The Pennsylvania State University

The Graduate School

WHEN PREDICTIONS AREN'T *PERFEKT*:

THE JOINT ROLES OF PREDICTION, CORRECTIVE FEEDBACK, AND PREDICTION ERROR IN L2 LEARNING

A Dissertation in
German Applied Linguistics and Language Science

by
Valerie Keppenne

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2023

The dissertation of Valerie Keppenne was reviewed and approved by the following:

Carrie N. Jackson
Professor of German and Linguistics
Dissertation Advisor
Chair of Committee
Acting Head of the Department of Germanic and Slavic Languages & Literatures

Michael T. Putnam
Professor of German and Linguistics

Janet Van Hell
Professor of Psychology and Linguistics

Kevin McManus
Associate Professor of Applied Linguistics

Katharina Schuhmann
Junior Professor of German as a Foreign Language
Carl von Ossietzky University of Oldenburg, Germany
Special Member

ABSTRACT

In contexts of first language (L1) acquisition, prediction and adaptation are often viewed as fundamental to the learning process (Rabagliati et al., 2016). One mechanism that has been proposed to specifically account for such adaptive behavior is error-based implicit learning (e.g., Chang et al., 2006). According to this account, speakers constantly predict upcoming language input. When these predictions are not met, it leads to adjustments in the linguistic system in order to avoid future errors, i.e., speakers learn. To what extent similar mechanisms apply in L2 learning is uncertain, as the degree to which L2 learners have been shown to predict upcoming input is modulated by grammatical structure, proficiency, and other individual variables (for a review, see Kaan, 2014; Kaan & Grüter, 2021). Generating fewer predictions means that opportunities for prediction error, and consequently for learning, may be compromised (Phillips & Ehrenhofer, 2015). Despite this, there is evidence that learning from prediction error can occur, particularly when learners are forced to make predictions (Grüter et al., 2021). However, it is unclear whether error-based learning is possible when the target structure is novel to L2 learners, especially in lower-proficiency learners, as some baseline knowledge of the target structure may be necessary (McDonough & Fulga, 2015; but see Bovolenta & Marsden, 2022; Brod, 2021).

Largely separate from frameworks of error-based learning, many studies investigating the differential effects of corrective feedback on L2 learning have shown that, like prediction, the effectiveness of corrective feedback can be modulated by factors such as the target structure and L2 proficiency (e.g., Ammar, 2008; Ammar & Spada, 2006; Li, 2014). An understanding of how learners process feedback and compute prediction error may potentially explain why certain types of feedback are more beneficial than others, pointing towards the need to study error-based learning and corrective feedback in tandem to better understand the mechanisms underlying L2 learning. In two experiments, this dissertation unites classroom-based corrective feedback research and psycholinguistic research on predictive processing to study error-based learning in order to advance our understanding of the underlying learning mechanisms in late L2 acquisition. By encouraging learners to predict and by tracking their linguistic development, this dissertation sheds light on the role of prediction for L2 learning at different proficiencies, and how learners process feedback and use prediction errors to update their developing L2 linguistic system.

Experiment 1 aimed to investigate the role of two different types of implicit feedback after prediction for the learning of German present perfect in learners of different proficiencies, since previous studies have shown both prediction and feedback effects to be modulated by target structure and learner proficiency. In German, the present perfect is formed with one of two auxiliary verbs: *haben* ‘to have’ is used with transitive verbs, such as *essen* ‘to eat’, while *sein* ‘to be’ is used with verbs that involve movement from A to B, such as *laufen* ‘to run’, or a change in state. The alternation thus has a semantic component that is difficult for L2 learners to acquire. The training task administered to participants created a context that encouraged them to engage in prediction in conjunction with receiving feedback to learn this alternation and the targeted past participles. In a pretest/posttest/delayed posttest design in a remote, web-based setup, L2 German learners (*N*=46) of different proficiencies completed two training blocks in which they predicted a past action based on the auxiliary in a sentence fragment. For instance, they read *Tina hat gestern Abend spät* ___. ‘Tina has ___ late last night.’ and chose between two pictures that either matched or mismatched the fragment (based on the auxiliary). They then completed the sentence by typing the participle of the corresponding main verb, and received semantic- or form-focused feedback (regardless of response correctness). No explicit instruction was provided. Alongside the pre-, immediate, and delayed posttests, an interim test was administered between training blocks. These outcome measures included a written picture description task and a picture matching task to target production and comprehension of the present perfect. The results indicated that participants generally increased in auxiliary and participle accuracy over the course of the training, and that feedback ordering had little effect on accuracy or the trajectory of learning. The results also showed that learners of intermediate proficiency increased most in accuracy across testing times. On the other hand, analyses of auxiliary accuracy split by learners with and without prior knowledge of the target feature show that guessing — creating weak expectations based on no prior knowledge — paired with implicit feedback did not lead to significant learning.

