivista korjaavaa palautetta saaneessa ryhmässä oppijoiden taidot muodostaa L2 englannin wh-kysymyksiä olivat merkittävästi paremmat jakson lopussa kuin niiden oppijoiden, jotka saivat vain oikein/väärin-palautetta.

Artikkelissa II tavoitteena oli selvittää, miten oppijoiden käsitykset korjaavasta palautteesta vaikuttivat heidän kokemuksiinsa ja reflektointiinsa dynaamisesta palautteesta, ja miten näiden kokemusten reflektointi sosiaalisessa vuorovaikutuksessa muutti näitä samoja käsityksiä. Tulokset osoittivat, että käsitystensä vuoksi jotkut oppijat jättivät palautteen huomioimatta. Toisaalta oppijat alkoivat omaksua uusia käsityksiä korjaavasta palautteesta vuorovaikutustilanteiden aikana.

Tutkimukset, joita raportoidaan artikkelissa III pyrkivät etsimään tukea Bauerin ja Nationin (1993) esittämän johtimien oppimisjärjestyksestä. Artikkelin IV tavoitteena oli selvittää mitä johto-oppi käsitteenä pitää sisällään. Näin ollen nämä kaksi artikkelia pyrkivät paikkaamaan sananjohto-oppiin liittyvässä aiemmassa tutkimuksessa olleita aukkoja. Artikkeli V raportoi tapaustutkimusta, joka pohjaa väitöskirjaani sisältyviin aikaisempiin tutkimuksiin, ja sen tavoitteena on antaa ehdotuksia siitä, millaista adaptiivinen korjaava palaute voi olla dynaamisessa L2 englannin sananjohto-opin testissä. Väitöskirjan yhteenvedossa erillisessä luvussa pohditaan lopuksi tutkimuksessa laaditun dynaamisen testin validiutta.

Tutkimuksen teoreettinen anti on ennen kaikkea siinä, että se antaa kvantitatiivista tietoa oppijoiden taitojen mukaan linjatun korjaavan palautteen hyödyistä. Tulokset myös osoittavat, että oppijoiden suoritusta ei tietokoneella annetun dynaamisen arvioinnin aikana ohjaa ainoastaan oppijan saama apu (esim. korjaava palaute), vaan myös heidän käsityksensä.

Tutkimuksen käytännön hyödynnettävyys on tulosten valossa se, että tutkimuksen aikana laadittua arviointi/tutorointi -työkalua tai muuta vastaavaa työvälinettä voidaan käyttää luokassa tiedon saamiseen oppijoiden avuntarpeesta, oppijoiden taitojen kehittämiseen ja helpottamaan oppijoiden korjaavan palautteen hyödyllisyyttä koskevien käsitysten muuttumista.

Avainsanat: dynaaminen arviointi, korjaava palaute, sosiokulttuurinen teoria, käsitykset, englanti toisena/vieraana kielenä

REFERENCES

Ableeva, R. (2010). *Dynamic assessment of listening comprehension in second language learning* (Unpublished doctoral dissertation, The Pennsylvania State University, University Park, PA). Retrieved from http://proquest.umi.com/pqdweb?did=2209252761&Fmt=7&clientId=23378&RQT=309&VName=PQD

Ableeva, R. (2012, March). *Transfer tasks: Diagnosing second language development*. Paper presented at AAAL 2012 Annual Conference, Boston, MA, USA.

Alanen, R. (2003). A sociocultural approach to young language learners' beliefs about language learning. In P. Kalaja & A. M. F. Barcelos (Eds.), *Beliefs about SLA: New research approaches* (pp. 55–85). Amsterdam: Kluwer Academic.

Alanen, R. (2013). Noticing and mediation: A sociocultural perspective. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 315–325). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Alanen, R. & Huhta, A. (2009, September). *L2 learners' performance across L2 writing tasks: Comparing tasks and language proficiency across CEFR levels*. Paper presented at TBLT 2009, Lancaster, UK.

Alanen, R. & Kalaja, P. (2010, March). *The emergence of L2 English questions across CEFR proficiency levels*, Paper presented at AAL 2010, Atlanta, USA.

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

Alderson, J. C. (2007). The Challenge of (Diagnostic) Testing: Do we know What we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner & C. Doe (Eds.), *Language Testing Reconsidered* (pp. 21–39). Ottawa: University of Ottawa Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., Haapakangas, E.-L., 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

Alderson, J. C., & Huhta, A. (2011). Can research into the diagnostic testing of reading in a second or foreign language contribute to SLA research? E*UROSLA Yearbook, 11*, 30–52. doi: 10.1075/eurosla.11.04ald

Alderson, J. C., Huhta, A., Nieminen, L., Ullakonoja, R., & Haapakangas, E.-L. (2013, May). *What is the impact of diagnostic language tests?* Paper presented at The Tenth Annual Conference of EALTA, Istanbul, Turkey.

Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the Zone of Proximal Development. *The*

Modern Language Journal, 78(4), 465–483. doi: 10.1111/j.1540-4781.1994.tb02064.x

Ammar, A. (2003). *Corrective feedback and L2 learning: Elicitation and recasts*. (Unpublished Doctoral Dissertation), McGill University, Montreal.

Ammar, A., & Spada, N. (2006). One size fits all?: Recasts, prompts, and L2 learning. *Studies in Second Language Acquisition, 28*(04), 543–574. doi: doi:10.1017/S0272263106060268

Amrhein, H. R., & Nassaji, H. (2010). Written corrective feedback: What do students and teachers think is right and why? *Canadian Journal of Applied Linguistics, 13*(2), 95–127. Retrieved from https://journals.lib.unb.ca/index.php/CJAL/article/view/19886/21712

Antón, M. (1999). The discourse of a learner-centered classroom: Sociocultural perspectives on teacher-learner interaction in the second-language classroom. *The Modern Language Journal, 83*(3), 303–318. doi: 10.1111/0026-7902.00024

Antón, M. (2009). Dynamic assessment of advanced second language learners. *Foreign Language Annals, 42*(3), 576–598. doi: 10.1111/j.1944-9720.2009.01030.x

Aro, M. (2009). *Speakers and doers: Polyphony and agency in children's beliefs about language learning* (Unpublished Doctoral dissertation, University of Jyväskylä, Jyväskylä). Retrieved from https://jyx.jyu.fi/dspace/handle/123456789/19882

Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing, 9*(3), 227–257. doi: 10.1016/S1060-3743(00)00027-8

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Barbour, A. (2003, February). *Interpersonal feedback: Origins and applications*. Paper presented at the Annual Meeting of the Western States Communication Association, Salt Lake City.

Barcelos, A. M. F. (2003). Researching beliefs about SLA: A critical review. In P. Kalaja & A. M. F. Barcelos (Eds.), *Beliefs about SLA: New research approaches* (pp. 7–33). Amsterdam: Kluwer Academic.

Barcelos, A. M. F., & Kalaja, P. (2013). Beliefs in second language acquisition: Teacher. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. New York: Blackwell Publishing.

Bauer, L, & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*(2), 102–118. doi: 10.1016/j.jslw.2007.11.004

Bitchener, J., & Knoch, U. (2009). The relative effectiveness of different types of direct written corrective feedback. *System, 37*(2), 32–329. doi: 10.1016/j.system.2008.12.006

Brown, A. V. (2009). Students' and teachers' perceptions of effective foreign language teaching: A comparison of ideals. *The Modern Language Journal, 93*(1), 46–60. doi: 10.1111/j.1540-4781.2009.00827.x

Bruton, A. (2000). What exactly are positive and negative evidence in SLA? *RELC Journal, 31*(2), 120–133.

Budoff, M., & Friedman, M. (1967). "Learning potential" as an assessment approach to the adolescent mentally retarded. *Journal of Consulting Psychology, 28*(5), 434–439. doi: 10.1037/h0040631

Cabrera, A. F. (2007, September). *An empirical study of effective corrective feedback strategies with implications for technological applications in applied linguistics.* Paper presented at the BAAL 2007, The University of Edinburgh.

Carroll, S. E. (1998). On Processability Theory and second language acquisition. *Bilingualism: Language and Cognition, 1*(01), 23–24. doi:10.1017/S1366728998000030

Cazden, C. B. (1972). *Child language and education*. New York: Holt, Rinehart and Winston, Inc.

Chapelle, C. A. (2008). Technology and second language acquisition. *Annual Review of Applied Linguistics*, *27*, 98–114. doi: 10.1017/S0267190508070050

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Chomsky, N. (2002). *Syntactic structures* (2 ed.). Berlin: Mouton de Gruyter.

Chuenjundaeng, J. (2006). An investigation of SUT students' receptive knowledge of English noun suffixes (Unpublished master's thesis). Suranaree University of Technology. Retrieved from: http://sutir.sut.ac.th:8080/sutir/bitstream/123456789/1585/2/jitlada_fulltext.pdf

Clahsen, H., & Neubauer, K. (2010). Morphology, frequency, and the processing of derived words in native and non-native speakers. *Lingua, 120*(11), 2627–2637. doi: 10.1016/j.lingua.2010.06.007

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

de Bot, K. (1992). A Bilingual Processing Model: Levelt's 'Speaking' Model Adapted. *Applied Linguistics, 13*(1), 1–24. doi: 10.1093/applin/13.1.1

Dempsey, J. V., Driscoll, M. P., & Swindell, L. K. (1993). Text-based feedback. In J. V. Dempsey and G. C. Sales (Eds.), *Interactive instruction and feedback* (pp. 21–54). Englewood Cliffs, NJ: Educational Technology Publications.

Dempsey, J. V., & Wager, S. U. (1988). A Taxonomy for the Timing of Feedback in Computer-Based Instruction. *Educational Technology, 28*(10), 20–25.

Di Biase, B., & Kawaguchi, S. (2002). Exploring the typological plausibility of Processability Theory: language development in Italian second language and Japanese second language. *Second Language Research, 18*(3), 274–302. doi: 10.1191/0267658302sr204oa

Diab, R. L. (2005). Teachers' and students' beliefs about responding to ESL writing: a case study. *TESL Canada Journal, 23*(1), 28–43. doi: 10.18806/tesl.v23i1.76

Donato, R. (1994). Collective scaffolding in second language learning. In J. P. Lantolf & G. Appel (Eds.), *Vygotskian Approaches to Second Language Research* (pp. 33–59). Norwood, NJ: Ablex Publishing Corporation.

Dörnyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. L. Erlbaum Associates.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Dufva, H. (2003). Beliefs in dialogue: a Bakhtinian view. In P. Kalaja & A. M. F. Barcelos (Eds.), *Beliefs about SLA: New research approaches* (pp. 131–151). Amsterdam: Kluwer Academic.

Dunn, W. E., & Lantolf, J. P. (1998). Vygotsky's zone of proximal development and Krashen's i+ 1: Incommensurable constructs; incommensurable theories. *Language Learning*, *48*(3), 411–442. doi: 10.1111/0023-8333.00048

Dyson, B. (2008). What we can learn from questions: ESL question development and its implications for language assessment. *Prospect Journal, 23*(1), 16–27.

Egi, T. (2007a). Interpreting recasts as linguistic evidence: The roles of linguistic target, length, and degree of change. *Studies in Second Language Acquisition, 29*, 51–537. doi: 10.1017/S0272263107070416

Egi, T. (2007b). Recasts, learners' interpretations, and L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 249–267). Oxford: Oxford University Press.

Egi, T. (2010). Uptake, modified output, and learner perceptions of recasts: Learner responses as language awareness. *The Modern Language Journal, 94*(1), 1–21. doi: 10.1111/j.1540-4781.2009.00980.x

Ellis, R. (2009a). A typology of written corrective feedback types. *ELT Journal, 63*(2), 97–107. doi: 10.1093/elt/ccn023

Ellis, R. (2009b). Corrective feedback and teacher development. *L2 Journal, 1*(1), 3–18. Retrieved from: http://escholarship.org/uc/item/2504d6w3

Ellis, R. (2010). Epilogue: A Framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition, 32*(Special Issue 02), 335–349. doi: doi:10.1017/S0272263109990544

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition, 28*(2), 339–368. doi: 10.1017/S0272263106060141

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. doi: 10.1016/j.system.2008.02.001

Eslami, E. (2014). The effects of direct and indirect corrective feedback techniques on EFL students' writing. *Procedia - Social and Behavioral Sciences, 98*, 445–452. doi: 10.1016/j.sbspro.2014.03.438

Fazio, L. L. (2001). The effect of corrections and commentaries on the journal writing accuracy of minority- and majority-language students. *Journal of Second Language Writing, 10*(4), 235–249. doi: 10.1016/S1060-3743(01)00042-X

Ferris, D. R. (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:10.1016/S1060-3743(99)80110-6

Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition*, *32*(02), 181–201. doi: 10.1017/S0272263109990490

Ferris, D. R., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*(3), 161–184. doi: 10.1016/S1060-3743(01)00039-X

Feuerstein, R., & Falik, L. H. (1999). Cognitive modifiability: A needed perspective on learning for the 21st century. *College of Education Review (in press–and other papers)*.

Feuerstein, R., & Feuerstein, S. (1999). Mediated learning experience: A theoretical review. In R. Feuerstein, P. S. Klein, & A. J. Tannenbaum (Eds.), *Mediated learning experience: Theoretical, psychosocial and learning Implications* (pp. 3–55). London: Freund.

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979) *The dynamic assessment of retarded performers: The learning potential assessment device: Theory, instruments, and techniques*. University Park Press, Baltimore.

Feuerstein, R., Rand, Y., & Rynders, J. E. (1988). *Don't accept me as I am. Helping retarded performers excel*. New York: Plenum.

Finnish National Board of Education (2004). *National Core Curriculum for Basic Education*. Vammala: Vammalan Kirjapaino OY.

Floropoulou, C. (2002). *Foreign language learners' attitudes to self-assessment and DIALANG: A comparison between Greek and Chinese Learners of English* (Unpublished master's thesis). Lancaster University, UK.

Friedline, B. E. (2011). *Challenges in the second language acquisition of derivational morphology: from theory to practice* (Doctoral dissertation). University of Pittsburgh, USA. Retrieved from http://d-scholarship.pitt.edu/8351/

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing, 20*(4), 384–408. doi: 10.1191/0265532203lt265oa

Ginsburg, H., & Opper, S. (1979). *Piaget's Theory of Intellectual Development*. Englewood Cliffs : Prentice-Hall.

Glahn, E., Hakansson, G., Hammarberg, B., Holmen, A., Hvenekilde, A., & Lund, K. (2001). Processability in scandinavian second language acquisition. *Studies in Second Language Acquisition, 23*(03), 389–416. doi: doi:10.1017/S0272263101003047

Glutting, J. J. and McDermott, P. A. (1990). Principles and problems in learning potential. In C.R. Reynolds and R.W. Kamphaus (Eds.), *Handbook of* *psychological and educational assessment of children. Intelligence and achievement*. New York: Guilford.

Guthke, J. (1982). The learning test concept—An alternative to the traditional static intelligence test. *The German Journal of Psychology 6*(4), 306–324.

Guthke, J., & Beckmann, J. F. (2000). The learning test concept and its application in practice. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 17–69). Amsterdam: JAI/Elsevier.

Hattie, J. A. (1999, June). *Influences on student learning* (Inaugural professorial address, University of Auckland, New Zealand). Retrieved from http://www.arts.auckland.ac.nz/staff/index.cfm?P=8650

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

Haywood, H. C., & Lidz, C.S. (2007). *Dynamic Assessment in Practice: Clinical and Educational Applications.* Cambridge : Cambridge University Press.

Hedgcock, J., & Lefkowitz, N. (1994). Feedback on feedback: Assessing learner receptivity to teacher response in L2 composing. *Journal of Second Language Writing, 3*(2), 141–163. doi: 10.1016/1060-3743(94)90012-4

Hew, S. H., & Ohki, M. (2013). Effect of animated graphic annotations and immediate visual feedback in aiding Japanese pronunciation learning: A comparative study. *CALICO Journal*, *21*(2), 397–419.
doi: 10.1558/cj.v21i2.397-419

Horn, G. M. (2011). *Lexical-Functional Grammar*. Berlin, DEU: Walter de Gruyter.

Huang, L. (2009). Washback on teacher beliefs and behaviour: Investigating the process from a social psychology perspective (Unpublished doctoral dissertation Lancaster University, UK). Retrieved from http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.538590

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Huhta, A. (2007). Itsearviointi osana kielitestiä – tutkimus käyttäjien suhtautumisesta tietokoneen antamaan palautteeseen kielitaidon itsearvioinnista [Self-assessment as part of a language test—a study on user reactions to computerised feedback on self-assessment]. In O.-P., Salo, P. Kalaja, & T., Nikula (Eds.), *AFinLA yearbook 2007*. Publications of the Association Finlandaise de Linguistique Appliquée 65. 371–388.

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

Huhta, A. (2010). *Innovations in diagnostic assessment and feedback: An analysis of the usefulness of the DIALANG language assessment system* (Unpublished Doctoral dissertation University of Jyväskylä, Jyväskylä).

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. doi: 10.1191/0265532206lt331oa

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, 25, 228–242. doi: 10.1017/S0267190505000127

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 03–073. doi: 10.1177/0265532208097336

Johnson, R. B. and Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches.* SAGE Publications.

Kalaja, P., & Barcelos, A. M. F. (2013) Beliefs in second language acquisition: Learner. In C.A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. New York: Blackwell Publishing.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). Westport, CT: American Council on Education and Praeger.

Kang, H.-S. (2009). The relative efficacy of explicit and implicit feedback in the learning of a less-commonly-taught foreign language. *IRAL - International Review of Applied Linguistics in Language Teaching, 47*(3–4), 303–324. doi: 10.1515/iral.2009.013

Kepner, C. G. (1991). An Experiment in the Relationship of Types of Written Feedback to the Development of Second-Language Writing Skills. *The Modern Language Journal, 75*(3), 305–313. doi: 10.2307/328724

Kern, R. G. (1995). Students' and teachers' beliefs about language learning. *Foreign Language Annals, 28*, 71–92. doi: 10.1111/j.1944-9720.1995.tb00770.x

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-Analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

Kozulin, A., & Garb, E. (2002). Dynamic Assessment of EFL Text Comprehension. *School Psychology International, 23*(1), 112–127. doi: 10.1177/0143034302023001733

Kozulin, A., & Garb, E.. (2004). Dynamic assessment of literacy: English as a third language. *European Journal of Psychology of Education, 19*(1), 65–77. doi: 10.1007/bf03173237

Krashen, S. D. (ed.). (2009). *Principles and Practice in Second Language Acquisition*. Retrieved from http://www.sdkrashen.com/content/books/principles_and_practice.pdf

Kulhavy, R. W., & Wager, W. (1993). Feedback in Programmed Instruction: Historical Context and Implicatoins for Practice. In J. V. Dempsey & G. C. Sales (Eds.), *Interactive Instruction and Feedback* (pp. 3-19). Englewood Cliffs, NJ: Educational Technology Publications.

Lalande, J. F., II. (1982). Reducing Composition Errors: An Experiment. *The Modern Language Journal, 66*(2), 140–149. doi: 10.2307/326382

Lantolf, J. P. (2000). Second language learning as a mediated process. *Language Teaching*, 33, 79–96. doi: 10.1017/S0261444800015329

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: bringing the past into the future. *Journal of Applied Linguistics, 1*(1), 49–72. doi: 10.1558/japl.v1i1.49

Lantolf, J. P., & Poehner, M. E. (2008). Dynamic assessment. In E. Shohamy & N. Hornberger (Eds.), *Encylopedia of Language and Education. Vol. 7. Language Testing and Assessment* (pp. 273–284). Springer.

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

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

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

Lee, Y.-W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*(3), 172–189. doi:10.1080/15434300902985108

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

Leung, C. (2007). Dynamic assessment: Assessment for and as teaching? *Language Assessment Quarterly, 4*(3), 257–278. doi: 10.1080/15434300701481127

Li, S. (2010). The Effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*(2), 309–365. doi: 10.1111/j.1467-9922.2010.00561.x

Lightbown, P. M., & Spada, N. (1999). *How Languages Are Learned*. Oxford: Oxford University Press.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16. doi: 10.3102/0013189x029002004

Loewen , S. , & Nabei , T. (2007). Measuring the effects of oral corrective feedback on L2 knowledge . In A. Mackey (Ed.), *Conversational Interaction in Second Language Acquisition: A Collection of Empirical Studies* (pp. 361–377). Oxford: Oxford University Press.

Loewen, S., & Philp, J. (2006). Recasts in the adult English L2 classroom: Characteristics, explicitness, and effectiveness. *The Modern Language Journal, 90*(4), 536–556. doi: 10.1111/j.1540-4781.2006.00465.x

Long, M.H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bahtia (Eds.), *Handbook of second language acquisition* (pp. 413–468). New York: Academic Press.

Long, M. H., Inagaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models and recasts in Japanese and Spanish. *The Modern Language Journal, 82*(3), 357–371. doi: 10.2307/329961

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351–372. doi: 10.1177/026553220101800403

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition, 26*(3), 399–432. doi: 10.1017/S0272263104263021

Lyster, R., & Izquierdo, J. (2009). Prompts versus recasts in dyadic interaction. *Language Learning, 59*(2), 453–498. doi: 10.1111/j.1467-9922.2009.00512.x

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition, 19*(01), 37–66.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA. *Studies in Second Language Acquisition, 32*(Special Issue 02), 265–302. doi: 10.1017/S0272263109990520

MacKey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *The Modern Language Journal, 82*(3), 338–356. doi: 10.1111/j.1540-4781.1998.tb01211.x

Mäntylä, K., & Huhta, A. (2013). Knowledge of word parts. In J. Milton. & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 45-59). Palgrave Macmillan.

Marslen-Wilson, W. (2007). Morphological processes in language comprehension. In G. Gaskell (Ed.), *Oxford Handbook of Psycholinguistics* (pp. 175–93). Oxford, UK: Oxford University Press.

McDonough, K. (2005). Identifying the Impact of Negative Feedback and Learners' Responses on ESL Question Development. *Studies in Second Language Acquisition, 27*(1), 79–103. doi: 10.1017/S0272263105050047

Mercer, S. (2011). Language learner self-concept: Complexity, continuity and change. *System, 39*(3), 335–346. doi: 10.1016/j.system.2011.07.006

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

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, 45(1), 35–44.

Meyer, J.P. (2013*). jMetrik (version 3.0.1)* [Computer software]. Retrieved from http://www.ItemAnalysis.com

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: an exploratory study. *System, 28*(2), 291–304. doi: 10.1016/S0346-251X(00)00013-0

Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM, 33*(3), 338–348. doi: 10.1145/77481.77486

Mory, E. H. (2003). Feedback research revisited. In D.H. Jonassen (Ed.), *Handbook of Research for Educational Communications and Technology* (2 ed., pp. 745–783). Mahwah, NJ: Lawrence Erlbaum.

Nassaji, H. (2003). L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly, 37*(4), 645–670. doi: 10.2307/3588216

Nassaji, H., & Swain, M. (2000). A Vygotskian perspective on corrective feedback in L2: The effect of random versus negotiated help on the learning of English articles. *Language Awareness, 9*(1), 34–51. doi: 10.1080/09658410008667135

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning Strategies in Second Language Acquisition*. Cambridge: Cambridge University Press.

Oxford, R. L. (1993). Research on second language learning strategies. *Annual Review of Applied Linguistics, 13*, 175–187. doi: 10.1017/S0267190500002452

Oxford, R. L. (2011). *Teaching and Researching Language Learning Strategies*. Pearson Education.

Pica, T. (1994). Questions from the language classroom: Research perspectives. *TESOL Quarterly, 28*(1), 49–79. doi: 10.2307/3587198

Pienemann, M. (Ed.) (2005). *Cross-linguistic aspects of Processability Theory*. Amsterdam: John Benjamins Publishing Company.

Pienemann, M., Johnston, M., & Brindley, G.. (1988). Constructing an Acquisition-Based Procedure for Second Language Assessment. *Studies in Second Language Acquisition, 10*(02), 217–243. doi: doi:10.1017/S0272263100007324

Poehner, M. E. (2005). *Dynamic assessment of oral proficiency among advanced L2 learners of French* (Unpublished doctoral dissertation). Pennsylvania State University, University Park, PA.

Poehner, M. E. (2008). Dynamic assessment: *A Vygotskian approach to understanding and promoting L2 development*. Berlin: Springer.

Poehner, M. E. (2011). Validity and interaction in the ZPD: interpreting learner development through L2 Dynamic Assessment. *International Journal of Applied Linguistics, 21*(2), 244–263. doi: 10.1111/j.1473-4192.2010.00277.x

Poehner, M. E. (2012). The zone of proximal development and the genesis of self-assessment. *The Modern Language Journal, 96*(4), 610–622. doi: 10.1111/j.1540-4781.2012.01393.x

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA). *Language Teaching Research, 17*(3), 323–342. doi: 10.1177/1362168813482935

Polio, C., Fleck, C., & Leder, N. (1998). "If I only had more time:" ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing, 7*(1), 43–68. doi: 10.1016/S1060-3743(98)90005-4

Põhikooli riiklik õppekava õigusakt; Lisa 2 [Basic School National Curriculum Act: Annex 2] (2010). Pub. L. No. RT I 2010, 6, 22. Retrieved from https://www.riigiteataja.ee/aktilisa/1281/2201/0017/13275423.pdf.

Purpura, J. (2014). Language learner styles and strategies. In M. Celce-Murcia, D. Brinton, & A. Snow (Eds.), *Teaching English as a Second or Foreign Language* (4th ed.) (pp. 532–549). Boston, MA: National Geographic Learning/Cengage Learning.

Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing 21*, 249–258. doi: 10.1191/0265532204lt283ed

Ringbom, H. (1987). *The role of the first language in foreign language learning*. Clevedon: Multilingual Matters.

Ringbom, H. (1990). On the relation between second language comprehension and production. In J. Tommola (Ed.), *Foreign language, comprehension and production: AFinLA yearbook* (pp. 139–148). Helsinki: Suomen soveltavan kielitieteen yhdistys.

Rosa, E. M., & Leow, R. P. (2004). Computerized task-based exposure, explicitness, type of feedback, and spanish L2 development. *The Modern Language Journal, 88*(2), 192–216. doi: 10.1111/j.0026-7902.2004.00225.x

Rose, H. (2012). Reconceptualizing strategic learning in the face of self-regulation: throwing language learning strategies out with the bathwater. *Applied Linguistics, 33*(1). 92–98. doi: 10.1093/applin/amr045

Saito, H. (1994). Teachers' practices and students' preferences for feedback on second language writing: A case study of adult ESL learners. *TESL Canada Journal, 11*(2), 46–69. doi: 10.18806/tesl.v11i2.633

Schachter, J. (1991). Corrective feedback in historical perspective. *Second Language Research, 7*(2), 89–102. doi: 10.1177/026765839100700202

Schimmel, B. J. (1988). Providing meaningful feedback in courseware. In D. H. Jonassen (Ed.), *Instructional Designs for Microcomputer Courseware* (pp. 183–196). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy, (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 199–227). Cambridge: Cambridge University Press.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition, 19*(1), 17–36.

Schmitt, N., & Zimmermann, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36*(2), 145–171. doi: 10.2307/3588328

Schulz, R. A. (1996). Focus on form in the foreign language classroom: Students' and teachers' views on error correction and the role of grammar. *Foreign Language Annals, 29*(3), 343–364. doi: 10.1111/j.1944-9720.1996.tb01247.x

Schulz, R. A. (2001). Cultural differences in student and teacher perceptions concerning the role of grammar instruction and corrective feedback: USA-Colombia. *The Modern Language Journal, 85*(2), 244–258. doi: 10.1111/0026-7902.00107

Sheen, Y. (2010). Introduction: The role of oral and written corrective feedback in SLA. *Studies in Second Language Acquisition, 32*(Special Issue 02), 169–179. doi: doi:10.1017/S0272263109990489

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System, 37*(4), 556–569.
doi: 10.1016/j.system.2009.09.002

Shohamy, E. (2001). *The Power of Tests. A critical Perspective on the Uses of Language Tests*. Pearson Education.

Silva, R., & Clahsen, H. (2008). Morphologically complex words in L1 and L2 processing: Evidence from masked priming experiments in English.

Bilingualism: Language and Cognition, 11(2), 245–260. doi: 10.1017/S1366728908003404.

Sereci, S. G. (2003). Computerized Adaptive Testing: An Introduction. In J. E. Wall & G. R. Walz (Eds.), *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators* (pp. 685–694). Greensboro, NC: ERIC Counseling and Student Services Clearinghouse.

Spada, N., & Lightbown, P. M. (1993). Instruction and the Development of Questions in L2 Classrooms. *Studies in Second Language Acquisition, 15*(02), 205–224. doi: 10.1017/S0272263100011967

Spada, N., & Lightbown, P. M. (1999). Instruction, first language influence, and developmental readiness in second language acquisition. *The Modern Language Journal, 83*, 1–22. doi: 10.1111/0026-7902.00002

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: the nature and measurement of learning potential.* Cambridge: Cambridge University Press.

Stobart, G. (2008). *Testing Times: The Uses and Abuses of Assessment.* Oxon: Routledge.

Stone, D., & Heen, S. (2014). *Thanks for the feedback: The science and art of receiving feedback well*. Penguin UK.

Sullivan, L. E. (2009). Mixed methods research. In *The SAGE glossary of the social and behavioral sciences* (Vol. 3, pp. 327–327). Thousand Oaks, CA: SAGE Publications Ltd. doi: 10.4135/9781412972024.n1625

Teddlie, C., & Tashakkori, A. (2010). Overview of contemporary issues in mixed methods research. In A. Tashakkori, & C. Teddlie (Eds.), *SAGE handbook of mixed methods in social & behavioral research.* (2nd ed., pp. 1–43). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781506335193.n1

Teo, A. (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning & Technology*, 16(3), 10–20. Retrieved from http://llt.msu.edu/issues/october2012/action.pdf

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan.

Thouësny, S. (2011). Dynamically assessing written language: To what extent do learners of French language accept mediation? In S. Thouësny & L. Bradley (Eds.), *Second language teaching and learning with technology: views of emergent researchers* (1st ed., pp. 169–188). Dublin: Research-publishing.net.

Truscott, J. (1996). The Case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327–369. doi: 10.1111/j.1467-1770.1996.tb01238.x

Truscott, J. (1999a). The case for "The Case Against Grammar Correction in L2 Writing Classes": A response to Ferris. *Journal of Second Language Writing, 8*(2), 111–122. doi: 10.1016/S1060-3743(99)80124-6

Truscott, J. (1999b). What's wrong with oral grammar correction. *Canadian Modern Language Review/ La Revue canadienne des langues vivantes, 55*(4), 437–456.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*(4), 255–272. doi: 10.1016/j.jslw.2007.06.003

Tseng, W.-T., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. *Applied Linguistics, 27*(1), 78–102. doi: 10.1093/applin/ami046

Tzuriel, D. (2005). Dynamic assessment of learning potential: A new paradigm. *Erdelyi Pszichologiai Szemle (Transylvanian Journal of Psychology), Special Issue,* 1, 7–16.

Tzuriel, D., & Shamir, A. (2002). The effects of mediation in computer assisted dynamic assessment. *Journal of Computer Assisted Learning, 18*(1), 21–32. doi: 10.1046/j.0266-4909.2001.00204.x

Valsiner, J. (2001) Process structure of semiotic mediation in human development. *Human Development, 44*, 84–97. doi:10.1159/000057048

Varnosfadrani, A. D., & Basturkmen, H.. (2009). The effectiveness of implicit and explicit error correction on learners' performance. *System*, 37(1), 82–98. doi: 10.1016/j.system.2008.04.004

Vasilyeva, E., Puuronen, S., Pechenizkiy, M., & Rasanen, P. (2007). Feedback adaptation in web-based learning systems. *International Journal of Continuing Engineering Education and Life-Long Learning, 17*, 337–357. doi: 10.1504/IJCEELL.2007.015046

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes/ L. S. Vygotsky; edited by M. Cole ... [et al.].* Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986). *Thought and language.* Cambridge, MA: MIT Press.

Vygotsky, L. S. (1987). *The collected works of L. S. Vygotsky, Vol. 1, Problems of general psychology: including the volume Thinking and speech.* R. W. Rieber & A. S. Carton (Eds.). New York: Plenum Press.

Vygotsky, L. S. (1998). The problem of age. In R. W. Rieber (Ed.), *The collected works of L.S. Vygotsky. Vol. 5. Child psychology* (pp. 187–205). New York: Plenum.

Weir, C. (1993). *Understanding and developing language tests*. New York: Prentice Hall.

Weir, C. (2005). *Language testing and validation. An evidence-based approach*. Basignstoke: Palgrave Macmillan.

Wells, G. (1998). Using L1 to master L2: A response to Anton and DiCamilla's 'Socio-Cognitive Functions of L1'. *Canadian Modern Language Review, 54*(3), 343–353. doi: 10.1111/0026-7902.00019

Wertsch, J. V. (1991). A sociocultural approach to socially shared cognition. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 85–100). Washington, DC: American Psychological Association.

Wertsch, J. V. (1998). *Mind as action*. New York: Oxford University Press.

Whyte, M. M., Karolick, D. M., Neilsen, M. C., Elder, G. D., & Hawley, W. T. (1995). Cognitive styles and feedback in computer-assisted instruction. *Journal of Educational Computing Research, 12*(2), 195–203. doi: 10.2190/M2AV-GEHE-CM9G-J9P7

Yang, R. (2003). *Investigating how test-takers use the DIALANG feedback (*Unpublished Master's thesis). Lancaster University, UK.

Yang, Y., & Lyster, R. (2010). Effects of form-focused practice and feedback on Chinese EFL learners' acquisition of regular and irregular past tense forms. *Studies in Second Language Acquisition, 32*(Special Issue 02), 235–263. doi: 10.1017/S0272263109990519.

Yoshida, R. (2008). Learners' perception of corrective feedback in pair work. *Foreign Language Annals, 41*(3), 525–541.
doi: 10.1111/j.1944-9720.2008.tb03310.x

Yoshida, R. (2010). How do teachers and learners perceive corrective feedback in the japanese langauge classroom? *The Modern Language Journal, 94*(2), 293–314. doi: 10.1111/j.1540-4781.2010.01022.x

Zourou, K. (2008, December). *Towards a Typology of Corrective Feedback Moves in an Asynchronous Distance Language Learning Environment*. Paper presented at the Media in Foreign Language Teaching and Learning, CLaSIC conference, Singapore.

APPENDIX 1: TEST SPECIFICATIONS

An interventionist sandwich format dynamic diagnostic test of learners' ability to form L2 English wh-questions with auxiliaries

Adapted from Bachman & Palmer (1996).

1 Design statement

1 Test purpose

A Inferences

1 test-takers' (hereinafter, also referred to as *learners*) ability to form/use wh-questions with auxiliaries (hereinafter, stage 5 questions);
2 the amount of help learners need to self-correct;
3 self-diagnosis of mistakes in formulating stage 5 questions.

B Decisions

1 Stakes: a low-stakes test (as conventionally defined):

a. results should help teachers in their classroom instruction;
b. results should help learners to develop their ability to form stage 5 questions.

2 Individuals affected:

a. learners;
b. teachers.

3 Specific decisions:

a. diagnosis for teachers:
   - problems that learners have in production of stage 5 questions;
   - the amount and detail of CF that learners require to self-correct their mistakes (if they are able to) at the time of assessment.

b. diagnosis for learners (self-diagnosis):
   - increase awareness of mistakes they make when forming L2 English stage 5 questions;
   - learn to formulate correct stage 5 questions.

2 Description of Target Language Use (TLU) domains/task types

A Tasks identification

1 TLU domains:

a. inquiries in E-mails (i.e., including those outside the classroom);
b. classroom tasks (i.e., language instructional).

2 Considerations for task-types selection:

a. suitability of the tasks: question words (removed from the test following the Pilot study), the word order of stage 5 questions, specific problems with modal auxiliaries and auxiliaries *do*, *did*, and *does*;

b. use and familiarity of the tasks to learners.

B TLU tasks

	Task 1	**Task 2**
	Writing a semi-formal E-mail where one asks questions with the aim of getting more information	**Written production in the classroom with the following evaluation the aim of which is to enable self-diagnosis and promote learners' ability to form stage 5 questions**
Setting		
Physical characteristics	Location: varied (wherever a computer with the Internet is available; not excluding the formal school setting)	Classroom.
Time	Individually variable	Individually variable
Input format		
Channel	Visual	Visual and aural
Form	Text	Both written and oral feedback
Language	Source	Both target and source
Type	Task	The task and the feedback
Expected response format		
Channel	Written	Written/oral
Language	Target	Target
Length	Production (an E-mail)	Varied (but limited production is more probable)
Speededness	Unspeeded	Varied (depending on the situation)
Language characteristics		
Grammatical	Morphology, syntax, and vocabulary, but also genre specificities	Morphology, syntax, and vocabulary
Textual	Typewritten	Typewritten/written/spoken

Input-response relation		
Reactivity	Non-reciprocal	Reciprocal/adaptive
Scope	Wide	Narrow
Directness	Direct	Direct

C Description of Task types

Pretest/posttest	Task type 1	Task type 2
	An E-mail to a pet-shop	Gap-filling (separate sentences)
Setting		
Physical characteristics	Location: a computer lab; but also, possibly, outside school, e.g., a computer at home)	the same
Time	Individually variable	Individually variable
Input format		
Channel	Visual	Visual
Form	Text	Text
Language	Both target and source	Both target and source
Type	Task and prompts	Task and prompts
Expected response format		
Channel	Visual	Visual
Language	Target	Target
Length	Production (a short E-mail)	Limited production: two word answer (auxiliary and the main verb)
Speededness	Unspeeded (check that 1 academic hour is enough to complete the tasks)	Unspeeded
Language characteristics		
Grammatical	Morphology and syntax	Morphology and syntax
Textual	Typewritten	Typewritten
Input-response relation		
Reactivity	Non-reciprocal	Non-reciprocal
Scope	Wide	Narrow
Directness	Direct	Direct

Test	Task type 3	Task type 4	Task type 5
	Matching*	Ordering	Multiple-choice
Setting			
Physical characteristics	Location: a computer lab; but also, possibly, outside school, e.g., a computer at home)	the same	The same
Time	Varied	Varied	Varied
Input format			
Channel	Visual	Visual	Visual
Form	Text (graphics)	Text (graphics)	Text (graphics)
Language	Both target and source	Both target and source	Both target and source
Type	Matching the corresponding words, selected-response	Ordering, selected-response	Selected-response
Expected response format			
Channel	Visual	Visual	Visual
Language	Target	Target	Target
Length	Short	Short	Short
Speededness	Unspeeded (check that 1 academic hour is enough to complete the tasks)	Unspeeded	Unspeeded
Language characteristics			
Grammatical	Vocabulary: general	Syntax	Syntax and morphology
Textual	Typewritten	Typewritten	Typewritten
Input-response relation			
Reactivity	Reciprocal/ adaptive	Reciprocal/ adaptive	Reciprocal/ adaptive
Scope	Narrow	Narrow	Narrow
Directness	Direct	Direct	Direct

*Currently excluded from the procedure due to it being uninformative (too easy) as regards learners' abilities (based on the Pilot study results).

3 Test-takers' characteristics

A personal characteristics

1. Age: 7-9-graders (i.e., 12-16 yrs).
2. Sex: males and females.
3. Mother tongues: Finnish, Russian (Estonian).
4. Type of education: general (school)
5. Prior experience with similar tests: familiarity with the modality / task types (needs confirmation). Highly probable: not familiar with adaptive CF.
6. Type and amount of preparation for the test: none.

B Topical knowledge

1 Knowledge of the genre, i.e., semi-formal/informal E-mail (not a necessary requirement).

C test-takers levels

1 General level: A1-B1 (a median of A2) on the CEFR scale (presumably).

E Affective responses

1. Adaptive CF group:
   a. expected to be positive since then learners will be able to see that they benefitted from the dynamic part of the test;
   b. some learners' affective responses can be less positive due to their beliefs/values, especially if the procedure is considered to be a test by them, e.g., a belief that a test/exercise should be all about getting a correct response.

2 Static CF group: while the learners in this group will receive feedback on their performance when working on the tasks, which learners perceive as useful, see the discussion on the learners' perceptions of the usefulness of corrective feedback, the affective responses are expected to be less positive as fewer learners will probably consider that they benefitted from taking the test.

4 Construct definition

A: the assessed level (stage 5 questions, i.e., wh-questions with auxiliaries, e.g., *Where can I find photos of the pets?*):

1 Knowledge of syntax:
   a. production/formation of wh-questions with auxiliaries (consider both accurate and inaccurate stage 5 questions with the **correct** word order).

2 Morphological knowledge:
   a. use of correct inflexions in stage 5 questions, both auxiliaries and main verbs (both regular and irregular).

3 Vocabulary knowledge: question words*:
   a. modal auxiliaries *can* and *must*; auxiliary *do*.

*Excluded from the assessed construct following the Pilot study (too easy).

B: the previous level(s) (stages 3 and 4 questions, include stage 4 questions to the pretest/posttest):

1 Knowledge of syntax:

a. production/formation stage 3, i.e., questions of questions with, wh-fronting (e.g., **Where I can find photos of the pets?*) but also do-fronting (e.g., *Do you also have talking parrots?*) and other-fronting;

b. production/formation of stage 4 questions, i.e., yes/no questions and questions with pseudo-inversion (e.g., *Where is the pet shop?*).

2 Morphological knowledge:

a. N/A, as the idea is to see whether the test-takers are able to form stage 4 (and 3) questions, both correct and incorrect, while not being able to form stage 5 questions.

3 Vocabulary knowledge:

a. question words*; modal auxiliaries *can* and *must*; auxiliary *do*.

*Not applicable any longer.

C: the feedback

1 Adaptive CF group: the feedback adaptation during the dynamic part of the assessment is operationalised based on Aljaafreh and Lantolf (1994: 471):

a. implicit indication that something is wrong, e.g., *think more carefully*;

b. the location of the error is narrowed down though not identified exactly, e.g., *look at this part of the sentence*;

c. the location of the error is narrowed down further, the nature of the error is identified, and elicitations and/or metalinguistic clues are provided, e.g., *which ending do you need here*;

d. the location of the error is identified; examples of the correct structure are given, e.g., *look at the following sentence and compare it to yours*;

e. the correct response is provided accompanied with the explicit explanation of what was wrong, e.g., *you needed to use the auxiliary* ***do****; the correct sentence is….*

2 Static CF group (KOR feedback):

a. explicit (but not detailed) indication that the response is correct, e.g., *your answer is correct*;

b. explicit (but indirect and not detailed) indication that the response is incorrect, e.g., *your answer is incorrect*.

3 No feedback: not applicable currently, but can be a possibility in the future.

4 Learner profiles for teachers (based on learners' pretest-intervention-posttest performance):

a. learners' mistakes;

b. the feedback they require to self-correct (if at all), i.e., to be able to form correct stage 5 questions as per the construct definition;

c. to consider, learners' current ability level operationalised following the developmental stages 3-5 in question formation;
d. an explanation of question development stages.

5 Evaluation of qualities of usefulness

A Reliability

1 Minimum acceptance level:

a. purpose: low-stakes, so a moderate level of reliability is enough;
b. construct(s): narrow and exact, so a moderate-to-high level is expected;
c. how to specify:
    - scoring criteria;
    - rubrics;
    - time allocated;
    - development as a result of the dynamic test (which is also a part of construct validation). The latter, in practice, means the unreliability (as conventionally defined) of the treatment (i.e., the dynamic part) and posttest as well as, possibly, lack of test-retest reliability.

2 Logical evaluation:

a. comparability/variability across subjects, test versions, and administrations:
    - the difference between the dynamic and static test versions should largely be due to the feedback provided to test-takers;
    - however, some learners' characteristics, such as their educational beliefs, e.g., that their mistakes should be corrected explicitly could result in them not paying attention to feedback and resorting to other strategies during the treatment, which would negatively affect the reliability of scores (but also the **construct validity**, the **interactivenes**, and the **impact** of the procedure);
b. test rubric:
    - the same instructions should be given to all the learners regardless of the group, L1, etc.;
    - this should include the equivalence/adequacy of translations of the instructions into Russian/Estonian but also the interface translation (see **usability**);
c. the characteristics of the expected responses of both the pretest/posttest and the test parts are consistent with the purposes of these parts;
d. the order of the feedback messages is based on Aljaafreh and Lantolf's (1994) results;
e. scoring:
    - the scoring rubric designed to clearly and unambiguously define what and how to score the constructed response items/tasks (pretest/posttest), e.g., *this test-taker can write wh-questions with auxiliaries*

using the correct word order or *this test-taker can write correct wh-questions with aux* ***do***;
- the anonymity of test-takers is ensured during the scoring of the pretest/posttest tasks, i.e., the only information about the test-taker (in addition to his/her performance) displayed to the rater is the test-taker's ID;

f. each test-taker logs into the system with his/her username and password:
- each username is linked to a unique test-taker's ID with which all the data recorded by the system is associated.

3 Empirical evidence

a. reliability estimates:
- comparability with other test versions/participants;
- internal consistency/reliability of the pretest;
- differences in the performance on the treatment tasks between the dynamic and static test groups;
- the pre-/posttest performance is not significantly different in the static CF group;
- the posttest performance of the adaptive CF group is significantly better than the pretest performance;
- The development of the adaptive CF group is significantly higher than that of static CF group;

a. time allocated: test-takers performance, interviews with test-takers;

b. scoring criteria clarity/quality: interviews with raters and teachers (also see **Section 5F**).

B Construct validity

1 Minimum acceptance level:

a. purpose:
- low stakes (high stakes?), some evidence required;

b. construct definition:
- evidence related to three components of language knowledge need to be collected;

c. domains of generalisation evidence that the interpretations of the performance are generalizable to:
- situations in which learners are expected to ask for (further) information in writing (specifically in a semi-formal E-mail); as well as
- educational tasks in which learners are expected to formulate stage 5 questions in writing.

2 Logical evaluation:

a. the language construct to be assessed is clearly defined (developmental stages in question formation—stage 5 as being assessed, stage 4 as an indication of the previous stage);

b. the construct as defined in the test is relevant for the test purpose, i.e., diagnosis (including self-diagnosis) and development, as
   - the definition incorporates development in a clear and unambiguous way and
   - common mistakes in syntax, vocabulary and grammar that learners make when forming stage 5 questions were established by means of analysis the performance samples collected in the CEFLING project;

c. the pretest/posttest and the treatment tasks fully reflect the construct, in that
   - they elicit the production/formation of wh-questions with modal auxiliaries, as well as auxiliaries *do*, *does*, and *did*;
   - the feedback, the tasks, the instructions, etc. take into account the specific genre, which is a semi-formal E-mail;
   - the vocabulary used in the exercises should be familiar to the learners. The following textbooks used in grades eight and below were studied for this purpose:
     Folland T., Haavisto A., Heinonen M., Nieminen A., Woods M.(2009). *Smart Moves 1: Exercises*. Otava.
     Kangapusta, R., Lehtonen, E., Peuraniemi, J., Westlake, P., & Haavisto, A. (2002). *Key English 7: Workbook*. Helsinki: WSOY.
     Kangaspunta, R., Lehtonen, E., Peuraniemi J., & Westlake, P. (2005). *Key English 8: Workbook*. Helsinki: WSOY.
     Westlake, P., Pitkänen, E.-L., Satamo, S., Lintunen, M.-L. (2000). Yes Beginnings Ratkaisut (2nd ed.). Sanoma Pro Oy.
   - translations should be provided for less frequent words/phrases in the help menu.

d. the scoring criteria for the pretest/posttest tasks reflect the construct (also see **Section 5A**);

e. the performance log obtained should indicate
   - whether learners are able to produce stage 5 questions (both accurate and inaccurate) and whether they can also produce stage 4 questions (both accurate and inaccurate);
   - the exact mistakes learners make when forming stage 5 questions and the amount of assistance they require to self-correct these mistakes;

f. the setting (i.e., a test-like situation) might result in that learners use strategies they use for taking conventional tests during the dynamic test part:
   - care should be taken not to call the procedure a test. Instead in the instructions to the test-takers, it is to be called the exercises that aim to help the learners see their problems and learn something;

  - the same should be done also in the instructions/introduction of the test to teachers;

g. instructions provided in learners' L1(s) should be understandable by eight-graders, as no special terms are used there;

h. the characteristics of the expected response that can result in different performance are those reflecting the construct being evaluated, i.e., learners' ZPDs and the specific problems they have with forming stage 5 questions.

i. The computerised modality might result in the construct irrelevant variance. Thus:
  - the task types should be selected such as the learners were familiar with them—for this purpose both the textbooks (see **Section 5B**) and computerised (web-based) exercises, including those available for the textbooks used at grade 8 in Finland were studied. The examples of the Websites studied included (but were not limited to):
    http://www.easyenglish.com
    http://www.tolearnenglish.com
    http://www.ecenglish.com/learnenglish
    http://a4esl.org
    http://www.learnenglish-online.com/
    http://www.englishexercises.org
    http://www3.otava.fi/smartmoves/
    http://wsoypro.fi
  - in the instructions (also available from the help menu), the mechanics of doing the tasks are detailed;
  - and practice tasks are supplied following the instructions (also see the **usability** section).

3 Empirical evidence

a. adequate construct definition:
  - interviews/questionnaires with teachers/learners but also less formal interviews with colleagues regarding their experience of taking the pretest and the dynamic/static intervention parts of the *Questions test*;
  - establishing the correspondence of the mistakes made in the pretest tasks with those that the treatment tasks were designed to cover;
  - analysis of the distractors in the multiple-choice items (i.e., that all of them were selected by learners);
  - evidence for development of the learners' ability to form stage 5 questions (but not stage 4 questions) in the adaptive feedback group;

b. consistency of the pretest/test items studied quantitatively;

c. evidence for the lack of construct irrelevant variance: interviews and questionnaires with test-takers and teachers (also see **Section 5F**).

C Authenticity

1 Minimum acceptance level:

a. considerations:
  - the potential impact of the *Questions test* on instruction is high, so the level of authenticity should be high;
  - domains of generalisation are rather specific, so a high level of authenticity should be required;

b. specification of the authenticity:
  - perceptions of test-takers and teachers regarding the correspondence between the pretest tasks and the 'real-life' TLU tasks;
  - perceptions of test-takers and teachers regarding the correspondence between the pretest/posttest and the test tasks and (remediation of) classroom instruction.

2 Logical evaluation

a. the description of the test tasks and the TLU tasks includes information about the setting, input, expected response and the relationships between the latter two;

b. taken together, the tasks correspond to the TLU domain tasks, as can be deduced from the tables in **Sections 2B** and **2C**.

3. Empirical evidence

a. the extent of correspondence between the TLU and pretest/test tasks:
  - developers', raters', teachers', and learners' responses to questionnaires and reports during interviews;
  - less formal/systematic communication, e.g., in the case of the developers, E-mails, meetings, etc.;

a. the extent to which the pretest/posttest and the treatment tasks are perceived as authentic:
  - same as previous.

D Interactiveness

1 Minimum acceptance level

a. considerations:
  - the involvement of learners' characteristics that are a part of the construct, as defined for the present procedure, i.e., it should be high;

b. levels:
  - language ability: high during the treatment, low during the pretest;
  - topical knowledge: low;
  - (meta)cognitive strategies: high;
  - affective response: high;

c. specification of the (extent of) interactiveness:
  - involvement of learners' characteristics (mostly qualitative analyses).

2 Logical Evaluation

a. topical knowledge beyond some limited specifics of the genre:
   - the appropriateness of some questions in a semi-formal E-mail is not required;
   - moreover, the latter will be taught in the treatment phase if required;

b. the characteristics of the test tasks are in line with the task items;

c. some metacognitive strategies, especially if test-takers consider the treatment to be a test, can result in a decreased interactivity between the feedback (and, consequently, the treatment items) and the test-taker during the treatment part (also see the **construct validity** section);

d. the processing of the pretest/posttest/treatment items involves a narrow range of abilities;

e. none of the language functions other than the demonstration of the language ability are involved in the construction of responses:
   - the feedback is given in learners' L1s;
   - the instructions are given in learners' L1s;
   - the vocabulary of the items should be selected such as they are familiar to the learners (see **Section 5B**);

f. the performance on the treatment tasks is dependent on all the previous items/tasks; the aim is that learners gradually realise what the correct responses should be and why;

g. the pre-/posttest tasks are designed so that, to the extent possible, no response on the previous items should influence the responses on the following items;

h. the involvement of metacognitive strategies in the pretest should be minimal (although not fully excluded);

i. during the treatment, test-takers are expected to employ different strategies, both cognitive and metacognitive, to make use of the feedback they receive, such as (but not limited to) making connections between their sentence and the ones shown to them, use context clues, identifying structures, evaluating, etc.:
   - arguably, during the posttest, they will probably employ a number of strategies to successfully solve the tasks due to the influence of the treatment;

j. learners' beliefs can also result in that during the treatment, they use strategies they would employ during a conventional assessment, such as achievement testing (**Section 5D**);

k. adaptive CF group learners' beliefs and values can result in less than positive affective responses (also **section 3E**); however, the affective responses should nevertheless be more positive as compared to the static CF group treatment;

l. the epistemological basis of the treatment procedure presupposes that the affective responses, except for the possible cases discussed above, should result in a situation where learners perform at the best of their abilities on the treatment tasks.

3 Empirical evidence:

a. feedback is not skipped by the test-takers:
   - time spent on (reading) the feedback during the treatment (recorded by the system);
   - test-takers questionnaire/interview responses;

a. the extent to which teachers/learners consider the language abilities, the (meta)cognitive strategies, the topical knowledge and affective responses to be involved in the learners' interaction with the tasks:
   - interviews/questionnaires.

E Impact

1 Acceptance level:

a. considerations:

- low/no-stakes decisions (as conventionally defined); high-stakes decisions if learners' development is considered important;
- possible effects of misevaluation: wrong feedback given by the teacher which is either not understood by the learner or is not required by the learner to improve his/her abilities, thus hindering his/her development;

a. level: high;

b. specification:

- development of learners' abilities operationalised as the difference between their pretest and the posttest scores (also **Section B**);
- possible change in learners' perception of the test, i.e., it should not always be about getting the answers right / good marks, but also about understanding their mistakes;
- possible change of learners' beliefs about the usefulness of corrective feedback;
- possible change of teachers' perceptions of their learners' abilities;

2 Logical evaluation:

a. the extent to which the experience of taking the test affects learners' language use (e.g. topical knowledge, perception of the target language use situation, areas of language knowledge, and use of strategies) should be at least noticeable;

a. due to the epistemological basis and consequently the design of the procedure, the adaptive CF provided to the learners should be relevant, complete, and meaningful;

b. the decision criteria are not applied in the same way to the dynamic and the static assessment groups, but are applied uniformly within the groups;

c. due to the increased precision of the data regarding learners' performance, these data should be relevant to the decisions being made;

d. learners are to be informed about the procedures with an important exception:

- for learners, the procedures will not be referred to as a test, as the latter can evoke certain beliefs/attitudes and, consequently, result in a different performance (also **Section 5B**) and a different impact on learners;

e. teachers will be fully informed about the procedures;

f. the areas of the language ability to be assessed are consistent with those that are included in the teaching materials (see **Section 5B**);

g. care was taken so that the test tasks characteristics would be, to the extent possible, consistent with the teaching/learning activities;

h. the purpose(s) of the test, i.e., diagnosis, self-diagnosis, but especially learners' development, should be fully consistent with teachers' goals; however, depending on teachers' beliefs, the way development is operationalised in the test can be incongruent with the way development is perceived by them;

i. the interpretations made on the basis of learners' performance, should be consistent with the values of the educational system, e.g., learners' independence as the ultimate goal of education;

j. the values and goals of the test developer coincide with those of society and the education system;

k. potential consequences of the test for teachers (on a micro level) include:
   - more fine-grained diagnosis of their learners' abilities;
   - change in teachers' beliefs regarding the purposes/applications of assessment;

l. The potential consequences for learners include:
   - the development of their ability to form stage 5 questions;
   - a change in their perceptions of / beliefs about the usefulness of different CF types, above all, implicit ones;
   - connected to the previous, one of the consequences can be raising awareness of utilising different strategies when self-correcting;

m. the most desirable consequence of using the test for the purpose it is designed for would be a remediation of classroom instruction following the test administration, including:
   - a change based on the specific results of the procedure;
   - more globally, a change in classroom feedback practices due to adopting a sociocultural perspective on development;

n. both changes can, in addition to the experience of the procedure and observation of the learners working on the test (both should be made available) be encouraged by detailed learner performance profiles compiled based on the data recorded by the *ICAnDoiT* system;

o. the least desirable negative consequence of taking the test can be a feeling of discouragement by the learners whom implicit feedback does not help to notice the gap between their response and the expected response:
   - this can result in that they do not benefit from the procedure and ignore any future feedback from the teacher whenever they consider it to be useless even if it is provided within their ZPD;

- similarly, if learners have prior beliefs resulting in that they consider implicit feedback useless, the result might be the same.

3 Empirical Evidence:

a. for test-takers:
    - test-takers will be informed about the purpose of the test while not being told it is a test so that their beliefs about the educational assessment are not evoked if possible (interview, questionnaires, and observations);
    - the development of their ability to form stage 5 questions (quantitative analysis of the learners' performance);
    - changes in their perception of the usefulness of corrective feedback (interviews);

b. for teachers:
    - a more fine-grained diagnosis of the learners' abilities (the teachers' evaluation of their learners' performance reports; possible: learner questionnaires / interviews are drawn upon in teacher interviews)
    - a change in the teachers' perceived usefulness of the procedure as a means of diagnosis and promotion of their learners' development (teacher interviews /think-aloud protocols);
    - a change in the teachers' evaluation of their learners' abilities (interviews following teachers' observation of their learners working on the tasks).

F Usability (based on Fulcher, 2003; Molich & Nielsen, 1990)
1 Acceptance level:

a. considerations:
    - depending on the elements of the interface, the potential effect of lack of usability on other aspects of usefulness is from small to high;
    - lack of usability of the system for test designers can result in decreased practicality and the possibility that it is not used at all;

b. level: high;

a. specification: quantitative and qualitative data regarding learners' experiences with the *ICAnDoiT* system.

2 Usability checklist:

a. the same functionality across the major browsers;
b. visibility of the system status;
c. match between the system and the real world;
d. user control;
e. consistency;

f. prevention of errors resulting from accidentally wrong choices (as opposed to errors pertaining to a lack of knowledge of the assessed construct);
g. recognition of what to do rather than recalling;
h. flexibility (and efficiency) of use;
i. aesthetic and minimalist design;
j. text size and font;
k. colours and graphics;
l. fields for constructed response items/tasks;
m. error messages to diagnose and recover from errors;
n. help and documentation.

3 Design decisions / logical evaluation

a. ensuring the same functionality and visual look (no additional software required) across:
   - Internet Explorer;
   - Google Chrome and;
   - Mozilla Firefox;
b. The system informs test-takers of what is going on through the feedback (not to confuse with the CF learners receive) in form of:
   - system messages;
   - instructions, etc.;
c. no specialised language is used in the part of the system that learners have access to):
   - the metaphors used as parts of the interface, e.g., the help button, are easily recognisable and are also explained in the instructions;
   - the emoticons serving a part of the feedback were designed to stress the hierarchy of the feedback messages;
d. user control:
   - users can log out from the system without having to finish the test; the next time they log into the system, they are asked to finish the unfinished test (start from the item they finished at) before they can move to another test;
   - the logout button is always present in the upper left-hand corner;
   - in response to the Pilot study results, no user progress bar is implemented in the system, as there will be variable test lengths due to algorithms used;
e. consistency in the use of language (e.g., button labels, etc.);
f. prevention of involuntary errors:
   - users cannot go back to previous items and change their answers to them, but the risk of sending an unfinished/unintended answer is minimised, as
   - the OK button is placed so that it is hard to press it accidentally;

- checks are implemented for the pretest tasks controlling that an answer is final before the user moves to the other task (e.g., *are you ready?*);

g. system users (i.e., test-takers, teachers, and system administrators) do not have to remember what to do next:
   - as regards test-takers, the help menu contains both exercise instructions and glosses where required; in the E-mail writing task, a part of the instructions is presented above the field where candidates are supposed to write their E-mail;
   - extensive hints are given to system administrators/teachers regarding various fields in the system, i.e., what should be written there and how;

h. flexibility:
   - learners can skip the practice tasks should they wish to;
   - depending on test designers' needs, the complexity of the interface (which, nevertheless, is counterbalanced with detailed instructions) varies, i.e., most of the test/item settings are optional and are used depending on the test design;

i. only the required information is provided to test-takers:
   - in some cases, e.g., the help button, easily recognisable metaphors are used;
   - system messages are concise;
   - navigation controls are kept to the minimum;
   - no scrolling is required, except for in the E-mail writing task;

j. text:
   - upper-case text is avoided;
   - a possibility to use bold text is added;
   - the text size is sufficiently large on a computer screen;
   - the default font type is *Arial*;
   - the font colour is black on the white background except for a limited number of cases to indicate a successful/failed action (see point **k**);

k. colours/graphics:
   - green background (just one line) is used to indicate a successful action (e.g., login) and red, a failure to do so;
   - graphics are used in the CF provided to learners;
   - adding a picture is possible (though optional) in the E-mail reading task type;

l. constructed-response task-types considerations:
   - the size of the fields was fixed in the gap-filling task of the pretest;
   - as it is hard to account for the length of the response in the E-mail writing task, instead of providing a large size field, vertical scrolling (scrollbar) is enabled for it;

m. an error message is displayed if a learner submits a blank response, giving clear and concise information about what is expected from the learner before returning the learner to the item;

n. help:
    - the help menu details what is expected to do in the task, and how to do the task;
    - glosses are provided;
    - a user manual is designed for the teachers / test designers*.

* Presently, this is replaced with instructions built into the system.

4 Empirical evidence (currently, only for test-takers)

a. a multi-stage usability study:
    - a semi-formal trying out of the system and the test with 1-2 learners and several colleagues followed by modifications in the interface (interviews, think-aloud protocols);
    - a pilot study with a larger group of learners (and teachers) aiming to collect both qualitative and quantitative data about the usability of the system and the test followed by modifications in the interface (interviews, questionnaires, think-aloud protocols);
    - a study with another group of learners; modifications introduced if required (interviews, questionnaires, think-aloud protocols).

G Practicality

1 Specification/evaluation:

a. the amount of resources required for the design and the implementation stages was not high, as the design team included two people only with occasional input from other people, e.g., when providing adequate translations for the available languages;

b. on the other hand, in terms of the time spent on the project, the limited human resources resulted in a higher amount of hours spend on the test design (up to 40 hours a week per person);

c. specific design decisions directed to increasing the practicality of test compilation stage included:
    - specific instructions for test designers accompanying the fields to be filled;
    - automatic generation of the feedback messages for all the task items (owing to intelligent algorithms) after the feedback messages are entered once (reducing the time of adding tests to the system substantially);

d. the piloting phase required slightly more human resources (e.g., appointing an interviewer), but these were not high either;

e. the administration stage:
    - little resources, as after the test is compiled, a potentially large number of learners can take it simultaneously (depending on the server).

6 BLUEPRINT

A Test structure

1 Number of parts/tasks: 3 parts (pretest-test-posttest):

a. pretest/posttest tasks: Advertisement reading, E-mail writing, gap-filling

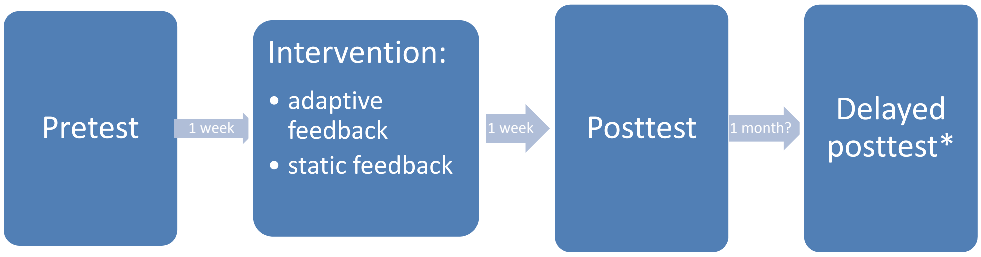
b. test-tasks: E-mail reading, matching (currently excluded from the procedure), ordering (the second ordering task for a higher-ability), multiple-choice—3 exercises (problems with, *do*, *does*, and *did*);

c. two types of intervention are designed:

- adaptive CF group: adaptive CF is presented in response to learners' item-by-item performance (**Section 4C**);
- static CF group: KOR feedback is presented in response to learners' item-by-item performance.

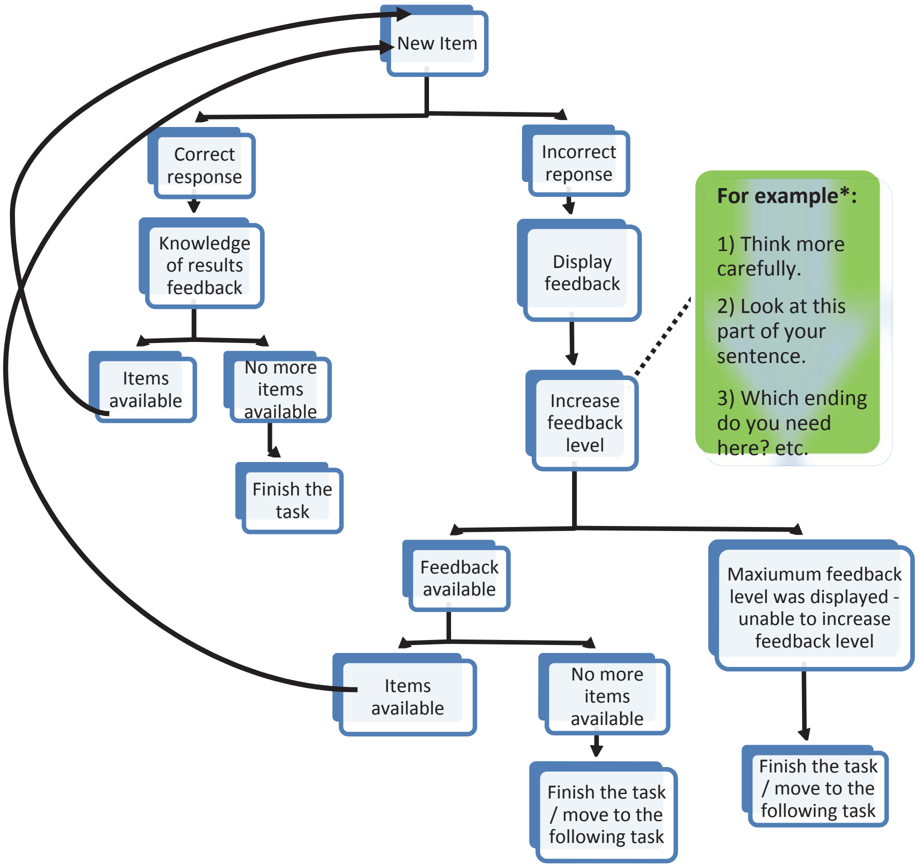
2 Flowcharts representing the whole procedure and the items in each task in the intervention phase:

a. the whole procedure:

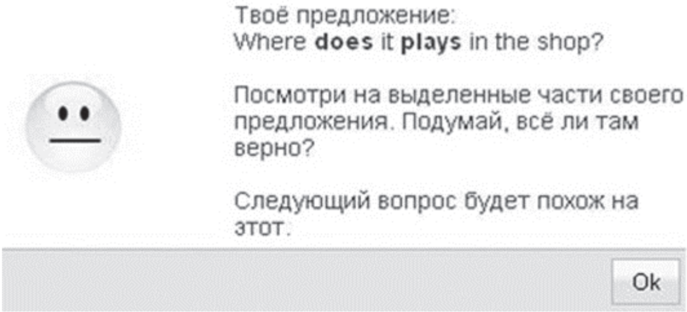


*A delayed posttest was not implemented in the actual study (Article I).

b. Treatment, experimental group:


New Item
Correct response
Incorrect reponse
Knowledge of results feedback
Display feedback
Items available
No more items available
Finish the task
Increase feedback level
For example*:
1) Think more carefully.
2) Look at this part of your sentence.
3) Which ending do you need here? etc.
Feedback available
Maxiumum feedback level was displayed - unable to increase feedback level
Items available
No more items available
Finish the task / move to the following task
Finish the task / move to the following task

* For clarity sake, it should be noted that the current version of the feedback is more complex than that. For example, the feedback that the participants in Leontjev (2014) in addition to the feedback message per se, included the sentence the test-taker formulated, the message telling him/her that the following item will be similar to the one the feedback is provided for, and an emoticon different for each feedback message as in the figure below:

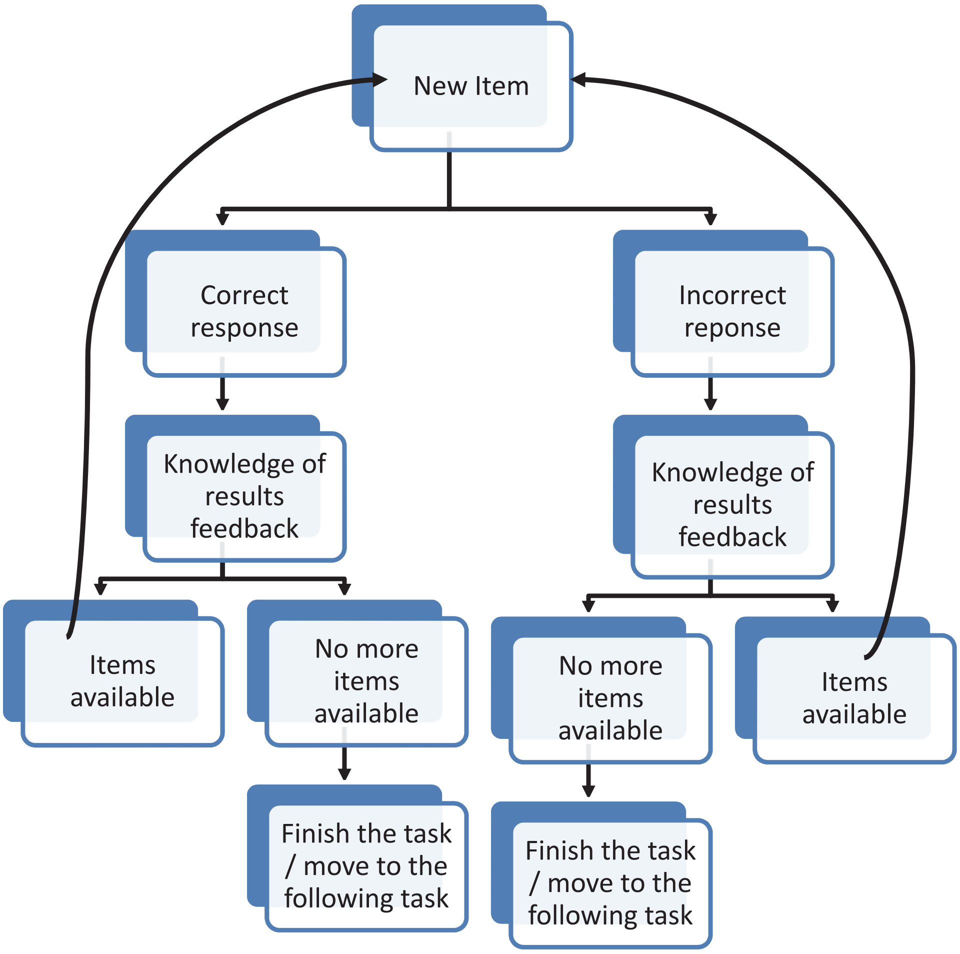


Твоё предложение:
Where **does** it **plays** in the shop?

Посмотри на выделенные части своего предложения. Подумай, всё ли там верно?

Следующий вопрос будет похож на этот.

c. Treatment, control group:



New Item
Correct response
Incorrect reponse
Knowledge of results feedback
Knowledge of results feedback
Items available
No more items available
No more items available
Items available
Finish the task / move to the following task
Finish the task / move to the following task

3 Relative importance of the parts:

a. the treatment procedure is the most important, as it is there where most diagnostic inferences are made.

4 Salience of the parts:

a. while the pretest, the treatment and the posttest are designed to evaluate the same construct and the performance on all the three parts should be evaluated as a whole, the pretest/posttest parts are clearly distinct from the treatment by design.

B Scoring

1 Pre-/posttest:

a. criterion-referenced:
   - only the accurately formed stage 5 (and stage 4 questions) are considered correct;

b. scoring rubrics reflect the assessed construct:
   - question words;
   - word order;
   - wh-questions with *do*;
   - wh-question with *does*;
   - and wh-question with *did*.

2 Treatment:

a. scoring is done automatically; the following is recorded:
   - test-takers' responses;
   - the correctness of the responses;
   - the feedback that test-takers receive for every item;
   - the time spent on solving each of the items;
   - and the time spent on (reading) the feedback.

C Communicating the instructions

1 Instructions to teachers:

a. teachers will be informed about the purposes of the test;

b. teachers will be instructed about what is expected from them while monitoring test-takers' performance, including:
   - that they are allowed to help the test-takers with their vocabulary queries, but not grammar;
   - that the test-takers will be required to work on their own and not use the sources other than those provided (i.e., the Internet grammar Websites, google translate, etc. will not be allowed to use);
   - and that they should not refer to the procedure as a test, but a set of exercises helping learners to find out about their problems in formulating questions in English;

c. teachers will also be instructed that if the procedure is used for learners' self-diagnosis and development, it should rather be emphasised that cheating on the tasks would be a self-deceit.

2 instructions to test-takers:

a. test-takers will be informed that they will complete a number of tasks so that they themselves are able to find out how well they can form questions in English;
b. test-takers will be told to turn to the persons monitoring their performance should they have any questions but also make use of the help menu in the system;
c. the instructions will be given predominantly in the / by the *ICAnDoiT* system.

D Administration

1 Preparation of the setting:

a. making sure that all of the computers function correctly;
b. although a similar usability is expected regardless of the Internet browser used, the same Internet browser should be used by all test-takers in the same session;
c. making sure that the Internet browser functions as expected on all the computers;
d. making sure that the teachers/proctors remember the instructions and are able to react in the expected way to the test-takers' queries / inappropriate behaviour.

E Try-Out

a. determination of appropriate time allocation;
b. collecting and analysing the data for the a posteriori evaluation of the qualities of usefulness;
c. changes (if required);
d. further validation.

ORIGINAL PUBLICATIONS

I

THE EFFECT OF AUTOMATED ADAPTIVE CORRECTIVE FEEDBACK: L2 ENGLISH QUESTIONS

by

Dmitri Leontjev 2014

APPLES : Journal of applied language studies, 8(2), 43-66

Reprinted with kind permission by APPLES : Journal of applied language studies

Apples – Journal of Applied Language Studies
Vol. 8, 2 , 2014, 43– 66

The Effect of Automated Adaptive Corrective Feedback: L2 English questions

Dmitri Leontjev, University of Jyväskylä

The research on the amount and the types of corrective feedback beneficial for learning a second or foreign language has produced inconsistent results. Interestingly, studying corrective feedback from the perspective of a sociocultural theory of learning has the potential to resolve these differences although so far, these studies have been largely qualitative. The present study attempts to contribute to the existing research on corrective feedback from this perspective by comparing the effects of two types of automated corrective feedback on learning: adaptive feedback (i.e., feedback incrementally adapting to learners' abilities by becoming more explicit and detailed) and knowledge of response feedback. The participants were learners of English randomly assigned to two groups, receiving either adaptive feedback (experimental group) or knowledge of response feedback (control group). The aim was to establish whether adaptive corrective feedback had a positive effect on learning, the target being L2 (second or foreign language) English questions. The findings indicate a significantly higher positive effect of the adaptive corrective feedback. Furthermore, the experimental group considered the feedback to be significantly more useful for learning than the control group although there was not a clear difference between the two groups' perceived usefulness of the feedback for getting the answers right during the intervention. It is argued that adaptive corrective feedback can raise learners' awareness of their mistakes, and it is suggested that it can facilitate individualised approach to learners. Further research is suggested.

Keywords: feedback, testing/assessment, second language (L2) learning, sociocultural theory, computer-assisted language learning

1 Introduction

It has been generally assumed that corrective feedback plays an important role in learning a second or foreign language (e.g., Bitchener 2008; Carroll & Swain 1993; Ferris 1995). At the same time, there is much less consensus as to the type and the amount of corrective feedback, both on written and spoken performance, that is more beneficial for learning (e.g., Ellis 2009; Pica 1994). This is especially the case with studies comparing the effect of explicit (i.e., overt corrective feedback) with that of implicit feedback (i.e., feedback that does not overtly state that the performance is incorrect). Hence, while some studies (e.g., Ellis et al.

Corresponding author's email: dmitri.leontjev@jyu.fi
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies, University of Jyväskylä

http://apples.jyu.fi

2006; Nassaji 2009) demonstrated the superiority of explicit corrective feedback, others (e.g., Iwashita 2003; Kang 2009) did not find any clear difference between the two kinds of feedback. Furthermore, it should not be forgotten that there are researchers who challenge the effectiveness of corrective feedback. Truscott (1996, 1999), for example, claimed that the evidence for the beneficial effect of correction had been inconsistent and suggested that corrective feedback can be detrimental for language learning, especially if it is provided regardless of learners' developmental readiness to understand their mistakes (1996: 344).

Interestingly, studies considering corrective feedback from the perspective of a sociocultural theory of learning (e.g., Aljaafreh & Lantolf 1994; Nassaji & Swain 2000) can potentially resolve these differences. These studies build on the Vygotskian concept of Zone of Proximal Development (ZPD), formulated as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky 1978: 86). In other words, according to the theory, learning is a result of collaboration between the tutor and the learner within the latter's ZPD, which involves graded support, known as mediation, provided by the tutor.

In the present study, I will refer to corrective feedback provided within learners' ZPD as to adaptive corrective feedback, the latter defined by Vasilyeva et al. (2007: 11) as feedback dynamically adjusting to users' abilities, characteristics, and/or performance. The reason for not using the term mediation is because the latter can include, but is not limited to, different forms of corrective feedback (e.g., Ableeva 2010; Poehner 2008). I will use the attributive static to refer to feedback/assessment not considering learners' ZPD.

However, studies looking into adaptive feedback/mediation in L2 teaching and learning are not numerous and often are predominantly descriptive. The study of Aljaafreh and Lantolf (1994) serves as an excellent example of presenting the process of negotiating corrective feedback in learners' ZPD until it matches their abilities. Tracing the influence of corrective feedback on three learners' L2 development, the authors designed a Regulatory Scale consisting of thirteen feedback messages gradually becoming more explicit and detailed. Importantly, they demonstrated that any feedback can be useful if it is provided within a learner's ZPD. While the contribution of the study is undoubted, the study design was largely descriptive.

Nassaji and Swain (2000) addressed this limitation, conducting a quasi-experimental case study of two L1 Korean learners of English, the first of whom was given adaptive corrective feedback in response to her mistakes in the use of the English articles, and the other given random feedback that did not take her ZPD into account. Having collected and analysed both qualitative and quantitative data, the authors concluded that the adaptive feedback was more beneficial when compared with the feedback that disregarded the learner's ZPD. They also found that in the case where the feedback was provided in a random manner, explicit feedback was more helpful. Yet, as this was a pilot study with only two participants, their results lack generalizability.

Adapting the amount of assistance to learners' abilities also lies in the core of dynamic assessment, which is based on the concept of ZPD and combines assessment and instruction into a single process. The key difference of dynamic assessment from its static counterparts is that during the former, the learners are provided with different kinds of mediation helping them to perform beyond the level they would be able to while working independently (Leung 2007; Poehner 2008). Dynamic testing/assessment seems to be a reasonable basis for accumulating empirical data on the effect of adaptive corrective feedback, as it allows for collecting experimental data by means of validated instruments. However, as with most of the research adopting the sociocultural paradigm, there appears to be a lack of quantitative studies in the field of dynamic testing/assessment (see section 2.1 for a discussion).

The lack of experimental evidence about the effect of adaptive corrective feedback is understandable considering the qualitative tradition in the sociocultural research, which conventionally aims at interpreting development rather than measuring it. On the other hand, studies confirming the positive effect of adaptive corrective feedback experimentally could strengthen the argument for its usefulness. Moreover, such studies have the potential to alleviate some of the criticism, especially directed towards dynamic assessment (see e.g., Poehner 2008 for a discussion). What is more, as regards classroom instruction, procedures allowing to trace the development of learners as a group could be helpful for language teachers, for example, for finding out whether a certain structure that they have been teaching is within most of their learners' ZPD.

The present study seeks to add to the body of research on corrective feedback from a sociocultural perspective by finding out whether adaptive corrective feedback provided during a computer-based dynamic test is more effective for learning than static implicit feedback (see section 3.1 for the specific research questions). On the basis of the previous (mostly qualitative) research, I could tentatively hypothesise that L2 English learners receiving adaptive feedback are more likely to develop their L2 ability than learners receiving static (implicit) corrective feedback. While recognising the value of qualitative analyses that dominate these studies, in the present study, I will place the emphasis on experimental evidence for the beneficial effect of corrective feedback provided within learners' ZPD. An obstacle for collecting such data has been the impracticality of assessing a number of learners in face-to-face sessions (which is a common way adaptive feedback / mediation is provided to learners); yet, a recent advancement in dynamic assessment addresses this issue. I will discuss this (and other research relevant to the study) in some detail in the section to follow. I will then describe the present study, introduce the data analyses, and report on the findings. I will also suggest further research to reinforce the findings of the study.

2 Background

In this section, I will present a review of the research on computerised dynamic assessment, learners' preferences and perceived usefulness of corrective feedback (which, I will argue, is important to take into account in computerised dynamic assessment), and the development of L2 English questions (which were selected as the target of the intervention).

2.1 Dynamic Assessment

There are two major approaches to dynamic assessment: interventionist approach and interactionist approach. The difference between them lies in the way mediation is provided during these two types of assessment. During the former, the mediation is standardised and is given in a predefined order, often in the form of corrective feedback ranging from implicit to explicit types. In the latter approach, the required mediation emerges during the interaction between the learner and the examiner (Poehner 2008).

There have also been several successful attempts at creating computerised dynamic tests where mediation is provided automatically. The drawbacks of computerised delivery include the impossibility of establishing how learners would respond if other mediation was provided (Poehner 2008: 177) and the difficulty of tracing learners' reciprocity to mediation (see Poehner (2005) for a discussion of the latter). Its advantages, however, which include the possibility of assessing a large number of learners simultaneously, (re-) assessing the learners under uniformed conditions, and generating learners' performance reports automatically, make computerised dynamic assessment an interesting research tool.

However, not many implementations of computerised dynamic assessment have been reported in the literature. The rare examples include a computerised version of Guthke and Beckman's (2000) Leipzig Learning Test, a test for diagnosing children's learning problems, and Teo's (2012) computer-based dynamic test of learners' metacognitive reading strategies. As regards L2 computerised dynamic assessment, there seems to be only one computer-based dynamic assessment system that addresses learners' problems with L2 grammar and only to the extent it is required for listening and reading comprehension (Ableeva 2010, 2012).

These tests are designed following the interventionist approach to dynamic assessment, which is close to psychometrically oriented non-dynamic tests. This approach, especially the sandwich test format, in which treatment is conducted between an unmediated pretest and a posttest (Poehner 2008) and which, consequently, favours experimental research designs, seems to be promising for the purpose of collecting evidence on the effect of adaptive corrective feedback.

However, there are not many studies on the influence of mediation in computerised dynamic assessment that are supported with quantitative data. In Teo's (2012) study mentioned earlier, the learners' abilities before and after the intervention were compared statistically, but the author did not contrast the effect of adaptive with that of static corrective feedback. Ableeva (2010) also conducted several quantitative analyses of her data, which revealed the positive effect of the mediation. Other than that, the reports have been largely descriptive.

2.2 Learners' Perspective on Corrective Feedback

Constructing learners' ZPD is a dialogical activity. Thus, learners' reciprocity to mediation is an integral part of the sociocultural perspective on development. In his study, Poehner (2005) designed a Learner Reciprocity Typology—a scale in which he arranged the learners' reciprocal moves from being unresponsive to mediation due to being other-regulated to incorporating it to rejecting it due to being fully self-regulated, which, he claimed, also reflected learners' development.

Nevertheless, it seems that learners' expectations of corrective feedback can also influence their responsiveness to and, ultimately, the usefulness of the latter. It has been found that while learners generally consider corrective feedback useful, especially feedback on their lexical, structural, and grammatical errors (Amrhein & Nassaji 2010; Hyland 2001; Leki 1991), teachers' practices, including feedback, may not be effective if they do not meet learners' expectations and preferences (e.g., Schulz 2001).

Speaking of the findings regarding learners' preferences of corrective feedback, they are somewhat varied. Amrhein and Nassaji (2010) found that both high-achieving and low-achieving learners are in favour of more explicit feedback types whereas teachers generally prefer more implicit feedback. Hyland (2001), on the other hand, points out that some learners also acknowledge the usefulness of implicit feedback for developing their language skills. However, by and large, the research demonstrates that if feedback is focused on grammatical and structural errors, then learners are generally in favour of more explicit corrective feedback (Ashwell 2000; Leki 1991). Amrhein and Nassaji (2010: 116) note that by doing so learners, especially high-achieving ones, make their lives easier, placing the responsibility of correcting their mistakes on teachers.

There is, thus, a possibility that learners can attribute different meanings to feedback usefulness—usefulness for learning and usefulness for getting the correct answers effortlessly. More importantly, this suggests that learners' rejection of feedback might not always be the manifestation of their abilities but also root in their preferences of corrective feedback. The latter is especially important for computerised dynamic assessment, where it is hard to trace learners' responsiveness to mediation.

2.3 Stages of Acquisition and Corrective Feedback

Alternatively, learners' development can be seen from a different perspective—as stages of acquisition. The stages in question development identified in the context of Pienemann's Processability Theory (Pienemann 2005) can serve as an illustration of this perspective (Table 1)[1].

Table 1. Stages in question development (adapted from Pienemann 2005; Spada & Lightbown 1999)

Stage 1	Single words, phrases: *How are you?*
Stage 2	SVO: *The tea is hot?*
Stage 3	Fronting: Do: **Do he work? Does he work?* Wh-: **Where the station is?* Other: **Is the boy is beside the bus?*
Stage 4	Inversion: Yes/No: *Has he seen you? *Have he seen it?* Pseudo Inversion: *Where is John?*
Stage 5	Do/Aux 2nd: *Why did he sell that car?*
Stage 6	Cancel Inversion: *I wonder where he has gone?*

According to this theory, a learner cannot, for example, move to stage 3 of question development before stage 2 questions have emerged in his/her interlanguage, and learners move through the same developmental stages regardless of their L1. Yet, one reservation should be made. This order refers to oral production. Alanen and Kalaja (2010), who studied the L2 English performance of 250 L1 Finnish grade 7-9 learners as a part of the CEFLING project (www.jyu.fi/cefling), found the same stages in writing. However, while learners tend to use more questions at higher stages as their proficiency grows (Alanen & Kalaja 2010), it seems that they do not adhere to the developmental stages as rigidly as in spoken language (e.g., Spada & Lightbown 1999).

A number of studies have also demonstrated that corrective feedback can influence the way learners use L2 English questions (e.g., McDonough 2005; White et al. 1991), especially if an opportunity for production of modified output is provided. This makes L2 English questions an interesting treatment target in studies comparing the effects of different kinds of corrective feedback.