In Experiment 2, I expanded on the findings of Experiment 1 by investigating the developmental trajectory of whether and how learners predictively process the auxiliary online, how they process feedback, and how they compute prediction error online during learning. Beginning to intermediate L2 German learners (*N*=13) with previous exposure to the target structure completed a similar training task as in Experiment 1, including four training blocks with feedback. During the auditory presentation of sentence fragments during training, as well as during feedback presentation, participants' eye movements and pupil dilation were recorded to track the emergence of predictive looks towards target pictures as well as the evolving processing of feedback and prediction error. Training was preceded by a pretest and followed by immediate and delayed posttests, including the same outcomes measures as in Experiment 1. The results showed that participants improved on productive measures of auxiliary and participle accuracy, but not on receptive measures of auxiliary accuracy, neither during training nor on the outcome measure targeting comprehension. Furthermore, the tracking of eye movements during training did not reveal the emergence of predictive looks to target images. On the other hand, the time course of the pupil response during feedback presentation did evolve over the course of the training task, and pupil responses differed depending on the correctness of learners' predictions.

Taken together, the results of these two experiments suggest that error-based learning is possible in L2 learning, albeit with some constraints related to the target linguistic feature, general L2 proficiency, and prior knowledge of the target feature. Furthermore, the importance of online predictive processing for learning in error-based learning frameworks is put into question, as learning occurred even in the absence of predictive looks during training. Finally, the usefulness of pupillometry to investigate feedback processing and its evolution over the course of a learning task is demonstrated. As such, this dissertation significantly contributes to investigations of L2 learning mechanisms and has a few pedagogical implications as well.

TABLE OF CONTENTS

LIST OF FIGURES ········ ix
LIST OF TABLES ········ xi
ACKNOWLEDGEMENTS ········ xiii

Chapter 1. INTRODUCTION ········ 1

1.1 Description of the Problem ········ 1
1.2 Layout of the Dissertation ········ 2

Chapter 2. BACKGROUND ········ 3

2. Error-Based Learning Mechanisms ········ 3
2.1 Error-Based Learning in Language ········ 3
2.1.1 Structural Priming ········ 5
2.1.2 Errorful Generation ········ 6
2.1.3 Feedback ········ 7
2.1.3.1 Surprisal ········ 7
2.1.3.2 Curiosity ········ 8
2.2 Error-Based Learning in an L2 ········ 11
2.3 Corrective Feedback in L2 Learning ········ 16
2.3.1 Theoretical Underpinnings for the Effectiveness of Corrective Feedback ········ 17
2.3.2 Types of Corrective Feedback ········ 19
2.3.3 Effects of Corrective Feedback in L2 Learning ········ 22
2.3.3.1 The Role of Proficiency ········ 24
2.3.3.2 The Role of Target Structure ········ 24
2.3.4 Intersections between Error-Based Learning and Corrective Feedback ········ 25
2.4 Pupillometry ········ 28
2.5 The Present Study ········ 29
2.5.1 Motivations and Goals ········ 29
2.5.2 The Target Structure ········ 30

Chapter 3. EXPERIMENT 1 ········ 32

3.1 Introduction and Research Questions ········ 32
3.2 Overview of Experimental Approach ········ 32
3.2.1 General Design and Time Course of Study ········ 32
3.2.2 Participants ········ 34
3.2.2.1 Screening and Proficiency Measures ········ 35
3.2.2.2 Materials and Data Processing ········ 35
3.2.2.3 Results ········ 36
3.2.3 Training ········ 36
3.2.3.1 Materials ········ 36
3.2.3.2 Training Procedure ········ 37
3.2.3.2.1 Training Task ········ 37

3.2.3.2.2 Feedback Conditions 39
3.2.4 Outcome Measures and Materials 41
3.2.4.1 Picture Description Task 41
3.2.4.2 Picture Matching Task 42
3.2.4.3 Debriefing 42
3.2.5 Norming Study 43
3.3 Analysis and Results 43
3.3.1 Data Scoring 44
3.3.1.1 Picture Matching Task 44
3.3.1.2 Picture Description Task 44
3.3.1.3 Training 44
3.3.2 Learning and the Role of Feedback Ordering 45
3.3.2.1 Analyses 45
3.3.2.2 Results 45
3.3.2.2.1 Picture Matching Task 45
3.3.2.2.2 Picture Description Task (Auxiliaries) 50
3.3.2.2.3 Picture Description Task (Participles) 55
3.3.2.3 Post-hoc Training Analysis 59
3.3.2.3.1 Analysis 59
3.3.2.3.2 Results 59
3.3.2.3.2.1 Auxiliaries 59
3.3.2.3.2.2 Participles 62
3.3.2.4 Summary: Learning and the Role of Feedback Ordering 65
3.3.3 Learning and the Role of Proficiency 65
3.3.3.1 Analyses 65
3.3.3.2 Results 66
3.3.3.2.1 Picture Matching Task 66
3.3.3.2.2 Picture Description Task (Auxiliaries) 70
3.3.3.2.3 Picture Description Task (Participles) 73
3.3.3.3 Summary: Learning and the Role of Proficiency 76
3.3.4 Exploratory Analysis: The Role of Prior Knowledge 76
3.3.4.1 Analyses 77
3.3.4.2 Results 77
3.3.4.3 Summary: The Role of Prior Knowledge 79
3.3.5 Exploratory Analysis: The Role of Awareness 79
3.3.5.1 Analyses 80
3.3.5.2 Results 81
3.3.5.2.1 Picture Matching Task 81
3.3.5.2.2 Picture Description Task 83
3.3.5.3 Summary: The Role of Prior Awareness 84
3.4 Discussion 85
3.4.1 The Role of Feedback Ordering in Error-Based L2 Learning 85
3.4.2 The Role of Proficiency in Error-Based L2 Learning 88
3.4.3 The Role of Prior Knowledge and Awareness 90
3.4.4 Implications for Error-Based L2 Learning 92
3.4.5 Limitations and Future Studies 94