3 Methodology

3.1 Research Questions

The present study adds to the existing research on corrective feedback by examining the adaptive corrective feedback provided automatically in a web-based assessment/tutoring system, with the goal of establishing its effect and its perceived usefulness. Specifically, the study aims at finding answers to the following questions:

- Do L2 English learners receiving adaptive corrective feedback improve their ability to form questions significantly more than learners receiving knowledge of response feedback?
- Do learners receiving adaptive corrective feedback consider it more beneficial than learners receiving knowledge of response feedback a) for getting their answers right and b) for learning?

3.2 Design

To answer the research questions, a randomised pretest/posttest control group study was conducted. L2 English questions were found suitable to serve as the content of the exercises for the following reasons:

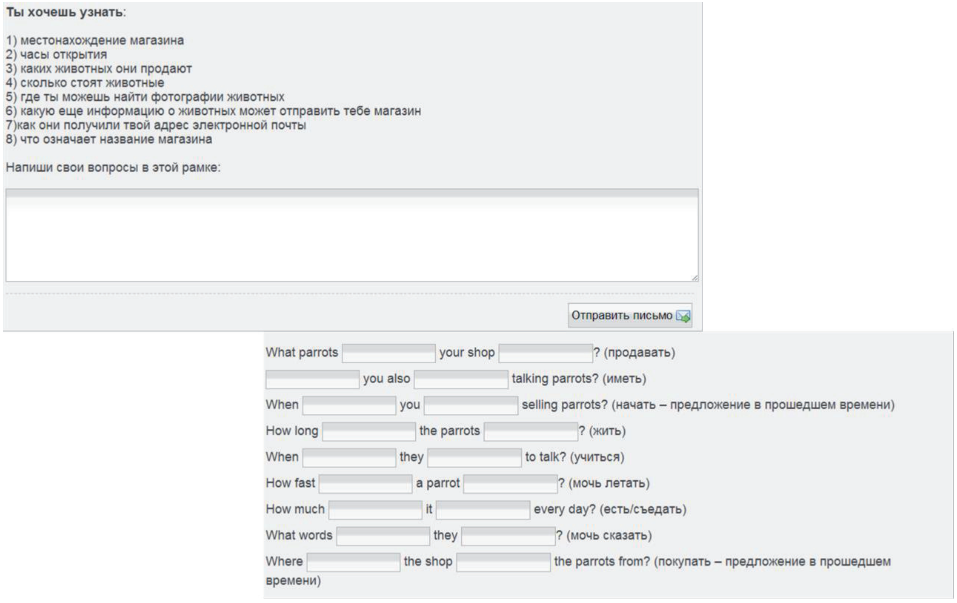
- feedback is found to influence the rate of their acquisition;
- learners generally consider feedback on grammar useful;
- the incremental development of questions allowed for tracing changes in the participants' performance in a more exact and a meaningful way;
- the stages in the development of L2 English questions seem to be the same regardless of learners' mother tongue.

To single out the typical errors the learners made, I examined Alanen and Kalaja's (2010) data. The analysis revealed a number of typical errors the learners made when formulating stage 5 questions, i.e., wh-questions with auxiliaries (see Table 1). Thus, I was able to focus the content of the exercises to stage 5 questions only. Nevertheless, to be able to trace the learners' development more clearly, it was decided to include several items eliciting the use of stage 4 questions (e.g., ___*you also* ___*talking parrots?*) into the pre-/posttest exercises (see section 3.3).

The independent variable in the study was the group the learners belonged to, either the experimental group (receiving the adaptive corrective feedback) or the control group (receiving the knowledge of response feedback). The number of stage 5 (and stage 4) questions correctly formed during the pre-/posttest and the learners' self-reports regarding the perceived usefulness of the feedback were the dependent variables.

3.3 Materials

The exercises in the pre-/posttest and the intervention were based on the imaginary situation where the learners received an E-mail from a pet shop, got interested in it, and decided to buy a puppy. It was expected that doing so would make the exercises resemble a real problem-solving communicative activity, thus adding to the authenticity of the exercises (see Bachman & Palmer 1996). In addition, it allowed for contextualising the sentences with pronouns as subjects in the exercises. Two exercises were designed for the pre-/posttest (Figure 1).



Ты хочешь узнать:

1) местонахождение магазина
2) часы открытия
3) каких животных они продают
4) сколько стоят животные
5) где ты можешь найти фотографии животных
6) какую еще информацию о животных может отправить тебе магазин
7)как они получили твой адрес электронной почты
8) что означает название магазина

Напиши свои вопросы в этой рамке:

What parrots ____ your shop ____? (продавать)
____ you also ____ talking parrots? (иметь)
When ____ you ____ selling parrots? (начать – предложение в прошедшем времени)
How long ____ the parrots ____? (жить)
When ____ they ____ to talk? (учиться)
How fast ____ a parrot ____? (мочь летать)
How much ____ it ____ every day? (есть/съедать)
What words ____ they ____? (мочь сказать)
Where ____ the shop ____ the parrots from? (покупать – предложение в прошедшем времени)

Figure 1. Pre-/posttest exercises (see Appendix 1 for a translation of the prompts)

The first exercise was writing an E-mail according to the prompts (provided in the learners' L1). It was selected as it was one of the task types used to collect the CEFLING project data (Alanen & Kalaja 2010). Six out of eight prompts elicited the production of stage 5 questions and two prompts, either stage 5 or stage 4 questions. The second exercise was a gap filling exercise in which each item had two gaps, one after the question word and the other after the subject.

The exercise contained nine items, one eliciting the use of stage 4 and eight eliciting the use of stage 5 questions.

The intervention exercises, which targeted the use of stage 5 questions only, were the following (the sample items presented in Figure 2):

- two ordering exercises to assess the learners' problems with the word order in stage 5 questions—the first with pronouns and the second with nouns as subjects (as Spada and Lightbown (1999) found that the former were easier to produce than the latter), and
- three ordered multiple-choice exercises (pronouns as subjects) aiming to discover the learners' problems with the use of auxiliaries *do*, *does*, and *did* and the use of the correct forms of lexical verbs in stage 5 questions.

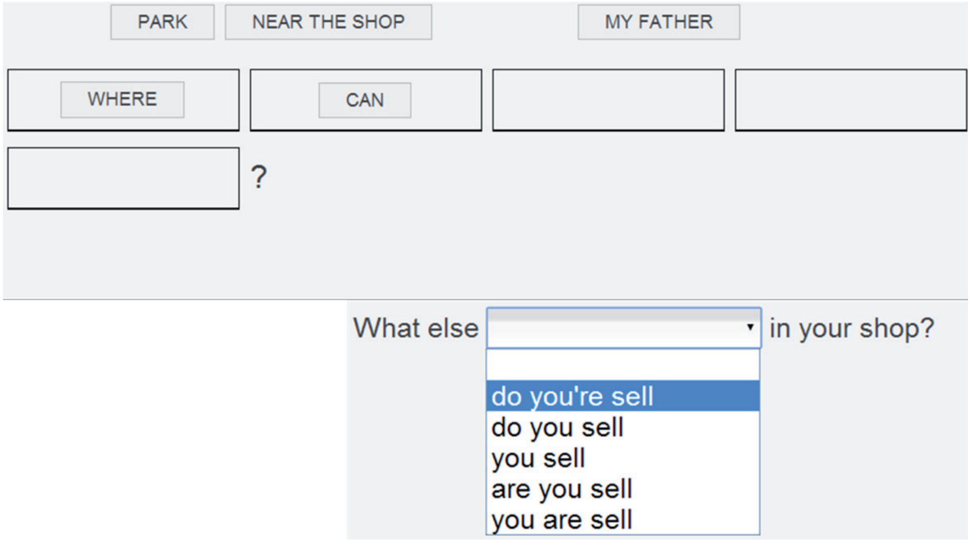


Figure 2. Intervention exercises: example items

In total, there were five exercises designed for the intervention, seven items in each (Appendix 2).

The presentation of the items and the feedback to the learners was designed in the following way, similar for all the intervention exercises in both groups:

1. an item was presented to a learner;
2. following the learner's response, feedback was displayed to him/her;
3. the learner was then presented with the next item, which had the same structure as the previous item.

There was, thus, a difference between the adaptive feedback (mediation) used in the present study (see Table 3) and the way mediation is commonly provided in dynamic assessment, i.e., learners go back to the same item until they are able to self-correct or are provided with the correct answer. The reason for doing so was primarily to make the learners realise that the pattern of stage 5 questions is the same/similar with different question words, lexical verbs, and auxiliaries.

The experimental group feedback was designed to follow the implicit-to-explicit adaptation similar to Aljaafreh and Lantolf's (1994) Regulatory Scale and looked as follows, the numbers indicating the levels of the feedback progression from implicit “think more carefully” to explicit explanation and overt correction (Table 2):

Table 2. Adaptive corrective feedback in the study

Level	Description	Example
0.	An indication that the response is correct	Your sentence: When does he come to work? Correct!
1.	An implicit hint that there might be something wrong with the answer	Your sentence: When did it appeared in your shop? Think more carefully. Try to complete the next question—it will be similar to this one.
2.	The location of the error is narrowed down	*Your sentence:* How long **does it sleeps** in the shop? Look at the highlighted part of your sentence. Think, is everything correct there? The following question will be similar to this one.
3.	The location of the error is further narrowed down, the nature of the error is identified, and metalinguistic clues or elicitations are provided	Your sentence: How often **do** you’**re** clean the shop? You used the correct helping word **do**. But do we need the verb **are** here? The following question will be similar to this one.
4.	Examples of the correct structure are given	Your sentence: How many times must **eat the puppy** every day? Not quite right. Look at the following examples: How are they different from your sentence? How **could** you **do** that? What **might** you **answer** him? Where **could** he **go**? The following question will be similar to this one.
5.	The correct response is provided with the explicit indication of what was wrong	Your sentence: When **you’re took** the picture of the puppy? Sorry, you need **did** before the word **you**; the verb **are** is not needed; and you had to use **take** instead of **took**. The correct answer is:

For the control group, the simple knowledge of response feedback was designed, i.e., the indication of whether their performance on the items was correct or not.

To administer the exercises, a web-based system called ICAnDoiT (Interactive Computer-Adaptive Diagnostic and Tutoring system) was designed. It served as a tool providing learners with instantaneous corrective feedback gradually attuning to their abilities. Additionally, it allowed for recording of the learners’ performance, including the mistakes they made, the feedback they received, etc. The ICAnDoiT system was created as a part of my on-going Ph.D. research and is currently hosted at https://solki4.cals.jyu.fi/icandoit/htdocs/. The usefulness of the system is that it allows language tests to be compiled using a variety of predefined task types with the possibility of adding feedback (both dynamic and static) to learners' item-by-item performance. In the following, the current state of validation process of the system will be outlined. A full account of the validation process will be given in a future report.

The exercises were piloted among 19 L1 Finnish learners of English (grade 8, average 14 years of age) in December 2010. The aim of the pilot study to establish the validity of the procedure. Additionally, the questionnaire used in the present study was piloted. As the major aim of the pilot study was to pilot the exercises, the study did not include the posttest and no control group was assembled.

The piloting resulted in a number of changes, such as modification/addition of several items in the pre-/posttest exercises. The feedback messages were also slightly modified to stress the similarity between the items. The pilot study also confirmed that the exercises elicited the production of wh-questions.

To reinforce the usability of the system, the system interface was designed according to the blueprint provided by Fulcher (2003). This was followed by a three-phase usability check, which used questionnaire replies, think-aloud protocols, and interviews as data. All in all, the usability study allowed for eliminating several usability problems, such as the difficulty to understand the mechanics of the ordering exercises.

A more comprehensive account of the piloting will be given in a future paper.

3.4 Participants and Data

The participants in this study were L1 Russian learners of English, average 14 years of age, studying at grade 8 in a school in Estonia (n = 64). The learners were from six different groups taught by two teachers. Each learner was randomly assigned to either the experimental (n = 35) or the control (n = 29) group.

However, the reported numbers refer to those who completed the intervention exercises. Since some learners were missing during the pretest, others during the posttest, and some cheated (as observed by either me or the teachers monitoring their performance), there were fewer learners whose performance on the exercises was analysed—26 and 21 learners respectively. As regards cheating, it was an extraneous variable that could introduce construct-irrelevant variance. Therefore, I decided to remove the performance of the learners who cheated from the analyses.

In Estonia (and in Finland), learners' first foreign language proficiency is expected to be at level B1.2 by the end of grade 9 (the end of lower-secondary school). Judging by the descriptors (Põhikooli riiklik õppekava õigusakt: Lisa 1 2010 [Basic School National Curriculum Act: Annex 1]), by the end of grade nine, learners are expected to ask wh-questions (e.g., when asking for directions). This reinforced the possibility that wh-questions should be within some of the participants' ZPD. Moreover, before the intervention, I asked the teachers whether by the time of the study, the learners had been taught to form questions in English (including wh-questions with auxiliaries), which they confirmed.

Judging by the teachers' reports and the state curriculum, I assumed that these questions were at least in some of the learners' ZPD.

The data come from the learners' performance on the exercises they took in the ICAnDoiT system. Additionally, the learners completed an online questionnaire (Appendix 3) which aimed at discovering their experiences with the feedback during the intervention. The questionnaire was conducted in the learners' mother tongue.

3.5 Procedure and Scoring

Before the pretest, it was explained to the learners that they were to complete several exercises so that they could see how well they were able to form questions in English. The learners were also advised to consult the help menu or ask for help from the persons monitoring their performance if they did not know any of the words in the exercises. They were given help only on vocabulary, not grammar. To save time, in the first exercise of the pretest, the learners were instructed to write the questions only. All the learners were working on the same exercises. The only difference was in the feedback the two groups received during the intervention.

The following sample from a learner's performance log demonstrates how the feedback incrementally adapted to the experimental group learners' abilities, gradually becoming more explicit and detailed (Table 3). The feedback this learner and the others received was originally in their L1 and was translated into English for the present article.

The log illustrates that the learner had a certain consistent error in wh-questions with the auxiliary *does*. After being provided with metalinguistic clues regarding the nature of his mistake, he was able to select the correct option and use it consistently until the end of the exercise.

After completing the intervention exercises, the learners were asked to fill in the questionnaire. Two academic periods were allocated for this part of the study. The procedure was not speeded, as all the learners managed in less time than that.

The posttest was conducted a week after the intervention and was the same as the pretest. One academic period was allocated for the posttest, but the learners completed the exercises in less time than that. Those who were missing during the pretest and the intervention were asked to complete the intervention exercises and the questionnaire while the others were working on the posttest.

Two scoring schemes were used to score the learners' responses to the pre-/posttest items. In the first scoring scheme, I gave each correctly formed stage 5 question one point. In the second, I awarded each correct stage 5 question two points and each correct stage 4 question, one point. It was decided to use a composite score (i.e., the sum of the learners' scores on the two pretest/posttest exercises) in the statistical analyses for the following reasons:

- both exercises assessed the same construct;
- this allowed for estimating the learners' abilities more precisely.

Table 3. A learner's performance on the *does*-exercise

Selected option	Feedback
Where **do it plays** in the shop?	Your sentence: Where do it plays in the shop? Think more carefully! Try to complete the next question—it will be similar to this one.
Why **do it looks** sad in the photo?	Your sentence: Why **do it looks** sad in the photo? Look at the highlighted part of your question. Think, is everything correct there? The following question will be similar to this one.
When **do it goes** to sleep?	Your sentence: When **do** it **goes** to sleep? You used the correct helping verb. But think about the word **it**. What do you have to add to the helping verb **do**? What should happen to the verb **goes**? The following question will be similar to this one.
When **does it close** on holidays?	Correct!
What **does it like** to eat?	Correct!
When **does he come** to work?	Correct!
How long **does it sleep** at night?	Correct!

4 Results

The exercises were designed based on the performance of the Finnish learners of English and were also piloted among them. Thus, ensuring the comparability of the pilot study group with the present study participants was necessary for reinforcing the construct validity of the exercises for the present study group.

For comparing the present study and the pilot study participants' performance, an independent samples t-test was conducted on the square-root transformed variable (percent correct on the two pretest exercises). It demonstrated that the performance of the present study participants (*M* = 4.01, *SD* = 2.85, *n* = 47) was not statistically different from the pilot study participants' performance (*M* = 3.77, *SD* = 2.34, *n* = 19), *t*(64) = 0.32, *p* =.748. Moreover, the present study learners made similar mistakes as the Finnish learners had made in the exercises, so the designed exercises (including the distractors in the multiple-choice exercises) and the feedback addressed their problems equally well.

This was followed by a modern item analysis of the present study participants' pretest performance conducted using *Winsteps* Rasch analysis software. It showed that there were no outfitting items in both the scoring that only took into account stage 5 questions (0.59 ≤ infit MNSQ ≤ 1.4) and the partial credit scoring (0.55 ≤ infit MNSQ ≤ 1.36). The person separation statistics of the two variables were 1.4 (Cronbach's alpha .84) and 1.49 (Cronbach's alpha .86) respectively, which is satisfactory (e.g., Fisher 2007). In other words, taken together, the pretest exercises could distinguish between high (or rather middle) and low performers.

Most of the following statistical analyses were conducted using IBM SPSS software. The results are presented in two sections, the first comparing the performance of the two groups and the second, the experiences of the two groups with the feedback in the study. Exact statistics will be provided whenever possible.

4.1 The effect of the adaptive feedback as contrasted with the knowledge of response feedback

To establish whether the adaptive feedback had any effect on the learners' ability to produce stage 5 questions, the differences in the learners' scores on the pretest and the posttest were compared. The descriptive statistics for the pretest and the posttest scores are reported in Table 4 and illustrated in Figure 3 below.

Table 4. Learners' pre-/post-test performance: descriptive statistics

	Pre-test			Post-test		
Groups	**Mean**	**SD**	**Median**	**Mean**	**SD**	**Median**
Experimental (*n*=26)	4	4.14	2.5	5.35	4.17	4.5
Experimental, partial credit (*n*=26)	8.81	8.93	5.5	11.92	8.87	10
Control (*n*=21)	4.19	3.26	4	3.9	2.81	3
Control, partial credit (*n*=21)	9.38	7	9	9	8	6

However, before studying the changes in the learners' performance after the treatment, I decided to reinforce the condition that the two groups' ability to form L2 English questions did not differ significantly before the treatment. Due to the verisimilitude of the figures obtained on the two scoring schemes, only the results on the partial credit scores are reported.

As the dependent variable was not normally distributed, a Mann-Whitney U test was used. It demonstrated that the experimental group learners (*Mdn* = 5.5) did not perform significantly differently from the control group learners (*Mdn* = 9), *Z* = -.59, *p* = .561. To corroborate the finding, a differential item functioning analysis was conducted. It confirmed that the learners in both groups performed similarly on all of the items, the highest *Welch's t* value being for item 11 (the second exercise), *t*(33) = 1.49, *p* = .15.

To establish whether the difference in performance between the two groups was statistically significant, I conducted an independent-samples t-test on the gain scores variables (the difference between the posttest and the pretest scores), which were normally distributed (e.g., for the partial credit scoring, *W*(26) = .973, *p* = .696 for the experimental group and *W*(21) = .953, p = .386 for the control group).

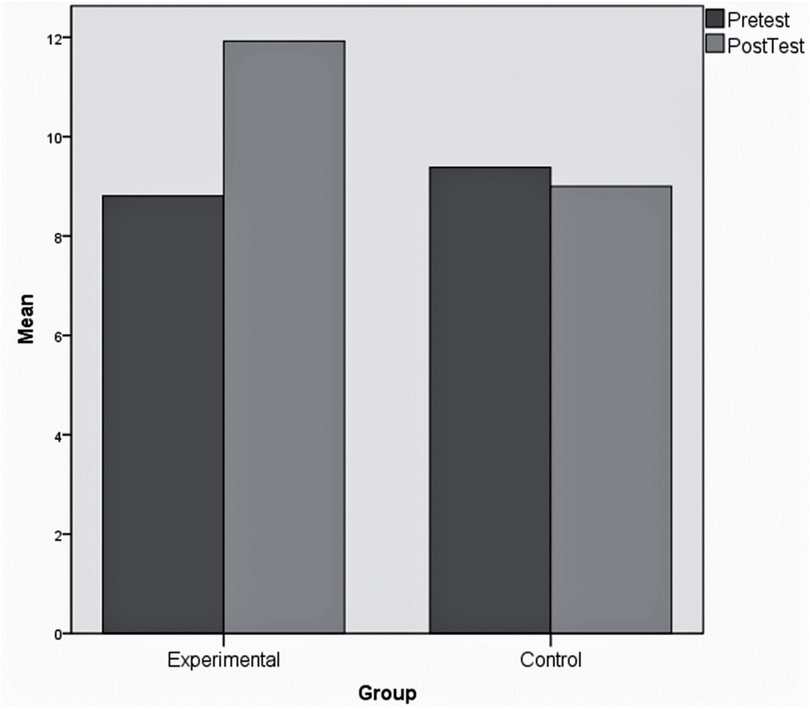


Figure 3. Learners' performance on the pretest and the posttest (partial credit scoring)

The t-test demonstrated that the experimental group (*M* = 1.35, *SD* = 2.64) increased the number of correctly formed wh-questions with auxiliaries significantly more than the control group (*M* = -.28, *SD* = 2.47), *t*(45) = 2.17, *p* = .035), where there was, in fact, a small decrease. The mean difference of 1.63 scale units, 95% CI [0.12, 3.15] indicated a moderate effect size, *d* = 0.64. The difference between the experimental group (*M* = 3.12, *SD* = 5.48) and the control group (*M* = -.38, *SD* = 5.95) was slightly less pronounced for the partial credit scoring scheme, *t*(45) = 2.09, *p* = .042. The mean difference of 3.50 scale units, 95% CI [.13, 6.86], indicated a moderate effect as well, *d* = 0.61[2].

However, the significant difference between the two groups does not imply that the knowledge of response feedback was useless for all the control group learners. The following Table 5 illustrates the opposite.

It can be deduced from the learner's performance log that he was not sure at first about the word order in the wh-questions with modal auxiliaries. By the end of the exercise though, he was able to use the correct word order more consistently. Moreover, during the postest, he formed all questions with the model auxiliary correctly as compared to half of the questions of the same kind formed correctly during the pretest.

The experimental group learner's performance log presented in Table 3 reveals a somewhat different picture in that he appeared to be more other-regulated that the control group learner whose performance is illustrated in Table 5. Interestingly, during the pretest, the experimental group learner was not able to form any stage 5 questions with *does*. During the post-test, however, he formulated two wh-questions with *does* correctly. I will elaborate on these two examples in the **discussion** section.

Table 5. A control group learner's performance on the first ordering exercise

Selected option	Feedback
WHERE / MUST / IT / SLEEP / IN THE FLAT?	Correct!
CAN / I / TEACH / MY DOG? / WHAT TRICKS	Your question is formed incorrectly.
WHY / MUST / YOU / CLOSE / THE SHOP EARLY?	Correct!
CAN / I / BUY / WHAT TOYS / FOR MY PET?	Your question is formed incorrectly.
WHEN / CAN / I / GET / MORE PHOTOS?	Correct!
WHERE / CAN / I / LEARN / MORE ABOUT DOGS?	Correct!
HOW LONG / CAN / IT / STAY / ALONE?	Correct!

What is interesting to note about the learners' performance is that the intervention seems to have also resulted in a somewhat higher number of level 4 questions (both correct and incorrect) produced by the learners. A closer look at the learners' performance reveals a rather interesting trend. Three out of four experimental group learners who failed to produce any questions higher than stage 3 (e.g., **What animals shop sells?)* during the pretest produced at least one stage 4 question (e.g., *Where's the shop located?*) during the posttest. One of those three also managed to produce three stage 5 questions. The fourth learner produced four correct stage 5 questions but no stage 4 questions during the posttest. It is hard to say to what extent knowledge of response feedback can facilitate the same development, as there was only one control group learner who produced one question at stage 4 and one at stage 5 during the posttest while having failed to produce any questions at these stages during the pretest. Not much can be said about the same trend in formulating stage 5 questions, as three experimental group learners out of six who failed to form any stage 5 questions during the pretest formed at least one (either correct or incorrect or both) during the posttest and two out of three control group learners were able to do the same.

4.2 Learners' Self-Reports

Twenty-eight experimental group and twenty-three control group learners completed the questionnaire. To compare the two groups' self-reports, their responses to one Likert-scale and two dichotomously scored items were analysed (see Appendix 3). The Likert-scale item asked the learners to rate the extent to which the feedback helped them to find the correct answers during the intervention. The first dichotomous item asked them whether they had learned anything having completed the intervention exercises. The second dichotomous item asked them whether the feedback had helped them to learn it.

A Mann-Whitney U test demonstrated that the experimental group (*Mdn* = 3.5) did not rate the usefulness of the feedback for completing the intervention exercises differently from the control group (*Mdn* = 3), *Z* = -0.59, *p* = .963. Moreover, although a higher proportion of the experimental group learners (64%) thought that they had learned something compared with the control group (48%), the difference was not statistically significant either, as demonstrated by a Chi-square test, $X^2(1, n = 51) = 1.40$, $p = .238$.

On the other hand, 14 learners from the experimental group (50%) answered positively when asked whether it was the feedback that had helped them to learn something, whereas only five learners from the control group (about 21%) were of the same opinion. A Chi-square test indicated that the difference was statistically significant, $X^2(1, n = 51) = 4.31$, $p = .038$, $\varphi = -.29$.

To interpret these results, I also looked at the learners' responses to the open-ended questions in the questionnaire. The qualitative analysis of the responses revealed some recurring patterns exemplified in Table 6.

Table 6. Learners' reported reasons for the feedback usefulness

Experimental	Control
• They showed me when I can use *can*.	• It showed that the answer was incorrect.
• They helped me by giving examples.	• I realised I was doing right and continued.
• They hinted that the word was in the wrong place.	• I don't know.
• I didn't remember the rule, and the feedback helped me to.	• I DON'T KNOW.
• Everything was explained: why the sentence was incorrect and how to correct it.	• They were of no use. I often didn't even look at them.
• Because I understood my mistake.	

5 Discussion

One of the aims of the present study was to determine whether the adaptive corrective feedback was more likely to facilitate learning than the try-again feedback (provided irrespective of the learners' ZPD). The findings demonstrate that the feedback adapting to the learners' abilities resulted in a significant increase in the number of correctly formed wh-questions with auxiliaries in the experimental group as compared with the control group who received the static implicit feedback group (where, in fact, there was a small decrease). There was at least a short term moderate positive effect of the adaptive feedback. The findings, therefore, confirm the hypothesis that adaptive feedback provided automatically can facilitate learning. This adds to the findings of the earlier studies regarding the influence of corrective feedback negotiated within learners' ZPD.

The analysis of the performance of those learners who failed to produce any stage 4 and stage 5 questions during the pretest revealed that after the intervention, stage 4 questions emerged in their performance. Certainly, from the point of view of Processability Theory, the emergence of stage 4, and not stage 5, questions indicated that some learners were simply not ready to advance to the latter higher level of question development. The intervention, however, was not designed to facilitate the development of stage 4 questions. Therefore, this issue deserves further examination, the more so as this part of the analysis looked at a very limited number of cases.

There is also some indication that in the control group, the learners' improvement in many cases might have to do with the increase in accuracy rather than the emergence of the correct structure(s) in their unassisted performance. Stage 4 questions in the learners' performance, which I discussed in the previous paragraph, can serve as an example of that. Another example can be the qualitative difference between the pre-/posttest performances of the two learners whose treatment performance is described in Tables 3 and 5. The evidence for that, however, is rather inconsistent. A future study can explore this possibility.

Importantly, the figures also reveal that implicit feedback, favoured by some teachers according to Amrhein and Nassaji (2010), is not always facilitative for learning. The implicit feedback not being helpful could be explained by the finding of Nassaji and Swain (2000), who discovered that the learner given the feedback irrespective of her ZPD was more likely to benefit from more explicit feedback. Thus, it would be interesting to conduct a similar study where the control group received explicit feedback (e.g., explicit correction and/or explicit explanation of the error) to compare the effect of adaptive feedback with that of explicit corrective feedback.

Alternatively, it could have been the learners' preferences for different feedback types that resulted in a higher acceptance of the adaptive feedback. This could have added to the facilitative effect of the adaptive feedback in the experimental group and hindered the usefulness of the feedback in the control group. It is also worth noting that the control group learners, even when considering the feedback helpful, were often unsure of the reason(s) for that. Thus, it seems that feedback adapted to learners' abilities might be accepted more readily than static implicit feedback.

On the other hand, the results demonstrated that the experimental group learners did not consider the feedback any more useful for getting their answers right during the treatment than the control group (probably because it did not give away the correct answers in most of the cases). Thus, it seems that learners do indeed attach different meanings to the word *usefulness*. More importantly, there is a possibility that the learners' perceived usefulness of the feedback could have negatively influenced the utility of certain feedback types which otherwise matched their abilities. That is to say, some learners skipped the feedback messages they *considered* useless and not because those feedback messages did not match their abilities. However, the data in the present study do not allow for drawing any conclusions in this regard. This would also be an interesting question to address in a further study.

The above interpretation does not mean that teachers should avoid giving implicit feedback to their learners—doing so would deprive learners of an important step on their way of becoming self-regulated in the use of a second/foreign language. On the contrary, the performance of some learners (including some of the control group learners) demonstrated that they did not need explicit and detailed feedback to self-correct during the treatment and increase their scores on the posttest exercises. Amrhein and Nassaji (2010) rightfully note that learner autonomy is one goal of pedagogy, and by preferring explicit correction, learners may unnecessarily place the responsibility of correcting their mistakes onto teachers, which contradicts this goal. Rather, from the perspective of a sociocultural theory of learning, the findings should be interpreted so that adapting the feedback to the learners’ ZPD was beneficial to a larger number of learners than providing the static feedback that disregarded the learners' ZPD.

6 Conclusion

The present study aimed at finding out whether adaptive corrective feedback had a facilitative effect on learning (in this case, L2 English questions), and whether this effect was significantly different from that of the knowledge of response feedback. Additionally, it compared the self-reports of the two groups of learners on the perceived usefulness of the feedback.

The study demonstrated that the learners who had received adaptive corrective feedback during the intervention produced significantly more correctly formed L2 English wh-questions with auxiliaries than the control group. The learners also tended to accept the adaptive feedback as useful for learning more readily than the knowledge of response feedback. The latter, however, might have also derived from the learners’ preference for more explicit feedback types as the previous research suggests.

The findings of the study have several implications. Adaptive corrective feedback provided to learners while they practice on a second/foreign language should allow them to self-diagnose their problems as well as to learn something. The finding that the adaptive feedback helped the learners to become aware of their mistakes and produce more correct responses during the posttest suggests that a similar procedure has implications for teaching. Learner profiles, similar to the one presented in the study (Table 3), would allow teachers to see the typical mistakes their learners make but also help them with the difficult task of finding out how much help their learners currently need with certain mistakes. Additionally, as I have suggested at the beginning of the paper, teachers would be able to see whether the required structure is within (most of) their learners’ ZPD or more teaching is required.

Amrhein and Nassaji (2010) suggest that teachers should change their learners’ feedback preferences if these preferences are not beneficial for their learners. One way the assessment/tutoring system used in this study, or a similar one, could help teachers achieve this goal is that they could discuss the performance profiles with their learners, so that the latter would see how implicit feedback had helped them. What is more, the experience of automated adaptive feedback might influence learners’ beliefs about the efficacy of different feedback types without teachers having to follow it up with discussions, which would save teachers time and effort. Whether this experience alone or followed with discussions could change learners’ preferences of corrective feedback seems to be an interesting topic to explore.

There are, however, several limitations to the study that might affect the generalizability of its results. Despite the decent overall number of participants, the fact that not everyone completed the pretest, the posttest, and the questionnaire resulted in a smaller number of cases in the analyses and might have affected the findings. Moreover, the pretest and the posttest contained only two exercises (17 items it total). Finally, due to the school schedule, a delayed posttest could not be conducted. Therefore, it is impossible to tell whether the adaptive feedback led to a long-lasting learning effect. At the same time, the posttest was conducted a week after the intervention, so the learning effect lasted for at least a week.

A similar study with a larger number of participants, more exercises/items in the pre-/posttest as well as with a delayed posttest could reinforce the findings of the present study. Additionally, further studies comparing adaptive corrective feedback with other types of corrective feedback, such as explicit correction or random feedback, should allow for creating a more comprehensive picture demonstrating whether corrective feedback negotiated within learners' ZPD is indeed superior to static corrective feedback. The no-feedback condition for the control group might also be used to address Truscott's (1996) claim about the negative effect of corrective feedback.

Nevertheless, despite the limitations of the study, it is hoped that it has provided useful insights into the applications of adaptive corrective feedback (that is to say, mediation) its effect on learning, and its usefulness as perceived by learners. I also hope that the study stimulates research on the effect of corrective feedback as seen from a sociocultural perspective. Collecting more experimental data would enable meta-analyses of the effectiveness of adaptive corrective feedback, thus strengthening the argument for its usefulness.

Endnotes

1. There are apparent epistemological differences between the paradigms underlying the concept of universal developmental stages and the sociocultural perspective on development. Specifically, while the former presupposes a uniform order of acquisition and, consequently, that instruction can only be effective when learners are developmentally ready to advance, according to the latter it is instruction that directs the development to follow, and there are, in effect, no prescribed developmental stages (e.g., Leung 2007). Resolving these differences, however, is beyond the scope of the present paper.
2. The shape of the distribution in the control group was slightly not symmetric. Thus, I supplemented the analysis with a Mann-Whitney U test, which showed that the difference between the gain scores on the stage 5 questions only scoring was statistically significant, $Z = 2.04$, $p = .040$, $r = .30$. That is to say, it confirmed the result obtained on the t-test as far as the stage 5 questions only (which were the target of the intervention) were considered. The difference in the gain scores obtained on the partial credit scoring was not significant, $Z = -1.85$, $p = .06$. What is more, Wilcoxon signed-rank tests demonstrated that the improvement after the treatment was significant in the experimental group, e.g., for the partial credit scoring, $Z = -2.65$, $p = .007$, r = .37, but not in the control group, $Z = -.26$, $p = .805$.

References

Ableeva, R. 2010. *Dynamic assessment of listening comprehension in second language learning.* Unpublished doctoral dissertation, Pennsylvania State University, University Park, PA. [Retrieved June 12, 2013]. Available at http://etda.libraries.psu.edu/paper/11063/6495.

Ableeva, R. 2012. *Transfer Tasks: Diagnosing Second Language Development*. Paper presented at AAAL 2012 Annual Conference, Boston, MA, USA.

Alanen, R., & P. Kalaja 2010. *The emergence of L2 English questions across CEFR proficiency levels*. Paper presented at AAAL 2010, Atlanta, Georgia, USA.

Aljaafreh, A. & J. P. Lantolf 1994. Negative feedback as regulation and second language learning in the Zone of Proximal Development. *The Modern Language Journal*, *78*(4), 465–483.

Amrhein, H. R. & H. Nassaji 2010. Written corrective feedback: What do students and teachers think is right and why? *Canadian Journal of Applied Linguistics*, *13*(2), 95–127.

Ashwell, T. 2000. Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing*, *9*(3), 227-257.

Bachman, L. F., & A. S. Palmer 1996. *Language testing in practice*. Oxford: Oxford University Press.

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

Carroll, S. & M. Swain 1993. Explicit and implicit negative feedback. *Studies in Second Language Acquisition, 15*(3), 357-386.

Ellis, R. 2009. A typology of written corrective feedback types. *ELT Journal*, *63*(2), 97–107.

Ellis, R., S. Loewen, & R. Erlam 2006. Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition, 28*(2), 339-368.

Ferris, D. R. 1995. Student Reactions to Teacher Response in Multiple-Draft Composition Classrooms. *TESOL Quarterly, 29*(1), 33-53.

Fisher, W. P., Jr. (2007) Rating Scale Instrument Quality Criteria. *Rasch Measurement Transactions*, *21*(1), 1095.

Fulcher, G. 2003. Interface design in computer-based language testing. *Language Testing, 20*(4), 384-408.

Guthke, J. & J. F. Beckmann 2000. The learning test concept and its application in practice. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 17–69). Amsterdam: JAI/Elsevier.

Hyland, F. 2001. Providing Effective Support: Investigating Feedback to Distance Language Learners. *Open Learning: The Journal of Open, Distance and E-Learning*, *16*(3), 233-247.

Iwashita, N. 2003. Negative Feedback and Positive Evidence in Task-Based Interaction: Differential Effects on L2 Development. *Studies in Second Language Acquisition, 25*(1), 1-36.

Kang, H-S. 2009. The relative efficacy of explicit and implicit feedback in the learning of a less-commonly-taught foreign language. *IRAL - International Review of Applied Linguistics in Language Teaching*, *47*(3-4), 303-324.

Leki, I. 1991. The preferences of ESL students for error correction in college-level writing classes. *Foreign Language Annals*, *24*(3), 203–218.

Leung, C. 2007. Dynamic assessment: Assessment for and as teaching? *Language Assessment Quarterly*, *4*(3), 257-278.

McDonough, K. (2005). Identifying the impact of negative feedback and learners' responses on ESL question development. *Studies in Second Language Acquisition, 27*(1), 79-103.

Nassaji, H. 2009. Effects of recasts and elicitations in dyadic interaction and the role of feedback explicitness. *Language Learning*, *59*(2), 411-452.

Nassaji, H. & M. Swain 2000. A Vygotskian perspective on corrective feedback in L2: The effect of random versus negotiated help on the learning of English articles. *Language Awareness*, *9*(1), 34–51.

Pica, T. 1994. Questions from the language classroom: Research perspectives. *TESOL Quarterly, 28*(1), 49-79.

Pienemann, M. 2005. An introduction to Processability Theory. In M. Pienemann (Ed.), *Cross-linguistic aspects of Processability Theory*. Amsterdam: John Benjamins Publishing Company. 1–60.

Poehner, M. E. 2005. *Dynamic assessment of oral proficiency among advanced L2 learners of French*. Unpublished doctoral dissertation, Pennsylvania State University, University Park, PA.

Poehner, M. E. 2008. *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. Berlin: Springer.

Põhikooli riiklik õppekava õigusakt: Lisa 1. 2010. Pub. L. No. RT I 2010, 6, 22. [Retrieved May 27, 2014]. Available at https://www.riigiteataja.ee/aktilisa/1281/2201/0017/13275423.pdf.

Schulz, R. A. 2001. Cultural differences in student and teacher perceptions concerning the role of grammar instruction and corrective feedback: USA-Colombia. *The Modern Language Journal*, *85*(2), 244–258.

Spada, N. & P. M. Lightbown 1999. Instruction, first language influence, and developmental readiness in second language acquisition. *The Modern Language Journal*, *83*(1), 1–22.

Teo, A. 2012. Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning & Technology*, *16*(3), 10-20. [Retrieved March 10, 2013]. Available at http://www.llt.msu.edu/issues/october2012/action.pdf.

Truscott, J. 1996. The case against grammar correction in L2 writing classes. *Language Learning*, *46*(2), 327-369.

Truscott, J. 1999. The case for "The case against grammar correction in L2 writing classes": A response to Ferris. *Journal of Second Language Writing*, *8*(2), 111–122.

Vasilyeva, E., S. Puuronen, M. Pechenizkiy, & P. Rasanen 2007. Feedback adaptation in Web-based learning systems. *International Journal of Continuing Engineering Education and Life-Long Learning*, *17*(4-5), 337–357.

Vygotsky, L. S. 1978. *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

White, L., N. Spada, P. M. Lightbown, & L. Ranta 1991. Input enhancement and L2 question formation. *Applied Linguistics, 12*(4), 416-432.

Appendix 1. The pretest/posttest exercises (the prompts translated into English)

Exercise 1
You are interested in:

1) location of the shop
2) opening hours
3) what pets they sell
4) how much the pets cost
5) where you can find the pets' photos
6) what other information about the pets the shop can send you
7) how they got your E-mail address
8) what the name of the shop means

Exercise 2*

1) What parrots ______ the shop ______? (to sell)
2) ______ you also ______ talking parrots? (to have)
3) When ______ you ______ selling parrots? (to begin / to start—the sentence is in the past tense)
4) How long ______ the parrots ______ ? (to live)
5) When ______ they ______ to talk? (to learn)
6) How fast ______ a parrot ______ ? (can fly)
7) How much ______ it ______ every day? (to eat)
8) What words ______ they ______ ? (can say)
9) Where ______ the shop ______ the parrots from? (to buy—the sentence is in the past tense)

* The task was preceded by the instructions where the learner was asked to imagine that his/her grandfather wanted to buy a parrot and asked the learner to forward his questions to the pet shop.

Appendix 2. The intervention exercises

Task 1*

1. WHEN/CAN/I/GET/MORE PHOTOS?
2. HOW LONG/CAN/IT/STAY/ALONE?
3. WHERE/MUST/IT/SLEEP/IN THE FLAT?
4. WHAT TOYS/CAN/I/BUY/FOR MY PET?
5. WHERE/CAN/I/LEARN/MORE ABOUT DOGS?
6. WHY/MUST/YOU/CLOSE/THE SHOP EARLY?
7. WHAT TRICKS/CAN/I/TEACH/MY DOG?

Task 2*

1. WHAT ELSE/MUST/MY FAMILY/KNOW/ABOUT DOGS?
2. WHERE/CAN/MY FATHER/PARK/NEAR THE SHOP?
3. WHAT/CAN/THE PUPPY/DO/IN MY FLAT?
4. HOW/CAN/MY GRANDPA/TEACH/PARROTS TO TALK?
5. HOW MANY TIMES/MUST/THE PUPPY/EAT/EVERY DAY?
6. WHEN/CAN/PUPPIES/GO/OUTSIDE?
7. WHY/MUST/PARROTS/LIVE/IN A CAGE?

Task 3**

1. How often [do you clean] the shop?
   a. do you clean
   b. do you're clean
   c. you are clean
   d. are you clean
   e. you clean
2. What else [do you sell] in your shop?
3. How [do I choose] the dog food?
4. What [do you feed] the puppies?
5. When [do I take] the puppy to the doctor?
6. Why [do you leave] the pets alone at night?
7. How often [do I wash] my puppy?

Task 4**

1. When [does it close] on holidays?
   a. does it close
   b. do it closes
   c. does it closes
   d. do it close
   e. it closes
2. How long [does it sleep] at night?
3. What [does it like] to eat?
4. When [does it go] to sleep?
5. When [does he come] to work?
6. Where [does it play] in the shop?
7. Why [does it look] sad in the photo?

Task 5**

1. Why [did I get] only one E-mail?
   a. did I get
   b. I was get
   c. did I'm get
   d. I'm got
   e. did I got
2. How [did you find] my E-mail address?
3. When [did you take] the picture of the puppy?
4. How many puppies [did you sell] last month?
5. Why [did he open] a pet shop?
6. When [did it appear] in your shop?
7. Where [did it live] before the pet shop?

*The order in which sentence parts, as separated with a "/", were displayed to the learners was randomised every time each item was retrieved from the item bank; the parts were never displayed in the correct order.

**The options, as presented for item 1, had the same structure in every item; the order of the options was randomised every time each item was retrieved from the item bank. The correct option is provided in the square brackets.

Appendix 3. Questionnaire items discussed in the study (English translation)

Please tell us how useful the hints were for you (how well they helped you to do the exercises). Choose only one option:

1. very useful (they helped me a lot)
2. quite useful (they helped me quite a lot)
3. not really useful but not useless either (they helped me a little)
4. quite useless (they did not help me much)
5. useless (they were of no help to me)

Did you learn anything after completing the exercises?

□yes □no

Please tell us what you learned:

Do you think the hints you received helped you to learn?

□yes □no

Please tell us how exactly the hints helped you to learn:

Received October 17, 2013
Revision received June 10, 2014
Accepted June 13, 2014

II

EXPLORING AND RESHAPING LEARNERS' BELIEFS ABOUT THE USEFULNESS OF CORRECTIVE FEEDBACK: A SOCIOCULTURAL PERSPECTIVE

by

Dmitri Leontjev

The article will be published in an upcoming issue of *ITL - International Journal of Applied Linguistics* in Volume 167 (2016) - a journal of John Benjamins Publishing Company, https://www.benjamins.com/#catalog/journals/itl/main

Exploring and Reshaping Learners’ Beliefs About the Usefulness of Corrective Feedback: A Sociocultural Perspective

A number of studies have shown that learners’ beliefs about the *usefulness of* corrective feedback for improving their L2 (a second of a foreign language) use influences the extent to which learners can utilize that same feedback. It seems, then, that changing some of these beliefs could benefit the L2 learning process. The present article reports on two small-scale studies, both drawing on a sociocultural perspective on the development of beliefs. Changes in learners’ beliefs about corrective feedback were observed both within a period of six months (Case study) and over the course of one research interview (Group study). The studies exemplify how the interplay of one’s own and other’s experience, others’ mediation, and authoritative voices facilitated these changes.

Keywords: learners’ beliefs; social interaction; dynamic assessment; feedback; sociocultural theory

INTRODUCTION

Teachers’ beliefs and practices can influence learners’ beliefs about learning a second or foreign language (L2) (Aro, 2009; Barcelos & Kalaja, 2013; Diab, 2005; Kern, 1995). However, teachers and learners do not necessarily share the same beliefs about learning an L2 (Barcelos & Kalaja, 2013; Brown, 2009; Kern, 1995), which can result in miscommunication, poor motivation, and non-participation in classroom activities (e.g., Barcelos & Kalaja, 2013; Kern, 1995). Similarly, a number of studies have demonstrated that learners and teachers can have different beliefs about corrective feedback (CF) in particular (Brown 2009; Diab, 2005; Hedgcock & Lefkowitz, 1994; Saito, 1994). For example, believing that it promotes learners' autonomy, teachers sometimes give implicit CF (i.e., not overtly stating that there is a mistake) (Amrhein & Nassaji, 2010; Yoshida, 2010). At the same time, learners, some studies suggest, consider explicit CF (e.g., overt correction or explicit explanation) more useful than implicit feedback (Amrhein & Nassaji, 2010; Ashwell, 2000; Leki, 1991), for example, because learners believe that teachers have a responsibility to correct their errors (Amrhein & Nassaji, 2010). Particularly as regards younger learners, the research findings on computerised CF, that is, feedback provided automatically to learners via computer, appear to be similar to the findings regarding CF in the classroom. Specifically, learners appear to consider implicit computerised CF rather useless for learning (e.g., Cornillie, Clarebout, & Desmet, 2012).

Learners' beliefs seem to be especially important when they interact with computerised feedback, as there is no educator to account for the way learners approach such feedback (e.g., skipping it because of not being able to understand it), adjusting the feedback accordingly. Heift (2002), for example, found that when given control, especially low-achieving learners extensively peeked at correct responses rather than read explanations. Pujolà (2001) found that learners who seemed to arrive at the correct responses by chance, generally, did not read the explanations of why their responses were correct. Thouësny (2011) studied the way learners approached CF in a computerised dynamic assessment where they could choose whether to access feedback or not. She found that learners either did not access or did not fully read about 47% of all the feedback provided to them.

Some studies have shown that learners are more likely to pay attention to corrective feedback when they believe it is useful; thus, feedback that learners perceive as useful could be more effective (Kern, 1995; Leki, 1991; Schulz, 2001). Thus, it has been suggested that teachers appraise their learners' perceived usefulness of corrective feedback or, better, work to change their beliefs when these are counterproductive (Amrhein & Nassaji, 2010; Brown, 2009; Hyland, 2003; Leki, 1991; Saito, 1994; Schulz, 2001). Several studies have investigated how teachers might do this.

Using questionnaire responses as data, Plonsky and Mills (2006) demonstrated that learners expressed significantly more belief in the usefulness of their teacher's feedback after he had explained his approach to correcting mistakes. Similarly, Sato (2013) demonstrated that training learners in providing CF to each other strengthened their belief in the usefulness of peer feedback.

However, there still appears to be no clear understanding of how these changes occur. Moreover, there are still relatively few longitudinal studies focusing on this issue. Finally, there seem to be no studies specifically tracing how exactly social interaction and experience lead to changes in learners' beliefs about the usefulness of CF. In the present study, I will define learners' *beliefs* as personal knowledge and assumptions (e.g., those of corrective feedback) which can either be relatively fixed and then used as means to mediate learners' actions, expectations, and strategies, or in a state of flux and constantly reshaped in social interaction and with new experience (Alanen, 2003; Aro, 2009; Barcelos, 2003; Dufva, 2003).

The two small-scale studies discussed in the present article were conducted to better understand the process through which learners' beliefs regarding the usefulness of corrective feedback develop and transform. I next present the research that motivated the two studies. I will argue that a sociocultural perspective on the development of beliefs is an appropriate theoretical framework for promoting our understanding of how learners' beliefs regarding CF develop. I will then present the two studies, discuss the findings, and sketch some directions for further research.

BACKGROUND

Researching Learners' Beliefs From a Sociocultural Perspective

Learners' beliefs have been studied from different perspectives, which can be broadly classified into *cognitivist* and *contextual*. In the following, I will give a brief account of the cognitivist *perspective* and discuss the *contextual* perspective in detail, the latter being the main theoretical grounding of my work.

Cognitivist (or *normative* and *metacognitive*) approaches to defining and studying learners' beliefs are informed, above all, by the Cartesian school of thought, which considers the human mind to be autonomous and almost unaffected by social phenomena (see Barcelos, 2003; Dufva, 2004). This view clearly influenced the studies by Horwitz (e.g., 1985) and Wenden (e.g., 1987), who used questionnaires to discover, for example, how learners' beliefs relate to their learning behaviour.

However, *contextual* approaches to studying beliefs (Barcelos, 2003), stating that beliefs are dynamic and are influenced by social factors have been gaining more prominence (Alanen, 2003; Aro, 2009; Barcelos, 2003; Barcelos & Kalaja, 2013; Dufva, 2003; Kalaja & Barcelos, 2013; Mercer, 2011). Judging by Barcelos' (2003) discussion, these approaches appeared as an alternative to cognitivist approaches, which aimed at generalisability in findings about beliefs, aiming instead at a deeper understanding of beliefs in contexts. In these approaches, beliefs are seen as dependent on and influenced by *contexts* that are socially constructed and dynamic. Within these contexts, beliefs emerge, transform, and in turn, construct these contexts (Barcelos, 2003). Among the contextual approaches, *dialogical* (e.g., Dufva, 2003) and *sociocultural* (e.g., Alanen, 2003) approaches are especially interesting, as they highlight both the individual and the social in the development of beliefs. Especially in the sociocultural approaches, the development of beliefs is perceived as a movement from other-regulation, when beliefs are in a state of flux and are constantly co-constructed and reconstructed in social interaction which mediates (i.e., guides) their development, to self-regulation, that is, appropriation of socially constructed knowledge to a private knowledge reservoir (Vygotsky, 1978). I next summarise two studies that used a combination of sociocultural and dialogical approaches to study learners' beliefs. These studies were used as models for the two studies reported on in the present article.

Building on the works of Bakhtin (e.g., 1981; 1986), Bråten (e.g., 1991a; 1991b), Cole (e.g., 1996), Kozulin (1998), Wertsch (e.g., 1991; 1998), and others, Alanen (2003) presented a neo-Vygotskian approach to the study of beliefs about L2 learning, which can be summarised as follows:

- beliefs are cultural artifacts that mediate human behaviour, constructed through social interaction;
- beliefs are experiential;
- significant others also shape learners' beliefs;
- beliefs are situational, that is, the context in which they emerge should be considered when studying them;
- the unit of analysis for the study of beliefs is *mediated action*, a system in which the relation between the subject and the object is mediated by a material or symbolic tool;
- dialogic speech is important for belief construction and is a type of mediated action;
- co-constructed beliefs may become a part of learners' knowledge through appropriation (one starts using an other's words to convey one's intentions);
- agency in utterances can thus be a sign of belief appropriation.

A term that needs elaboration is *mediated action*. An important premise of a sociocultural perspective on interaction is that not only every action, including dialogic speech, is mediated but also that agents and mediational means are interdependent (Wertsch, 1998). That is to say, to be able to fully understand how beliefs develop, instead of concentrating on separate elements presumably promoting the development of beliefs, these elements should be studied as a system. Following Wertsch (1991; 1998), Alanen

(2003) suggested that in mediated action that is dialogic speech it is important to consider not only what the interlocutors say, but also which (and whose) words they use, who uses these words, and in which order the words appear. One example from Alanen's study was how a learner's immediate repair of her own utterance after another learner responded differently illuminated the process of co-construction of beliefs in social interaction.

Using Bakhtinian dialogical and Vygotskian sociocultural frameworks, Aro (2009) studied the development of fifteen learners' beliefs about learning English. The premise for combining the two perspectives was that while dialogical perspective emphasises both the importance of the social and the individual in cognition and metacognition, it does not explicitly discuss the development of beliefs focusing rather on their nature. On the other hand, development is emphasised in the sociocultural perspective. Thus, the two perspectives complement each other.

Stressing the importance of appropriation, Aro discussed the results in terms of genres in the Bakhtinian sense (utterances typical for certain contexts) and in terms of polyphony, that is, a multitude of *voices* (the speaking consciousness: e.g., a child, a learner, a teacher) in learners' reflections. While some beliefs were appropriated early, remaining almost intact, she found that others changed with the learners' experience. It is worth mentioning that different learners' beliefs became more similar over time, suggesting the influence of authoritative voices, such as teachers'. She also found that the way the interviewer's questions were formulated invoked different beliefs—questions containing the second-person singular resulted in learners reflecting on their own experience, whereas questions about 'people' did not. In addition, teacher's voice was transparent in the learners' utterances, which contributed to the formation of their beliefs. Dialogic approaches to the study of beliefs emphasise, thus, that there are always others that learners have interacted with who have contributed to shaping their beliefs.

In fact, Alanen's (2003) study can also be considered a combination of these two perspectives. For example, Alanen observed how one interviewer told a learner that the learner could use English in Singapore, where her godmother lived. This resulted in that a year and a half later, during a research interview with another interviewer, this learner used the first interviewer's words when asked whether she would like to study English, saying *I would! Because my godmother lives in Singapore!* (Alanen 2003: 75).

In light of the above, it is important to note that the interviewer's/researcher's role in the contextual approaches is that of an active participant in the interaction, jointly creating the context with learners (Alanen, 2003; Dufva, 2003). That said, it should not be forgotten that beliefs are above all experiential. In the following section, I will suggest how experience of dynamic assessment, can contribute to a transformation of learners' beliefs about corrective feedback.

Dynamic Assessment

Dynamic assessment (DA) builds on the concept of the *Zone of Proximal Development* (ZPD), defined as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978: 76). Application of the ZPD concept to educational assessment (including assessment of L2) resulted in a shift away from the traditional assessment paradigm, which is often perceived to be in opposition to instruction, toward the view that assessment should facilitate learners' development by simultaneously assessing and promoting their abilities (e.g., Poehner, 2008). At the core of DA lies mediation (assistance provided within the learners' ZPD), which includes *adaptive corrective feedback*, that is, corrective feedback that adjusts dynamically to learners' performance (Vasilyeva et al., 2007).

The study of Aljaafreh and Lantolf (1994) serves as an illustration of dynamic assessment, showing how learners' development was facilitated by CF gradually adapted to match their abilities. Aljaafreh and Lantolf (1994) reported on an interactionist DA, where mediation depends on learners' reciprocity and emerges in interaction between the learner and the mediator, the latter adjusting the following mediational moves based on the learner's reaction. However, this gradual adaptation from implicit to explicit and detailed mediation is equally applicable to interventionist dynamic assessment, where feedback is standardised, consisting of a battery of predefined mediational moves, often in the form of CF arranged by its explicitness, which are provided to learners one by one until the learners are able to self-correct or are provided with the correct response (e.g., Poehner, 2008).

Computerised dynamic assessment is a relatively recent development in interventionist DA (Poehner, 2008). Thus, only a few dynamic tests of L2 exist. Poehner and Lantolf (2013), for example, demonstrated that computerised DA promoted learners' L2 Chinese and French listening and reading abilities. Teo (2012) went further and collected learners' reflection on the way DA mediated their performance, which suggested that learners realised that the adaptive CF helped them to find out which strategies helped them read between the lines and which were not useful. However, Teo did not collect the learners' reflections on the usefulness of different CF types they were given during the DA.

Nevertheless, it can be suggested that experience with dynamic assessment may allow learners to realise that they do not always need explicit CF for their learning to progress. That said, learners enter DA with their own beliefs, which can guide their DA performance, which can also mediate, that is, guide, their DA performance (e.g., Thouësny, 2011).

It should be noted that while experience is important in belief formation, it is no less important what is noticed in this experience and what mediates what is noticed (see Alanen, 2013). Thus, mediating learners' reflection of their DA experience to help them notice (and understand) the way CF helped them during the DA has the potential to transform their beliefs about corrective feedback. That said, to my knowledge this has not been addressed in previous research, which inspired the present study.

METHODOLOGY

In the present paper, I will address the following questions:

1) How are learners' beliefs about the usefulness of corrective feedback transformed by their experience of dynamic assessment?
2) How are learners' beliefs about the usefulness of corrective feedback co-constructed in social interaction?

The primary data come from a series of semi-structured interviews conducted (a) before and after human-mediated dynamic assessment sessions focusing on learning

English as a foreign language (Case study) and (b) after one session of computerised dynamic assessment (Group study). This constrained the social interaction analysed in the study to that happening during the research interviews. The first study was a Case study of one learner of English while the second one was a short-term study of a group of learners. To enhance confidence in the interpretation of the results, and overall, to provide a richer picture, the data from the dynamic assessment sessions were used to compare and contrast the data from the interviews, as will be detailed in the corresponding sections to follow. In the approach adopted in the study, the interviewer was considered to be an active participant in the interaction, his contribution being equally relevant for the construction of the learners' beliefs. In both studies, utterance (e.g., *I think these hints are very helpful.*) as a type of mediated action was used as the unit of analysis (cf. Wertsch, 1991; 1998). The aim of the studies was to find out what it was that mediated learners' actions in the context of utterance. For example, I noted cases of others' mediation similarly to the way Alanen (2003) did it (see the **Background** section). While acknowledging the inseparable relationship between agents and mediational means, in my analysis I chose to concentrate on the latter. The interviews were consequently transcribed and their structure analysed.

Following Aro's (2009) finding regarding the triggering of learners' self-beliefs, the interviewer used the second person (singular or plural) when addressing the participants. The changes in learners' beliefs will be traced by noting, for example, paralinguistic features, (e.g., changes in the learners' intonation or hesitation) or degree of agency (e.g., learners saying *'I think'* as contrasted with *'the learner should'*) in the learners' utterances, which were identified and interpreted by two people separately and later agreed upon, in addition to studying what was said by the learners (and the interviewer). Transcription markings are presented in **Appendix A**. I conducted the interviews myself and report the data separately for the two studies. I was not the teacher of any of the participants in the two studies. However, I conducted the dynamic assessment with the participant in the Case study.

I followed Alanen's (2003) suggestion to study the data chronologically, noting how what had been reported at earlier points of the interviews influenced the learners' later utterances. The following section will provide an overview of the Case study.

CASE STUDY

Participant

The participant in the Case study was an L1 Russian learner of English in grade ten (16 years old) at an upper-secondary school in Estonia (hereafter, referred to as M). By the time of the study, M had studied English for about seven years, both at school and in private language courses. M was chosen for the study since I assumed that his English proficiency would be around level B1.2 of the adapted Common European Framework of Reference for Languages (CEFR). This assumption is based on the fact that the Estonian State Curriculum specifies that learners' proficiency in English as the first foreign language should be around level B1.2 at the end of grade 9 (Põhikooli riiklik õppekava õigusakt: Lisa 1, 2010). Judging by previous studies (e.g. Nation, 2001), learners at lower-intermediate level of L2 proficiency (which is roughly equivalent to level B1 on the CEFR scale) are more likely from the instruction in word derivation, the target of the dynamic assessment sessions (see below), than at lower levels of L2 proficiency.

Data and Procedure

The data in the Case study come from (1) three human-mediated dynamic assessment sessions, comprising a set of exercises on word derivation mediated by the interviewer, and (2) three semi-structured interviews conducted one week before, one week after, and six months after the last DA sessions. The DA sessions were administered within a period of three weeks, with about one-week intervals.

The interview topics included:

- feedback from the school teacher of English and the learner's perceived usefulness of this feedback (the second interview did not include this topic);
- usefulness of corrective feedback of different degrees of explicitness;
- usefulness of adaptive corrective feedback in general;
- learning an L2.

Conducted in Russian, the interviews were translated into English for the present article (as also in the Group study). Printed sample feedback messages were presented to the learner as exhibits to reduce potential confusion between different feedback types.

There were the following types of exercises in the Case study:

- classification exercises, e.g., *which of these words are adverbs; what parts of speech are the rest of the words: momentary, literacy, ability, hyperactively*;

- suffix and prefix elicitation exercises, e.g., *on the basis of the word in the brackets, form a word that fits the sentence*: *They want to raise ………… (aware) of the problem*;

Adaptive corrective feedback was provided to the learner based on Aljaafreh and Lantolf's (1994) Regulatory Scale, and included the following:

1) implicit indication that there is a mistake, e.g., *look at number one again*;
2) the mistake is located, e.g., *she showed her disapprove<u>ment</u>*?
3) attention is directed to syntactic function, e.g., *which part of speech do we need*?
4) the meaning of the affix is either hinted or revealed, e.g., *this suffix means a quality*;
5) example sentences are given containing words formed with the affix;
6) the correct response and explicit explanation are provided.

The following **excerpt 1** illustrates how M's performance was mediated during the DA sessions (in the excerpt, the item *He is very brave. He is known for his .... (fear).* from the third session is discussed; hereafter, M = the participant, I = the interviewer). Since both English and Russian were used in the dynamic assessment sessions, the Russian transcription and the English translation will both be given, but in the rest of the excerpts (and also in the Group study), only the English translation will be used.

(1) *I:* Posmotri na sed'moe…

Look at the seventh…

M: (5.2)

I: …predlozhenie. Kakuyu chast' rechi nam nuzhno obrazovat'? He is known for ↑hi:s-

… sentence. Which part of speech do we need to form? He is known for ↑hi:s-

M: Besstrashie – sushtshestvitel'noe.

Fearlessness—a noun.

I: Tak. A u tebya chto?

Right. And what do you have?

M: Ah (0.6) prilagatel’noe.

Ah (0.6) an adjective.

I: Tak. Chego-to ne hvatayet. To est’ (.) u tebya prilagatel’noe ‘fearless’. Oznachazushtshee ↑chto?

Right. Something is missing. That is (.) you have the adjective ‘fearless’. Which means ↑what?

M: Besstrashnyi.

Fearless.

I: To est’ tebe ostalos’ dobavit’ suffiks kotoryi delaet ego sushtshestvitel’nym.

So, what you need to add is a suffix that makes it into a noun.

M: (4.0).

14 *I:* Podumai chto oznachaet slovo. Besstrashie – eto chto?

Think what the word means. What is fearlessness?

15 *M:* Kachestvo.

A quality.

16 *I:* Zamechatel'no!

Great!

17 *M:* (16.5) Fearnessless?

18 *I:* Tol'ko naoborot.

Yes, but vice versa.

19 *M:* ((laughter)) Fearlessness. ((laughter))

So, generally, first, the interviewer asked the learner to look at the item, giving him some time to respond (line 2). If there was no response, the interviewer gave the learner a hint (line 3) regarding the part of speech. If the learner was still unsure, the interviewer prompted him further (lines 6-18).

Findings

As I will demonstrate, M's beliefs about the usefulness of corrective feedback were transformed over the period of time the interviews were conducted. To support these findings, I will present excerpts from the dynamic assessment since I believe the assessment sessions, taken as a whole, facilitated these changes.

Changes in M's Beliefs

During the first interview, M was asked how his English teacher at school gave feedback on errors. In M's words: "*the teacher first shows where the mistake is (.) and if we don't understand (0.4) she (.) our teacher, starts explaining the specific rule word or situation."* I take his statement as a starting point: this is what M believed about how corrective feedback was given at the onset of the study. However, M doubted the efficacy of some of his teacher's feedback (**excerpt 2**). **Excerpt 2** also exemplifies how the concept of mediated action was applied during the analysis.

(2) *I:* Can you learn something from these hints? Let's say when the teacher shows where the mistake is.

M: (2.0) If he [in Russian, the grammatical gender of the word *teacher* is masculine] explains why it is incorrect—what the rule is—then I'll remember it.

I: OK. What if he doesn't?

M: We::ll then I'll try to understand why it is a mistake (.) but I can be wrong, or maybe I won't understand at all.

The mediated action in **excerpt 2** is the product of the interaction between M and the interviewer, the latter mediating M's utterance by asking "*What if he doesn't"*. However, it is also constructed by M's experience with his teacher's feedback practices. It appeared that this experience strongly influenced the way M reported on the usefulness of implicit CF; thus, the interviewer's mediation did not lead a noticeable change in his belief. However, during the interview a week after the DA, M changed his opinion (**excerpt 3**).

1 (3) *I:* Which of these hints was the most useful?
2 *M:* I think when you hinted (.) there I still had to think.
3 *I:* Uhu:
4 *M:* But (.) already in the <u>right</u> direction.
5 *I:* Hinted that there is a mistake or hinted where the mistake is?
6 *M:* Well, hinted that there is a mistake and hinted about the rule (1.2) something like
7 that.
8 *I:* And these hints are useful for what? To learn something or to find a mistake?
9 *M:* I guess both.
10 *I:* Let's say I tell you that you that you have a mistake here. Can this be useful and why?
11 *M:* Well, it is useful that you tell me there is a mistake. (0.6) I try to find it, and exclude some options that do not fit, including the one that I wrote.

First, M is invited to reflect on his experience with DA (line1). With reference to his DA experience, M reports that when the interviewer/researcher hinted, M had to think, which was useful. The interviewer then mediates M's reflection by providing alternatives (lines 5 and 7). In line 6, M accepts only the first alternative. The interviewer, however, then formulates the question in such a way that M has to refer to his DA experience (line 9). This all creates a context which is different from **excerpt 2** (where M reflected on the school teacher’s feedback), and as a result a different way in which M reported on the same CF as in **excerpt 2**.

By contrast, during the first interview, M considered feedback that gave examples of the correct structure to be the most useful, saying, *“the teacher gives me an example of some other sentence with a similar meaning or a word in which (.) eh I had the mistake. Then he asks me what the difference is between my sentence and (.) the sentence that the teacher gave me.”*

During the second interview, M still considered this type of feedback useful, but seemed to have changed his mind regarding what he felt it was useful for (**excerpt 4**).

(4) *I:* You remembered this hint. Why?
M: Because it helped a lot.
I: Did it help you to find your mistakes or to learn something?
M: I think above all to find the mistakes.

Here, as in **excerpt 3**, the interviewer mediated M’s response by providing two options. Nevertheless, as the context was the DA, it helped M to formulate his response.

Interestingly, six months later, M still considered the feedback hinting about meanings (which M referred to as ‘rules’) and the feedback hinting that there is a mistake to be useful because “*in the first (.) and the one where it is shown that there is a mistake (0.6) you have to think there*’, and ‘*where the rule is shown, there you [me/one] (.) too have to* ↑*think. And there (1.4) we::ll, it looks like the* <u>*first*</u> *one where it is shown that there is a* *mistake.*" The rising intonation at the end of what otherwise seems to be a confirmatory sentence might be interpreted as hesitation. However, I would suggest that it indicates M's desire to continue his thought that the first three corrective feedback types were similar in that they made him think. Importantly, this time, no explicit mediation by the interviewer was observed.

The most notable change across the interviews was the way in which M's opinion developed about the usefulness of examples of correct structures. During the last interview, which took place six months after the DA, M said, "*where [the teacher] gives examples of correct sentences is (.) in my opinion (.) <u>useless</u> becau:se (.) I will correct the mistake but (0.6) I might not understand (.) the rule or remember the mistake*." Thus, it seems that, in contrast to the first interview, where, judging from his use of the word *understand* (see **excerpt 2**), M doubted that implicit feedback would result in *awareness with understanding* (Alanen, 2013), during the last interview, M had similar doubts about feedback providing examples of correct structures.

In the following section, I will elaborate on how dynamic assessment mediated the way M reported on his beliefs during the last two interviews.

Dynamic Assessment

During the dynamic assessment, there were several episodes when M was able to self-correct after implicit CF. Consider **excerpt 5**, for example. During the first session, M did not know the meaning of the suffix *-ess*; I had to give him the correct answer and explain the meaning of the suffix. During the second session, the process was quite different:

(5) *I:* Yea:h. Almost right. You missed one letter in number nine.

M: "r"? I knew it!

I: Huntress. You wrote it right. You remembered that it [[is a suffix-]]

M: [[of the feminine gender.]]

The recognitional overlap at the end of the exchange suggests that M did not require the interviewer's mediation, as he was able to do it himself. Another example is **excerpt 1**, where M was directed to the meaning of the word *fearlessness* (lines 14-16) and, as a result, was able to recall the suffix *-ness* (line 17).

Episodes like these seem to have led M to classify references to hints about 'rules' and mistakes as the most useful CF types, during the second and the third interviews.

GROUP STUDY

Two group interviews were analysed in order to understand how social interaction brings about changes in learners' beliefs over a short period of time.

Participants

The participants in the Group study were 6 L1 Russian learners of English at grade 8 (14-15 years of age) from a secondary school in Estonia. By the time of the study, they had been studying English for about six years.

The participants in the study were selected from a larger group participating in a study aiming at establishing the effect of dynamic assessment on learners' ability to form

L2 English questions (see Leontjev, 2014). The larger group from which the present participants were sampled was recruited such that wh-questions with auxiliaries were within the learners' ZPDs. This was done by asking their teacher whether the learners were familiar with the trained structure. What is more, all of the participants were able to form several correct questions during the DA (albeit some learners with rather explicit assistance), which confirmed that the structure was within the participants' ZPD.

The selection of the participants in the present study was based on (a) their teachers' evaluation of their abilities and (b) the learners' performance on two unmediated exercises which showed their unassisted ability to form wh-questions with auxiliaries (the target of the DA). The first exercise measuring their unassisted performance was E-mail writing according to the prompts given and the second, a gap-filling exercise. More details about the original group from which the participants in the present study were sampled and the exercises can be found in Leontjev (2014).

The learners were interviewed in two groups, as mentioned above, formed based on the learners' unassisted performance and their teacher's evaluation of their abilities. The first group included two high-achieving learners (coded *HA1* and *HA2*) and one middle-achieving learner (*MA1*). The second group had two low-achieving learners (coded *LA1* and *LA2*) and one middle-achieving learner (*MA2*). The low-achieving learners were selected among those whose unassisted performance was in the lower tertile, the middle-achievers, in the middle tertile, and the high-achievers, in the top tertile in their group. The teacher confirmed that HA1 and HA2 were, indeed, high-achievers as regards their performance in the class, the two low-achievers were among the low-achieving learners in the class, and MA1 and MA2 were among the averagely performing learners. It should be noted that the high-achieving learners only occasionally required CF during the dynamic assessment, and when they did, it did not reveal much detail about their mistakes. The low-achieving learners, on the other hand, received the CF almost on every item in the test and experienced feedback of different explicitness and level of detail. The middle-achievers' experience with the DA was somewhere between the little implicit feedback that the high-achievers received and extensive CF of various explicitness and level of detail received by the low-achievers.

One high-achieving learner missed the day of the interview and had to be replaced by another high-achiever (and coded as HA1 instead of the missing learner) who, however, had knowledge of results feedback (i.e., either telling that his response was correct or that it was wrong) and not adaptive feedback as the rest of the participants did, as this high-achiever was from the control group in Leontjev (2014). While HA1's experience with the CF in the study seems to be different at the first glance, it was rather similar to HA2's. HA1 received the CF (i.e., this is wrong) thrice, and so did HA2, two of them *think more carefully*, and one, *look at this part of your sentence*. Further details on the participants are presented in **Appendix C**.

Data and data analysis

The data consist of two small-group interviews with the six learners. Additional data collected in the study come from a questionnaire (**Appendices B and D**), observation of the learners working on the computerised DA tasks and their performance logs (**Appendix C**), and an interview with their teacher of English. The learner interview topics included:

- teacher's feedback practices and learners' perceived usefulness of it;
- feedback in the dynamic assessment;
- learning an L2.

To help the interviewees recall their experience, they were presented with screenshots of sample feedback messages and exercise items. For verification, learners' interview data were triangulated with the data accumulated from the learners' performance on the dynamic test, the observation of the learners working on that same test (**Appendix C**), and the interview with their teacher. The interviews were transcribed, noting what mediated the learners' utterances (the teacher's voice, DA experience, other participants' mediation, etc.) and what, in addition of the utterances themselves (e.g., lack of hesitation or degree of agency) could be later interpreted with reference to the beginning of appropriation of the learner's beliefs. In one case (**excerpt 14**), both coders were not sure whether one of the utterances produced by LA2 was the manifestation of this learner's agency. Thus, a third person was asked to help with establishing whether it was so.

The overall procedure was as follows:

1) two unmediated exercises;
2) a computerised DA (immediately following the unmediated exercises);
3) a questionnaire (immediately following the DA);
4) interviews (on the following day after the DA).

All the exercises and the questionnaire were completed online.

During the dynamic assessment, the learners completed five exercises:

- two ordering exercises intended to diagnose problems with the word order of wh-questions with auxiliaries, e.g., *park/where/near the shop/my father/can ?*
- three ordered multiple-choice exercises evaluating problems with *do*, *does*, and *did* in wh-questions with auxiliaries, e.g., *what else [do you sell] in your shop?*

Based on Aljaafreh and Lantolf's (1994) Regulatory Scale, the feedback messages in the exercises (originally in Russian) had growing explicitness and detail, as indicated by numbers from 1 to 5 (also see **Appendix B**):

0) An indication that the response is correct.

1) An implicit hint that the response is incorrect, e.g., *Think more carefully*.

2) The part of the sentence containing the error is highlighted, e.g., *When **he comes** to work? Look at the highlighted part of your sentence*.

3) Metalinguistic clues and/or elicitations are given, e.g., *Where does it **plays** in the shop? You used the correct helping verb **does**. But something should be changed in the verb **plays** in your question.*

4) Examples of the correct structure are given, e.g., *Not quite right. Look at the following examples… How are they different from your sentence?*

5) The correct response with explicit explanation is provided, e.g., *How did you **found** my E-mail address? Unfortunately, it is incorrect. You used the verb **did** in the Past*

*Simple tense. Great! But since **did** is already in the past tense, you shouldn't have used **found** in the past tense. The correct answer is…*

With each incorrect response from the learners, the level of the feedback message that followed was increased (i.e., after the first mistake, feedback level 1 was displayed, after the second, feedback level 2, and so on). The feedback level was reset in each new exercise. I will refer to feedback given in the Group study in terms of these five levels.

Findings

In what follows, I will present excerpts from the two interviews in chronological order to illustrate the changes in the learners' beliefs that emerged during the interviews. It should be noted that the social interaction was constrained by the activity of the interview. Thus, the learners in the Group study, in most cases, did not respond to each other's utterances directly. However, in the analysis, it was noted how what was said by one participant mediated the way other participants constructed their utterances. Before turning to the interviews, I will briefly summarise the learners' responses to the questionnaire (**Appendix B**) and their teacher's feedback practices.

Questionnaire

I used the learners' responses to the questionnaire as an indication of the beliefs about feedback with which they entered the interviews. These responses are summarised in

Appendix D.

By and large, the high-achieving learners entered the interview with varying perspectives regarding the usefulness of feedback. The low-achievers, however, presented somewhat more homogeneous perspectives. Similarly to the findings of previous research (e.g., Amrhein & Nassaji, 2010), LA2 and MA2 considered explicit feedback to be the most useful, finding implicit feedback the least useful. LA1 thought all feedback in the dynamic test was useless. Notably, and perhaps relevant to these beliefs, two of the three low-achieving learners did not benefit from more implicit feedback during the dynamic assessment, as it did not help them to self-correct their mistakes. However, this does not apply to MA2, who was sometimes able to self-correct with less explicit feedback (**Appendix C**).

Teacher's Interview

The teacher's interview seemed somewhat contradictory. On the one hand, she confirmed that she often directed learners to the correct answer without overtly correcting, adding, "*that's my way*" because "*if we have some additional information or some hints, it makes our brain work.*" On the other hand, she added that "*the most usual way (1.1) first of all, I explain and then say the correct [answer].*" She also believed that her low-achieving learners expected her to provide only overt correction. Thus there is a possibility that the teacher's feedback practices were different with less-able learners, whose mistakes she corrected explicitly.

High-achieving Interviewees

MA1's experience mediated by the interviewer resulted in an interesting development in the way he reported on his beliefs, which influenced the rest of the discussion, including the interviewers' questions later during the interview (**excerpt 6**).

(6) *I:* Which one wasn't understandable?

MA1: This one, the second.

I: What was it that was not understandable?

MA1: Because it's simply eh (.) 'think more carefully'.

I: OK, <u>think</u> more carefully.

MA1: Well (.) as soon as I <u>thought</u> (.) I immediately understood ((laughter)).

The final MA1's utterance was in part constructed by the interviewer, who stressed the word *think* when repeating MA1's previous utterance. MA1 then not only repeated the word *think*, but also used the same intonation as the interviewer. In fact, the whole exchange can be seen as the interviewer's scaffolding the learner's response. That MA1 used the interviewer's words to formulate his utterance can be interpreted as the beginning of appropriation of the belief that such feedback can be useful for him (initiated by MA1's DA experience and mediated by the interviewer).

MA1's reflection notwithstanding, HA2 continued to report that the think-more-carefully feedback was useless. Thus, the interviewer turned to HA2's experience with this feedback (**excerpt 7**).

(7) *I:* OK. Can you remember (.) while you were doing the exercises did you get this feedback (.) and could you find the correct answer after it?

HA2: (1.0) Well yes (2.0) because in the beginning (.) I had like (0.8) <u>two</u> options that could fit.

I: Aha.

HA2: And since the first one was wrong, it could only have been the second one.

This reflection on his own experience could have been a reason for a change in HA2's report. Specifically, towards the end of the interview, the interviewer covered all the sample feedback messages but the most explicit and asked the learners whether it would have been useful if they had been always provided with correct responses during the test. To this, HA2 responded, *"I don't think so. Because it's better to understand the rule."* This was a change from his questionnaire responses (**Appendix D**), where he did not discuss the usefulness of the CF in terms of helping him understand the rules (something that was brought up by MA1 in **excerpt 6**). What is more, at the end of the interview, when the interviewer switched the context to the feedback from the teacher, HA2 reported that the teacher "*may at first hint that there is a mistake (.) the learner will try to guess it himself (1.1). It is more useful*."

Low-achieving Interviewees

Judging by the questionnaire responses of the low-achieving group learners, I expected that peer utterances would result in little changes in the way learners talked about their beliefs about CF. However, it was one of the learners whose utterances guided the utterances of other learners. To start with, soon after the beginning of the interview, LA1 seized the initiative, discussing why she felt feedback was useless for her (**excerpt 8**).

(8) *LA1:* There if you do one time correctly (.) the rest are the same. It happens with 'did' and without the ending (.) for example with 'does'. (0.4) And then everything else was correct. And ↑these ones ((points at the ordering exercises)) are a <u>real idiocy</u>.

What it more, the interviewer then decided to change the topic and asked the learners if they found any mechanics of doing the exercises that could be made better. However, instead of responding to the interviewer's question, LA1 continued to criticise, saying that "*the hints were really not understandable*," as a matter of fact, interfering with MA2's response.

However, she was also the first to acknowledge the usefulness of implicit feedback from the teacher (**excerpt 9**).

(9) *I:* What do you [2nd person pl.] think can such feedback be useful to you? When the teacher does not correct you, but (.) e:r (.) tries to help you to find the correct answer, so that you yourself find the mistakes?

LA1: <u>Of course</u> they are useful.

I: How?

LA1: Because we think ourselves (0.4) we start (1.2) something like <u>these</u> rules

((points in the direction of the screenshots of the feedback and the exercises)) (0.6)
so:mehow (0.4) connections.

LA's last utterance is quite interesting. The unfilled pauses could be due to the increased cognitive load (e.g., Goldman-Eisler, 1960). That the utterance was rather fragmentary also suggests this interpretation. LA1 making a connection between the teacher's feedback and the feedback in the exercises (something like these rules), or the interplay of her two discordant beliefs—that such feedback is useful and not useful—might have increased the load. The teacher's voice (see the teacher interview) may also be reflected in LA1's utterances, triggered perhaps by the interviewer's use of the second-person plural at the onset of the exchange. That is to say, LA1 might have been more likely to see the advantages of corrective feedback when the implied agents were other, perhaps, more able learners, whom the teacher directed to correct responses without revealing them (as emerging from teacher interview). In any case, this exchange became rather important for the rest of the interview, as this connection between the teacher's feedback and the feedback in the study helped the other learners to construct their utterances about the usefulness of feedback in the study.

In the questionnaire, LA2 rated the last two levels as the most useful. In fact, also immediately before the episode quoted in **excerpt 9**, both MA2 and LA2 reported that feedback that showed them the correct response was the best. However, in **excerpt 10**, LA2 reports on these two feedback levels somewhat differently.

(10) *LA2:* And for me it was that one.

I: The <u>second</u> to last?

LA2: Yes. (0.6) Well (.) the last one and this one.

I: M: the last one and the second to last.

LA2: Yeah, you [one/me] have to (.) ↑think a bit there.

I: Aha. And here you have to think <u>less</u>? ((points at level 5 feedback))

LA2: Yes.

At first, LA2 pointed to level 4 feedback (i.e., examples of correct structure), only mentioning overt correction after some hesitation. LA2 then reported that one had to *think a bit* when level 4 feedback was displayed, again seeming rather hesitant, as the rising intonation suggests. LA2's utterance "*you have to think a bit"* is very similar to LA1's "*we think ourselves*" in **excerpt 9**. This choice of words is not coincidental, considering that this LA2's utterance appeared only about three minutes after LA1's utterance in **excerpt 9.** This is all the more interesting because in the questionnaire LA2 discussed the usefulness of the feedback in terms of whether it explicitly revealed what his mistakes were (useful) or not (useless). That is to say, LA2 used LA1's words to construct his utterance. The sequence I see is the following: the teacher's feedback that does not reveal the correct answer is useful because it makes them think (as LA1 reported); the feedback in the study also makes them think (the connection that LA1 made between the teacher's CF and the feedback in the study); one has to think more when one is given fewer details about the error. The interviewer mediated this emerging belief by stressing that the learners had to think less when provided with overt correction.

In **excerpt 11**, changes in the beliefs of two learners emerge.

(11) *I:* OK (.) let's go back to what you [2nd person pl.] learned yesterday. What do you [2nd person pl.] think, could the hints, which you got yesterday (1.1) help you to learn that? That is, for you: the sentences in the past tense, for example?
LA2: Yes (.) they could.
I: Uhu.
LA2: Well, at first I did not understand, but then-
I: Uhu.
LA2: I read the hints and it (.) became a little more understandable.
(2.0)
I: What about you?
LA1: No.
I: Why?
LA1: Because I told you that the construction was the same (.) and I put it because I knew it was right (0.6). I did not think (0.5) didn't read the sentences.
I: Right. (0.6) What about you?
MA2: Well, there were similar sentences. Sometimes I simply pressed OK (.) and it was right.
I: So, the hints did not help you?

19 *MA2*: We:ll (0.8) some of them ↑helped.
20 *I*: Uhu.
21 *MA2*: The ones that (0.8) were (.) the second to last.

First of all, LA2, having reported in the questionnaire that the feedback did not help him, now confirms that it actually did (lines 6-8). As LA1 reminds us of her strategy (lines 13-14), MA2 reveals that he, too, sometimes answered randomly (lines 16-17; see also **Appendix C**). He, however, adds that the feedback did help him. Importantly, he now only refers to level 4 feedback (lines 19-21; cf. **Appendix D**).

LA2 now seemed to have abandoned his view that overt correction is always the best, but LA1 also realised something about effortlessly getting the correct answers (**excerpt 12**).

1 (12) *I:* What if you [2nd person pl.] (.) instead of all this, you [2nd person pl.] had been
2 given the last feedback only? Would it have been more useful for you [2nd person
3 pl.] (.) what do you [2nd person pl.] think?
4 *LA2:* No it wouldn't.
5 *I:* Why?
6 *LA2:* Well (.) like (.) the:n you [one/me] (.) don't try to understand why it is so.
7 Well it does not make you [one/me] think about it.
8 *I:* Uhu.

9 LA1: For the test result (.) if you need to know the correct answer (.) then it would
10 be yes (.) more useful.
11 *I:* And for you?
12 *LA1:* °No°.

LA2 again used the verb *think* (line 7) to refer to level 4 feedback, this time, however, without hesitating (differently from **excerpt 10**). It is also interesting that LA2 used the word *understand* as a positive aspect of feedback (line 6); that is to say, L2 appreciated the awareness with understanding that more implicit feedback made possible, which was a change from his questionnaire responses (**Appendix D**). There is also (arguable) evidence for LA2's agency. The sentence he formed (lines 6-7) was a mononuclear impersonal sentence (in the original: "*ne pytaesh'sya ponyat"*). These are often used to refer to self in the Russian language.

While LA1 in lines 9-10 responded to the interviewer's question, she also responded to LA2's utterance. In lines 9 and 10, it then appears that (a) LA1 assumed CF not revealing the correct answers could have been useful for her during the DA (which line 12 also suggests) and (b) that she probably considered the DA to be a conventional test. LA1's "*no"* (line 12) was uttered in a soft, quiet voice. Perhaps, she did not want to admit that the feedback she had skipped might have been useful for her. This does not mean, however, that she fully abandoned her prior belief—at the end of the interview, she added that, above all, she needed "*a good graduation diploma, and then knowledge*." Though her existing belief in the superiority of good marks over knowledge was still strong, her negation indicates that it had weakened. In this respect, LA2's negation (line 4) is different from LA1's, as he was confident in his response and justified it (lines 6-7). This can be interpreted with reference to a different degree of appropriation of their beliefs, LA2 having appropriated his belief to a greater extent than LA1 did.

OVERALL DISCUSSION

The present study aimed at discovering how learners' beliefs were transformed by their experience of dynamic assessment and were co-constructed in social interaction unfolding during research interviews, as well as, what, in addition to the learners' DA experience, mediated these changes.

The findings of the two studies indicate that the learners' beliefs about the usefulness of corrective feedback transformed (or started transforming) through social interaction and experience. In the Case study, what the learner reported about corrective feedback differed across the interviews. That is, M started appreciating implicit feedback because it made him think (in M's words) rather than more explicit feedback that he thought only helped him to self-correct. In the Group study, the participants' responses to the questionnaire were different from what they reported during and especially at the end of the interviews.

With reference to the first research question, it appears that the recent experience of DA influenced the way the learners reported on the usefulness of corrective feedback. Evident in the Case study above all, this is also seen in the Group study. Importantly, the results of the Case study suggest that the beliefs that emerged from the learner's experience of dynamic assessment persisted and even developed further over the following six months. That said, it should not be assumed that the DA experience was uniform for all the high-achieving and all low-achieving interviewees (**Appendix C**), especially considering the nature of DA, where mediation (in the case of the present study, CF) is attuned to each learners' ZPD but also that learners entered the DA with their own beliefs and expectations. This was particularly evident in the case of LA1. It appears that due to the teacher's intervention during the DA (**Appendix C**), LA1 perceived the procedure as a common test (see **excerpt 12**). This resulted in that she rejected the feedback, instead memorising her responses when these were correct by chance. It can even be suggested that at times (e.g., **excerpt 8**), instead of responding to the interviewers' questions (which should be expected in an interview), she decided that her task was to criticise the DA procedure since she considered it to be useless (probably having certain expectations of what a test should be like).

What is more, it was not the experience per se, but the experience mediated in the social interaction unfolding during the interviews that brought about changes in the learners' utterances. In the following, the results contributing to the response to the second research question will be discussed.