Chapter 4. EXPERIMENT 2 ······ 95

4.1 Introduction and Research Questions ······ 95
4.2 Overview of Experimental Approach ······ 95
4.2.1 General Design and Time Course of Study ······ 95
4.2.2 Participants ······ 96
4.2.2.1 Screening and Proficiency Measures ······ 97
4.2.2.2 Materials and Data Processing ······ 97
4.2.2.3 Results ······ 97
4.2.3 Training ······ 97
4.2.3.1 Materials ······ 98
4.2.3.2 Training Procedure ······ 99
4.2.3.2.1 Training Blocks and Trial Setup ······ 99
4.2.3.2.2 Audio Files ······ 100
4.2.3.2.3 Procedure ······ 101
4.2.4 Outcome Measures and Materials ······ 101
4.2.4.1 Picture Matching and Picture Description Tasks ······ 101
4.2.4.2 Debriefing ······ 101
4.2.4.3 Measures of Learning Aptitude ······ 102
4.2.4.3.1 Serial Reaction Time (SRT) Task ······ 102
4.2.4.3.2 LLAMA F ······ 102
4.3 Analysis and Results ······ 103
4.3.1 Data Scoring ······ 103
4.3.2 Improvement on Outcome Measures ······ 103
4.3.2.1 Analyses ······ 103
4.3.2.2 Results ······ 104
4.3.2.2.1 Picture Matching Task ······ 104
4.3.2.2.2 Picture Description Task (Auxiliaries) ······ 106
4.3.2.2.3 Picture Description Task (Participles) ······ 108
4.3.2.3 Summary: Improvement on Outcome Measures ······ 112
4.3.3 Improvement during Training: Behavioral Data ······ 112
4.3.3.1 Analyses ······ 112
4.3.3.2 Results ······ 112
4.3.3.2.1 Auxiliaries ······ 112
4.3.3.2.2 Participles ······ 114
4.3.3.3 Summary: Improvement during Training: Behavioral Data ······ 118
4.3.4 Eye Movements during Picture Choice ······ 118
4.3.4.1 Analyses ······ 118
4.3.4.2 Results ······ 119
4.3.4.3 Summary: Eye Movements during Picture Choice ······ 121
4.3.5 Pupil Size during Feedback Presentation ······ 121
4.3.5.1 Analyses ······ 121
4.3.5.2 Results ······ 122
4.3.5.3 Summary: Pupil Size during Feedback Presentation ······ 123
4.3.6 Relationship between Pupil Size Changes an Outcome Measures ······ 124

4.3.6.1 Relationship between Learning Aptitude and Outcome Measures ··· 124
4.4 Discussion ··· 125
4.4.1 Improvement during Training and on Outcome Measures ··· 125
4.4.2 Predictive Eye Movements in an Error-Based Learning Task ··· 126
4.4.3 Pupil Dilations during Feedback Processing ··· 128
4.4.4 The Relationship between Feedback Processing and Learning ··· 129
4.4.5 Implications for Error-Based L2 Learning ··· 130
4.4.6 Limitations and Future Studies ··· 131

Chapter 5. CONCLUSIONS ··· 133

5.1 Overview of Major Findings ··· 133
5.2 Implications for Error-Based L2 Learning and Corrective Feedback ··· 134
5.3 Pedagogical Implications ··· 137
5.4 Open Questions and Directions for Future Research ··· 138
5.5 Final Remarks and Conclusion ··· 139

REFERENCES ··· 140

Appendix A Language Background Questionnaire Experiment 1 and 2 ··· 158
Appendix B Proficiency Test Experiment 1 and 2 ··· 174
Appendix C Materials for Experiment 1 and 2 ··· 179
Appendix D Experiment 1 Session 2 Task Instructions ··· 182
Appendix E Debriefing Questionnaires Experiment 1 and 2 ··· 201
Appendix F Untrained Generalization Items ··· 205
Appendix G Descriptive Statistics for Pupil Size in Experiment 2 ··· 221
Appendix H LLAMA F and SRT ··· 223

LIST OF FIGURES

Figure 2.1	Types of corrective feedback on the implicit-explicit continuum (Lyster et al., 2013)	22
Figure 3.2.1	Experimental design with feedback condition counterbalanced within groups and across blocks of training	33
Figure 3.2.2	Timeline of testing and training in Experiment 1	34
Figure 3.2.3	Context encouraging prediction provided to participants at the beginning of the training session	38
Figure 3.2.4	Sample trial for the picture choice task	39
Figure 3.2.5	Sample of the a) semantics-focused feedback presentation and b) form-focused feedback presentation	40
Figure 3.3.1	Accuracy for the Picture Matching Task by testing time	46
Figure 3.3.2	Accuracy for the Picture Matching Task by group and testing time	47
Figure 3.3.3	Graphic representation of the marginal interaction effect between testing time and group	48
Figure 3.3.4	Accuracy of *sein* auxiliaries for the Picture Description Task by testing time	51
Figure 3.3.5	Accuracy of *sein* auxiliaries for the Picture Description Task by group and testing time	52
Figure 3.3.6	Graphic representation of the significant interaction effect between testing time and group	53
Figure 3.3.7	Accuracy of participles on the Picture Description Task by testing time	55
Figure 3.3.8	Accuracy of participles on the Picture Description Task by group and testing time	56
Figure 3.3.9	Graphic representation of the significant interaction effects between testing time and group	57
Figure 3.3.10	Accuracy of auxiliaries during training by group and trial	60
Figure 3.3.11	Accuracy of participles during training by group and trial	62
Figure 3.3.12	Accuracy for the Picture Matching Task by testing time	66
Figure 3.3.13	Visualization of interaction terms between proficiency and testing time on the Picture Matching Task	68
Figure 3.3.14	Accuracy of *sein* auxiliaries for the Picture Description Task by testing time	70
Figure 3.3.15	Visualization of interaction terms between proficiency and testing time for auxiliaries on the Picture Description Task	71
Figure 3.3.16	Accuracy of participles for the Picture Description Task by testing time	73
Figure 3.3.17	Visualization of interaction terms between proficiency and testing time for participles on the Picture Description Task	74
Figure 3.3.18	Auxiliary accuracy for the Picture Matching Task by course level and testing time	78
Figure 3.3.19	Auxiliary accuracy (*sein* only) for the Picture Description Task by course level and testing time	79
Figure 3.3.20	Relationship between course level and awareness status	81
Figure 3.3.21	Auxiliary accuracy on the Picture Matching Task by testing time and awareness level	83

Figure 3.3.22 Auxiliary accuracy (*sein* only) on the Picture Description Task by testing time and awareness level ··· 84
Figure 4.2.1 Timeline of testing and training in Experiment 2 ··· 96
Figure 4.2.2 Sample trial in Experiment 2 ··· 100
Figure 4.3.1 Accuracy for the Picture Matching Task by testing time ··· 105
Figure 4.3.2 Graphic representation of individual participant's accuracy across testing times ··· 106
Figure 4.3.3 Accuracy of *sein* auxiliaries for the Picture Description Task by testing time ··· 107
Figure 4.3.4 Graphic representation of individual participant's accuracy across testing times ··· 108
Figure 4.3.5 Accuracy of participles on the Picture Description Task by testing time· 109
Figure 4.3.6 Graphic representation of individual participant's accuracy across testing times ··· 110
Figure 4.3.7 Accuracy of auxiliaries during training, by training block ··· 113
Figure 4.3.8 Accuracy of auxiliaries during training by trial number and training block ··· 114
Figure 4.3.9 Accuracy of participles during training, by training block ··· 115
Figure 4.3.10 Accuracy of participles during training by trial number and training block ··· 116
Figure 4.3.11 Time course of looks to the target and foil pictures as a function of training block and auxiliary (from onset of auxiliary, i.e., 0 ms, to 3000 ms post-onset) ··· 120
Figure 4.3.12 Percent change in pupil size from baseline by training block, auxiliary, and error type ··· 123
Figure F.1 Accuracy for untrained generalization items on the Picture Matching Task by testing time ··· 206
Figure F.2 Accuracy for untrained generalization items on the Picture Matching Task by group and testing time ··· 207
Figure F.3 Auxiliary accuracy for untrained *sein* generalization items on the Picture Description Task by testing time ··· 210
Figure F.4 Auxiliary accuracy for untrained *sein* generalization items on the picture description task by group and testing time ··· 211
Figure F.5 Accuracy for untrained generalization items on the Picture Matching Task by testing time ··· 214
Figure F.6 Visualization of interaction terms between proficiency and testing time for untrained generalization items on the Picture Matching Task ··· 215
Figure F.7 Auxiliary accuracy for untrained *sein* generalization items on the Picture Description Task by testing time ··· 217
Figure F.8 Visualization of interaction terms between proficiency and testing time for untrained generalization items on the Picture Description Task ··· 218

LIST OF TABLES

Table 2.1.	Types of corrective feedback	21
Table 3.2.1	Participant demographics	36
Table 3.3.1	Means for auxiliary accuracy on the Picture Matching Task with bootstrapped 95% CIs	46
Table 3.3.2	Model output for auxiliary accuracy in the Picture Matching Task as a function of group	49
Table 3.3.3	Means for *sein* auxiliary on the Picture Description Task with bootstrapped 95% CIs	50
Table 3.3.4	Model output for auxiliary accuracy (*sein* only) on the Picture Description Task as a function of group	54
Table 3.3.5	Means for participles on the Picture Description Task with bootstrapped 95% CIs	55
Table 3.3.6	Model output for participle accuracy on the Picture Description Task as a function of group	58
Table 3.3.7	Means for auxiliary accuracy during training with bootstrapped 95% CIs	59
Table 3.3.8	Model output for auxiliary accuracy during training as a function of group (by trial)	61
Table 3.3.9	Means for participle accuracy during training with bootstrapped 95% CIs	62
Table 3.3.10	Model output for participle accuracy during training as a function of group (by trial)	64
Table 3.3.11	Model output for auxiliary accuracy on the Picture Matching Task as a function of proficiency	69
Table 3.3.12	Model output for auxiliary accuracy on the Picture Description Task as a function of proficiency	72
Table 3.3.13	Model output for participle accuracy on the Picture Description Task as a function of proficiency	75
Table 3.3.14	Counts of participants below, at, or above chance on the Picture Matching Task by awareness level and testing time	82
Table 3.3.15	Counts of participants below, at, or above chance on the Picture Description Task by awareness level	83
Table 4.2.1	Participant demographics	97
Table 4.3.1	Descriptive statistics for auxiliary accuracy on the Picture Matching Task	104
Table 4.3.2	Descriptive statistics for auxiliary accuracy (*sein* only) on the Picture Description Task	106
Table 4.3.3	Descriptive statistics for participle accuracy on the Picture Description Task	108
Table 4.3.4	Results from the Friedman and Wilcoxon signed-rank tests for the Picture Matching and Picture Description Tasks	111
Table 4.3.5	Descriptive statistics for auxiliary accuracy during training	113
Table 4.3.6	Descriptive statistics for participle accuracy during training	114