The interviewer, who also was the mediator during the DA in the Case study, was the most evident source of mediation during the interviews, being a significant other (researcher). It appeared that he mediated the learners' utterances in a variety of ways, for example, providing two options (e.g., **excerpts 3** and **4**) or eliciting the learners' DA experience (e.g., **excerpt 7**). In fact, the sole presence of the interviewer during the final interview of the Case study was enough to construct the context, where M discussed the usefulness of his school English teacher's feedback with reference to his DA experience.

Regarding the latter, it is possible that, in line with the previous studies (e.g., Aro, 2009), and considering the teacher's own beliefs about CF emerging from the interview, teacher's voice was present in LA1 utterance in **excerpt 9** ("*we think ourselves"*). Even if the teacher was not the 'speaking entity' in LA1's utterance, at least the teacher's (an authoritative other's) feedback practices led to LA1's report. That is, LA1's belief in the usefulness of her teacher's feedback mediated her reflection.

However, as the Case study demonstrated, it is not always beneficial to draw on teachers' feedback. M appeared to have negative experiences with his teacher's feedback in locating a mistake. This could have been the reason why during the last interview, when there was no explicit mediation by the interviewer, M never mentioned this CF type among the ones that could be useful for him.

What other learners reported also mediated the learners' utterances. One example of that is LA1' utterance in **excerpt 9** discussed earlier. After LA1 reported that teacher's feedback made them think, the focus of the learners' discussion of the usefulness of the feedback changed from whether or not it revealed the correct answers and explained everything to whether or not it made them think. Similarly, it seems that what MA1 reported in **excerpt 7** mediated HA2's utterances towards the end of the interview. Some of the changes in the way learners reported on the usefulness of corrective feedback can be interpreted with reference to mediation of noticing (Alanen, 2013). That is to say, other participants' mediation helped the learners to recall something from their experience (e.g.,

excerpt 7) or make connections between the CF they received during the DA and the feedback from their teacher (e.g., LA1's report helping to construct LA2's utterance).

That said, especially as regards the Group study, it would be inaccurate to talk about the transformation of the learners' beliefs about corrective feedback. Rather, the changes in the learners' utterances can be perceived as the beginning of the process of appropriation of these beliefs, different for different learners, at different times during the interviews, and in different interviews. For example, LA2's absence of hesitation in **excerpt 12** while present in **excerpt 1** when he reported on that CF is useful when it makes one think can be interpreted as a change in the degree of appropriation of this belief. Similarly, the degree of appropriation of LA2's "*no it wouldn't*" and the following elaboration in **excerpt 12** seems to be higher than LA1's "*no*" in the same excerpt due to the lack of hesitation and an arguable degree of agency in the former. Finally, that there was no mediation of M's beliefs by the interviewer (apart from the latter being present) during the last interview in the Case study also suggests that M's beliefs about CF were appropriated more during the last interview.

With reference to appropriation, LA1's case is especially interesting, as it appears that she had two beliefs: that the teacher's implicit feedback was useful and that good marks and knowing the correct responses were important (both, judging by the degree of agency in her utterances, appropriated considerably). However, she appeared to have used the latter to mediate her selection of strategy during the DA and not the former (see also, e.g., Mercer, 2011). Possibly, she turned to this belief to mediate her experience (also during the interview) because she considered the procedure to be a conventional test, since the situation, for her at least, was indeed test-like. A further reason for LA1's performance can be her frustration at not being able to find the correct answers in the ordering exercises (**excerpt 8**). That LA1 reported on both of these beliefs during the interview can be explained by the natural polyphony of voices and beliefs mediating her (and others') reflections (Aro, 2009; Bakhtin, 1986; Dufva, 2003).

LIMITATIONS AND IMPLICATIONS

The two studies reported here aimed to add to the understanding of how recent experience and social interaction between the interviewer and the learners, but also utterances of other learner participants change learners' beliefs about the efficacy of corrective feedback.

There are, however, several limitations to the studies, the biggest of which concerns the absence of longitudinal data in the Group study. It is therefore impossible to confirm or refute that the learners appropriated the beliefs that emerged during the research interviews. It is also difficult to say how the interaction would have unfolded and what beliefs would have emerged should there have been less guidance by the interviewer. It should also be noted that while the task that the learners were given was to answer the interviewer's questions, they could have perceived their task as, for example, having to respond 'correctly', that is, to tell the interviewer what they thought he wanted to hear (see also Alanen, 2003). Finally, caution must be exercised when extending the findings from these two studies to other contexts.

Notwithstanding these limitations, the findings suggest that learners' beliefs about corrective feedback can change through social interaction and recent experience.

Specifically, learners’ own experience, other participants’ reflections, the interviewer’s questions, and voices of significant others (i.e., the teacher) influenced the ways in which the learners reflected on their beliefs about the usefulness of different types of corrective feedback. The findings also suggest that beliefs about corrective feedback emerge and start transforming (and being appropriated) both with time (the Case study) and over as short a period of time as one research interview (the Group study).

The implications of these findings stretch beyond the immediate pedagogical context, namely, to change learners’ beliefs about corrective feedback through discussion (see Amrhein & Nassaji, 2010). The findings imply that learners might sometimes skip feedback provided to them during computerised dynamic assessment if they believe it is useless (see also Thouësny, 2011), which may hinder the reliability and validity of DA. Specifically, when learners skip feedback because they believe it is useless, they may lose opportunities for development, which decreases the usefulness of DA. Furthermore, any inferences made from the performance of such learners (e.g., the amount of assistance they require with certain structures) can be unreliable. However, discussions similar to the ones presented here might potentially serve as a remedy for this problem. The implied reason that LA1 skipped the feedback has further implications for computerised dynamic assessment. Manipulating the starting level of the complexity of feedback, that is, making it more explicit and detailed for these learners, should reduce the possibility that less able learners get frustrated and skip the feedback they receive. Future studies can shed more light on the ways learners’ reciprocity (including when it is guided by their beliefs) can be accounted for in computerised DA.

Acknowledgments

I am grateful to Professor Riikka Alanen, Dr. Tatjana Rynkänen, and the two anonymous reviewers for their input and insights.

References

Alanen, R. (2003). A sociocultural approach to young language learners' belief about language learning. In P. Kalaja & A. M. F. Barcelos (Eds.), *Beliefs about SLA: New research approaches* (pp. 55–85). Amsterdam: Kluwer Academic.

Alanen, R. (2013). Noticing and mediation: A sociocultural perspective. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 315–325). Honolulu, HI: University of Hawaiʻi, National Foreign Language Resource Center.

Aljaafreh, A., & Lantolf, J.P. (1994). Negative feedback as regulation and second language learning in the Zone of Proximal Development. *The Modern Language Journal, 78*(4), 465–483. DOI: 10.1111/j.1540-4781.1994.tb02064.x

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. Retrieved from https://journals.lib.unb.ca/index.php/CJAL/article/view/19886/21712

Aro, M. (2009). *Speakers and doers: Polyphony and agency in children's beliefs about language learning* (Doctoral dissertation, University of Jyväskylä, Jyväskylä, Finland). Retrieved from https://jyx.jyu.fi/dspace/handle/123456789/19882

Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing, 9*(3), 227–257. DOI: 10.1016/S1060-3743(00)00027-8

Bakhtin, M. (1981). *The dialogic imagination: Four essays by M. M. Bakhtin.* Austin, TX: University of Texas Press.

Bakhtin, M. (1986). *Speech genres and other late essays.* Austin, TX: University of Texas Press.

Barcelos, A. M. F. (2003). Researching beliefs about SLA: A critical review. In P. Kalaja & A. M. F. Barcelos (Eds.), *Beliefs about SLA: New research approaches* (pp. 7–33). Amsterdam: Kluwer Academic.

Barcelos, A. M. F., & Kalaja, P. (2013). Beliefs in second language acquisition: Teacher. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. New York: Blackwell Publishing.

Bråten, I. (1991a). Vygotsky as precursor to metacognitive theory: I. The concept of metacognition and its roots. *Scandinavian Journal of Educational Research, 35*(3), 179–192. DOI: 10.1080/0031383910350302

Bråten, I. (1991b). Vygotsky as precursor to metacognitive theory: II. Vygotsky as metacognitivist. *Scandinavian Journal of Educational Research*, *35*(4), 305–320. DOI: 10.1080/0031383910350406

Brown, A. V. (2009). Students' and teachers' perceptions of effective foreign language teaching: A comparison of ideals. *The Modern Language Journal, 93*(1), 46–60. DOI: 10.1111/j.1540-4781.2009.00827.x

Cole, M. (1996). *Cultural psychology: A once and future discipline.* Cambridge, MA: Harvard University Press.

Cornillie, F., Clarebout, G., & Desmet, P. (2012). Between learning and playing? Exploring learners' perceptions of corrective feedback in an immersive game for English pragmatics. *ReCALL: The Journal of EUROCALL, 24*(03)*,* 257–278. DOI: 10.1017/S0958344012000146

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment* [electronic version]. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_en.pdf

Diab, R. L. (2005). Teachers' and students' beliefs about responding to ESL writing: A case study. *TESL Canada Journal, 23*(1), 28–43. Retrieved from http://www.teslcanadajournal.ca/index.php/tesl/article/view/76

Dufva, H. (2003). Beliefs in dialogue: A Bakhtinian view. In P. Kalaja & A. M. F. Barcelos (Eds.), *Beliefs about SLA: New research approaches* (pp. 131–151). Amsterdam: Kluwer Academic.

Dufva, H. (2004). The contribution of the Bakhtin Circle to the psychology of language. In M. Nenonen (Ed.), *Papers from the 30th Finnish Conference of Linguistics* (pp. 21–26). Joensuu, Finland: University of Joensuu.

Hedgcock, J., & Lefkowitz, N. (1994). Feedback on feedback: Assessing learner receptivity to teacher response in L2 composing. *Journal of Second Language Writing, 3*9(1), 141–163. DOI: 10.1016/1060-3743(94)90012-4

Heift, T. (2002). Learner control and error correction in ICALL: Browsers, peekers, and adamants. *Calico Journal, 19*(2), 295–313. Retrieved from http://www.jstor.org/stable/24149363

Horwitz, E. K. (1985). Using student beliefs about language learning and teaching in the foreign language methods course. *Foreign Language Annals*, *18*(4), 333–340. DOI: 10.1111/j.1944-9720.1985.tb01811.x

Hyland, F. (2003). Focusing on form: student engagement with teacher feedback. *System, 31*(2), 217–230. DOI:10.1016/S0346-251X(03)00021-6

Kalaja, P., & Barcelos, A. M. F. (2013) Beliefs in second language acquisition: Learner. In C.A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. New York: Blackwell Publishing.

Kern, R. G. (1995). Students' and teachers' beliefs about language learning. *Foreign Language Annals, 28*(1), 71–92. DOI: 10.1111/j.1944-9720.1995.tb00770.x

Kozulin, A. (1998). *Psychological tools: A sociocultural approach to education.* Cambridge, MA: Harvard 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. 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. Retrieved from http://apples.jyu.fi/ArticleFile/download/459

Mercer, S. (2011). Language learner self-concept: Complexity, continuity and change. *System, 39*(3), 335–346. DOI: 10.1016/j.system.2011.07.006

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Plonsky, L., & Mills, S. V. (2006). An exploratory study of differing perceptions of error correction between a teacher and students: Bridging the gap. *Applied Language Learning, 16*(1), 55–74. Retrieved from http://www.dliflc.edu/wp-content/uploads/2014/04/all16one.pdf

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. Berlin: Springer.

Poehner, M. E, & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA). *Language Teaching Research, 17*(3), 323–342. DOI: 10.1177/1362168813482935

Pujolà, J. (2001). Did CALL feedback feed back? Researching learners' use of feedback. *ReCALL: The Journal of EUROCALL, 13*(1), 79–98. DOI: 10.1017/S0958344001000817

Põhikooli riiklik õppekava õigusakt: Lisa 1 [Basic School National Curriculum Act: Annex 1] (2010). Pub. L. No. RT I 2010, 6, 22. Retrieved from https://www.riigiteataja.ee/aktilisa/1281/2201/0017/13275423.pdf

Saito, H. (1994). Teachers' practices and students' preferences for feedback on second language writing: A case study of adult ESL learners. *TESL Canada Journal, 11*(2), 46–69. Retrieved from http://www.teslcanadajournal.ca/index.php/tesl/article/view/633

Sato, M. (2013). Beliefs about peer interaction and peer corrective feedback: Efficacy of classroom intervention. *The Modern Language Journal, 97*(3), 611–633. DOI: 10.1111/j.1540-4781.2013.12035.x

Schulz, R. A. (2001). Cultural differences in student and teacher perceptions concerning the role of grammar instruction and corrective feedback: USA-Colombia. *The Modern Language Journal, 85*(2), 244–258. DOI: 10.1111/0026-7902.00107

Teo, A. (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning & Technology, 16*(3), 10–20. Retrieved from http://llt.msu.edu/issues/october2012/action.pdf

Thouësny, S. (2011). Dynamically assessing written language: To what extent do learners of French language accept mediation? In S. Thouësny & L. Bradley (Eds.), *Second language teaching and learning with technology: Views of emergent researchers* (pp. 169–188). Dublin: Research-publishing.net.

Vasilyeva, E., Puuronen, S., Pechenizkiy, M., & Rasanen, P. (2007). Feedback adaptation in web-based learning systems. *International Journal of Continuing Engineering Education and Life-Long Learning, 17*(4-5), 337–357. DOI: 10.1504/IJCEELL.2007.015046

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wenden, A. (1987). Metacognition: An expanded view of the cognitive abilities of L2 learners. *Language Learning, 37*(4), 573–597. DOI: 10.1111/j.1467-1770.1987.tb00585.x

Wertsch, J. V. (1991). A sociocultural approach to socially shared cognition. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 85–100). Washington, DC: American Psychological Association.

Wertsch, J. V. (1998). *Mind as action*. New York: Oxford University Press.

Yoshida, R. (2010). How do teachers and learners perceive corrective feedback in the Japanese language classroom? *The Modern Language Journal, 94*(2), 293–314. DOI: 10.1111/j.1540-4781.2010.01022.x

APPENDIX A

Transcription Symbols.

Symbol	**Meaning**
text	a stressed word or a part of it
↑	noticeably rising intonation[1]
((text))	non-verbal behaviour, e.g., laughter, gestures, etc.
A: [[text]]	
B: [[text]]	overlapping utterances
(.)	pause of 0.2 seconds or less
(0.0)	timed pause
:	elongation of the preceding sound
-	an utterance is cut off
°text°	uttered in a noticeably quieter, softer voice
[text]	comment

[1] Punctuation markers are not deliberately used to indicate intonation in the transcript although question marks show a somewhat rising intonation and full stops, unless otherwise indicated, show a somewhat falling intonation.

APPENDIX B

Questionnaire Items Discussed in the Study (English Translation).

Please tell us how useful the hints were for you (how well they helped you to complete the exercises).

□ Very useful (they helped me a lot)

□ Quite useful (they helped quite a lot)

□ Not really useful, but not entirely useless either (they helped me a little bit)

□ Quite useless (they did not really help me)

□ Useless (they did not help me)

Did you learn anything after completing the exercises?

□yes □no

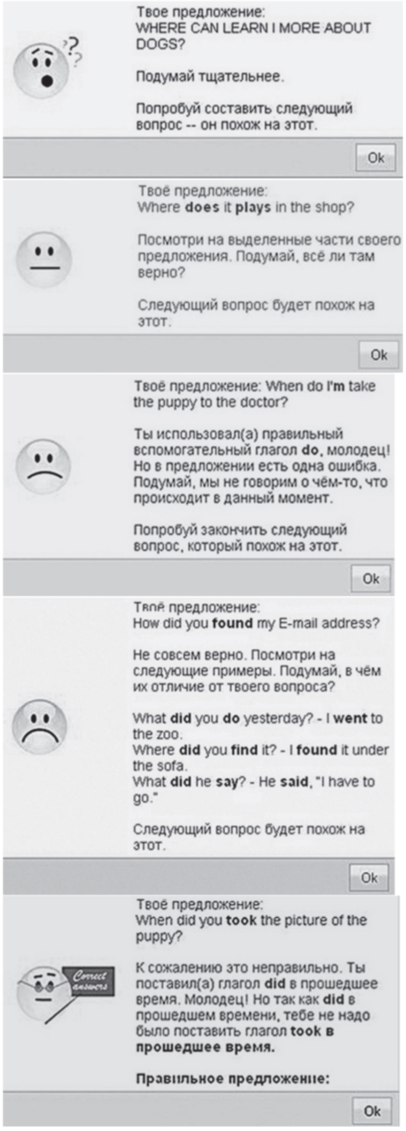
Please tell us what you learned:

Do you think the hints you received helped you to learn it?

□yes □no

Please tell us how exactly the hints helped you to learn:

Below are the hints similar to the ones you probably saw while doing the exercises. Click on the hints and give <u>each</u> of them a mark from '1' to '5' depending on its usefulness for you. Give a '1' to the most <u>useless</u> and a '5', to the most <u>useful</u>.



Твое предложение:
WHERE CAN LEARN I MORE ABOUT DOGS?

Подумай тщательнее.

Попробуй составить следующий вопрос -- он похож на этот.

Твоё предложение:
Where **does** it **plays** in the shop?

Посмотри на выделенные части своего предложения. Подумай, всё ли там верно?

Следующий вопрос будет похож на этот.

Твоё предложение: When do **I'm** take the puppy to the doctor?

Ты использовал(а) правильный вспомогательный глагол **do**, молодец! Но в предложении есть одна ошибка. Подумай, мы не говорим о чём-то, что происходит в данный момент.

Попробуй закончить следующий вопрос, который похож на этот.

Твоё предложение:
How did you **found** my E-mail address?

Не совсем верно. Посмотри на следующие примеры. Подумай, в чём их отличие от твоего вопроса?

What **did** you **do** yesterday? - I **went** to the zoo.
Where **did** you **find** it? - I **found** it under the sofa.
What **did** he **say**? - He **said**, "I have to go."

Следующий вопрос будет похож на этот.

Твоё предложение:
When did you **took** the picture of the puppy?

К сожалению это неправильно. Ты поставил(а) глагол **did** в прошедшее время. Молодец! Но так как **did** в прошедшем времени, тебе не надо было поставить глагол **took в прошедшее время.**

Правильное предложение:

Please tell us why do you think the hint that you gave the highest mark was the most useful for you?

Why was the hint that you gave the lowest mark useless for you?

APPENDIX C

Interviewees' Performance/Observation.

Code	Sex	Performance	Observation
HA1	M	From the knowledge of results feedback group. Required feedback once in the first exercise and twice in the second exercise.	I observed him working through the exercises, as he looked at the items quickly, thought for several seconds, and selected the correct options.
HA2	M	Required level 2 feedback in the second ordering exercise to self-correct and level 1 feedback, in the *did*-exercise.	I watched him carefully studying different options in the exercises before submitting his answers.
MA1	M	Required feedback level 2 to 4 in the exercises.	I noticed him pressing the *OK* button on a level 2 feedback window immediately after it appeared. By and large, though, he seemed to be reading the feedback.
LA1	F	Was not able to find any correct answers in the ordering exercises, but in the	At first, she tried to find out the correct answers from her classmates, but the teacher reminded her that she had to work

		multiple-choice exercises, only required feedback level 1-3. The reason for this is discussed in the article.	independently, which she, after this point, did. Spent more time on reading feedback after she responded correctly.
LA2	M	Needed level 5 feedback in three out of five exercises. Needed levels 3 and 4 feedback in the other two.	When I observed him, he seemed rather concentrated not looking around or asking anything either from his teacher, me, or his classmates.
MA2	M	Required from no feedback up to level 5 feedback in the exercises.	On two occasions, I noticed that he skipped a feedback message, pressing the OK button immediately after the feedback message appeared.

APPENDIX D

Summarisation of the Participants' Questionnaire Responses.

Code	Overall usefulness of the feedback in the study	Useful feedback	Useless feedback	Comments
HA1	Useful, but only for finding correct answers, not for learning.	Both telling that the answer was correct and incorrect.	None	Received knowledge of results feedback.
HA2	Not useful.	Level 5	Levels 1 and 2	Reported that getting explicit explanations would remind him about the rules.
MA1	It helped him to find the correct answers in all the exercises and helped him to learn how to form questions.	All the feedback	None	-
LA1	Useless.	None	All the	Reported that she did not

			feedback	understand the meaning of the hints.
LA2	Somewhat useful, but did not help him to self-correct/learn anything.	Levels 4 and 5	Levels 1 and 2	Reported that the levels he rated as the most useful showed him his mistakes, whereas the useless did not.
MA2	Useful, as it helped him to self-correct in half of the exercises. Did not learn anything.	Levels 4 and 5	Level 1	The useful levels explained what his mistake was, whereas the useless did not.

III

L2 ENGLISH DERIVATIONAL KNOWLEDGE: WHICH AFFIXES ARE LEARNERS MORE LIKELY TO RECOGNISE?

by

Dmitri Leontjev

The article will be published in an upcoming volume 6 issue 2 of *Studies in Second Language Learning and Teaching,* 2016. Copyright SSLLT Journal © 2016.

L2 English Derivational Knowledge: Which Affixes Are Learners More Likely to Recognise?

Abstract

Knowledge of derivational morphology is considered an important aspect of vocabulary knowledge both in L1 (mother tongue) and L2 (second or foreign language) English language learning. However, it is still not clear whether different derivational affixes vary in their (learning) difficulty. The present study examines whether Bauer and Nation's (1993) teaching order of L2 English affixes can account for the difficulty learners have with recognising the affixes. The participants in the study were L1 Estonian and Russian learners of English at upper-secondary schools in Estonia (n=62). Their performance was measured on a word segmentation task. There were significant differences in the number of affixes the learners were able to successfully recognise at different levels, as classified by Bauer and Nation (1993). By and large, with the exception of no significant difference between level 5 and level 6 affixes, the higher the affix level was, the less likely the learners were to recognise the affixes at this level. I argue that these results can support the order proposed by Bauer and Nation. The implications of the finding for teaching and further research are also discussed.

Keywords: derivational morphology, affix difficulty, L2 English teaching

1. Introduction

A number of studies have revealed that L2 (second or foreign language) inflectional morphology poses problems for learners (e.g., Clahsen et al., 2010; Felser & Clahsen, 2009; Jiang 2004; Lardiere, 1998). At the same time, while L2 learners (and native speakers alike) face even bigger problems with derivational morphology, (Friedline, 2011; Schmitt & Meara, 1997; Schmitt & Zimmermann, 2002; Silva & Clahsen, 2008), not many studies on learners' word derivation knowledge and its acquisition have been conducted.

Friedline (2011, p. 60) suggests that the reason for the small number of studies on word derivation has been, until recently, the predominance of theories that argue for a clear dichotomy in morphology, such as Split Morphology Hypothesis (Perlmutter, 1988), which states that whereas inflection is rule-based, derivation only occurs in the lexicon. Nevertheless, more recent advances in morphology research, especially in the field of psycholinguistics (e.g., Alegre & Gordon, 1999; Clahsen & Neubauer, 2010), suggest that at least some derived words can be processed within the same rule-based system as (some) inflected words. Therefore, more research into word derivation and its acquisition is necessary, the more so as many questions, including how exactly learners acquire L2 word derivation knowledge, remain unanswered.

As regards L2 inflection, some relatively early morphological studies (e.g., Bailey, Madden, & Krashen, 1974), but also later studies (e.g., Pienemann, 1998), sought an answer to the question of whether there is a universal order of acquisition of L2 inflectional morphemes. With a similar objective in mind, using the research findings on the English affixes available at that time, Bauer and Nation (1993) classified L2 English affixes (both derivational and inflectional) into seven levels. The levels ranged from considering each form a different word (level 1), to classical roots and affixes (level 7). Later, Nation (2001) refined the classification, adding a number of affixes to the levels and limiting the list to derivational affixes only.

Bauer and Nation (1993) suggested that the levels could be used as a framework for teaching/learning affixes for reading in English. They further proposed that the levels could reflect what should be included in word families at different levels of learners' morphological awareness and be used as a reference point in empirical research on the development of word derivation knowledge. Nevertheless, up until the present time, this order has not been unambiguously confirmed or rejected empirically both as a difficulty order and the order in which learners do indeed acquire derivational affixes, or at least some of their aspects.

The present study endeavours to find evidence for Bauer and Nation's (1993) proposal aiming at confirming that the levels they defined reflect the increasing difficulty learners have with recognising the affixes, that is, the question the study aims to answer is:

- Does the difficulty learners have with recognising derivational affixes differ significantly across the affix levels as classified by Bauer and Nation (1993), increasing as the level grows?

I will discuss Bauer and Nation's (1993) study in some detail and other research relevant for the present study in the following section. I will then present the study and the analyses, report on the findings, and suggest some research to follow which could reinforce the findings.

2. Background

In the present section, I will provide further details on Bauer and Nation's (1993) and Nation's (2001) teaching order of L2 English affixes as well as discuss the studies that used their classification or tried to challenge it. I will also discuss some (further) factors that can offer an explanation for the difficulty learners have with word derivation. Hereinafter in the paper, the levels will be referred to as Bauer and Nation's levels.

2.1. Bauer and Nation's Affix Levels

Bauer and Nation (1993) based their classification of affixes on the following criteria:

- frequency,
- productivity,
- predictability of the meaning of the affix,
- regularity of written/spoken form of the base,
- regularity of spelling / phonological form of the affix,
- regularity of function.

It is evident, and the authors themselves acknowledged it, that the criteria are not unique to Bauer and Nation's (1993) study. Similar criteria were found to explain the acquisitional order of inflectional affixes (e.g., Goldschneider & DeKeyser, 2001), but were also used much earlier, for example, by Thorndike (1942). As the levels were defined with recognition / understanding during reading in mind, the priority was given to the written forms. The levels as identified by Bauer and Nation (1993) are presented in **Table 1**.

TABLE 1. Difficulty order of L2 English affixes (Bauer & Nation, 1993; Nation, 2001).

Level 1	A different form is a different word.
Level 2	Regularly inflected words are part of the same family, e.g., -ed, -ing, -s, etc.
Level 3	The most frequent and regular derivational affixes: -able, -er, -ish, -less, -ly, -ness, -th (fourth), -y, non-, un- (unusual)*.
Level 4	Frequent and regular affixes, e.g., -al (coastal), -ation, -ful, -ism, -ist, -ity, -ise (-ize), -ment, -ous, in-*.
Level 5	Infrequent but regular affixes, e.g., -age, -al (arrival), -ance, -ant, -ship, en-, mis-, un- (untie), etc.
Level 6	Frequent but irregular affixes, e.g., -ee, -ic, -ify, -ion, -ition, -pre-, re-, etc.
Level 7	Classical roots and affixes, e.g., -ate, -ure, etc.

*All with restricted uses; see **Appendix 1** in Bauer and Nation (1993) for details.

Both Bauer and Nation (1993) and Nation (2001) stressed that there was no empirical evidence for the order. On the other hand, the authors encouraged researchers to use the levels as a reference for affix difficulty in their studies. I will discuss the studies that utilised Bauer and Nation's levels in the following subsection.

2.2. Studies Using Bauer and Nation's Levels

Bauer and Nation's levels have been used to operationalise L2 English affix difficulty in several studies. Schmitt and Meara (1997), for example, used the levels when creating their instruments in a longitudinal study of 95 learners of English. Being the first to test the interplay between different aspects of vocabulary knowledge empirically, the authors used word derivation knowledge as one of these aspects. There was a significant, albeit small, increase in the participants' suffix knowledge over the course of the academic year (5% in the productive measure and 4% in the receptive one). The authors, however, did not find any noticeable differences between the suffixes in terms of their difficulty, which could be because they used only two or three different suffixes at each level and only one level 7 suffix.

Similarly, Schmitt and Zimmermann (2002) used the levels to control for the difficulty of the word forms across the word classes in their instrument. However, the aim of their study was to find out which parts of speech learners were the most likely to produce. Thus, the authors did not present any data that could allow for making assumptions regarding the difficulty their participants had with affixes at different Bauer and Nation's levels nor discussed their data in terms of a potential implicational order of derivational affixes.

The authors also considered possible reasons for the difficulty that word derivation poses to L2 learners. They drew on the work of Jiang (2000), according to whom syntactic and especially morphological specifications are integrated into the lexical entry during the last stage of learning a word. Drawing on the morpheme acquisition studies (e.g., Larsen-Freeman, 1976), natural language acquisition studies (e.g., Lardiere, 1998), and psycholinguistic research (e.g., Gollan, Foster, & Frost, 1997), Jiang (2000) also claimed that by the time this latter stage is reached, many words have become fossilised.

This was an important point raised. As a matter of fact, there is psycholinguistic research demonstrating that, at least in oral-aural processing, L2 learners often process meaning before they process form and often rely on lexical and semantic cues rather than morphological and syntactic cues during lexical processing (e.g., VanPatten, 1996). Jiang's lexical development model can serve an explanation for that finding. On the other hand, there is also research (e.g., Clahsen & Neubauer, 2010) showing that L2 learners rely on frequency when processing derived words, that is, processing more frequent words as wholes, which can explain the usual superiority of processing meaning over processing form discussed by VanPatten (1996) and expands on Jiang's (2000) model. Specifically, more frequent L2 words may be stored and processed as wholes (perhaps due to fossilisation, according to Jiang), whereas attempts are made to analyse less frequent words, which also presents a difficulty for learners in the light of Jiang's discussion. Clahsen and Neubauer's (2010) finding is also in line with the Declarative-Procedural model (e.g., Ullman, 2004), according to which, there are two systems involved in processing: procedural, which is rule-based and includes the processing of both inflection and derivation, and declarative for storing/retrieving frequent lexical entries as wholes. These studies present a rather strong case for controlling for frequency in word derivation research, suggesting that the more frequent morphologically complex L2 words are, the less likely they are to be analysed by learners.

Another study that used Bauer and Nation's levels was conducted by Hayashi and Murphy (2010). Their study aimed at comparing the ability to derive words of L1 (mother tongue) Japanese learners of English (*n* = 22) and adult native speakers of English (*n* = 20). The study also aimed at finding a relation between learners' size of vocabulary and their morphological awareness. The authors used affixes from different Bauer and Nation's levels as a way to establish the frequency and productivity of the affixes they used in the instruments—a word segmentation task as a measure of receptive morphological awareness and an affix elicitation task as a productive measure of it.

The authors did not elaborate on their decision to use a word segmentation task as a measure of receptive morphological awareness. However, Friedline (2011), for example, used a similar format in one of his instruments, asking the participants to write the base forms of the given derived words. He drew, above all, on the findings of Carlisle (2000) and Carlisle and Fleming (2003), which confirmed the prediction made in Schreuder and Baayen's (1995) model of morphological processing that children are able to define novel morphologically complex words in their mother tongue when they have access to corresponding bases and bound morphemes. Despite the lack of research confirming whether the same is true for L2 English, word segmentation/decomposition task types seem to be useful for establishing whether L2 learners have access to / can recognise affixes and bases in English morphologically complex words.

Hayashi and Murphy (2010) also considered semantic transparency of the items, that is, the degree to which the meaning of a whole morphologically complex word can be understood from the meaning of its parts, as one of the factors. They checked whether semantic transparency influenced their participants' morphological awareness, as previous research (e.g., Marslen-Wilson, 2007) has demonstrated that semantic transparency, among other factors, influences the processing of morphologically complex words. Having completed the qualitative evaluation of the participants' performance, Hayashi and Murphy (2010) discovered that semantic transparency influenced the way the learners performed on the word segmentation task. Specifically, they found that all of their participants were able to correctly separate affixes in the items *disorder* (level 7), *enable* (level 5), *rewrite* (level 6) and *childhood* (level 5), which were formed with help of affixes at different Bauer and Nation's levels, but which, arguably, were all semantically transparent. Judging by this finding, but also by the previous studies, semantic transparency should be taken into consideration in word derivation research, especially if it aims at establishing an implicational order of derivational affixes.

To my knowledge, there are two studies that tried to find a difficulty order (or an order of acquisition) of L2 English derivational affixes. One of them was the study of 403 Japanese learners of English conducted by Mochizuki and Aizawa (2000). The authors evaluated the learners' knowledge of suffixes and prefixes on two non-word tasks, operationalising suffix knowledge as the ability to identify the part of speech formed with the help of the suffixes and prefix knowledge as the receptive knowledge of the meaning of the prefix. They also had an interesting way of defining affix acquisition, suggesting that affixes known by more learners are acquired earlier whereas those known by fewer learners, later. The affix order they established had several discrepancies with Bauer and Nation's levels, for example, suffix -*er* (level 3) being more difficult that suffix -*ation* (level 4).

One of the issues that Mochizuki and Aizawa's (2000) study had was the authors' operationalisation of suffix knowledge. It is logical to assume that syntactic function of affixes should be a part of learners' word derivation knowledge. However, limiting word derivation knowledge to syntactic function only seems to be an overgeneralisation. Moreover, the way the authors defined affix acquisition should rather be considered the order of difficulty the learners had with the affixes. Finally, as the authors mentioned, the order they established could have been affected by English loan words in Japanese.

The second study that aimed at finding a difficulty order of derivational suffixes (among other research questions) was Chuenjundaeng's (2006) Masters thesis. For this purpose, the researcher used an instrument consisting of two translation tasks including sixteen base and sixteen derived forms (eight base and eight derived forms per task), the base form in one task being the derived form in the other and vice versa. The second task was set a week after the first one. The suffixes that the author selected for the instrument were -*er*, -*tion*, -*ment*, and -*ity*. The Thai learners of English (n = 167) were asked to provide a definition / translation of the words in the tasks in their mother tongue. Their responses to each item were classified into four categories: '1' when they provided definitions for both the base and the derived form, '2' when they provided definition for only the base form, '3' when it was only the derived form, and '4' when they failed to define both the base and the derived form. The responses in category 1 were awarded the score of two, that is, one score for both the base and the derived form each. The responses in categories 2 and 3 were awarded the score of one, that is, one point for either the base or the derived form. The score of zero was given for the items in the last category. The author used the composite score on categories 1 and 3 as an indication of the learners' knowledge of the derived words formed with the four affixes she studied. Based on the results, the author identified the following increasing difficulty order of the suffixes: *-tion* (the total score on categories 1 and 3 being 216), *-er* (the total score of 206), *-ity* (the total score of 154), and *-ment* (the total score of 143).

There are discrepancies between the order found by Chuenjundaeng (2006) and that found by Mochizuki and Aizawa (2000). On the other hand, it is hard to say whether the difficulty order found by Chuenjundaeng (2006) agrees with Bauer and Nation's levels or not, as all the suffixes that the author selected except for *-er* were at level 4 of Bauer and Nation's (1993) classification. Moreover, the difference between the scores on *-er* and *-tion* was small, and the author treated *-ation*, *-ion*, and *-ition* as allomorphs of the same suffix at level 4 of Bauer and Nation's (1993) classification. The latter is not entirely incorrect.

Indeed, Bauer and Nation (1993) themselves discussed the issue of the suffix *-ation* and its allomorphy and admitted the problem of determining whether *-ation*, *-ion*, and *-ition* should be considered the allomorphs of the same suffix or not. They, however, decided that only *-ation* should be included at level 4.

In the lack of evidence for (or against) the order proposed by Bauer and Nation (1993), the present study sets to determine whether Bauer and Nation's levels reflect the increasing difficulty learners have with L2 English derivational affixes. However, unlike the two studies discussed above, instead of studying learners' performance on separate affixes, I will consider the affixes at each Bauer and Nation's level as a group.

3. Methodology

3.1 Materials

To answer the research question, I analysed the learners' performance on a word segmentation task, a task type also used by Hayashi and Murphy (2010) and somewhat similar to Friedline's (2011) decomposition task. The purpose of the task was to find out how likely learners were to recognise affixes at different Bauer and Nation's levels. However, instead of trying to challenge Bauer and Nation's (1993) classification by studying separate affixes, as the previous studies did, I looked at the affixes at each of Bauer and Nation's level collectively, studying affixes at each of the levels as a group.

Another difference from the previous studies, specifically, Friedline (2011) and Hayashi and Murphy (2010), concerned the items used in the instrument and the learners' selection procedure. What neither of the authors of the two studies did was controlling for the possibility that the items might have been known by their participants and thus exhibiting the frequency effect. Moreover, Friedline (2011) did not control for the potential effect of the semantic transparency. Hayashi and Murphy (2010) did account for that. They, however, did not establish whether the words they selected as the items in their instruments were semantically transparent to their participants and instead rated the items themselves. As with any judgmental phenomenon, not only could dissimilar ratings have been produced by other raters, but it was also not known whether the leaners were actually able to discern the meanings of the words that the authors rated as semantically transparent from the meanings of the bases and the affixes these words were composed from. That is to say, it is not known whether these words were semantically transparent to the participants in their study.

I addressed the issue differently, and instead of producing figures for the frequency and the semantic transparency of the items and controlling for these while analysing the learners' performance, I made sure that the learners did not know the words selected for the task before I started the analyses.

For the item selection, I used Affix Levels @ Frequency Tester instrument from Compleat Lexical Tutor website (http://www.lextutor.ca/cgi-bin/morpho/fam_affix/index.pl). This instrument classifies words, or, rather, base words and word families, in the British National Corpus into frequency bands by thousand most frequent words / word families. The instrument contains the first twenty thousand most frequent word families, breaking them into twenty frequency bands. It then separately lists derived words formed with affixes at different Bauer and Nation's levels at each of the frequency bands. That is to say it allows for singling out words of certain frequency (or, rather, frequency of their bases) formed with affixes at particular Bauer and Nation's levels.

For the word segmentation task, I decided to select words formed with affixes at Bauer and Nation's level 3 to level 6. A total of 12 words per affix level were randomly selected such that there were three words formed with affixes at each of Bauer and Nation's levels selected at each of 5,000, 6,000, 7,000, and 8,000 most frequent base words / word families. There were, however, several constraints to the otherwise random selection.

First of all, there were not more than three words formed with the same affix and before the selection started, words that might be known to the participants were removed. Thus, for example, words denoting languages, such as *Croatian*, were excluded. Secondly, words containing two suffixes, such as *momentousness* were excluded as well. At the same time, the instrument included four words formed with both a prefix and a suffix.

A further 6 words not including any derivational affixes were selected to serve as distractors. All in all, the instrument included a total of 50 items, of which 44 were formed with help of a total of 48 affixes, of them 10 prefixes. With the exception of level 5, there were two prefixes per Bauer and Nation's level. Since there are considerably more prefixes at level 5 than at the rest of Bauer and Nation's level, having four items formed with prefixes at level 5 reflected this overall tendency. After the selection, the order of the items was randomised. The items in the present version of the instrument are presented in **Table 2**.

To make sure that the learners did not know any of the words in the task, and thus to account for a possible frequency effect, the participants were asked to supply translations or definitions for the items. Arguably, this also allowed for control of whether the items were semantically transparent for the learners, on the assumption that they would supply a definition or translation for any item which meaning they could deduce from the meaning of the bases and the affixes.

Identifying bases could compromise the results, as this could have allowed the learners to separate affixes without actually having to recognise them. Therefore, it was decided to exclude the performance of those learners whose translations or definitions indicated that they identified any of the bases the items.

Asking the participants to find prefixes and suffixes, invariably meant that they had to refer to their metalinguistic knowledge to complete the task, and metalinguistic knowledge has been found to present a problem even to native speakers (e.g., Alderson, Clapham, &

Steel, 1997). On the other hand, the advantages of the format, above all, the possibility to control for the influence of the participants' vocabulary knowledge, and the suitability of the task type for answering the research question, that is, finding out how well learners are able to recognise derivational affixes at different Bauer and Nation's levels outweighed this limitation in my opinion.

The instrument was piloted among five learners of English whose proficiency on the Common European Framework of Reference (CEFR) was at about B1 level, as they studied at grades 9 to 11 of Estonian schools (see Põhikooli riiklik õppekava õigusakt, 2010). The reason for selecting learners at this particular level of proficiency for the piloting (and also for the present study) was Nation's (2001) suggestion that the best time for starting teaching L2 affixes to learners would be when they are at lower-intermediate level of their L2 proficiency, that is at about level B1 on the CEFR scale (Council of Europe, 2001).

A major aim of the piloting was to establish whether learners would be able to recognise the affixes without being able to define the words or their bases. Based on the results of the piloting, several items were replaced, as one or several learners provided translations or definitions for these words. Among the items that had to be replaced were three words at level 6. The problem I faced with at this stage was that at the base frequency bands selected for creating the instrument, there were no alternatives to the selected words, as the rest either contained the suffix -ion, which three other items in the instrument already contained, or were easy to define (e.g., *atomic* or *combative*). Thus I decided to select two items at frequency bands 9 and 10 instead (i.e., among the 9,000 and 10,000 most frequent word families) and selected *mortify* as a new item at frequency band 7 which is not listed in the Affix Levels @ Frequency Tester instrument, but nevertheless belongs to this frequency band, as the Affix Levels @ Frequency Tester instrument does not include words containing bound base morphemes. For a similar reason, *unambiguous* replaced the item *unconstitutional*. The items used in the present version of the word segmentation task are presented in **Table 2** and **Appendix A**.

TABLE 2. Items in the word segmentation task.

	Frequency bands				
	5,000	**6,000**	**7,000**	**8,000**	**9,000/10,000**
Level 3	indiscreet**ly** **un**ambiguous* void**able**	bland**ness** croft**er** obscene**ly**	Brisk**ness** decipher**able** lush**ness**	brim**less** stout**ly** **un**shackle	
Level 4	boast**ful** **in**discreetly unambigu**ous***	enshrine**ment** **in**apt reaffirm**ation**	Arson**ist** frugal**ity** solemn**ise**	discern**ment** pail**ful** slander**ous**	
Level 5	disciple**ship** **inter**lace moist**en**	**en**shrinement err**ant** **mis**apprehend	Defer**ence** **en**mesh repent**ant**	bestow**al** deflation**ary** tern**ary**	
Level 6**	exempt**ion** herald**ic** obstruct**ive**	digress**ion** prohibit**ive** **re**affirmation	Evict**ion** mort**ify** **re**coup	detain**ee**	regress**ive** cherub**ic**
Distractors	abolish, bulletin, comprise, magnitude, mediocre, scrutiny				

* Item *unambiguous* (with a bound base morpheme) replaced an item that was found unacceptable during the piloting.
** Two items from the 9,000 and 10,000 most frequent word families and item *mortify* (with a bound base morpheme) were added to replace three items that were found unacceptable during the piloting.

3.2 Participants

The participants in the study were seventy-six L1 Estonian and L1 Russian learners of English studying at grade 10 in Estonian schools. However, fourteen of them supplied more or less accurate translations or definitions to one or several items or their bases (e.g., *wetness* for the item *moisten*), so their results were not included in the analysis. The final sample included a total of sixty-two learners from six different groups taught by five different teachers.

Although the proficiency level of most participants was expected to be at about level B1 on the CEFR scale, this assumption was corroborated by asking the learners to self-assess their writing and reading proficiency using the CEFR descriptors from the self-assessment scale (available at http://www.keelemapp.ee/keelemapp/keelemapi-osad/). Furthermore, the teachers were also asked to assess their learners' reading and writing proficiency using the same scale. Then, the median across the four ratings was calculated and used as the measure of the learners' proficiency in the study. While such judgmental figures should not be considered very reliable, since it was just a background variable, it was found sufficient for the purposes of the present study. Moreover, the agreement between the ratings of the learners and the teachers as calculated by Kendall's tau b was substantial, $r_k = 603$, $p < .001$ for reading and $r_k = 492$, $p < .001$ for writing, which added to the reliability of the figure. Further details on the participants are presented in **Table 3**.

TABLE 3. Description of the participants.

		CEFR level*				**Median**
L1	**N**	**A2**	**B1**	**B2**	**C1**	**CEFR level**
Estonian	27	3	17	6	1	B1
Russian	35	12	14	9	-	B1

*The proficiency estimate was calculated as the median across the learners' self-assessment of their reading and writing proficiency on the CEFR self-assessment scale and the assessment of their reading and writing proficiency on the same scale completed by their teachers.