Table 4.3.7	Results from the Friedman and Wilcoxon signed-rank tests for the picture choice and participle writing during training	117
Table 4.3.8	Proportion of error type per training block	122
Table F.1	Means for auxiliary accuracy for untrained generalization items on the Picture Matching Task with bootstrapped 95% CIs	205
Table F.2	Model output for auxiliary accuracy for untrained generalization items on the Picture Matching Task as a function of group	208
Table F.3	Means for auxiliary accuracy for untrained generalization items on the Picture Description Task with bootstrapped 95% CIs	209
Table F.4	Model output for auxiliary accuracy for untrained generalization items on the Picture Description Task as a function of group	212
Table F.5	Model output for auxiliary accuracy for untrained generalization items on the Picture Matching Task as a function of proficiency	216
Table F.6	Model output for auxiliary accuracy for untrained generalization items on the Picture Description Task as a function of proficiency	218
Table G.1	Mean (baseline corrected) pupil size (pixels) by training block, auxiliary, and error type	221
Table G.2	Mean peak (baseline corrected) pupil amplitude (pixels) by training block, auxiliary, and error type	222
Table G.3	Mean peak latency (in ms) by training block, auxiliary, and error type	222
Table H.1	Results of the SRT and LLAMA F	223

ACKNOWLEDGEMENTS

This project would not have been possible without the help of so many people. First, my gratitude goes to my advisor, Carrie, for your personal and professional support throughout the years. Thank you for showing me how to become a better researcher and for your constant feedback throughout my projects, both personal and professional. Without your ability to focus my attention on what is most important in the moment, I probably would have spent way more time worrying about things I have no control over. I am also incredibly thankful for your support in figuring out my path after graduate school comes to an end. Thank you for making me a better researcher and teaching me to do research not just for the sake of it but to answer questions that actually matter. Of course, I would also like to thank the rest of my committee members, who guided me throughout my time in graduate school. Mike, thank you for the many hallway conversations that always made me feel better about myself. Katharina, thank you for your thorough feedback and critical questions, which pushed me to think deeper and further about many different topics. Janet, you were always able to point out additional things to think about, both in designing experiments as well as interpreting findings, and I want to thank you for that. And thank you, Kevin, for jumping in last minute to become part of the committee. Your feedback on all my early projects was very helpful in guiding me. Thank you also to Theres Grüter for listening to my ideas and providing your incredible expertise when this dissertation was still in its very early stages. Your contributions were invaluable and made this project much stronger.

I also need to thank not only the lab members for their feedback on all my projects and for the many Zoom game nights, but also all the research assistants that have worked on these projects with me throughout my years in the lab. You were incredible and working with you was a pleasure. Special thanks also go out to Maike Rocker for lending your voice to my experiments (more than once). And thank you to all the German instructors at Penn State and Texas Tech that let me recruit participants in their German classes.

Without funding, there would be no dissertation. I would therefore also like to thank the organizations that funded this research. In particular, the research and writing of the dissertation were made possible by the Adele Miccio Award from the Center for Language Science at Penn State, a *Language Learning* Dissertation Grant, and finally the Max Kade Graduate Fellowship and funding from the Max Kade Foundation. Thank you also to the Department of Germanic and Slavic Languages and Literatures and the Center for Language Science at Penn State for generously funding professional development opportunities and conference participation.

Last, but certainly not least, I would like to thank the friends and family that supported me along the way. I am so very fortunate to have had colleagues that turned into such wonderful friends (you know who you are). Thank you for providing advice, encouragement, and laughs whenever they were needed the most. Of course, thank you also to my family, Maman, Papa, et Alexis. Without your support throughout my life, I wouldn't be who I am and where I am today. I am also deeply indebted to my grandparents: Bonne-Maman, for inspiring me to become a scientist in spite of all the hurdles that might be put in my way; Papi, who instilled in me a love for languages and was always happy to discuss anything related to language with me; and Mamie, for teaching me about taking care of family in the best way possible. And finally, to my husband, Cam. There were many a night that you listened to me talk about L2 acquisition and eye-tracking: I am infinitely grateful for your patience, your ability to ground me, and for reminding me that there is more to life than writing a dissertation. Thank you for always being there when I needed you. Without you, this project would have been impossible.