3.3 Procedure

Before the start of the study, the learners were informed that the study aimed at finding out how well they could recognise suffixes and prefixes in English. They were then given a written description of the study, which also detailed that they were expected to assess their abilities with the help of a self-assessment scale and complete one exercise. The learners were also informed that in the study, group results rather than those of individual learners would be analysed. They then gave their permission to use their performance for calculating the group statistics.

In order for the learners not to get discouraged by failing to define all (or, at least, most) of the words in the task, they were instructed that they should not be worried if they did not know any of the words in the task, as the task was rather difficult. The same was restated in the written instructions (**Appendix A**).

It was also stressed that the learners were expected to work individually and that there was no purpose in cheating. In addition, the teachers were asked to help monitor the learners' performance. The task was not speeded, and, as the piloting had also established, it took the learners about twenty minutes to complete.

4. Results

In the present section, the results of the study will be presented. First, I will present the overall results, including the reliability estimate for the task. Following that, I will present the results that allowed me to find the answer to the research question posed in the study.

As has been mentioned earlier, the performance of 62 learners was analysed. The learners' performance was scored such that each correctly recognised affix was awarded one point.

The Cronbach's alpha for the 48 items was .89, which suggested that the internal consistency of the instrument was rather high. Moreover, none of the learners scored a zero, and there was no item in the task in which none of the learners was able to recognise the affix. This showed that there was no floor effect observed in the task. The mean number of affixes recognised by the learners was 21, *SD* = 8.2, with the weighted average of 34.85 in the 95th percentile, which means that those who scored the highest on the task were able to recognise affixes in about 73% of the items, that is, there was no ceiling effect either.

The L2 English proficiency of the two L1 groups of learners was roughly the same (see **Section 3.2**). Moreover, the L1 Estonian learners (*n* = 27) did not perform significantly differently from the L1 Russian learners (*n* = 35) on the segmentation task either, as demonstrated by an independent-samples t-test, *t*(60) = 0.61, *p* = .55. Thus, in the following, for the most part, the performance of the two groups will be considered together. I will, however, corroborate the main analysis by comparing the two L1 groups, too.

To discover whether the learners were able to recognise affixes at different Bauer and Nation's levels to a different degree, a composite score was calculated separately for the affixes at each of Bauer and Nation's levels, the maximum possible score being 12 at each level. The descriptive statistics are presented in **Table 4**. For the sake of comparison, I also supplied the means and the medians separately for each of the L1 groups.

TABLE 4. Affixes correctly recognised at different Bauer and Nation's (1993) levels (*n* = 62; *k* = 12 at each of the levels).

		Mean		**95 % CI of the mean**		**SD**	**Median**	
				Lower	**Upper**			
Level 3	Estonian (n = 27)	9.48	8.92	8.08	9.76	3.29	11	10
	Russian (n = 35)	8.49					9	
Level 4	Estonian (n = 27)	6.04	5.63	5.03	6.23	2.37	7	5
	Russian (n = 35)	5.31					5	
Level 5	Estonian (n = 27)	3.41	3.23	2.68	3.77	2.16	3	3
	Russian (n = 35)	3.09					3	
Level 6	Estonian (n = 27)	2.78	3.21	2.54	3.88	2.66	2	2
	Russian (n = 35)	3.54					2	

From the descriptive statistics, it can be deduced that with the exception of almost no difference between level 5 and level 6 affixes, the numbers of affixes recognised at different Bauer and Nation's levels were rather different, and the higher the level was, the less affixes were recognised. Specifically, on average, the learners recognised about 75% of all the affixes at level 3, about a half at level 4, and about a quarter at levels 5 and 6 respectively.

However, from the descriptive statistics, it was not clear whether the differences between the levels were statistically significant. Thus, a repeated measures ANOVA was conducted, the number of affixes recognised at different Bauer and Nation's levels forming the within-subjects factor.

The repeated measures ANOVA, with the Greenhouse-Geisser correction of the degrees of freedom applied as the sphericity assumption was violated, confirmed that there was a significant difference in the learners' ability to recognise derivational affixes at different Bauer and Nation's levels, $F(2.44, 149.08) = 117.66$, $p < .001$, $\eta_p^2 = .66$[1]. The effect size value indicated that the affix levels accounted for 66% of all the variance in the learners' performance, which is a very strong effect. Graphically the learners' performance is presented in **Figure 1**.

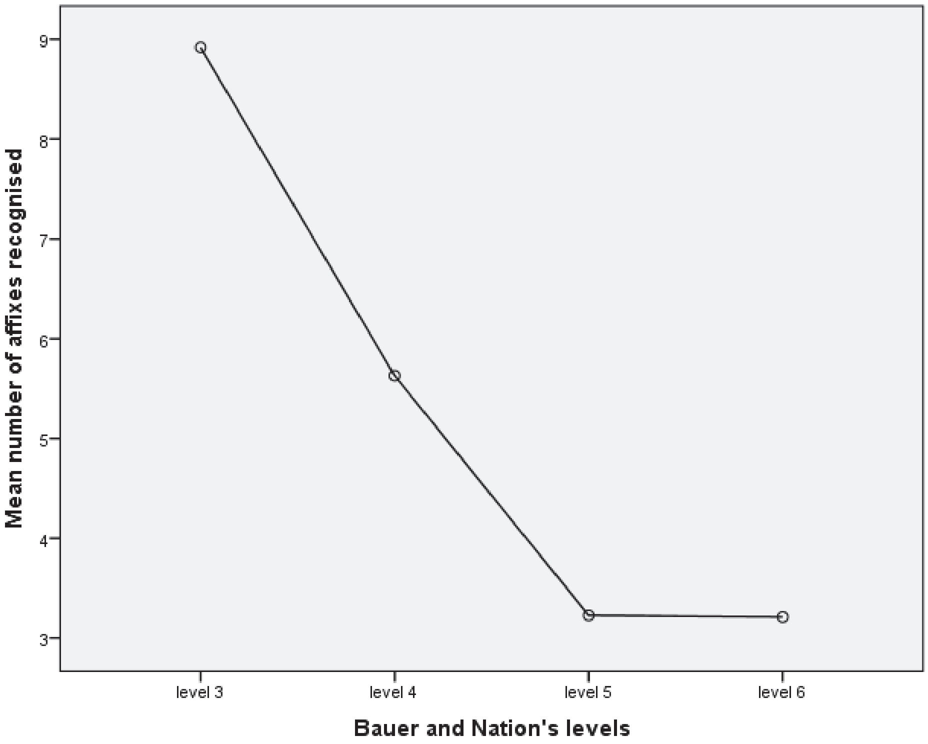


FIGURE 1. Mean number of affixes recognised at different Bauer and Nation’s (1993) levels.

I then compared the means at the affix levels pairwise, in essence conducting a series of post-hoc tests, using the Bonferroni correction to account for the family-wise error (**Table 5**).

TABLE 5. Pairwise comparisons.

Levels	**Mean difference**	**Significance***
Level 3 and 4	3.29	< .001
Level 3 and 5	5.69	< .001
Level 3 and 6	5.71	< .001
Level 4 and 5	2.4	< .001
Level 4 and 6	2.42	< .001
Level 5 and 6	0.02	n.s.

*The *p*-values were adjusted using the Bonferroni correction

The pairwise comparisons, thus, demonstrated that with the exceptions of no significant difference between levels 5 and 6, all the differences were significant. In fact, the trend was the same if the groups were compared separately, as can be seen in **Figure 2**.

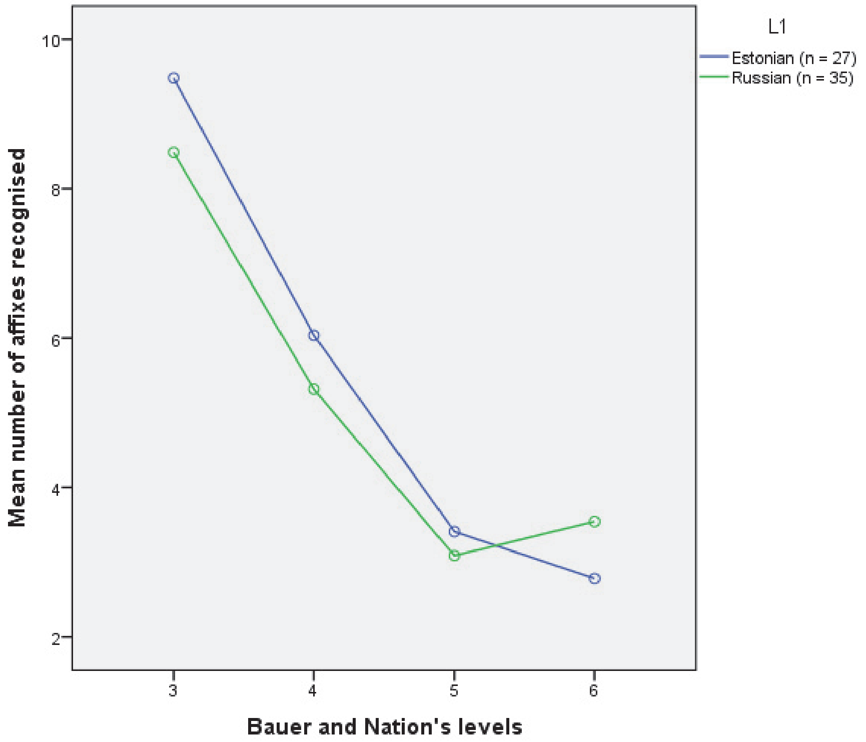


FIGURE 2. Mean number of affixes recognised at different Bauer and Nation's (1993) levels by L1 Estonian and L1 Russian learners.

Moreover, the repeated measures ANOVAs and the pairwise comparisons conducted separately for each of the L1 groups demonstrated the same as the analysis conducted for the whole sample. What is more, while, with the exception of the performance on level 6 affixes, the L1 Estonian group slightly outperformed the L1 Russian group (**Table 4**; **Figure 2**), there was no significant difference in the ability of either of the two groups to recognise affixes at any of Bauer and Nation's levels, as demonstrated by a series of the independent-samples t-tests. For example, the biggest mean difference of 0.99 between the performance of the two groups was in the ability to recognise the level 3 affixes, and the t-test demonstrated that this difference was not statistically significant, $t(60) = 1.18$, $p = .241$.

This being said, the results do not imply that the learners found all the affixes at level 3 easier to recognise than affixes at level 4 or all the affixes at level 6 harder to recognise than level 4 affixes. For example, many learners recognised prefix *re-* in *recoup* (45 learners, i.e., 73%) and reaffirmation (30 learners, i.e., 48%), which were much higher numbers than those who recognised the rest of the affixes at level 6 (ranging from 3 to 17 learners). This suggests that the prefix was rather easy to recognise. In fact, these numbers are comparable to those who recognised many of the level 4 affixes in the task (e.g., prefix *in-* in *inapt* recognised by 30 learners). Similarly, the numbers of learners who recognised suffix -*ful* in *pailful* (55 learners, i.e., 89%) and in *boastful* (54 learners, i.e., 87%) were higher than the number of learners who recognised, for example *-ly* in *stoutly* (51 learners, i.e., 82%), the latter being a level 3 affix.

What is more, the results do not reveal why substantially different numbers of learners recognised the same affix in different items, such as prefix *re-*, as illustrated in the previous paragraph. Other examples include suffix *-ly,* which was recognised in *stoutly by* 51 learners but only by 42 (68%) in *indiscreetly* or suffix *-ary* (level 5), which was recognised by 15 learners (24%) in *deflationary*, but only by 2 (3%) in *ternary*.

5. Discussion

The present study aimed at finding empirical evidence for (or against) the order of L2 English affixes proposed by Bauer and Nation (1993). Differently from the previous research, I did not challenge Bauer and Nation's levels by looking at separate affixes, but instead considered affixes at different levels as groups. The potential influence of frequency effect (e.g., Clahsen & Neubauer, 2010) was countered by making sure that the learners did not know the words in which they were asked to find the affixes. Arguably, this allowed for control of the influence of semantic transparency as well.

The results demonstrated that with the exception of no difference between the learners' ability to recognise level 5 and level 6 affixes, the higher Bauer and Nation's level was the less affixes on average the learners recognised at this level. That is to say, their ability to recognise the affixes, for the most part, followed the affix order proposed by Bauer and Nation (1993). What is more, the difficulty order of the affixes accounted for 66% of all the variance in the learners' performance, which is a large effect and should thus be considered a rather strong evidence for the order proposed by Bauer and Nation (1993).

The difference between level 5 and level 6 affixes was not statistically significant. The reason for that can, in part, be attributed to the fact that many learners recognised prefix *re-*. As a matter of fact, Mochizuki and Aizawa (2000) found that the meaning of prefix *re-* (i.e., its most common meaning of *again*) was recognised by the largest number of learners as compared to other prefixes, which can explain the ease of recognition of the prefix in the present study. What is more, Bauer and Nation (1993) accounted for the undoubted productivity of the prefix, that is, there is a possibility that the learners in the present study met this prefix quite often.

On the other hand, Bauer and Nation (1993) rightfully noted that because of the number of meanings prefix *re-* has in addition to *again* and *anew* and a number of tokens with *re-* that have become lexicalised, learners can end up misanalysing words containing *re-* if they learn the semantics of this prefix. I would, however, suggest that this assumption might be reconsidered in future. Prefix *re-* is present in many languages and its meanings, at least in some of them, are similar to those it has in English. Specifically, it is the case with Estonian and Russian. In Estonian, one can find it, for example, in *representatiivne* (representative), where it has an intensifying meaning, and in Russian in *репродукция* [*reproduktsiya*] (reproduction), where it has the meaning of *again*. Thus, there could have been the influence of the mother tongue that influenced the learners' performance on prefix *re-* as well. Further studies, for example, teaching experiments, can shed more light on whether it indeed makes sense to classify the prefix to an earlier stage.

Connected to the previous discussion, it should not be assumed that the learners found all the affixes at level 3 easier to recognise than the level 4 affixes and all the affixes at level 4 easier to recognise than those at levels 5 and 6. One illustration of the opposite is the learners' recognition of *re-* in the task. Another example could be the learners' performance on the items with *-ful*. As regards the latter suffix, it has been found by Mochizuki and Aizawa (2000) that more learners were able to indicate the syntactic role of suffix *-ful* than suffixes at level 3, such as *-ly* and *-er*, which can serve an explanation for the finding of the present study. What is more, the meaning of suffix *-ful* also seems easy to remember, as it is the same as that of the word *full*, a very frequent word. Presumably, these could be the reasons for the number of learners recognising the suffix in the study.

However, I would refrain from making any claims regarding the potential reclassification of these affixes. The results of the present study do not allow for establishing reasons for the learners' performance on affixes like *re-* or *-ful*. Moreover, it is hard to say whether comparable numbers of learners recognising the same affixes will be found in future studies. Judging by the previous studies that examined separate affixes and produced dissimilar orders of difficulty, this might not be the case. What is more, while the participants in the present study did not know the meanings of the words in the task, which reduced, if not excluded, the possibility of the effects of frequency and semantic transparency, it is not clear why different numbers of learners recognised the same affixes in different items. I would suggest that before trying to find answers to these questions, more research into L2 English word derivation is required, which should increase our understanding of what is included in L2 English word derivation knowledge and how it develops.

Thus, it is best to interpret the results of the present study in the most straightforward way. That is to say, the results indicate that learners are able to recognise significantly less affixes at higher Bauer and Nation's levels than at lower levels; or, perhaps, that learners are more likely to recognise affixes at lower Bauer and Nation's levels than at higher levels.

It should also not be forgotten that I only measured the learners' ability to recognise the affixes. Thus, the results could have been different should I have studied their ability to recognise/recall the meanings of the same affixes, for example.

6. Conclusion

The present study aimed at establishing whether the classification of the English affixes proposed by Bauer and Nation (1993) can indeed account for the difficulty L2 English learners have with recognising derivational affixes. The results demonstrate that with the exception of the lack of difference between level 5 and level 6 affixes, the learners were more likely to recognise affixes at lower levels than at higher levels of difficulty as defined by Bauer and Nation (1993), thus providing evidence for the validity of the levels.

The findings have several implications, both theoretical and practical. The empirical confirmation of the difficulty order reinforces Bauer and Nation's (1993) proposal to use the levels as a reference for affix difficulty in morphological research. Moreover, Bauer and Nation's levels could be used as a starting point for establishing an/the order of acquisition of L2 English derivational affixes, if any. As far as pedagogical implications are concerned, teachers of English could take the levels into consideration when instructing their learners to refer to morphological knowledge when inferring the meanings of unknown words in texts. That is to say, the levels should help L2 English teachers to find which affix properties as defined by Bauer and Nation (1993) make it more likely that their learners will recognise the affixes.

These implications are not new. In fact, Bauer and Nation (1993) discussed these as possible applications of their classification. I, however, argue that the findings reported in the paper present a stronger case for doing so. Having said that, I feel that several limitations of the study should be listed, so that further studies could account for them.

One of the limitations of the study has already been mentioned in **Section 3.1**. To complete the task, the learners were also required to demonstrate their metalinguistic knowledge, which learners often have problems with (Alderson, Clapham, & Steel, 1997). On the other hand, the word segmentation task is, arguably, the best for determining how well the learners recognise derivational affixes. I could have rephrased the instructions, and, similarly to Hayashi and Murphy (2010), instead of mentioning prefixes and suffixes in the instructions, asked the learners to break the words into meaningful units. This, however, would mean that the learners would also have to identify the bases, and I wanted to minimise this possibility. What is more, if the instructions had been phrased without mentioning prefixes and suffixes, the learners might have misinterpreted what they had been asked to do (cf. Hayashi & Murphy, 2010).

Another limitation concerns the inability to say whether the results would be exactly the same if other affixes had been used. Using other, perhaps, more easily recognisable affixes at level 5, such as *-hood*, *post-*, or *neo-*, could result in learners recognising more affixes at level 5. Finally, the way I controlled for semantic transparency might have been insufficient. Further studies, using other affixes, and perhaps, a larger number of affixes, and a better control for semantic transparency could confirm or disprove the findings of the present study.

Furthermore, future studies could also determine whether learners find it easier to recognise or recall meanings and/or syntactic roles of affixes at different Bauer and Nation's levels taken as a group. This would strengthen the case for Bauer and Nation's levels or present evidence against them. In any case, I hope that the present study stimulates the research on L2 English word derivation knowledge.

Notes

1. The shapes of the distributions of level 3, level 5, and level 6 affixes recognised by the learners were not symmetric. Thus, I supplemented the repeated measures ANOVA analysis with a Friedman's test, which does not assume normality. The results of the Friedman's test corroborated the results of the ANOVA confirming that it was robust to the deviations from normality present in the variables, X^2(3, n = 62) = 114.08, p < .001. The pairwise comparisons also confirmed the results, demonstrating that there were significant differences between all the affix levels except for no difference between levels 5 and 6.

References

Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research, 1*(2), 93-121.

Alegre, M., & Gordon, P. (1999). Rule-based versus associative processes in derivational morphology. *Brain and Language, 68*(1-2), 347-354.

Bailey, N., Madden, C., & Krashen, S. (1974). Is there a "natural sequence" in adult second language learning? *Language Learning, 24*(2), 234-243. doi: 10.1111/j.1467-1770.1974.tb00505.x.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing, 12*(3), 169-190. doi:

10.1023/A:1008131926604.

Carlisle, J. F., & Fleming, J. (2003). Lexical Processing of Morphologically Complex Words in the Elementary Years. *Scientific Studies of Reading, 7*(3), 239-253. doi: 10.1207/S1532799XSSR0703_3.

Chuenjundaeng, J. (2006). *An investigation of SUT students' receptive knowledge of English noun suffixes.* (Unpublished MA thesis), Suranaree University of Technology, Thailand.

Clahsen, H., & Neubauer, K. (2010). Morphology, frequency, and the processing of derived words in native and non-native speakers. *Lingua, 120*(11), 2627-2637.

Clahsen, H., Felser, C., Neubauer, K., Sato, M., & Silva, R. (2010). Morphological structure in native and nonnative language processing. *Language Learning, 60*(1), 21-43. doi: 10.1111/j.1467-9922.2009.00550.x.

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment* [electronic version]. Retrieved August 21, 2014, from http://www.coe.int/t/dg4/linguistic/Source/Framework_en.pdf.

Felser, C., & Clahsen, H. (2009). Grammatical Processing of Spoken Language in Child and Adult Language Learners. *Journal of Psycholinguistic Research, 38*(3), 305-319. doi: 10.1007/s10936-009-9104-8.

Friedline, B. E. (2011). *Challenges in the second language acquisition of derivational morphology: from theory to practice.* Unpublished Ph.D. dissertation, University of Pittsburgh.

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "Natural Order of L2 Morpheme Acquisition" in English: A Meta-analysis of Multiple Determinants. *Language Learning, 51*(1), 1-50. doi: 10.1111/1467-9922.00147.

Gollan, T. H., Foster, K. I., & Frost, R. (1997). Translation priming with different scripts:

Masked priming with cognates and noncognates in Hebrew-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 112-239.

Hayashi, Y., & Murphy, V. (2010). An investigation of morphological awareness in Japanese learners of English. *The Language Learning Journal, 39*(1), 105-120. doi: 10.1080/09571731003663614.

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics, 21*(1), 47-77. doi: 10.1093/applin/21.1.47.

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics, 25*(04), 603-634. doi: 10.1017/S0142716404001298.

Lardiere, D. (1998). Case and Tense in the 'fossilized' steady state. *Second Language Research, 14*(1), 1-26. doi: 10.1191/026765898674105303.

Larsen-Freeman, D. (1976). An explanation for the morpheme acquisition order of second language learners. *Language Learning, 26*(1), 125-34.

Marslen-Wilson, W. (2007). Morphological processes in language comprehension. In G. Gaskell (ed.) Oxford Handbook of Psycholinguistics, 175–93. Oxford, UK: Oxford University Press.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: an exploratory study. *System, 28*(2), 291-304.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Perlmutter, D. (1988). The Split Morphology Hypothesis: Evidence from Yiddish. In M. Hammond & M. Noonan (Eds.), *Theoretical Morphology* (pp. 79-100): Academic Press, Inc.

Pienemann, M. (1998). *Language processing and second language development:*

processability theory. Amsterdam ; Philadelphia: J. Benjamins.

Põhikooli riiklik õppekava õigusakt; Lisa 2 [Basic School National Curriculum Act: Annex 2.] (2010). Pub. L. No. RT I 2010, 6, 22. [Retrieved March 14, 2015]. Available at https://www.riigiteataja.ee/aktilisa/1281/2201/0017/13275423.pdf.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition, 19*(1), 17-36.

Schmitt, N., & Zimmermann, C. B. (2002). Derivative Word Forms: What Do Learners Know? *TESOL Quarterly, 36*(2), 145-171. doi: 10.2307/3588328.

Schreuder, R. & Baayen, R. H. (1995). Modeling morphological processing. In L.B. Feldman (Ed.), Morphological aspects of language processing. Hillsdale, NJ: Lawrence Erlbaum.

Silva, R., & Clahsen, H. (2008). Morphologically complex words in L1 and L2 processing: Evidence from masked priming experiments in English. *Bilingualism: Language and Cognition, 11*(2), 245-260. doi: 10.1017/S1366728908003404.

Thorndike, E. L. (1942). The Teaching of English Suffixes*. Teachers College. 43*(8), 657-658.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition, 92*(1–2), 231-270.

VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Norwood, NJ: Ablex.

APPENDIX A

The word segmentation task (originally, the instructions were in the participants' mother tongues; the correct responses are highlighted).

Some of the words you see below are formed with help of prefixes **OR** suffixes **OR** both prefixes and suffixes. **Circle** the prefixes and suffixes in these words.

In some words there are both **a prefix and a suffix**. In some words, there are **only prefixes**. In some, there are **only suffixes**. In some of the words, there is **neither a prefix nor a suffix**.

If you know the meanings the words, **write** what they mean (a translation or a definition). But even if you don't know the meaning of any of the words, don't worry. These are very difficult words. Even if you know the meanings of one or two, you have got a rather good vocabulary.

The first two words are **examples**.

FARMER _____talunik______

DEASHED _________________

unshackle _________________

mediocre _________________

pail**ful** _________________

slander**ous** _________________

recoup _________________

brisk**ness** _________________

unambigu**ous** _________________

indiscreet**ly** _________________

void**able** _________________

comprise _________________

herald**ic** _________________

lush**ness** _________________

regress**ive** _________________

repent**ant** _________________

decipher**able** _________________

evict**ion** _________________

bland**ness** _________________

stout**ly** _________________

frugal**ity** _________________

croft**er** _________________

prohibit**ive** _________________

deflation**ary** _________________

inapt _________________

brim**less** _________________

enshrine**ment** _________________

abolish _________________

boast**ful** _________________

discern**ment** _________________

obstruct**ive** _________________

bulletin _________________

interlace _________________

solemn**ise** _________________

err**ant** _________________

mort**ify** _________________

cherub**ic** _________________

moist**en** _________________

obscene**ly** _________________

digress**ion** _________________

magnitude _________________

scrutiny _________________

disciple**ship** _________________

defer**ence** _________________

detain**ee** _________________

tern**ary** _________________

reaffirm**ation** _________________

enmesh _________________

arson**ist** _________________

bestow**al** _________________

exempt**ion** _________________

misapprehend _________________

IV

WORD DERIVATIONAL KNOWLEDGE AND WRITING PROFICIENCY: HOW DO THEY LINK?

by

Dmitri Leontjev, Ari Huhta, and Katja Mäntylä

This article is to be published in the upcoming volume 58 of *System* (2016),
copyright Elsevier (2016), doi: 10.1016/j.system.2016.03.013

Word derivational knowledge and writing proficiency: How do they link?

Although word derivational (WD) knowledge, i.e., how new words are formed from existing words with help of derivational affixes, is considered important for learners of second or foreign languages (L2), there is still no clear answer as to what aspects comprise the construct of L2 English word derivational knowledge and how it develops. The present study adds to our knowledge on how the ability to derive English words develops among L2 English learners. More specifically, it sheds light on how word derivational knowledge relates to communicatively defined Common European Framework of Reference (CEFR) language proficiency levels regarding learners' writing skills. In the study, 117 10th grade learners of English in Estonia and Finland were administered two writing tasks as well as nine measures which were hypothesised to tap learners' word derivational knowledge. The findings indicated that the learners' performance on almost all WD measures were significantly and fairly strongly (at .4–.6 level) correlated with their writing proficiency. The findings also suggest that some aspects of WD ability develop rather steadily between CEFR levels, but others may increase more rapidly after level A2 or B1. These findings thus demonstrate a relationship between word derivational knowledge and language proficiency.

Keywords: *word derivation, L2 proficiency, CEFR, L2 writing*

1. Introduction

Studies that combine language testing and second language acquisition (SLA) research have become more common in the past few decades (e.g. Glaboniat et al. 2005; Bartning, Martin, & Vedder, 2010; Carlsen, 2013; see also Bachman and Cohen, 1998). One reason for this development is the introduction of the Common European Framework of Reference, CEFR, (Council of Europe, 2001). The development of CEFR has created an interest in Europe in how language learners' communicative ability in a foreign or second language (L2), as described in the CEFR levels, develops in terms of linguistic elements of proficiency, that is, vocabulary and structures (Bartning, Martin, & Vedder, 2010). Some of the questions that arose in relation to CEFR included finding out whether the CEFR levels can be distinguished with reference to particular linguistic features or their combinations or to what extent such patterns of linguistic features might depend on learners' first language (L1) or the language they are learning. An interest in finding answers to such questions has characterised the work of several projects across Europe and across several languages such as English (English Profile; e.g., Green, 2012; www.englishprofile.org), German (Profile Deutsch; Glaboniat et al., 2005), and Norwegian (Norsk profil; Carlsen, 2013). The European-wide SLATE (Second Language Acquisition and Language Testing in Europe; www.slate.eu.org) network brings together researchers who share an interest in examining the linguistic basis of the CEFR.

The CEFR has become central to European language education, and it is widely used for setting targets for language learning in curricula and for describing the level of language courses, textbooks and tests (Huhta, 2012; Martyniuk & Noijons, 2007). CEFR levels are also used for such high-stakes purposes as defining language proficiency requirements for citizenship (Extra, Spotti, & van Avermaet, 2009). Despite its widespread use, the CEFR has been criticised, for instance, for its uncertain basis on second language acquisition research. The framework scales that appear to describe stages of L2 development are not based on empirical research on how proficiency actually develops (Hulstijn, 2007). These criticisms notwithstanding, the fact that the CEFR does not describe the use of any particular language but a language in general means that there is a need to understand how learners coming from a particular L1 background develop in linguistic terms in a particular L2 they are learning.

Word derivation (WD) is a linguistic feature that has received relatively little attention is SLA research so far. Word derivation is the process of forming new words on the basis of existing words, such as *lucky*, *unlucky* and *luckless* from *luck.* It involves the addition of a morpheme such as a prefix or a suffix or both (in the above examples *un*- is an example of a prefix and -*y* and -*less* are examples of suffixes), or an infix (e.g., *Tenne-bloody-see*), which is very rare in English. It should be noted that derivation produces new lexemes and thus differs from inflection which produces grammatical variants of the same lexeme (e.g., *luckier, luckiest*).

The present study adds to our knowledge on how the ability to derive English words develops among L2 English learners. More specifically, we aim at shedding light on how word derivational knowledge relates to CEFR levels defined with reference to learners' writing skills.

Below we will first describe the nature of vocabulary and word derivational knowledge and then present a review of research on derivation and its development, after which we will introduce the current study.

2. Multidimensional and incremental nature of word derivational knowledge

Knowing a word can be defined in several ways. Different lexical models have been presented by, for example, Milton & Fitzpatrick (2013), Nation (2001) and Ringbom (1987). These models can be broadly classified as either dimensional or developmental (see, e.g., Read, 2000, for a discussion). In the following two sections, we will define the two approaches and outline research proposing a) multidimensional and b) incremental models of lexical development.

2.1Multidimensional nature of vocabulary and word derivational knowledge

The first approach to defining vocabulary knowledge seems to be influenced by the connectionist epistemology (e.g., Seidenberg & Gonnerman, 2000), according to which the development of L2 lexical knowledge happens in several knowledge domains, such as orthography, phonology, syntax, and semantics. It dates back to Richards' (1976: 83) influential discussion of the possible dimensions of lexical competence, i.e., knowledge of associations, syntactical properties of words, their form (including derivatives), constraints of use, among others.

One of the well-known dimensional vocabulary knowledge models has been proposed by Nation (e.g., 2001), who outlined three broad aspects of vocabulary knowledge, i.e., form, meaning, and use, and further classified them into subcomponents, e.g., spoken, written, and word parts in the *form* component, as well as differentiated between receptive and productive knowledge of these subcomponents. Ringbom's (1987; 1990) model of lexical knowledge (see **Figure 1**) is similar to Nation's (2001) model. The difference is that it also incorporates the development within each dimension. The developmental approach will be discussed in more detail in the following section.

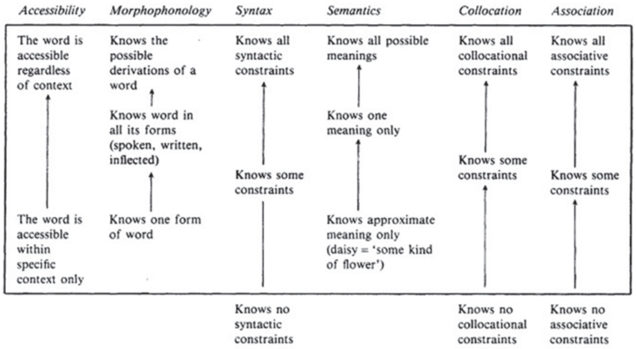

Accessibility
Morphophonology
Syntax
Semantics
Collocation
Association
The word is accessible regardless of context
The word is accessible within specific context only
Knows the possible derivations of a word
Knows word in all its forms (spoken, written, inflected)
Knows one form of word
Knows all syntactic constraints
Knows some constraints
Knows no syntactic constraints
Knows all possible meanings
Knows one meaning only
Knows approximate meaning only (daisy = 'some kind of flower')
Knows all collocational constraints
Knows some constraints
Knows no collocational constraints
Knows all associative constraints
Knows some constraints
Knows no associative constraints

Figure 1.
Ringbom's (1987) model of lexical knowledge.

No comprehensive dimensional model of word derivational knowledge appears to exist. However, research on L2 (and L1) English word derivational knowledge has found that many of the dimensions listed in the vocabulary knowledge models above are also relevant to WD knowledge. These include, for example, syntactic knowledge (e.g., Schmitt, 1998; Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002), knowledge of semantics of derivational affixes (e.g., Chuenjundaeng, 2006), and L1/L2 morphophonology / morpho-orthography (e.g., Alegre & Gordon, 1999; Friedline, 2011). Another dimension is accessibility/control, which has also been labelled as productive/receptive knowledge, or recognition/recall of vocabulary (e.g. Schmitt & Meara, 1997; Hayashi & Murphy, 2010).

2.2 Incremental development of vocabulary and word derivational knowledge

An alternative approach to defining vocabulary knowledge is the developmental one. As the name suggests, this approach stresses development and developmental stages. Research has shown that vocabulary knowledge develops incrementally and correlates positively with learners' proficiency (e.g., Nation, 2001; Schmitt, 1998; 2010). Similarly, learners' word derivational knowledge appears to develop incrementally, both in L1 and L2 English. For example, Tyler and Nagy (1989) found that while at grade four, learners were able to recognise frequent L1 English stems and derivatives, by grade eight, they increased their syntactic knowledge of derivational affixes. Later, Nagy, Diakidoy, and Anderson (1993) found significant differences in the knowledge of the meaning of frequent L1 English derivational affixes between grade four and upper-secondary school, most of the improvement occurring between grade two and seven.

The development of word derivational knowledge in L2 English acquisition is far less studied than in L1 English acquisition but the findings are similar. For example, Schmitt and Meara (1997) found that university students increased their knowledge of some derivational affixes after one academic year although the increase was modest at best and was not on a par with the increase of their general vocabulary knowledge. In his longitudinal study of four university learners, Schmitt (1998) also found proof for the incremental development of L2 English derivational knowledge although he could not find evidence for any particular order of acquisition.

Not surprisingly, links between learners' knowledge of derivational affixes and their language proficiency and vocabulary size/depth have been discovered although the findings vary. Mäntylä and Huhta (2013) found significant correlations between learners' L2 writing proficiency and their performance on three affix elicitation tasks. Friedline's (2011) cross-sectional study had mixed results as regards L2 proficiency and WD knowledge. Friedline discovered no relationship between language proficiency, as measured by the Michigan Test of English Language Proficiency and learners' performance on a lexical decision task (learners had to state how certain they were that the presented derivative was a real English word) or a word decomposition (learners had to write the base form of the given derived words) task. However, Friedline's (2011) results suggested that learners' proficiency related to their performance on the word-relatedness task where they had to rate their certainty in the relatedness of the pairs of words, e.g., *decorative–decoration*. Schmitt and Meara (1997) found a moderate correlation ($.27 \leq r \leq .41$) between learners' derivative suffix knowledge, both productive (learners listed all the suffixes that they thought could be attached to the base words given) and receptive (learners marked all the allowable suffixes that they thought could be attached to the base words) and their receptive vocabulary size, as measured by the Vocabulary Levels test

(Nation, 1990) but not between their suffix knowledge and language proficiency (i.e., learners' TOEFL scores). The correlation was higher for the receptive suffix knowledge measure. Mochizuki and Aizawa (2010) also found a moderate correlation ($.54 \leq r \leq .65$) between learners' vocabulary size and their knowledge of the meanings of prefixes and syntactic role of suffixes. Hayashi and Murphy (2010) found that the scores on the affix elicitation task were strongly correlated with both productive ($r = .832$) and receptive ($r = .842$) vocabulary size of the Japanese learners of English, but their results on the word segmentation task were not. More recently, Collins and Nation (2015), in an exploratory study, found that learners' ability to understand the meanings of derived words from the previously unknown word families after being provided with the L1 equivalent of the roots did not predict their scores on a vocabulary size test. However, understanding derived forms was a better predictor of their reading speed (in words per minute).

Although following the developmental paradigm, it is tempting to assume that some L2 (and L1) English derivational affixes are acquired earlier than other, their acquisitional order is yet to be discovered—and so is the acquisitional order of WD knowledge dimensions. In this respect, the teaching order of L2 English derivational affixes proposed by Bauer and Nation (1993) and further developed by Nation (2001) could be a starting point in the process of discovering one. According to them, English derivational affixes could be classified into difficulty levels based on their morphological and phonological properties: frequency, productivity, semantic transparency, regularity of written/spoken form of the bases they attach to, regularity of their spelling/spoken form, and regularity of function. Bauer and Nation (1993) suggested that affixes should be taught to L2 learners in this order. The affix levels as identified by Bauer and Nation (1993) are presented below in **Table 1**.

Table 1.

Teaching order of L2 English derivational affixes (Bauer & Nation, 1993; Nation, 2001).

Level 1	A different form is a different word.
Level 2	Inflectional categories: plural -s, past tense -ed, comparative -er, etc.
Level 3	The most frequent and regular derivational affixes: -able, -er, -ish, -less, -ly, -ness, -th (fourth), -y, non-, un-*
Level 4	Frequent and regular affixes, e.g., -al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in-*.
Level 5	Infrequent but regular affixes, e.g., -age -ance, -ship, mis-,etc.
Level 6	Frequent but irregular affixes, e.g., -ee, -ic, -ion, re-, etc.
Level 7	Classical roots and affixes, e.g., -ate, -ure, etc.

*All with restricted uses; see **Appendix 1** in Bauer and Nation (1993) for details.

It should be stressed, though, that Bauer and Nation's (1993) ranking of the affixes by their difficulty is rather arbitrary, and it is premature to consider this an/the order of acquisition. Moreover, to our knowledge, this difficulty order is yet to be corroborated empirically.

Overall, the studies examining learners' L2 English word derivational knowledge are few and result in mixed findings. These studies often consider only a limited number of dimensions of learners' WD knowledge, which adds to the difficulty of operationalising and generalising the complex construct of learners' L2 English WD knowledge. The present study endeavours to add to the existing body of research by studying whether learners' proficiency relates to their performance on a number of measures estimating hypothesised dimensions of L2 English WD knowledge as represented in Ringbom's (1987; 1990) model of lexical knowledge, and containing derivational affixes from different Bauer and Nation' (1993) levels. We used these levels to introduce variability in the difficulty of affixes, in the absence of an empirically validated order of difficulty.

3. Methodology

3.1 Research questions

In the context of Finnish and Estonian learners of English as a foreign language:

1. Do different aspects of word derivational knowledge relate to learners' writing proficiency?
2. If word derivation and writing are related, is the relationship stable (i.e., do derivation skills increase steadily from level to level) or does ability to derive words increase rapidly at a particular level?

Despite the conflicting results that the previous research on the relationship between learners' L2 English proficiency and word derivational knowledge has produced, informed by Ringbom's (1987; 1990) lexical knowledge model, we hypothesised that different aspects of learners' L2 English word derivational knowledge develop as their proficiency grows.

3.2 Tasks

All in all, two writing tasks as well as nine measures which were hypothesised to tap learners' word derivational knowledge were administered.

Two different writing samples were collected from each learner. The L1 Finnish participants completed the same writing assignments administered as a part of a previous research project in Finland. The writing samples collected from L1 Estonian and Russian participants were a part of their usual classroom assignments, and were thus different for learners at different schools or taught by different teachers. Despite that, the genres / task types were similar in most of the groups, those being argumentative texts (e.g., essays) and formal letters. Other task types, such as narratives (description of an event or a story) were also used. Regarding the genres, it should be noted that judging by the state curricula in Finland and Estonia, learners are expected to write, particularly in the upper-secondary school, and should be familiar with the genres they wrote in. What is more, in the Matriculation Examination in Finland and the English State Exam in Estonia learners are commonly asked to write these types of texts, and these are also covered in different coursebooks used in the schools in the two countries.

The writing samples were independently rated by two raters on the CEFR scale using the procedures and benchmark samples designed in the Finnish research project mentioned above. The ratings were analysed with the multi-faceted Rasch analysis program *Facets* (which we will discuss in more detail when presenting the results), which, to an extent, accounted for the different genres of the written performance samples in different groups. In addition, rating the learners' writing performance on the CEFR scale, which is task-independent, and rating two written performance samples per learner also minimised the possibility of genre affecting the ratings.

The word derivation measures in the study were designed to represent different dimensions of WD knowledge as appearing in Ringbom's model of lexical knowledge (see **Figure 1**). Most of the measures were designed specifically for the current study, as few appropriate measures existed. The lack of measures was particularly acute for measures of word derivation in context; to our knowledge, only Schmitt and Zimmerman (2002) and Mäntylä and Huhta (2013) have developed such measures. We should note that the measures in our study were, nevertheless, somewhat similar to those used in the previous research. Hence, a word segmentation task was also used by Hayashi and Murphy (2010). However, the measure used in the present study involved finding derived words in context rather than using single words as Hayashi and Murphy did. Overall, our measures evaluated both receptive and productive knowledge (active and passive recognition and recall), contained affixes at Bauer and Nation's (1993) levels 3 to 6, formed different parts of speech, and, as regards the base words used as items, belonged to the first five thousand most frequent lemmatised words in the British National Corpus (ref.). The word segmentation task was somewhat different in this respect, as it was based on three excerpts from authentic texts, which we slightly adapted for the purpose of the study. The frequency of the lemmatised items in the task ranged from the first to the twenty-first thousand (the latter being item *revengeful*) most frequent words, with most of the items (k = 31) falling to the first five thousand most frequent words. The items in the word segmentation task were formed with a total of 49 derivational affixes at Bauer and Nation's levels 1-7 (with only 2 items formed with level 7 affixes).

The aim of the word segmentation task (in which the learners were asked to find derived words in three coherent text excerpts and mark derivational affixes in them) was, above all, to study the accessibility dimension of the word derivational knowledge. However, it can be assumed that other types of knowledge, such as semantic and syntactic knowledge of derivational affixes, were also used by the learners when they worked on the word segmentation task. As to the other measures, the affix elicitation task (in which the learners were asked to form derived words from the words in bold, but also using L1 translations to complete the sentences) regarded the accessibility and semantics dimension. The non-word affix elicitation task (where the learners were required to add affixes to non-words to complete the sentences based on the definitions provided to them) aimed to tap into semantics of derivational affixes and, to an extent, to control for the learners' vocabulary knowledge. In the prefix elicitation task, the learners were asked to select prefixes among provided to complete the derived words in the sentences. The task, we hypothesised, above all, had the semantics dimension (but also, e.g., accessibility). In the grammar recognition task, the learners were required to select one word among the three provided (all having the same bases but different suffixes, forming different parts of speech). The task was to tap into the learners' syntactic knowledge, and so was the aim of the metalinguistic prompts task (although the latter lacked the accessibility dimension and required to demonstrate metalinguistic knowledge). In the metalinguistic prompts task, the learners were asked to write one noun, one verb, and one adjective formed from the given words. The meaning recognition and the passive recognition of the meaning tasks were expected to include the receptive semantics dimension. Finally, the free production task, we suggested, above all, tapped especially into the morpho-phonology dimension. Needless to say, the morpho-phonology dimension was present in all the other measures as well. More details on the measures are presented in **Appendix A**.

All the tasks except for the word segmentation task were administered in an online assessment system, which allowed the participants to complete them faster but also facilitated the coding and the analysis of the data. The afore-mentioned system was designed following the procedure discussed by Fulcher (2003). Specifically, it was designed to be based on a detailed system and test specifications, and included a multi-stage trialling of the interface and the tasks. For details regarding the description and the usability of the system, see Leontjev (2014).

The WD measures and the instructions, except for the three context-dependent measures (i.e., affix elicitation, non-word affix elicitation, and prefix elicitation tasks), designed and used earlier by Mäntylä and Huhta (2013), were piloted with 22 university students in Estonia (roughly at level B2 of their L2 English proficiency as estimated by their teacher), which allowed us to address some problems in the items (e.g., too difficult items) and in the task instructions. None of the pilot study participants took part in the main study.

3.3 Data and participants

The data come mainly from the learners' performance on the tasks. Additionally, some background data, such as the participants' age, were collected.

The participants in the study were a total of 117 L1 Finnish, Estonian, and Russian learners of English at grade 10, i.e., senior secondary school level (mean age = 16.7; range 15–18), 56 male, 58 female (the sex/gender data of three learners was not available), in Finland and Estonia. There were four different groups of learners taught by two teachers in Finland and five different groups of learners taught by four teachers in Estonia. Further details on the participants are presented in **Table 2**.

Table 2.

Learners' writing proficiency on the CEFR scale.

Country	**N**	**CEFR level***				**Median CEFR level by country**
		A1	**A2**	**B1**	**B2**	
Estonia	47	-	5	22	20	B1
Finland	70	1	11	38	20	B1

*The learners' writing proficiency level is a rounded fair average from Facets (see section 3.2).

The decision to select learners at grade 10 as the participants was rooted in both theoretical and practical considerations. As the study is a part of a larger project, we wanted at least some of the learner participants in the project (who were at grade 10 at the time of the data collection) to participate in the present study. Moreover, according to both the Finnish (Finnish National Board of Education, 2003) and the Estonian (Põhikooli riiklik õppekava õigusakt: Lisa 1, 2010) state curricula, the learners' proficiency in the first foreign language should be at level B1 of the CEFR by the beginning of the senior secondary school (i.e., at grade 10 in both countries), which made the groups more comparable (also see **Table 2**). According to the national curricula, the number of academic hours of L2 English instruction in the first nine years of school in the two countries are somewhat different, that is, 735 in Estonia and 608 in Finland. However, since the data collection in Finland took place about four months later than in Estonia, the amount of instruction in L2 English that the Finnish participants had received was quite comparable to that in Estonia. Furthermore, Nation (2001) suggested that learners can be taught derivational affixes at lower-intermediate level of L2 proficiency, which roughly corresponds to CEFR level B1 (Council of Europe, 2001). It is worth mentioning that we learned from the teachers of the participating learners that they taught their learners word derivation although not extensively and not systematically.

3.4 Procedure

Before the study, the participants granted their permission to use the data for research purposes. As an incentive, they were provided with detailed feedback regarding their performance on the tasks as well as pieces of advice on how to improve their knowledge of vocabulary and word derivation.

A total of two hours was allocated to completing the online tasks measuring the word derivational knowledge, but all learners managed to complete the tasks quicker than that, so the tasks were not speeded. Two groups completed the word segmentation task together with the other tasks (*n* = 37, in Finland). The rest completed the word segmentation task within a week after the online tasks. When working on the tasks, the learners were in a classroom. In Estonia, a researcher monitored the procedure alongside with the teachers in most of the groups. In Finland, only the teachers did. Detailed instructions were written for the teachers of the participating learners as regards prevention of / reporting on the cases of cheating and responding to the learners' queries during the data collection. The written performance samples were collected within a month and a half before or after the participants completed the WD tasks and these were not speeded either. The written performance samples were checked for plagiarism (also against the work of the other students) to make sure that the learners worked independently on the task.

4 Results

4.1 Reliability of the tasks

Generally, the word derivational knowledge measures used in the study were found to be reliable (internally consistent), for an exploratory study, $.85 \geq \alpha \geq .63$. However, the ten items in the meaning recognition task had a low internal consistency, $\alpha = .46$.

To accompany the reliability analysis, a modern item analysis of the measures was conducted using *Winsteps* Rasch analysis software. The results indicated that item two in the meaning recognition task was misfitting, *infit MnSq* = 1.57 (*Zstd* = 4.5), *outfit MnSq* = 1.92 (*Zstd* = 4.3). Thus, the task was analysed without this item. Even with this adjustment of the scale, the meaning recognition task had Rasch reliability of .45 and Cronbach's alpha of .56. The grammar recognition and the passive recognition of the meaning task also had somewhat poor Rasch reliability (.63 and .55 respectively), but their Cronbach's alpha coefficients were acceptable for an exploratory study (.73 and .63 respectively). We assume that the main reason for this was the low number of items in the tasks. Therefore, inferences based on the meaning recognition task in particular but also the grammar recognition and the passive recognition of the meaning tasks should be made with caution. The Rasch reliability (ranging from .67 to .93) and the Cronbach alpha coefficients (ranging from .79 to .95) of the rest of the tasks were acceptable. As regards the rater consistency in estimating the learners' proficiency, the Rasch analysis indicated that the ratings were consistent, *infit mean-square* figures being 1 and 0.93 for rater A and B respectively. It should be noted that while the average length of the produced written texts was 130 words, there was a great variation in length (min. = 26; max. = 569). However, even when we controlled for the length, the results of the analyses (see Section 4.2) were interpreted the same.

To reinforce our decision to consider the performance of the two countries (and different L1s) together, we also conducted a Mann-Whitney *U* test on the learners' estimated CEFR proficiency level variable (**Table 2**), the country being the independent variable. The analysis confirmed that the two countries did not differ significantly (*Z* = -1.652, *p* = .099). Moreover, apart from the word segmentation task and the non-word affix elicitation task, the learners in the two countries did not perform significantly differently. The difference in the two tasks was tiny, the country accounting for only 3% to 5% of the variance.

4.2. Word derivational knowledge and writing

In order to address the first research question concerning the relationship between L2 writing and word derivational knowledge, we computed the Spearman rank order correlations between the raw scores from the WD measures and the Facets fair averages based on the ratings of students' writing. Before that, the descriptive statistics are presented (**Table 3**).

Table 3.

Descriptive statistics for the measures used in the study.

Measure	**N**	**Mean**	**S.D.**	**Max. score**	**Total number of items**
Free production task	116	7.65	4.60	29	-
Metalinguistic prompts	114	5.38	4.62	18	-

task					
Affix elicitation	114	9.58	3.96	15	15
Non-word affix elicitation	114	4.18	3.44	11	13
Prefix elicitation	114	6.64	2.92	12	12
Grammar recognition	113	6.48	2.35	10	10
Meaning recognition	113	5.85	1.93	9	9
Passive recognition of the meaning	111	6.27	2.22	10	10
Word segmentation (# of words)	107	19.39	7.33	36	39
Word segmentation (# of affixes)	106	13.53	5.52	29	49
N listwise	98				

The number of cases differs for different measures because some learners skipped some tasks. The cut-off criterion for considering that the task was skipped was five seconds or less spent on the task. The total possible numbers of correct responses in the free production and the metalinguistic prompts tasks are not indicated in Table 3 and elsewhere in the manuscript. This is because in the free production task, by design, the number of words the learners were asked to form per item was not limited and in both tasks, learners were allowed to use inflectional affixes as well. **Table 4** presents the correlations for the entire group (number of learners varied from 106 to 117).

Table 4.

Correlation of the word derivation measures with the learners' writing proficiency (Facets fair averages).

Measure	Spearman rho	Significance
Free production task	.458	<.001
Metalinguistic prompts task	.465	<.001
Affix Elicitation	.585	<.001
Non-word affix elicitation	.410	<.001
Prefix elicitation	.581	<.001
Grammar recognition	.642	<.001
Meaning recognition	.578	<.001
Passive recognition of the meaning	.504	<.001

Word segmentation (# of words)	-.101	.300
Word segmentation (# of affixes)	-.001	.989

Correlational analysis of the relationship between writing in English and word derivation measures revealed that some of the latter had strong (over .5 or .6) correlation with writing and even the lower correlations were over .4. The only exception was the word segmentation task in which the learners had to mark in a text all derived words and all affixes that they could find. The number of words or affixes the learners could locate did not correlate at all with their writing proficiency.

To further investigate the relationship between writing proficiency and word derivation, we conducted a multiple linear regression with the Facets fair average for the learners' writing proficiency as the dependent variable and the WD measures as the supposed predictors[2]. The word segmentation task was excluded from the analysis. We also bootstrapped confidence intervals using Bias-Corrected and accelerated method and 2,000 resamples.

The results indicated that the linear combination of the prefix elicitation task (β = .366, $t(99) = 4.19$, $p < .001$), the grammar recognition task (β = .290, $t(99) = 3.20$, $p = .002$) and the meaning recognition task (β = .246, $t(99) = 3.12$, $p = .002$) significantly related to the learners' writing proficiency ($R^2 = .58$, $R^2_{adj}= .57$, $F(3,99) = 45.49$, $p < .001$), accounting for about 57–58% of the variance[3]. Since the meaning recognition task had a low reliability, we also ran a regression without this measure. This time, the affix elicitation task emerged a significant predictor, too (β = .260, $t(99) = 2.95$, $p = .012$). The variance that these three measures accounted for was similar to that in the first linear regression analysis, $R^2 = .58$, $R^2_{adj}= .56$.

Next, we examined whether learners' derivation ability increases as their writing proficiency grows and whether this increase is steady. For this, we divided the learners into three groups according to their writing proficiency by rounding the students' fair average scores from the Facets analysis into the nearest CEFR level. All learners except one could be placed at A2, B1, or B2 levels. The one student placed at A1 level was included in the nearest, A2, group. We then computed the percent correct scores separately for the three groups. The results presented in **Table 5** reveal that the changes were not that steady across the levels.

Table 5.

Mean percent correct at different CEFR proficiency levels across the measures.

Measure	**Proficiency on the CEFR scale**		
	A2	**B1**	**B2**
Free production	-	-	-
Metalinguistic prompts	-	-	-
Affix Elicitation	37	61	79
Non-word affix elicitation	14	30	43
Prefix elicitation	24	55	69
Grammar recognition	42	59	83
Meaning recognition	49	59	80
Passive recognition of the meaning	46	60	73
Word segmentation (# of words)	47	52	47
Word segmentation (# of affixes)	25	28	28

To study the differences statistically, we ran a number of ANOVAs, the raw scores in each of the measures being the dependent variable and the CEFR level, the between-subjects independent variable. We then supplemented the regular ANOVA analyses with the linear contrast analyses to establish whether the differences across the learners' proficiency levels related linearly to their WD knowledge as estimated by our measures. In the cases where the homogeneity of variance assumption was violated, we utilised Welche's F-test instead of the regular F-test. The following **Table 6** gives an overview of the results of the ANOVAs we obtained.

Table 6.

Relationship between the learners' CEFR level and their performance on the measures (analyses of variance).

Measure	F-test / Welche's F-test	Effect size
Free production	$F(2, 113) = 7.20, p = .001$	$\eta^2 = .11$
Metalinguistic prompts	Welch's $F(2, 45.29) = 11.66, p < .001$	$\eta^2 = .20$
Affix elicitation task	$F(2, 111) = 20.13, p < .001$	$\eta^2 = .27$
Non-word affix elicitation[4]	Welch's $F(2, 49.76) = 10.65, p < .001$	$\eta^2 = .13$
Prefix elicitation	Welch's $F(2, 36.38) = 22.04, p < .001$	$\eta^2 = .34$
Grammar recognition	$F(2, 110) = 35.60, p < .001$	$\eta^2 = .39$
Meaning recognition	$F(2, 110) = 22.37, p < .001$	$\eta^2 = .29$
Passive recognition of the meaning	$F(2, 108) = 10.59, p < .001$	$\eta^2 = .16$
Word segmentation, # of words	$F(2, 104) = 0.71, p = .492$	$\eta^2 = .01$
Word segmentation, # of affixes	$F(2, 103) = 0.31, p = .731$	$\eta^2 = .006$

The results of the analyses of variance revealed that the effect of the learners' proficiency varied from moderate to strong in all of the measures with the notable exception of the word segmentation task. Additionally, the trend analyses revealed that except for the word segmentation task, the linear trend accounted for most of the variance associated with the learners' writing proficiency. For example, in the meaning recognition task, where there was the biggest difference between the effect of the learners' proficiency and that of the linear trend associated with the proficiency, the linear trend accounted for 24% of the variance (cf. **Table 6**). The linear trend also transpires in the means plots, to which we added 95% confidence intervals (**Appendix B**).

We then followed the ANOVAs with the pairwise comparisons, in which, to account for unequal sample sizes, we used Hochberg's GT2 when the variances were homogeneous and Games-Howell post-hoc tests when they were not[5]. **Table 7** gives a summary of the pairwise comparisons.

Table 7.

Pairwise comparisons of the learners' performance on the measures across the CEFR levels.

Measure	**Mean difference and significance**		
	A2–B1	**B1–B2**	**A2–B2**
Free production	1.89, *p* = .313	2.55, *p* = .016	4.44, *p* = .002
Metalinguistic prompts	1.39, *p* = .278	3.86, *p* < .001	5.25, *p* < .001
Affix elicitation (*k* = 15)	3.70, *p* < .001	2.62, *p* = .001	6.32, *p* < .001
Non-word affix elicitation (*k* = 13)	2.09, *p* = .015	1.68, *p* = .055	3.77, *p* < .001
Prefix elicitation (*k* = 12)	3.61, *p* < .001	1.77, *p* = .005	5.37, *p* < .001
Grammar recognition (*k* = 10)	1.72, *p* = .003	2.44, *p* < .001	4.16, *p* < .001
Meaning recognition (*k* = 10)	0.94, *p* = .118	1.85, *p* < .001	2.79, *p* < .001
Passive recognition of the meaning (*k* = 10)	1.39, *p* = .053	1.29, *p* = .009	2.68, *p* < .001
Word segmentation (# of words) (*k* = 39)	1.71, *p* = .832	- 1.70, *p* = 621	0.01, *p* = 1.00
Word segmentation (# of affixes) (*k* = 49)	1.30, *p* = .831	0.01, *p* = 1.00	1.31, *p* = .848

The pairwise comparisons demonstrated that while the trend analysis indicated that the relationship was fairly linear across the measures, the differences between A2 and B1 and B1 and B2 were not as stable. In the affix elicitation, the prefix elicitation and the grammar recognition tasks, both the difference between A2 and B1 and between B1 and B2 were significant (hereinafter in the paragraph, at *p* < .05). In the metalinguistic prompts, the meaning recognition and the passive recognition of the meaning tasks, only the difference between B1 and B2 was significant. The results obtained on the pairwise comparisons for the passive recognition of the meaning are somewhat counterintuitive when percent correct figures across the CEFR levels are considered (see **Table 5**). Nevertheless, the results obtained on the ANOVA should be considered more reliable, as the latter takes into consideration variances and measurement errors. In the non-word affix elicitation task only the difference between level A2 and B1 was significant. It should also be noted that in the free production task, based on the results of the Hochberg's pairwise comparisons, there was a significant difference between levels B1 and B2. However, judging by the means plot with the confidence intervals added (**Appendix B**), only the difference between A2 and B2 was significant.

5. Discussion

The present study aimed at exploring whether learners' L2 English word derivational knowledge is related to their writing proficiency and whether this relation is steady across learners' proficiency levels on the CEFR scale. While comparing learners' proficiency with word derivational knowledge is not new, this study adds to previous research in that we analysed a more comprehensive number of aspects of word derivation.

While the results of the previous research regarding the relation of word derivational knowledge and learners' more general proficiency were mixed (Mäntylä & Huhta, 2013; Friedline, 2011; Schmitt & Meara, 1997), the present study indicated a rather strong correlation between the participants' proficiency and their word derivational knowledge.

The only exception was the word segmentation task, which did not correlate with the learners' writing proficiency. This finding was in line with the previous research, such as Friedline's (2011) and Hayashi and Murphy's studies (2010). We hypothesise that in the present study, this can be attributed to the effect of the task. The learners were asked to find derived words in a coherent text. Thus, they could not but process the text for its meaning. More able learners then were able to recognise more words, which may have interfered with their ability to analyse the words (Ullman, 2001). On the other hand, in Friedline (2011) and Hayashi and Murphy (2010), single words were used in segmentation/decomposition tasks, and not a coherent text. Thus, there seem to be other factors that affect learners' ability to analyse words. Psycholinguistic theories, such as declarative/procedural model (e.g., Ullman, 2001), predict that L2 learners are more dependent on declarative memory and thus are likely to store more L2 linguistic forms as entities in their memory than L1 linguistic forms, can explain this finding. Nevertheless, it is clear that more research into this is required.

The similarity of correlations between WD and writing proficiency in the present study and that of Mäntylä and Huhta (2013) is not surprising, at least for the affix elicitation, the prefix elicitation, and the non-word tasks, as these measures were used in both studies. The strong correlation found by Hayashi and Murphy (2010) between the participants' performance on the affix elicitation task and their vocabulary knowledge is also strengthened by our findings, as vocabulary knowledge is found to correlate with proficiency quite strongly (see, e.g., Schmitt, 2010).

However, the present study also found strong correlations between the learners' syntactic (grammar recognition and metalinguistic prompts tasks, i.e., both receptive and productive) and semantic (meaning recognition and passive recognition of the meaning) knowledge of derivational affixes and learners' proficiency. Interestingly, based on their results obtained on two tasks requiring learners to demonstrate syntactic knowledge of derivational affixes, Schmitt and Zimmerman (2002) also suggested a relationship between the derivational knowledge and proficiency, but did not test their assumption statistically which was done in the present study. On the other hand, there is a discrepancy between our findings regarding the correlation between the proficiency and learners' performance on the free production task and those of, for example, Schmitt and Meara (1997) on a similar task. In the latter study, however, the non-significant correlation might have been due to the small sample size (n = 28).

While the results clearly demonstrated the relation of the participants' writing proficiency and their word derivational knowledge, only some of the measures significantly predicted the learners' writing proficiency. However, they accounted for over 50 percent of the variance.

The likely reason for the grammar recognition task being a significant predictor of writing is that learners constantly refer to their syntactic knowledge when writing in L2, and the latter develops as their abilities in their L2 grow, or at least, the accuracy in and complexity of its use increase (see, e.g., Alanen & Kalaja, 2010). We hypothesise that the reason for the prefix elicitation task significantly predicting the learners' writing proficiency is that it, above all, required learners to refer to their semantic knowledge of the prefixes either as such or through analogy with words containing the same prefix. Also, the prefixes were to a large extent transparent. The meaning recognition task, which also predicted writing though the prediction was much smaller, was designed to tap learners' semantic knowledge of derivational affixes. Yet, prefixes are, arguably, more transparent, which may be why the prefix elicitation task predicted writing more strongly. As regards the affix elicitation task, it required the participants to demonstrate their vocabulary knowledge more than the rest of the tasks. In fact, Hayashi and Murphy's (2010) finding that learners' vocabulary knowledge strongly predicted their performance on an affix elicitation task indirectly speaks in favour of this interpretation. The question is, however, why other tasks, such as the metalinguistic prompts task, did not predict writing proficiency, although the latter required the learners to recall rather than recognise the affixes and their syntactic functions. We assume that in those tasks, there could have been other factors that interacted with the learners' performance. Specifically, in the case of the metalinguistic prompts task, the learners had to refer to their metalinguistic knowledge, which has been found to relate with language proficiency only weakly (e.g., Alderson, Clapham, & Steel, 1997).

We were also curious as to whether there is a possibility that learners' ability to derive words in English increases more rapidly at a particular level of their proficiency. While overall the linear trend in the measures across the CEFR proficiency levels was rather strong, as can be deduced from both the linear trend analyses and the means plots (**Appendix B**), in roughly half of the measures, there was a bigger change in the learners' performance between levels B1 and B2 than between levels A2 and B1. In two of the measures where it was not the case, that is, the measures also used in Mäntylä and Huhta's (2013) study (the latter arriving to a similar finding to ours), the differences were almost the same across the proficiency levels (see **Table 5** and **Appendix B**). The notable exception was the prefix elicitation task, where there was clearly a bigger increase between A2 and B1.

6. Conclusion

Taken together, the findings suggest that depending on the way word derivational knowledge is operationalised, the ability to derive words is either more or less stable or increases more after level B1. In addition, it seems that syntactic and semantic aspects of word derivational knowledge predict learners' proficiency stronger than others. The results, thus, not only demonstrate that there is a link between word derivational knowledge and writing proficiency, but also suggest that not all of its aspects develop steadily as learners' proficiency grows.

Next, we will discuss the limitations of this study. The first limitation concerns the way we operationalised learners' writing proficiency. The figures were based on two written performance samples (e.g., an essay or a formal letter) per learner rated by two raters only. Moreover, as regards the Estonian sample, at least for some groups, the task types were different. While the approach we selected, i.e., using a Rasch estimation of the learners' abilities, improves the quality of the ratings, having a third rater and asking the learners to complete exactly the same writing tasks would have increased the reliability of the scale representing their writing proficiency.

Moreover, this was a cross-sectional study, so it does not provide an ideal basis for interpreting results in terms of the development of the learners' ability to derive words in English. Rather the findings can be considered a starting point in accumulating evidence for the development of different aspects of word derivation that should be confirmed in longitudinal studies. In fact, the nature and quality of operationalising L2 English proficiency is likely to vary across studies (cf. Friedline, 2011; Schmitt & Meara, 1997), which makes a systematic comparison of the findings with the previous research even more difficult.

While we did not analyse the two countries separately due to the verisimilitude of their performance on the tasks, the participants' L1 might have still influenced the results somewhat. On the other hand, the sample sizes in different L1 groups were even more unequal than those across the proficiency levels, which could have complicated the analyses further and increased the possibility of making both Type I and Type II errors. Nevertheless, we think that future studies of L2 word derivational knowledge should take participants' L1 into consideration. An interesting possibility is that not only learners' mother tongue influences their word derivational knowledge (or some aspects of it) but also their second language(s) do, which could be addressed in a future study. Considering the participants L1's, since Finnish and Estonian are related languages, it is difficult to say whether and to what extent the results are generalisable to learners of other L1s.

Caution should also be exercised when generalising the results of the present study to other measures of word derivational knowledge, as the results could be due to the effect of the task type. Moreover, due to the complex nature of L2 English word derivational (and vocabulary) knowledge, we cannot claim that a certain aspect of word derivational knowledge is more difficult than another despite the results on the pairwise comparisons of the ANOVAs. At the very least, this would require controlling for frequency of the bases, the whole lemmas and their semantic transparency, as well as the length of entire items (e.g., words, phrases, sentences) before we could argue that the task/item difficulty was due to a certain aspect of word derivational knowledge or an interplay of several such aspects.

Finally, as also previous studies on WD have found, it is difficult to separate word derivational knowledge from general vocabulary knowledge. We addressed this challenge by having several different methods of study. Still, the differences in the learners' performance across the CEFR levels and their correlations with the learners' writing proficiency might have also been due to factors other than their word derivational knowledge. More sophisticated statistical analyses, such as Structural Equation Modelling, could shed more light on this issue in future studies.

Since word derivational skills enhance vocabulary learning, studying and understanding them is worthwhile. Still little researched questions include establishing the effect of the second language and the type of language teaching, doing a more systematic division of different aspects of word derivational knowledge into receptive and productive types, tracing the development of word derivational knowledge incrementally, and confirming and rejecting empirically the order of derivational affixes proposed by Bauer and Nation (1993). Besides that, more research is needed to be able to answer the questions of whether the development of L2 English word derivational knowledge is steady and if not, what point in learners' proficiency can be considered as crucial for its development.

7. Endnotes

1. It should be noted that the average length of the written performance samples in Estonia was significantly higher. However, as an ANCOVA demonstrated, length controlled for, there was no significant difference in the learners' proficiency between the two countries, $F(1, 114) = 2.71$, $p = .102$,$\eta_p^2 = .02$. Moreover, the correlational and the regression analyses conducted with the fair average variable (representing the learners' proficiency) residualised on the average length variable (i.e., the learners' proficiency with the variance introduced by the text length excluded) demonstrated that the figures were similar to those obtained on the original proficiency variable. Thus, regardless of the length of the text, the same conclusions were drawn from the analyses.

2. The figures were obtained using the list-wise deletion of cases. However, the pairwise deletion of cases resulted in the same significant predictors although their order was different.

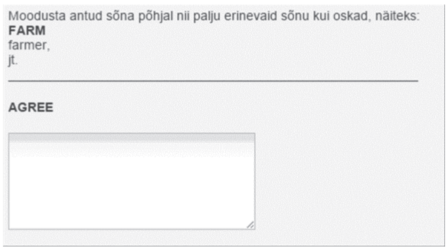
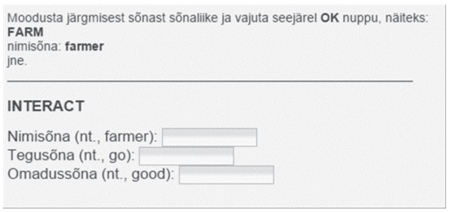
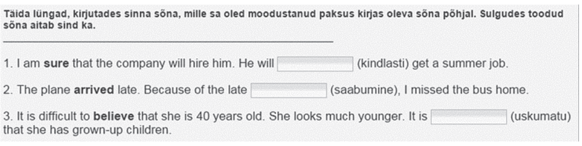
3. The writing proficiency variable was leptokurtic (i.e., peaked), which is not surprising considering that the sample selected for the study being represented by the learners studying in the same school year. To check whether there was any observable influence of the kurtosis of the DV on the results of the regression analysis, we randomly removed half of the cases at the B1 level (n = 16) from the sample, which resulted in the total of 89 cases in the analysis. The resulting variable was normally distributed ($S\text{-}W(89) = .98$, $p = .139$). The analysis conducted on this variable demonstrated that the same variables and in the same order significantly predicted the writing proficiency ($R^2 = .62$, $R_{adj}^2 = .61$, $F(3, 85) = 45.59$, $p < .001$).

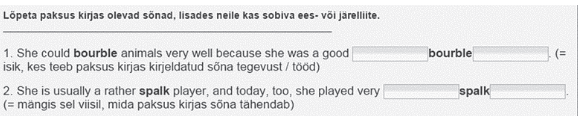
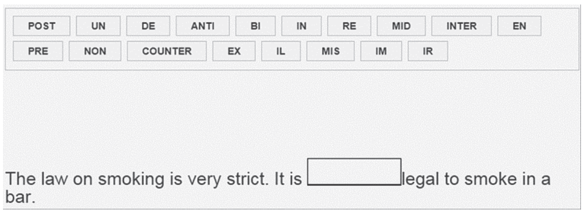
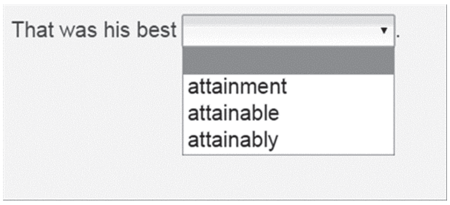
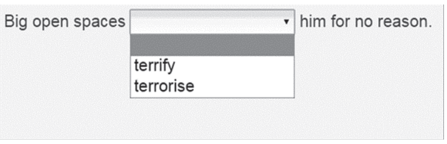
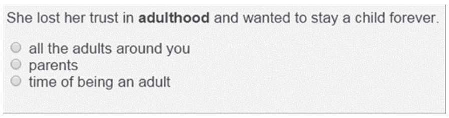
4. About 22% of the learners scored zero on the task, which can be considered a floor effect.

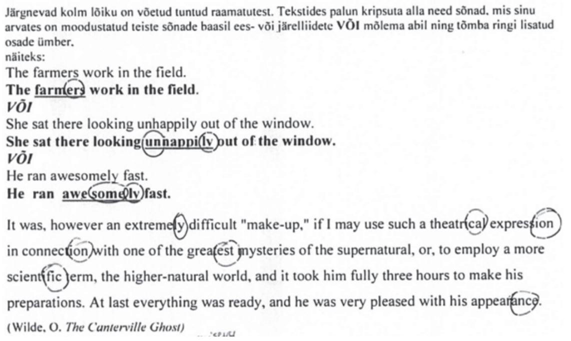
5. In fact, Welch's F-test results might also be not reliable when the distributions are differently skewed, as was the case with the non-word affix elicitation task. However, considering the *p*-value of the Welch's F-test being less than .001 (in effect, .00014), we think that its result was reliable enough. Thus the post-hoc pairwise comparisons could be conducted.

5. The meaning recognition task had a low reliability, and the affix elicitation task appeared as a significant predictor when the meaning recognition task was excluded.

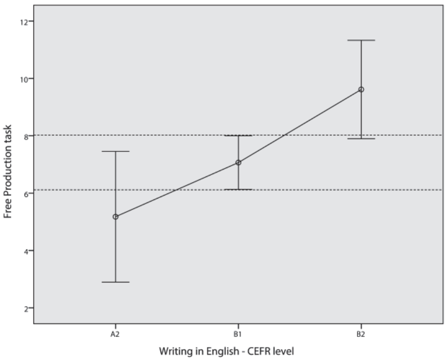
Appendix A. Measures used in / designed for the study and the rationale for their selection.

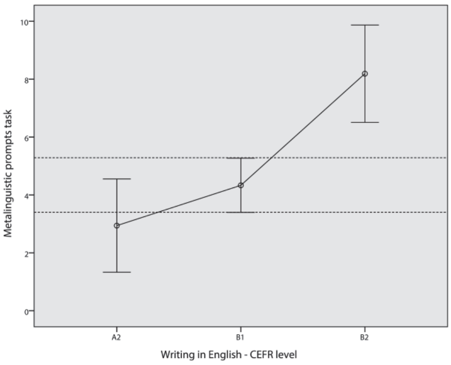
Name of the measure	Description	Sample item
Free production	Ability to produce different derivations formed from the base word (10 items); context-independent (see, Schmitt & Meara, 1997 for a similar measure).	 Moodusta antud sõna põhjal nii palju erinevaid sõnu kui oskad, näiteks: **FARM** farmer, jt. **AGREE**
Meta-linguistic prompts	Ability to produce derived words by using metalinguistic information (names of the parts of speech; 10 items, each requiring producing 3 parts of speech); context-independent; similar to Schmitt (1998)'s measure, although the modality of the latter was oral.	 Moodusta järgmisest sõnast sõnaliike ja vajuta seejärel **OK** nuppu, näiteks: **FARM** nimisõna: **farmer** jne. **INTERACT** Nimisõna (nt., farmer): Tegusõna (nt., go): Omadussõna (nt., good):
Affix elicitation	Ability to produce frequent, derived words in context (15 items); context-dependent (see, e.g., Friedline, 2011; Hayashi & Murphy, 2010, for similar measures).	 **Täida lüngad, kirjutades sinna sõna, mille sa oled moodustanud paksus kirjas oleva sõna põhjal. Sulgudes toodud sõna aitab sind ka.** 1. I am **sure** that the company will hire him. He will (kindlasti) get a summer job. 2. The plane **arrived** late. Because of the late (saabumine), I missed the bus home. 3. It is difficult to **believe** that she is 40 years old. She looks much younger. It is (uskumatu) that she has grown-up children.

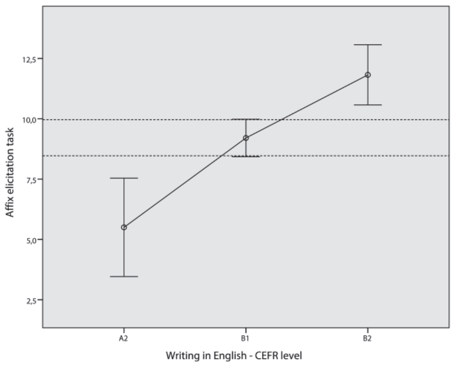
Non-word affix elicitation	Ability to produce derived forms of non-words (13 items); context-dependent. A modification of the affix elicitation task. The non-words were taken from the list developed for the Vocabulary Size Placement Test of Dialang (Alderson, 2005).	 **Lõpeta paksus kirjas olevad sõnad, lisades neile kas sobiva ees- või järelliite.** 1. She could **bourble** animals very well because she was a good □**bourble**□. (= isik, kes teeb paksus kirjas kirjeldatud sõna tegevust / tööd) 2. She is usually a rather **spalk** player, and today, too, she played very □**spalk**□. (= mängis sel viisil, mida paksus kirjas sõna tähendab)
Prefix elicitation	Ability to produce derived words by supplying the correct prefix, selecting the latter from the provided list (12 items); context-dependent. A modification of the affix elicitation task.	 POST UN DE ANTI BI IN RE MID INTER EN PRE NON COUNTER EX IL MIS IM IR The law on smoking is very strict. It is □legal to smoke in a bar.
Grammar recognition	Ability to recognize correct vs. incorrect derivation in terms of different parts of speech (and meaning); context-dependent (10 items). A measure similar to Akande's (2003) test of knowledge of inflectional affixes.	 That was his best ▾. attainment attainable attainably
Meaning recognition	Ability to recognize correct vs. incorrect derivation in terms of meaning; context-dependent (10 items). A modification of the grammar recognition task adapted for recognition of semantics of derivational affixes, i.e., same bases, different affixes in the options.	 Big open spaces ▾ him for no reason. terrify terrorise
Passive recognition of the meaning	Ability to recognize the meaning of derived words; context-dependent (10 items). Somewhat similar to Nation's (2008) (also Nation &	 She lost her trust in **adulthood** and wanted to stay a child forever. ○ all the adults around you ○ parents ○ time of being an adult

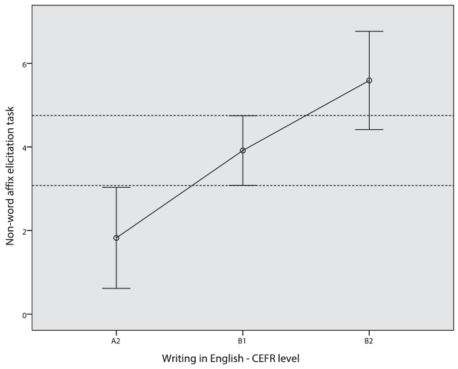
	Gu, 2007) Vocabulary Size Test adapted for word derivation, i.e., the options elicited different meanings of derivational affixes.	
Word segmentation	Ability to recognize derived forms / derivational affixes in text context; context-dependent (49 derivational affixes). Similar to Hayashi and Murphy's (2010) word segmentation task (although in the latter, single words were used)	 Järgnevad kolm lõiku on võetud tuntud raamatutest. Tekstides palun kripsuta alla need sõnad, mis sinu arvates on moodustatud teiste sõnade baasil ees- või järelliidete **VÕI** mõlema abil ning tõmba ringi lisatud osade ümber. näiteks: The farmers work in the field. **The farmers work in the field.** ***VÕI*** She sat there looking unhappily out of the window. **She sat there looking unhappily out of the window.** ***VÕI*** He ran awesomely fast. **He ran awesomely fast.** It was, however an extremely difficult "make-up," if I may use such a theatrical expression in connection with one of the greatest mysteries of the supernatural, or, to employ a more scientific term, the higher-natural world, and it took him fully three hours to make his preparations. At last everything was ready, and he was very pleased with his appearance. (Wilde, O. *The Canterville Ghost*)

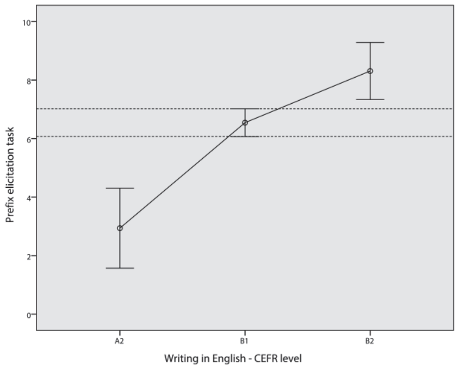
Appendix B. Differences across the CEFR levels in the measures: Means plots with 95% confidence intervals

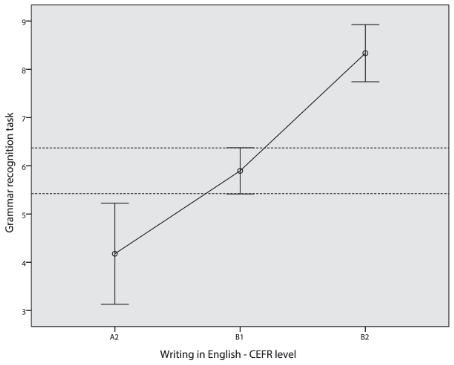


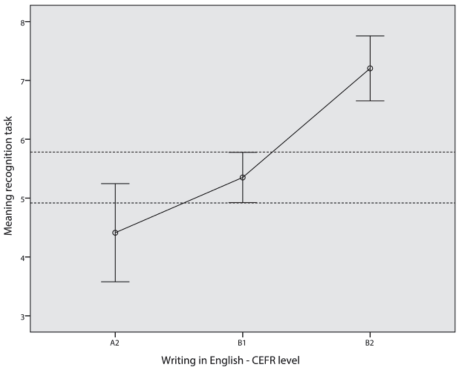


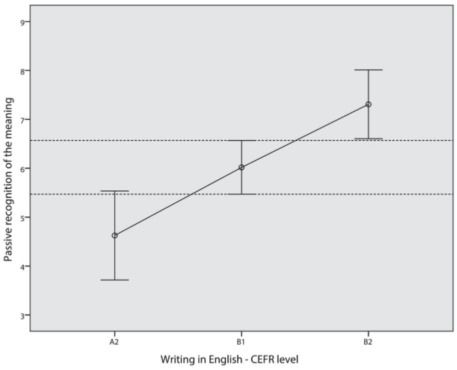

12,5
10,0
7,5
5,0
2,5
Affix elicitation task
A2
B1
B2
Writing in English - CEFR level

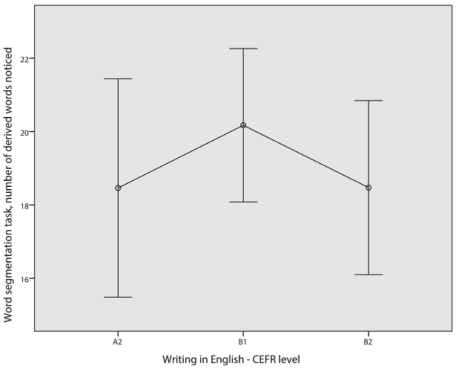

6
4
2
0
Non-word affix elicitation task
A2
B1
B2
Writing in English - CEFR level

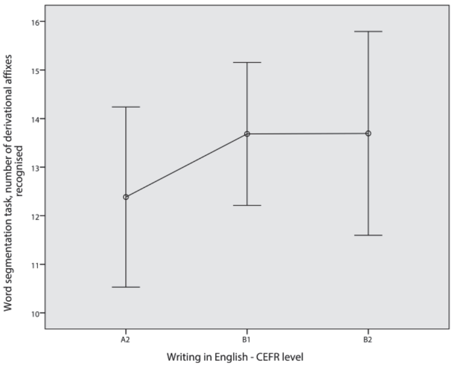

10
8
6
4
2
0
Prefix elicitation task
A2
B1
B2
Writing in English - CEFR level


9
8
7
6
5
4
3
Grammar recognition task
A2
B1
B2
Writing in English - CEFR level


Meaning recognition task
8
7
6
5
4
3
A2
B1
B2
Writing in English - CEFR level


Passive recognition of the meaning
9
8
7
6
5
4
3
A2
B1
B2
Writing in English - CEFR level


Word segmentation task, number of derived words noticed
22
20
18
16
A2
B1
B2
Writing in English - CEFR level


Word segmentation task, number of derivational affixes recognised
16
15
14
13
12
11
10
A2
B1
B2
Writing in English - CEFR level

References

Akande, A. T. (2003). Acquisition of the inflectional morphemes by Nigeria learners of English. *Nordic Journal of African Studies*, *12*(3), 310-326.

Alanen, R., & Kalaja, P. (2010). *The emergence of L2 English questions across CEFR proficiency levels*. Paper presented at AAAL 2010, Atlanta, Georgia, USA.

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

Alderson, J. C., Clapham, C., & Steel, D., 1997. Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research, 1*(2), 93–121.

Alegre, M., & Gordon, P. (1999). Rule-based versus associative processes in derivational morphology. *Brain and Language*, *68*, 347–354.

Bachman, L. & Cohen, A. (1998). *Interfaces between second language learning and language testing research.* Cambridge University Press.

Bartning, I., Martin, M. & Vedder, I. (Eds.) (2010) *Communicative proficiency and linguistic development: Intersections between SLA and language testing research.* EUROSLA Monograph series 1. European Second Language Association. Retrieved February 9, 2015, from http://eurosla.org/monographs/EM01/EM01home.html.

Bauer, L, & Nation, I. S. P. (1993). Word Families. *International Journal of Lexicography, 6*(4), 253–279.

Carlsen, C. (Ed.) (2013). Norsk profil. Det europeiske rammeverket spesifisert for norsk. Et første steg. Oslo, Novus forlag.

Chuenjundaeng, J. (2006). *An investigation of SUT students' receptive knowledge of English noun suffixes.* (MA thesis), Suranaree University of Technology.

Collins, B., & Nation, P. (2015). Testing receptive knowledge of derivational affixes. *Journal of Second Language Teaching & Research*, *4*(1), 6-23.

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment* [electronic version]. Retrieved February 10, 2015, from http://www.coe.int/t/dg4/linguistic/Source/Framework_en.pdf.

Extra, G., Spotti, M., & van Avermaet, P. (Eds.) (2009). *Language testing, migration and citizenship: Cross-national perspectives on integration regimes*. Continuum.

Finnish National Board of Education (2004). *National Core Curriculum for Upper Secondary Schools 2003* [Unofficial Translation]. Retrieved February 10, 2015, from http://www.oph.fi/download/47678_core_curricula_upper_secondary_education.pdf.

Friedline, B. E. (2011). *Challenges in the second language acquisition of derivational morphology: from theory to practice.* (Ph.D. dissertation), University of Pittsburgh.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing, 20*(4), 384–408.

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile Deutsch*. Langenscheidt.

Green, A. (2012). Language functions revisited. Theoretical and empirical bases for language construct definition across the ability range. Cambridge University Press.

Hayashi, Y., & Murphy, V. (2010). An investigation of morphological awareness in Japanese learners of English. *The Language Learning Journal, 39*(1), 105–120.

Huhta, A. (2012). Common European Framework of Reference. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell.

Hulstijn, J. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*, 663–667.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 36–55.

Leontjev, D. (2014). The effect of automated adaptive corrective feedback: L2 English questions. *APPLES: Journal of applied language studies, 8*(2), 43-66. Retrieved from http://apples.jyu.fi/ArticleFile/download/459.

Linacre, J. M. (2015). Facets computer program for many-facet Rasch measurement, version 3.71.4. Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2015). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com.

Mäntylä, K. & Huhta, A. (2013). Knowledge of word parts. In J. Milton. & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge*. Palgrave Macmillan, 45-59.

Martyniuk, W., & Noijons, J. (2007). *Executive summary of results of a survey on the use of the CEFR at national level in the Council of Europe member states*. Strasbourg: Council of Europe. Retrieved February 9, 2015, from http://www.coe.int/t/dg4/linguistic/Source/Survey_CEFR_2007_EN.doc.

Milton, J., & Fitzpatrick, T. (Eds.)(2013). *The Dimensions of Word Knowledge*. Palgrave Macmillan.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: an exploratory study. *System, 28*(2), 291–304.

Nagy, W. E., Diakidoy, I.-A. N., & Anderson, R. C. (1993). The Acquisition of Morphology: Learning the Contribution of Suffixes to the Meanings of Derivatives. *Journal of Literacy Research, 25*(2), 155–170.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston: Heinle & Heinle.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Põhikooli riiklik õppekava õigusakt; Lisa 2 [Basic School National Curriculum Act: Annex 2] (2010). Pub. L. No. RT I 2010, 6, 22. Retrieved February 10, 2015, from https://www.riigiteataja.ee/aktilisa/1281/2201/0017/13275423.pdf.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*(1), 77–89.

Ringbom, H. (1987). *The role of the first language in foreign language learning*. Clevedon: Multilingual Matters.

Ringbom, H. (1990). On the relation between second language comprehension and production. In J. Tommola (Ed.), *Foreign language, comprehension and production: AFinLA yearbook* (pp. 139–148). Helsinki: Suomen soveltavan kielitieteen yhdistys.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*(2), 281–317.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Palgrave Macmillan.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition, 19*(01), 17–36.