CHAPTER 1: INTRODUCTION

1.1 Description of the Problem

In contexts of a first language (L1), prediction, prediction error, and adaptation are often viewed as playing a key role in child language acquisition (Peter et al., 2015; Rabagliati et al., 2016; Zhang et al., 2018; see also Phillips & Ehrenhofer, 2015). Prediction thereby refers to the generation of an expectation about a future event; prediction error to the failure of this expected event to occur; and adaptation to the subsequent adjustment of the prediction model in order to avoid future prediction errors. A mechanism that can account for such adaptive behavior in language is error-based implicit learning (Dell & Chang, 2014). According to this account, speakers constantly predict upcoming language input. In fact, prediction has been proposed to be a hallmark of human cognitive activity, and there is ample evidence that the human brain is a "prediction machine" (Clark, 2013). When linguistic predictions fail, this leads to adjustments in the linguistic system to avoid future errors — speakers adapt and learn. To what extent error-based learning applies in second language (L2) learning, however, is uncertain, as the degree to which L2 learners predict upcoming input seems to be modulated by target structure, proficiency, and other individual variables (for a review, see Kaan, 2014; Kaan & Grüter, 2021). This means that opportunities for prediction error, and consequently for learning, may be compromised (Phillips & Ehrenhofer, 2015). Understanding whether and how this mechanism applies in L2 learning can improve language teaching, and is therefore particularly relevant as we increasingly use educational technology to learn and practice an L2. There is evidence that learning from prediction error can occur (Jackson & Hopp, 2020), particularly when L2 learners are forced to predict (Grüter et al., 2021). However, it is unclear if error-based learning is possible when the target structure is novel to L2 learners, especially in lower-proficiency learners, as some baseline knowledge of the target structure may be necessary (McDonough & Fulga, 2015; McDonough & Trofimovich, 2015; but see Bovolenta & Marsden, 2021; Brod, 2021; Potts et al., 2019; Potts & Shanks, 2014).

Largely separate from frameworks of error-based learning, many studies investigating the differential effects of corrective feedback on L2 learning have shown that, like prediction, the effectiveness of corrective feedback can be modulated by factors such as the target structure and L2 proficiency (e.g., Ammar, 2008; Ammar & Spada, 2006; Li, 2014). While research on corrective feedback mostly agrees that the positive impact of corrective feedback stems from learners noticing a mismatch between their output and the target form, understanding how learners process feedback and compute prediction error may explain why certain types of feedback are more beneficial than others, and shed light on the mechanisms that underly L2 learning. Currently, it is unclear how learners process feedback in the moment they receive it, as investigations of corrective feedback have mostly relied on offline measures that take place after the fact, such as modified output or stimulated recall interviews, to assess whether learners notice their error and learn from it (e.g., Egi, 2010). However, research has started to address this question using eye-tracking and the visual world paradigm (e.g., Andringa, 2020; Curcic et al., 2019). In addition, pupillometry — the measurement of changes in pupil dilation in response to a stimulus — can complement these measures, as it provides a means to investigate the computation of prediction error (Zhang et al., 2018), as well as the allocation of attention while encoding during learning (Miller & Unsworth, 2019).

In this dissertation, I address these open questions with the goal of creating a better understanding of the learning mechanisms that are at play in contexts of L2 learning as well as with the hopes of informing pedagogical language teaching practices. In two experiments, I investigate the learning of German *Perfekt*, primarily the auxiliary alternation (*haben* 'to have' vs. *sein* 'to be') and secondarily the past participle, by encouraging learners to engage in prediction followed by corrective feedback. Experiment 1 asks 1) if learners acquire the target structure, 2) if the type of corrective feedback and the order of its presentation plays a role, and 3) if proficiency interacts with learning of the target structure. Participants completed a training unit with two blocks of training. They received one of two kinds of corrective feedback in the form of the correct response, counterbalanced by training block: semantic-focused feedback targeting verb choice based on the auxiliary in the sentence or form-focused feedback targeting both verb choice and the past participle form of the main verb. Experiment 2 attempts to repeat the findings from Experiment 1 regarding the learning outcomes and in addition asks 1) if feedback processing in L2 learners can be detected via pupil dilation, 2) if pupillary responses correlate with the amount of learning, and 3) if learners develop predictive processing as measured via eye movements. Participants completed a similar training as in Experiment 1, but only received form-focused feedback. In addition, their eye movements and pupil dilation were tracked during training to explore the emergence of predictive processing and allocation of attention, alongside the computation of prediction error during feedback processing. In conjunction, the results of these two experiments can inform psycholinguistic theories of L2 learning as well as bolster and/or transform existing pedagogical approaches to L2 learning.

1.2 Layout of the Dissertation

This dissertation is divided into five chapters. Chapter 2 provides a literature review of prior research that is most relevant to the current dissertation. The review first introduces error-based learning mechanisms and the role of prediction and feedback, in particular as they pertain to L1 and L2 language learning. The review then introduces relevant research on corrective feedback in L2 learning, including the role of proficiency and the target structure for its effectiveness, before drawing connections between research on error-based learning and corrective feedback. Chapter 2 then introduces the motivations and goals of the two experiments conducted as part of the dissertation. Finally, Chapter 2 introduces the L2 structure used in the dissertation to investigate error-based learning in an L2 context, namely the German *Perfekt* and its auxiliary alternation. Then, Chapter 3 introduces Experiment 1 — investigating the role of feedback ordering, proficiency, as well as prior knowledge and awareness — including the methodology, results, discussion of findings, and limitations as well as avenues for future research. Chapter 4 introduces Experiment 2 — investigating the emergence of predictive behavior and the online signature of feedback processing within an error-based earning task — similarly including the methodology, results, discussion of findings, as well as limitations and future research. Finally, Chapter 5 summarizes the results of both experiments and discusses their theoretical and pedagogical contributions. Chapter 5 ends with a discussion of potential limitations and highlights directions for future research.