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36*, 145–171.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88.

Seidenberg, M. S., & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, *4*(9), 353–361.

Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language, 28*(6), 649–667.

Ullman, M.T. (2001). The neural basis of lexicon and grammar in first and second language: the declarative/procedural model. *Bilingualism: Language and Cognition, 4*(02), 105–122. doi: 10.1017/S1366728901000220

V

DYNAMIC ASSESSMENT OF WORD DERIVATIONAL KNOWLEDGE: TRACING THE DEVELOPMENT OF A LEARNER

by

Dmitri Leontjev

The article is published in *Eesti Rakenduslingvistika Ühingu aastaraamat [Estonian Papers in Applied Linguistics]* (2016)*, 12*, 141-160, doi: doi:10.5128/ERYa12.09

Dynamic assessment of word derivational knowledge: Tracing the development of a learner

Dmitri Leontjev

Abstract. The present paper reports on a case study that explored the applicability of dynamic assessment (DA) for promoting learners' word derivational knowledge in English as a second or a foreign language (L2). One learner's performance on tasks assessing his word derivational knowledge was measured four times. The first two measurements were conducted before and after three weekly human-mediated DA sessions and the last two, which took place a year and a half later, before and after three weekly computerised DA sessions. Think aloud protocols and interviews were used to trace changes in the learner's use of strategies and knowledge sources. The results revealed that following the dynamic assessment, the learner improved his performance and used strategies and knowledge sources more successfully. The findings have implications for designing dynamic tests of L2 English word derivational knowledge and for word derivational knowledge instruction.*

Keywords: sociocultural theory, mediation, inferencing strategies, knowledge sources, self-regulation, L2 learning, English

1. Introduction

Word derivation presents a problem to learners of English as a second or a foreign language, (Friedline 2011, Schmitt, Meara 1997). However, not much research on the acquisition of word derivation in English as a second or foreign language (henceforth L2) has been conducted. What is more, even less has been done as regards the way theoretical research findings can be applied in the L2 English classroom (Friedline 2011).

Nakayama (2008), for example, found that a systematic teaching of prefixes to Japanese learners of English was more effective for learning vocabulary than an unsystematic teaching, but only as regards immediate gains. The limitation of the study was that the author did not compare the groups prior to the intervention.

Friedline (2011) aimed at acquiring a better understanding of the construct of word derivational (henceforth WD) knowledge and the way it can be trained. He first established differences in performance on several word derivation tasks of native speakers of English and L2 English learners, as well as differences between the learners of different mother tongues (L1s) and levels of proficiency. He then investigated whether pushed output, that is, collaborative dialogue in which learners are directed to producing output (e.g., Swain 1998) would be more effective for improving learners' WD knowledge than simple input processing. Contrary to his hypothesis, there was no significant difference between the groups (although both improved their performance). He suggested that it could have been the influence of novelty effect, i.e., novelty of information increasing the possibility of its long-term storage (e.g., Tulving, Kroll 1995) that outweighed the effect of the treatment. Finally, using Activity Theory framework, Friedline studied how learners' beliefs, attitudes, and actions changed in the course of the study. finding that learners integrated morphology into their language learning strategies (LLS) repertoire.

These studies produced important insights into the ways learning of WD knowledge can be guided. However, a deeper understanding of how training promotes the development of learners' WD knowledge is required. The aim of the present study is to understand how

* I am grateful to the participant in the study and would like to thank two anonymous reviewers for their comments and suggestions.

dynamic assessment (DA), being a pushed output activity, directed one learner's performance and promoted his WD knowledge. Before presenting the study, I will outline some of the research on L2 English word derivation and learning strategies / learners' self-regulatory behaviour.

2. Background

2.1. Research on L2 English word derivation

Studies of L2 English word derivational knowledge are not numerous. However, some interesting findings have been produced. For example, evidence for its incremental development has been found (e.g., Schmitt 1998, Schmitt, Meara 1997). Following the developmental paradigm adopted in these studies, it is logical to assume that some affixes can be easier to learn than others. With this intention in mind, Bauer and Nation (1993) proposed a teaching order of derivational affixes based on a number of their morphological, phonological, and orthographical properties (Table 1).

Table 1. Affix difficulty order (Bauer, Nation 1993)

Level 1	A different form is a different word.
Level 2	Inflectional affixes.
Level 3	The most frequent and regular derivational affixes, e.g., -able, -er, -less, -ly, -ness, un-.
Level 4	Frequent and regular affixes, e.g., -al, -ation, -ful, -ism, -ize, -ment, in-.
Level 5	Infrequent but regular affixes, e.g., -age -ship, mis-,etc.
Level 6	Frequent but irregular affixes, e.g., -ee, -ion, re-, etc.
Level 7	Classical roots and affixes, e.g., -ate, -ure, etc.

While not much empirical evidence exists for the validity of the order, Leontjev (to be published), for example, found that for the most part (i.e., except for no significant difference between affix levels 5 and 6), the order holds as the order of difficulty learners have with recognising derivational suffixes.

In addition, a link between learners' L2 English WD knowledge and proficiency has been found (e.g., Leontjev et al. to be published, Mäntylä, Huhta 2013). It appears that this link depends on the operationalisation of learners' WD knowledge. Friedline (2011) found that learners' proficiency seemed to relate to learners' performance on a word relatedness task (asking to indicate whether two words, e.g., productive–production, are related) although this assumption was not tested statistically. On the other hand, he did not find a relationship between learners' proficiency and their performance on lexical decision (asking to rate the certainty that presented derived words were real) and word decomposition (asking to write base forms of the presented derived words) tasks. Mäntylä and Huhta (2013) found strong correlations between learners' proficiency and their performance on affix elicitation tasks. Finally, Leontjev et al. (to be published) demonstrated that both syntactic and semantic knowledge of derivational affixes strongly predicted learners' writing proficiency.

However, a question still remains how exactly derivational affixes should be taught. Friedine (2011) found that following the treatment, learners integrated morphology into their LLS repertoire, at the same time, each still using their own array of strategies. This suggests that as learners become more self-regulated in the use of word derivation, they adopt new techniques to regulate their learning, as will be also outlined in the following section.

2.2. Strategy use or self-regulatory capacity

It has long been noticed that learners regulate their learning by using a number of techniques, and that self-regulatory capacity increases as their abilities grow (Dörnyei 2005). It is no wonder that considerable research has been conducted targeting language learning strategies (see, e.g., Dörnyei 2005 for a discussion). Based on the proposed LLS taxonomies (Oxford 1990; O'Malley, Chamot 1991), LLS can be divided into the following categories:

- metacognitive (planning the learning process);
- cognitive (manipulating the material to be learned);
- social/affective strategies (interacting with peers and adjusting one's beliefs, feelings, and emotions).

As Dörnyei (2005) noted, the LLS-based paradigm has several issues, including the fuzziness of construct definition, classifications of LLS, and methods of study. Instead, a shift from LLS (i.e., product) to self-regulation (i.e., process) was proposed. Tseng et al. (2006) designed an instrument aiming to tap into learners' self-regulatory processes in vocabulary learning. The instrument was a Likert-scale type questionnaire, its items falling into one of five facets:

- commitment control (helping to preserve learners' commitment to original goal);
- metacognitive control (controlling the concentration on the task);
- satiation control (eliminating boredom);
- emotion control (generating emotions that help to implement to goal, e.g., self-encouragement);
- environmental control (minimising negative and making use of positive environmental influences, e.g., asking friends for help).

However, as, for example, Rose (2012) argued, the model suggested by Tseng et al. (2006) is compatible with LLS-based models and they should rather represent parts of the same construct. Rose (2012) also urged for a more qualitative research of strategic learning.

Nassaji's (2003) study would have been difficult to conduct only within the paradigm suggested by Tseng et al. (2006). The author explored the relationship between learners' vocabulary inferencing strategies and inferencing success. Interestingly, the author, in addition to strategies, also considered knowledge sources, defining the strategies as cognitive or metacognitive actions used to understand the problem and/or overcome it and knowledge sources as references to particular sources of knowledge (e.g., phonology). Importantly, instead of a questionnaire, the author used think aloud protocols and interviews as data collection tools. The author found that while morphological knowledge had the highest rate of success, not any one knowledge source or strategy alone resulted in successful inferencing, but rather combinations of these did. He concluded that it was not the quantity of strategies that mattered but their quality.

It should be noted that Bowles (2010), found that thinking aloud can, in some cases, facilitate learning. However, as the author noticed, the results vary, and generally, the effect of thinking aloud as compared to silent thinking is small.

As regards strategy instruction, several studies (e.g., Kozulin, Garb 2002, Teo 2012) aimed at discovering whether mediating learners' strategies in dynamic assessment improved their abilities. Next, I will discuss these studies in some detail.

2.3. Dynamic assessment of L2

Dynamic assessment (DA) developed at the crossroads of assessment and instruction as an alternative to conventional assessment, which DA proponents often refer to as static assessment (SA). It builds on Vygotskian concept of Zone of Proximal Development, which is “the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers” (Vygotsky 1978: 86). Application of this concept to assessment resulted in a view that no assessment can provide a full picture of learners’ development without incorporating their potential for development. In DA, this is achieved by providing guided support, known as mediation, which aims at both discovering learners’ potential development and promoting their abilities (Poehner 2008). Mediation in DA is often operationalised as a number of feedback messages gradually becoming more explicit and detailed until learners are either able to self-correct their mistakes or are provided with the correct response (Poehner, Lantolf 2013; Teo 2012).

It has been demonstrated that DA is successful in promoting learners’ L2 abilities (Leontjev 2014, Kozulin, Garb 2002, Poehner, Lantolf 2013, Teo 2012). Some of these studies reported on computerised DA (Leontjev 2014, Poehner, Lantolf 2013, Teo 2012). The advantages of the computerised modality include the possibility to assess several learners simultaneously. However, computerised DA is limited to *interventionist* approach, in which mediation is standardised and is provided in a predefined fashion. Often, the dynamic part in interventionist DA is conducted between a static pre- and posttest (the so-called sandwich format; Poehner 2008). In contrast, in *interactionist* DA mediation emerges in interaction between the learner and the assessor. When learners’ development within one or across several DA sessions is traced, transfer items, that is, items assessing the same feature, can be used to trace the increase in learners’ abilities (Poehner, Lantolf 2013).

Some of DA studies have an explicit focus on LLS. Kozulin and Garb (2002) studied the effect of mediating learners’ LLS. The authors found that DA improved learners’ reading comprehension. Unfortunately, they did not illustrate the actual mediation process nor reported on the learners’ use of strategies following the DA.

Teo (2012) studied learners’ LLS after a computerised DA. The author found that the computerised DA helped the learners to use a number of strategies appropriately, which, she argued, improved their inferential reading abilities. However, the author did not collect any data on the learners’ LLS use before the DA.

3. Methodology

3.1. Research question

The previous research has produced important insights learners’ L2 English word derivational knowledge and on the way dynamic assessment promotes the development of learners’ abilities. The present case study aims at combining these two strands of research by finding answers to the following question:

- How, if at all, does dynamic assessment promote L2 English learners’ ability to derive words?

The particular emphasis in the study will be on the way the participant regulated his learning prior to and following dynamic assessment.

3.2. Participant and data

The participant in the study was a L1 Russian learner (16 year old) studying English at grade 10 of an Estonian school at the onset of the study. Hereinafter in the paper, he will be referred to as M.

Nation (2001) suggested that L2 word derivation instruction is beneficial to learners at lower-intermediate level, which is roughly equivalent to level B1 on the Common European Framework of Reference scale (Council of Europe 2001), the L2 proficiency that learners in Estonia are expected to be by grade 10 (Põhikooli riiklik õppekava õigusakt: Lisa 2, 2010). Moreover, the results of Leontjev et al. (to be published) suggested that learners' WD knowledge increases after level B1 is reached. Thus, the participant was selected among ten-graders.

By the time of the study, M had been studying English for about seven years. He also revealed that at school, he was occasionally taught word derivation, which reduced the possibility of the novelty effect due to the introduction of WD during the treatment.

The data come from a) M's performance on four computerised static assessment (SA) sessions, each consisting of seven tasks requiring M to demonstrate different aspects of his WD knowledge, b) M's think aloud protocols collected when he was working on the first three items of each SA task, c) four interviews immediately following each SA session, and d) M's performance on three weekly human-mediated and three weekly computerised dynamic assessment sessions. The tasks were administered in an online tutoring/assessment system (see Leontjev 2014 for details). The procedure was the following:

1) two SA sessions, one preceding and one following three weekly human-mediated DA sessions, both SA followed up with an interview;
2) a year and a half gap;
3) two SA sessions, one preceding and one following three weekly computerised DA sessions, both SA followed up with an interview.

The decision to have a year and a half gap was due to a modest at best improvement in the knowledge of derivational affixes in the course of one academic year found by Schmitt and Meara (1997). Therefore, a larger gap was introduced to allow for a greater increase in M's word derivational knowledge. By the time of the third SA session, M was at the end of grade eleven and was 17 years old.

A combination of think aloud protocols and research interviews was used to establish strategies and knowledge sources that M used during the SA (cf. Nassaji 2003).

The task types in the SA were:

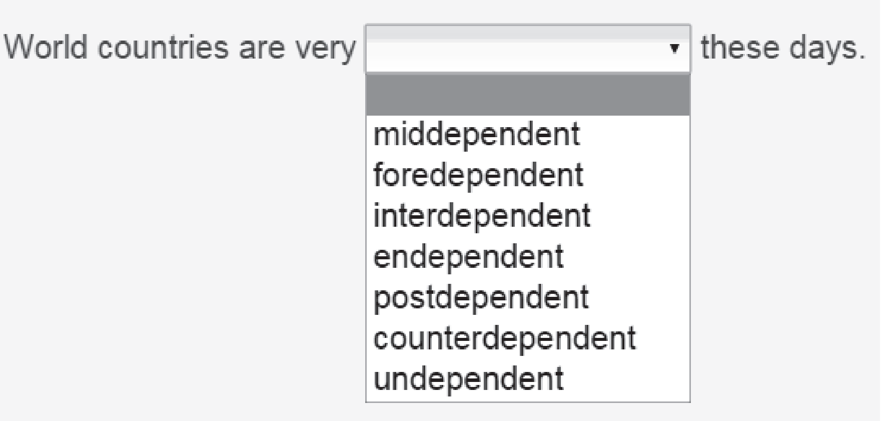
- free production (form as many words as possible from the given words);
- metalinguistic prompts (form different parts of speech from the given words);
- non-word affix elicitation (complete the non-words in the sentences using the context and the explanations);
- prefix elicitation (complete the words in the sentences using provided prefixes);
- grammar recognition (complete the sentences selecting one option among provided; same base, different affixes forming different parts of speech);
- meaning recognition (same as previous, but the options were the same parts of speech);
- passive recognition of the meaning (select the definition among provided to the highlighted words in the sentences).

Sample items from the SA tasks are presented in Appendix 1. For further details on the tasks, see Leontjev et al. (to be published). No feedback on M's performance was given to him before the end of the final SA session.

Both the human-mediated and the computerised DA were designed following the interventionist sandwich DA format (see Section 2.3). Based on the findings of Leontjev et al. (to be published), the mediation targeted M's use of syntactic or semantic knowledge of affixes, or both. The task types in the human-mediated DA were:

- classification exercises, e.g., *which of these words are adverbs; what parts of speech are the rest of the words:* ***momentary, literacy, ability, hyperactively***;

- affix elicitation exercises, e.g., *on the basis of the word in the brackets, form a word that fits the sentence*: ***They want to raise ……….. (aware) of the problem***; Multiple-choice task format was used in the computerised DA (Figure 1).



World countries are very ▾ these days.

Figure 1. Sample item from the computerised dynamic assessment

The difficulty of the items was operationalised as Bauer and Nation's (1993) affix levels (Table 1). Generally, in earlier DA sessions, affixes at lower Bauer and Nation's levels were used than in later sessions. However, several transfer items (i.e., different items with the same affixes) were included to later DA sessions to see if there was any progress in the use of these affixes. Thus separate higher levels affixes appeared at earlier DA session, and some lower level affixes were used in later sessions.

The desktop video recordings of M's SA performance were made. These were used during the interviews to help M recall what he had been doing while working on the tasks. The human-mediated DA sessions were audio recorded. Detailed logs of M's performance on the computerised DA were recorded by the online system.

3.3. Analysis

To determine whether there was any progress in M's unassisted performance, his correct responses on the SA tasks were counted across the sessions. Then, the assistance M required during the computerised and the human-mediated DA was compared across the DA sessions.

The video- and audio-recorded data were transcribed, coded, and analysed with the help of Atlas.ti qualitative analysis software. The coding was done by two coders (I being one of them) independently, and then agreed upon in the cases where dissimilar decisions were made. The coding was inspired by Nassaji's (2003) list of strategies and knowledge sources, but, above all, the codes emerged from the analysis of the transcript (Tables 2 and 3).

Table 2. M's strategies

Strategy	**Description**
Repetition	repeating any portion of the text.
Verifying	checking the appropriateness of the response against the wider context.
Self-inquiry	asking oneself questions.
Analysing	analysing a word morphologically.
Monitoring	showing awareness of the problem or the difficulty of the task.
Analogy	drawing on similarities with other words.

Table 3.M's knowledge sources

Knowledge source	Description
Syntactic knowledge: a) affixes b) words	knowledge of syntactic functions of affixes or words.
Semantic knowledge: a) affixes b) words	knowledge of the meanings of the affixes or words (either translations or definitions).
Mother tongue/English	L1/L2 analogy

The strategies identified in the analysis of the transcript were classified as either cognitive (e.g., analysing) or metacognitive (e.g., self-inquiry). This is because social/affective strategies were generally not present during the SA, the exception being one use of a social strategy (5). Analogy is present in both tables since both coders agreed that the knowledge source for analogy (i.e., English or L1) had to be specified. Since except in the last three SA tasks, base words were given, with several exceptions (e.g., M recognising *inter-* in *interactive* in the metalinguistic prompts task in the last interview), the analysing strategy was identified in the last three tasks only.

In addition, in line with Vygotsky's (1978) sociocultural theory, not only the DA, but also the previous SA sessions, the interviewer's utterances, and what M' himself had previously reported mediated his self-reports. Therefore, each interview, think aloud protocol, and the whole data set collectively were also analysed holistically.

4. Results

First, I will demonstrate that M's performance improved across the static assessment sessions. Then, I will trace how M's use of strategies and knowledge sources changed in the course of the study. Finally, I will demonstrate how dynamic assessment facilitated these changes and how both the interviewer and M's own verbalisation of his thinking guided M's performance.

4.1. Improvement of M's *performance*

Table 4 illustrates changes in M's performance. It should be noted that, as some of Bauer and Nation's level 2 affixes (e.g., *-ing*) can be both inflectional and derivational, I limited the figures to words formed with help of levels 3 to 7 (see Table 1). Although the number of base words in the free production task was 10, the number of words possible to form was not limited. In the metalinguistic prompts task, while there were 10 items, the total number of words possible to form was thirty.

Table 4. M's performance on the tasks across the four SA sessions

	1	**2**	**3**	**4**
Free production	12	26	23	25
Metalinguistic prompts	11	15	16	18
Non-word affix elicitation (k = 10[1])	4	7	9	10
Prefix elicitation (k = 12)	8	11	10	11
Grammar recognition (k = 10)	8	8	9	9

[1] Three items containing Bauer and Nation's level 2 affixes were removed from the scale.

Meaning recognition (k = 10)	6	7	9	8[2]
Passive recognition of the meaning (k = 10)	8	8	9	9

The biggest increase in M’s performance was between SA sessions 1 and 2. The computerised DA resulted in a smaller increase, which at least in one task, can be attributed to the ceiling effect. There seemed to be little or no progress in the grammar recognition and the passive recognition of the meaning tasks. However, as the qualitative analysis revealed, the way that M worked through the tasks (including the two tasks where there seemed to be no improvement) was different across the SA sessions (see Section 4.2). It is also interesting to note that during SA sessions 2 and 4, M correctly used several affixes (e.g., *en-, -en, -ic, -ist*) that were not trained in the DA. What is more, generally, M required less help with the transfer items during the DA (see, e.g., Example 4 and the following discussion).

4.2. M’s strategies and knowledge sources

Already during the first SA session, M reported to use a number of strategies, as is exemplified in Example 1 (see Figure 2.6 in Appendix 1 for a sample item) from the think aloud protocol (hereinafter in the transcript, interviewer = **I**; the transcription symbols are explained in Appendix 2). I found that the English translation was sufficient for reporting on M’ strategies and knowledge sources. Thus, the original Russian transcript will not be supplied. I will, however, note pauses, intonation, non-verbal behaviour, etc. I will supply line numbers in longer examples. Note that the parts that were originally in English are italicised.

(1)

1 **M:** *Bi:g open spaces* (2.0) [opens up the drop-down menu] *hi:m for no reason. (5.0)*
2 Hm: (1.6) Big open spaces in some way influence him (0.8) for no reason. That is,
3 here, by implication, fits a verb. That is (.) *terrorise* (1.8) or *terrif-*? *Terrify*. (2.2)
4 **I:** So, what’s with them?
5 **M:** What exactly? (4.2) A:h scare him? Big open spaces scare him (.) without any
6 special reason for that [selects ‘terrify’] (1.8).
7 **I:** Right. And why did you choose it?
8 **M:** Because it is a verb (3.1) [opens the drop-down menu; looks at the options] (6.8)
9 **I:** Continue (.) thinking.
10 **M:** Uf:: (4.2) *terrify* (4.2) *terrify terrorise* (5.6)
11 **I:** Speak out your thoughts.
12 **M:** I now think that *terrify* is an adverb (1.6) and *terrorise* is a verb (4.2).
13 **I:** The reason?
14 **M: (**4.1**)** Because *terrorise* has the ending (0.8) *aɪ es i:*, which is the ending of some
15 verbs, for example, (0.4) *rise.*

From Example 1, it transpires that M used verifying (lines 1 and 5-6), repetition (lines 3 and 10), self-inquiry (lines 3 and 5), analysing (line 14) and analogy (line 15). These strategies (except for analysing) were common during the first SA, and some of them could have been beneficial for finding the correct response. However, this task required M to think about the meanings of derivational affixes, but he resorted to his syntactic knowledge. In fact, throughout the first SA session, M referred to semantics of affixes only seven times, often failing to do so even in the tasks that were difficult to complete correctly otherwise.

[2] Considering that M understood his mistakes in the task before the end of the test (see Section 4.4), the score can be raised to 10.

In the following SA sessions, M reported on the semantics of affixes considerably more, that is, fourteen, twenty-two, and thirty times respectively. Interestingly M's use of this knowledge source was slightly different after the human-mediated and computerised DA. The usual way M referred to this knowledge source after the human-mediated DA is exemplified in (2).

(2)

M: Here we have two adjectives (3.8).

I: Right.

M: We:ll, -*able* means aptitude for something.

M first acknowledged that both options were adjectives and only then analysed one of the options, supplying the meaning of suffix -*able*. This pattern was rather frequent during the second SA. For example, in the meaning recognition and the passive recognition of the meaning tasks (where semantic knowledge was required), M used it seven times. M still used this pattern during the third SA four times.

However, the usual way M worked through the meaning recognition and the passive recognition of the meaning tasks was different during the last SA (see Example 3 from the interview).

(3)

I: *Clarity*. Why?

M: (2.2)

I: Why not *clarification*?

M: Because it is a process.

I: Uhu. And *clarity*?

M: It's like (0.7) well, like a quality.

I: Right. Do you know these words or where did you- (.) or what (.) process, quality?

M: Well, suffixes.

That is to say, during the last SA session, M did not rely on syntactic knowledge in these two tasks.

In addition, during the first SA, M only occasionally tried to analyse the words (all in all, seven times) whereas in later SA sessions, this number increased to ten, fifteen, and nineteen times respectively. This is not to say that M used only one strategy / knowledge source to solve each item. On the contrary, in most cases, it was a combination of several of them, as Example 1 illustrates. For example, in the third SA, M often combined other strategies with knowledge of meanings of words, successfully using it all in all 30 times, as compared to 15 and 20 during the first two SA sessions respectively). This suggests that M's vocabulary knowledge increased, which can explain the improvement between SA sessions 2 and 3 (Table 4). Interestingly, after the computerised DA, M's use of this knowledge source decreased to 24.

It should also be noted that in later SA sessions, M's certainty in his responses increased, as manifested in the decreased frequency of using monitoring, repetition, and self-inquiry. The analysis of M's DA performance sheds more light on these changes.

4.3. Dynamic assessment

The way mediation was provided to M during the human-mediated DA is illustrated in Example 4 from the third DA session.

(4)

1 **I:** Look at the seventh
2 **M:** (5.2)
3 **I:** sentence. Which part of speech do we need to form? *He is known for ↑hi:s-*
4 **M:** Fearlessness—a noun.
5 **I:** Right. And what do <u>you</u> have?
6 **M:** Ah (0.6) an adjective.
7 **I:** Right. Something is missing. That is (.) you have the adjective *'fearless'*. Which
8 means ↑what?
9 **M:** Fearless.
10 **I:** So, what you need to add is a suffix that makes it into a noun.
11 **M:** (4.0).
12 **I:** Think what the word means. What is fearlessness?
13 **M:** A quality.
14 **I:** <u>Great</u>!
15 **M:** (16.5) *Fearnessless*?
16 **I:** Yes, but vice versa.
17 **M:** ((laughter)) *Fearlessness*. ((laughter))

The interviewer, first, drew M's attention to the sentence with a mistake. He then elicited the syntactic function of the word and invited M to use the context (line 3). While M established that a noun was required (lines 4-6), he was still hesitant as to which suffix to use, so the interviewer asked M to think about the meaning of the word. That M first provided the meaning and then solved the item was not coincidental, as during the first human-mediated DA session, M was explicitly told that the meaning of *-ness* was that of *quality*. This was but one example of M requiring less help with transfer items.

A difference between the human-mediated and computerised DA was that during the computerised DA, M never selected an option which was a wrong part of speech. Therefore, the mediation during the latter did not elicit syntactic functions of the affixes (Table 5).

Table 5. Performance log from a computerised DA session (English translation)

Try	Mediation
1	Your answer: The reflectable surface of the lake shines in the sun. Think more carefully.
2	Your answer: The **reflectant** surface of the lake shines in the sun. Read your sentence carefully. Think what the suffix that we need can mean. Which suffixes among provided do you think can mean it.
3	Your sentence: The reflect**ory** surface of the lake shines in the sun. Suffix **-ory** means **serving for something** or **characterised by something**. The suffix that we need means **doing something specified**.
4	Your sentence: The reflective surface of the lake shines in the sun. Correct.

As is demonstrated in Example 4 and Table 5, both in human-mediated and computerised DA, the mediation, did not explicitly instruct M to analyse the words, but still elicited this strategy. Depending on M's responses, the instruction to use specific knowledge sources varied in explicitness.

While, as has already been mentioned, generally, M required less help with transfer items, in a small number of cases, M required more assistance with them. One such case was suffix -*ive*, with which M required level 1 feedback during the first computerised DA session and level 3 feedback, during the third session (see Table 5). In both cases, M's first choice was suffix -*able*, which suggests that he was still not fully self-regulated in its use as he was not in using suffix -*ive*. This is also evident during the last SA session, as will be discussed in Section 4.4.

It should also be noted that M required less help during the computerised DA as compared to the human-mediated DA.

4.4. Mediation during the static assessment

As mentioned in Section 3.3, the interviewer eliciting responses from M appeared to direct M's performance. One example of it was Example 1, where the interviewer was pushing M to verbalise his reasons for selecting *terrify*, which was the correct response, but, as became apparent, was selected for the wrong reasons. As such, it was not a typical think aloud procedure, and it resulted in that M selected the incorrect option. This might be considered a negative influence of the interviewer's intervention, but in fact, it resulted in a more accurate representation of M's ability.

However, it was not just the interviewer who mediated M's performance. In Example 5 from the last SA session, M was thinking aloud while solving the item *You must show* ***demonstrative*** *improvement of your work* from the meaning recognition task.

(5)

M: Here, it is again a difference in meaning. If you put *demonstratable* (.) it means that improvement is able to demonstrate itself. ↑M: (3.0) m: *demonstrative* (.) is demonstrative. (2.8) Here (.) it is °demonstrative° (6.1).

I: Right.

M: You know it, but I don't know. ((chuckle)) (4.0) I'm leaning towards ↑ demonstratable (3.2).

I: Right.

M: No, demonstrative ((chuckle)) (10.2) demonstrative.

Apparently, M was not sure which option was correct, as is manifested in his pauses, rising intonation, pronouncing the option *demonstrative* softly and quietly, acknowledging the interviewer as a master of the ability, and contrasting the latter's abilities with his own. Interestingly, in the following task (i.e., passive recognition of the meaning), M's performance on the item with *-ible* (*suggestible*) was different (6).

(6)

M: [selects 'can be easily changed by others'] (3.1)

I: And how (.) why the third option?

M: I finally remembered this (.) after the third practice, the third time taking this test [actually, the fourth], I remembered (0.6) what suffix -*able* means. It means that the children are subject to be influenced. Well (.) that's the meaning (1.8).

I: m:

> **M:** That is, it's not that they are able [to do something], but they are able to be influenced.

Without much thinking, M selected the correct option and produced a coherent explanation. What is more, during the interview which followed immediately after the SA session, M laughed when he saw the video recording of him working on the item (5) and told the interviewer that because he thought that *-able* had a different meaning, he actually made two mistakes in this task, one where he used *-ive* in place of *-able* and the other where he used *-able* instead of *-ive*. I will discuss this change with reference to self-mediation in Section 5.

5. Discussion

The present study endeavoured to find answers to (a) whether dynamic assessment (DA) can promote learners' L2 English word derivational (WD) knowledge and (b) how it can do it.

The results confirmed that DA, both human-mediated and computerised, improved the participant's WD knowledge operationalised as his scores on static assessment (SA) tasks and his performance on transfer items. The increase in M's performance after the computerised DA was smaller, but considering the fact that after the computerised DA, M performed at the ceiling on the non-word affix elicitation (perhaps, also meaning recognition) task and the fact that the difference between sessions 3 and 4 (i.e., due to the DA) was similar to or bigger than that between sessions 2 and 3 (a year and a half time), this was a noticeable increase. The relatively small increase in M's unassisted performance between SA sessions 2 and 3 can be explained with reference to Schmitt and Meara's (1997) finding that there was not much of improvement in their participants WD knowledge within one academic year. What is more, during the SA sessions that followed DA, M often recalled the meanings of the affixes that were taught to him during the dynamic assessment (e.g., Example 4), but also improved his performance in the use of affixes that he was not taught during the DA.

The analysis of the transcript revealed that, in line with the previous studies (e.g., Kozulin, Garb 2002, Teo 2010), in addition to the content knowledge, DA promoted M's use of strategies and knowledge sources. Specifically, owing to the mediation M received during the DA, he started analysing words morphologically and referred to semantics of derivational affixes more frequently than before the DA. The connection found between M's self-reports and the mediation in the DA also suggests that it was dynamic assessment that led to these changes.

That is to say, M learned to analyse words to get their meaning, paying attention to both the affixes and bases, but also realised that syntactic knowledge, while being useful, does not always help. Thus, in addition to learning some suffixes, M was able to recognise other suffixes, which improved his performance as well.

It is important to emphasise though that DA did not result in the emergence of new strategies in M's repertoire—all the strategies and knowledge sources that he used during later SA sessions had already been present during the first SA. What is more, during the later SA sessions, M successfully used strategies that were not elicited during the DA.

The latter can be interpreted with reference to the model of Tseng et al (2006). The mediation provided to M (Example 4; Table 5) reminded him of the goals by eliciting that he had to pay attention to affixes, thus also helping him to stay concentrated on the tasks. Importantly, at later DA sessions, less mediation was provided, which should have confirmed that the techniques that M had been using previously were successful and gave him more control in selecting these techniques. This resulted in that M became aware how certain strategies and knowledge sources helped *him* to improve his performance. In other words the change in his strategy use was qualitative rather than (or in addition to) quantitative (cf. Friedline 2010, Nassaji 2003).

The presence of the interviewer, who urged M to continue thinking aloud, also appeared to guide M's decisions. In Example 1, but also, as Example 6 demonstrates it, M appeared to consider the interviewer's utterances as indicative of (in)correctness of his reports. Thus, although being told that the role of the interviewer/researcher during the SA was to learn about M's thinking, M still perceived him as a person whom he could turn to for help. However, the cases which can be interpreted as the interviewer mediating M's performance this mediation actually resulted in performance which reflected M's WD knowledge more accurately.

Interestingly, following Example 5 (i.e., in the following task), M acknowledged the two mistakes he made in the previous task and corrected them. In other words, should M had been given a possibility to go back to the items, he would have had a perfect score on the meaning recognition task in the last SA session.

Vygotsky's (1978) understanding of Zone of Proximal Development offers an explanation for that. Vygotsky considered that development continues even after it switches from the interpersonal to the intrapersonal plane. An example he provided was a child verbalising his/her own following actions, thus guiding these actions. As the last SA session demonstrated, the DA alone was not enough for M to learn the correct meanings/use of -*able* and -*ive*. However, M's self-mediation resulted in that he was finally able use -*able* and -*ive* correctly. which he also confirmed during the final interview. This also suggests that static assessment was not that static for M after all. I will list this and other limitations in Section 6.

6. Conclusion

The present study aimed at understanding how (if at all) dynamic assessment can promote the development of L2 English word derivational knowledge. I initially hypothesised that DA should promote the use of certain strategies and increase the participant's overall self-regulatory capacity.

The results spoke in favour of the hypothesis. This is not to say that M did not have access to these strategies and knowledge sources prior to the DA. However, because of the DA, M started using certain strategies more frequently and learned which techniques helped him to solve the tasks requiring demonstration of L2 English WD knowledge, which generally allowed him to use these techniques in proper contexts.

These findings have several implications. First of all, they suggest that adapting feedback to learners' performance can promote their L2 English word derivational knowledge, making their learning more strategic. Furthermore, the study exemplifies how a dynamic test of WD knowledge can look like, which has implications both for test designers and for further research, including, but not limited to, quantitative studies aiming at establishing the effectiveness of DA in promoting learners' WD knowledge.

This said, the study has some limitations (which I would like to discuss next). Above all, as with all case studies, the findings lack generalisability. Further research should be conducted to confirm or disprove the findings of the present study. The second limitation arises due to the method selected for the study. Both M's thinking aloud (cf. Bowles 2010) and the interviewer's intervention, however small, mediated M's SA performance. Thus, it is not possible, for example, to ascertain whether M's performance would be the same should he have not been thinking aloud. On the other hand, limiting the data to interviews (i.e., a retrospective method) only would make the results less reliable, due to the lack of methodological triangulation. What is more, the study, above all, aimed at establishing how DA changed the way M approached the tasks rather than calculating reliable scores across the SA sessions. The last limitation arises from the difference in the tasks types and the modality of different DA sessions. The tasks in the human-mediated and the computerised DA were different. Therefore, although M required less help during the computerised DA, it cannot be assumed that it was only because of the development of his WD knowledge. Moreover, the modality of the assessment was different. Thus a definite conclusion cannot be made in this regard.

Despite these limitations it is hoped that the study produced interesting insights into the development of L2 English word derivational knowledge and ways that dynamic assessment can guide this development.

References

ATLAS.ti. Version 5.0. [Computer software] 2004. Berlin: Scientific Software Development.

Bauer, Laurie; Nation, I.S. Paul 1993. Word Families. – International Journal of Lexicography, 6 (4), 253–279. doi: 10.1093/ijl/6.4.253

Bowles, Melissa A. 2010. The Think-Aloud controversy in Second Language Research. Routledge.

Council of Europe 2001. Common European Framework of Reference for Languages: Learning, Teaching, Assessment [electronic version]. http://www.coe.int/t/dg4/linguistic/Source/Framework_en.pdf (30.6.2015).

Dörnyei, Zoltán 2005. The psychology of the Language Learner: Individual Differences in Second Language Acquisition. Mahwah, NJ: Erlbaum.

Friedline, Benjamin E. 2011. Challenges in the Second Language Acquisition of Derivational Morphology: From Theory to Practice. Doctoral dissertation. University of Pittsburgh.

Kozulin, Alex; Garb, Erica 2002. Dynamic assessment of EFL text comprehension. – School Psychology International, 23 (1), 112–127. doi: 10.1177/0143034302023001733

Leontjev, Dmitri 2014. The Effect of Automated Adaptive Corrective Feedback: L2 English questions. – APPLES: Journal of applied language studies*,* 8 (2), 43–66. http://apples.jyu.fi/article/abstract/301 (30.6.2015).

Leontjev, Dmitri (to be published). L2 English Derivational Knowledge: Which Affixes Are Learners More Likely to Recognise? – Studies in Second Language Learning and Teaching.

Leontjev, Dmitri; Huhta, Ari; Mäntylä, Katja (to be published). Word derivational knowledge and writing proficiency: How do they link? – System.

Mäntylä, Katja; Huhta, Ari 2013. Knowledge of word parts. – J. Milton, T. Fitzpatrick (Eds.). Dimensions of Vocabulary Knowledge. Basingstoke, UK: Palgrave, 45–59.

Nakayama, Natsue 2008. Effects of Vocabulary Learning Using Affix: Special Focus on Prefix. 63–78, http://www.kyoai.ac.jp/college/ronshuu/no-08/nakayama.pdf (30.6.2015).

Nassaji, Hossein 2003. L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. – TESOL Quarterly, 37 (4), 645–670. doi: 10.2307/3588216

Nation, I.S. Paul 2001. Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

O’Malley, J. Michael; Chamot, Anna U. 1990. Learning Strategies in Second Language Acquisition. New York: Cambridge University Press.

Oxford, Rebecca L. 1990. Language Learning Strategies: What Every Teacher Should Know. New York: Newbury House.

Poehner, Matthew E. 2008. Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting L2 Development. Berlin, Springer.

Poehner, Matthew E.; Lantolf, James P. 2013. Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). – Language Teaching Research, 17 (3), 323–342. doi:10.1177/1362168813482935

Põhikooli riiklik õppekava õigusakt; Lisa 2 [Basic School National Curriculum Act: Annex 2] (2010). Pub. L. No. RT I 2010, 6, 22. https://www.riigiteataja.ee/aktilisa/1281/2201/0017/13275423.pdf (2.7.2015).

Rose, Heath 2012. Reconceptualizing strategic learning in the face of self-regulation: Throwing language learning strategies out with the bathwater. – Applied Linguistics, 33 (1), 92–98. doi: 10.1093/applin/amr045

Schmitt, Norbert; Meara, Paul. 1997. Researching vocabulary through a word knowledge framework. – Studies in Second Language Acquisition, 19 (01), 17–36.

Swain, Merrill 1998. Focus on form through conscious reflection. – C. Doughty, J. Williams (Eds.). Focus on Form in Classroom Second Language Acquisition. New York: Cambridge University Press, 64–81.

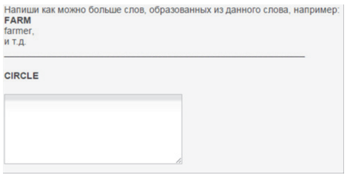
Teo, Adeline 2012. Promoting EFL students' inferential reading skills through computerized dynamic assessment. – Language Learning & Technology, 3, 10–20. http://llt.msu.edu/issues/october2012/action.pdf (20.9.2014).

Tseng, Wen-Ta; Dörnyei, Zoltán; Schmitt, Norbert 2006. A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. – Applied Linguistics, 27 (1), 78–102. doi: 10.1093/applin/ami046

Tulving, Endel; Kroll, Neal 1995. Novelty assessment in the brain and long-term memory encoding. – Psychonomic Bulletin & Review, 2 (3), 387–390. doi: 10.3758/BF03210977

Vygotsky, Lev S. 1978. Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA: Harvard University Press.

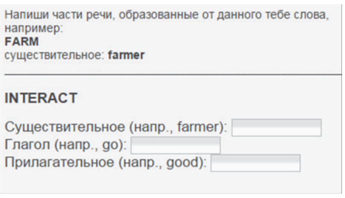
Appendix 1. Static assessment tasks



Напиши как можно больше слов, образованных из данного слова, например:
FARM
farmer,
и т.д.

CIRCLE

Figure 2.1. Sample item from the free production task

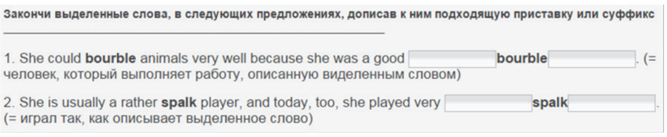


Напиши части речи, образованные от данного тебе слова, например:
FARM
существительное: **farmer**

INTERACT

Существительное (напр., farmer):
Глагол (напр., go):
Прилагательное (напр., good):

Figure 2.2. Sample item from the metalinguistic prompts task

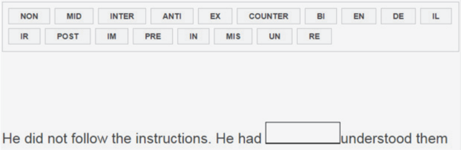


Закончи выделенные слова, в следующих предложениях, дописав к ним подходящую приставку или суффикс

1. She could **bourble** animals very well because she was a good **bourble**. (= человек, который выполняет работу, описанную виделенным словом)

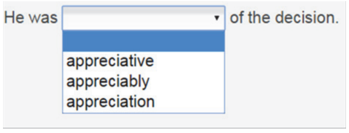
2. She is usually a rather **spalk** player, and today, too, she played very **spalk**. (= играл так, как описывает выделенное слово)

Figure 2.3. Sample item from the non-word affix elicitation task



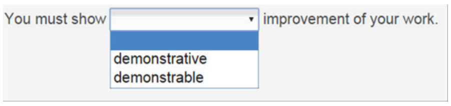
He did not follow the instructions. He had understood them

Figure 2.4. Sample item from the prefix elicitation task



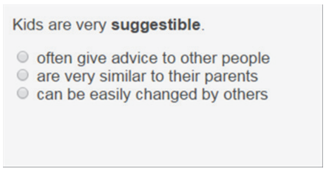
He was of the decision.

Figure 2.5. Sample item from the grammar recognition task



You must show improvement of your work.

Figure 2.6. Sample item from the meaning recognition task



Kids are very **suggestible**.

○ often give advice to other people
○ are very similar to their parents
○ can be easily changed by others

Figure 2.7. Sample item from the passive recognition of the meaning task

Appendix 2. Transcription Symbols

Symbol	**Meaning**
Text	originally in English
Text	stressed word or a part of it
↑	noticeably rising intonation
((text))	non-verbal behaviour, e.g., laughter, gestures, etc.
(.)	pause of 0.2 seconds or less
(0.0)	timed pause
:	elongation of the preceding sound
-	utterance is cut off
°text°	uttered in a noticeably quieter, softer voice
[text]	comment

Dmitri Leontjev (University of Jyvaskyla, Centre for Applied Language Studies) is a doctoral student at the University of Jyväskylä. His research interests include dynamic and diagnostic assessment of English as a second or a foreign language.
dmitri.leontjev@jyu.fi

Sõnatuletuse oskuste dünaamiline hindamine: ühe õpilase arengu jälgimine

Dmitri Leontjev
University of Jyvaskyla, Centre for Applied Language Studies

The aim of the study is to increase our understanding of how dynamic assessment can promote learners' ability to derive words. The development of one learner's word derivational knowledge was compared before and after human-mediated and computerised dynamic assessment. The study focused on the strategies and the knowledge sources that the learner used. Think aloud protocols and interviews were the primary research tools. In addition, to trace the learner's development over time, there was a gap of a year and a half introduced between the human-mediated and the computerised dynamic assessment.
The results demonstrated that the dynamic assessment promoted the learner's word derivational knowledge. The analysis revealed that the learner not only started using the strategies elicited during the assessment but generally became more successful in regulating his performance. That is, both the learner's word derivational knowledge and strategic learning were promoted.

Keywords: sociocultural theory, mediation, inferencing strategies, knowledge sources, self-regulation, L2 learning, English