CHAPTER 2: BACKGROUND

2. Error-based Learning Mechanisms

Imagine you are talking to your best friend about the food you had for breakfast this morning, and they say "I had a breakfast classic, a sandwich with peanut butter and …". Before they even finish their sentence, you already know that the next word is going to be "… jelly." We most certainly have all experienced a situation like this, in which we or someone else effortlessly predict the end of a sentence. In fact, prediction has been proposed to be a hallmark of human cognitive activity, and there is ample evidence that the human brain is a "prediction machine" (Clark, 2013): Our brain constantly predicts features of the environment we interact with, facilitating the processing of incoming information. But now imagine a situation in which you are having lunch with the same friend. You both have a similar palate and decide to get two sandwiches to share. After eating for a while, your friend says: "You know what's funny, we always finish each other's …". Of course, you expect the next word to be *sentence*. But then you hear: "… sandwiches." You may be surprised at this, because you were not expecting the sentence to end this way. The word you predicted to come up next did not match what your friend actually said – this mismatch resulted in a prediction error. The next time you are at lunch, you may therefore predict a different ending to a similar sentence.

This example illustrates that prediction not only facilitates the processing of incoming stimuli (as in the first case), but it also allows humans to efficiently evaluate their internal knowledge system and to adapt it in response to stimuli that do not match their expectations (as in the second case). This type of adaptive behavior – or learning – in response to prediction error is another vital part of the brain's function (Clark, 2013). One of the goals of such behavior is to minimize the discrepancy between predicted outcome and the outcome experienced, as experiencing prediction error can be cognitively costly (Husband & Boloventa, 2020; Zirnstein et al., 2018; but see Frisson et al., 2017) and disrupt the processing of incoming information. Optimizing the internal model of our environments is thus crucial in order to avoid prediction error, and the human brain therefore constantly tests its predictions against the experienced outcomes to be able to produce more accurate predictions in the future. Such adaptation to prediction errors can take different forms depending on the magnitude of a prediction error, i.e., whether you are surprised that your prediction was incorrect or whether you did not expect your prediction to be accurate in the first place. An error's surprisal value is modified by the frequency with which a similar error occurs (Holroyd et al., 2009) as well as the confidence with which a prediction was made (e.g., Brod, 2021; Potts & Shanks, 2014; Potts et al., 2019). Related to levels of confidence is also the degree of prior knowledge that contributes to the creation of an expectation (see Brod, 2021; Coumel et al., 2022). In turn, the magnitude of prediction error experienced has an effect on the extent to which the internal system subsequently adapts: larger prediction errors have been found to lead to stronger adaptation (Bultena et al., 2020; Butterfield & Metcalfe, 2001; 2006; Greve et al., 2017). Several explanations in various disciplines have been proposed to account for such behavior (e.g., Chang et al., 2006; Dell & Chang, 2014; Hebb, 1949; McClelland, 2006; McClelland & Rumelhart, 1981; Rescorla & Wagner, 1972). This dissertation will refer to such behavior as prediction-based or error-based learning.

2.1 Error-Based Learning in Language

Prediction is ubiquitous in language, and it can facilitate linguistic processing (Pickering & Garrod, 2007). Decades of research have found that language users not only rapidly integrate words in the input as they encounter them, but that they also predict specific features of the upcoming input. In this case, prediction is defined as the pre-activation of linguistic information prior to encountering the language input that carries that information (Pickering & Gambi, 2018). Many studies using online methodologies like electroencephalography (EEG) or eye-tracking in conjunction with the visual-world paradigm (VWP) have shown that L1 language users indeed appear to pre-activate information at various linguistic levels, such as semantics and morphosyntax (Altmann & Kamide, 1999; Hintz et al., 2017; Hopp, 2015a; Kamide et al., 2003; Knoeferle et al., 2005; Otten et al., 2007; Otten & Van Berkum, 2008; Wicha et al., 2004).[1] While language users thus frequently engage in prediction, and such prediction may facilitate processing, it is not a necessary prerequisite for language processing (Huettig, 2015). Indeed, the extent to which language users predict seems malleable based on the reliability of predictive cues (Brothers et al., 2017; Hopp, 2016) and is modulated by individual variables within a language user as well (e.g., Curcic et al., 2019; Favier et al., 2021; Huettig & Janse, 2016; Kaan, 2014; Kaan & Grüter, 2021).

Nevertheless, besides facilitating online language processing, prediction, and prediction error in particular, have been viewed as fundamental processes to language adaptation among adult L1 speakers and, increasingly, as playing a key role in child language acquisition (Peter et al., 2015; Rabagliati et al., 2016; Zhang et al., 2018; see also Phillips & Ehrenhofer, 2015). One mechanism that specifically accounts for adaptive behavior in L1 language processing is error-based implicit learning (Chang et al., 2006; Dell & Chang, 2014; Kaschak, 2007). For instance, the P-chain model by Dell and Chang (2014) – an extension of Chang et al.'s (2006) connectionist model of the development and use of abstract syntax – argues that language users are constantly engaged in predicting upcoming language input based on their prior language experience. Speakers then experience a prediction error when this prediction is not met, i.e., when there is a mismatch between the prediction and the actual linguistic input. Experiencing a mismatch between a prediction and the actual linguistic input is what guides the internal learning mechanisms to adjust linguistic representations and their weights to avoid future prediction errors. This process underlies the implicit adaptation to and learning of specific language structures. If, for instance, one hears a double object construction ("Tom gave Mary the book.") in a context for which one predicted a prepositional object construction ("Tom gave the book to Mary."), this then results in a prediction error. Consequently, the language user adjusts the weight previously assigned to double object constructions and prepositional object constructions, leading to an adjustment of the linguistic system: Double object constructions are now slightly more expected than they were before, which may result in smaller prediction errors the next time a similar prediction is formulated and a double object construction encountered. Similar predictions as in Dell and Chang's (2014) account are made by Jaeger and Snider's (2013) belief-updating model, where learners adjust their computation of a structure's probability if the input they encounter does not match their current statistical information stored (see also Reitter et al., 2011 for an alternative account). In summary, then, predicting language not only facilitates language processing but also contributes to the adaptive behavior of the linguistic system.[2]

[1] There is debate whether language users indeed predict at all linguistic levels, including phonology and lexical form (see Nieuwland et al., 2017). However, this discussion is beyond the scope of this dissertation.

[2] See also the plethora of research that has investigated rapid adaptation to accented speech, showing that listeners can quickly adapt to accented speech and learn to process it efficiently (e.g., Baese-Berk et al., 2013; Bradlow & Bent, 2008; Brown et al., 2020; Maye et al., 2008).

2.1.1 Structural Priming

One experimental research paradigm that has both contributed to the formulation of accounts of error-based learning as well as investigated some of the specific mechanisms that are involved is that of structural priming (Bock, 1986; Ferreira & Bock, 2006). Structural priming describes the heightened expectation of encountering a formerly unpredicted construction again in a similar context, as well as the increased probability of producing a formerly dispreferred construction after encountering it in the incoming input. Structural priming studies generally present participants with prime trials followed by target trials. In prime trials, participants see a picture and read or hear the sentence containing the structure of interest, or produce a sentence based on a sentence fragment that only allows for the structure of interest. On target trials in production priming studies, participants are then presented with a picture – and oftentimes a set of words to be used – and are asked to describe the picture with one sentence. Priming is observed when the participant re-uses the structure presented in the prime trial in the target trial. Indeed, structural priming has been observed across a variety of structures and languages (e.g., Chen et al., 2013; Kootstra & Doedens, 2016; Loebell & Bock, 2003) and both in and across comprehension and production (Bock et al., 2007; Branigan et al., 2005; Litcofsky & Van Hell, 2019; Segaert et al, 2012; for reviews, see Mahowald et al., 2016; Pickering and Ferreira, 2008).

Generally, experimental research on structural priming also distinguishes between immediate or short-term priming, cumulative priming, and longer-term effects. Immediate priming refers to a situation in which target trials immediately follow the prime trial without any intervening material. Cumulative priming occurs when the magnitude of priming, i.e., the probability to use the same structure as in the prime sentence, increases over the course of the priming task, either due to language users' own utterances or due to being exposed to more prime sentences containing the target structure (e.g., Hartsuiker & Westenberg, 2000; Kaan & Chun, 2018; Kaschak et al., 2011). Finally, long-term priming refers to instances when priming effects are maintained over a longer period of time. This can occur in different ways, one of which is when language users encounter one or more filler sentences that do not contain the target structure between the prime and the target trial (so-called long-lag priming; e.g., Bernolet et al., 2016). Long-term priming can also be measured by comparing the proportion of utterances containing the target structure in a baseline phase preceding the priming task with the proportion of similar utterances in a posttest phase immediately following the priming task (e.g., Hartsuiker & Westenberg, 2000). Other studies measure long-term effects by comparing baseline and posttest target utterances at longer intervals, such as a day or even weeks following the priming task (e.g., Branigan & Messenger, 2016; Kaschak et al., 2014; Shin & Christianson, 2012).

Another effect which is found in many priming studies is the so-called inverse frequency effect (Bernolet & Hartsuiker, 2010), which describes the finding that structures of lower frequency generally result in larger priming effects. Learning accounts that ascribe a critical role to surprisal predict this effect: Due to structures of lower frequency being less expected than structures of higher frequency, these result in larger prediction errors when encountered and, consequently, lead to greater priming and adaptation of the linguistic system, compared to structures with higher frequencies. Following a similar line of thinking, structures that are novel to speakers – and thus completely unexpected when encountered – should lead to very large prediction errors, and consequently, to heightened degrees of adaptation.

The fact that structural priming effects can appear immediately after prime sentences or remain stable for a longer period of time led Bernolet et al. (2016) to propose a multifactorial account of structural priming. Based on conflicting findings for immediate versus cumulative and long-term priming, they suggest that effects of immediate priming are conditioned by the explicit memory of target structures in prime sentences. On the other hand, Bernolet et al. suggest that cumulative priming effects, and similarly long-term priming effects, are due to implicit learning mechanisms that lead to the adaptation of linguistic representations in language users by generalizing over multiple different instances of priming. In three experiments, Bernolet et al. showed that priming effects were strongest when prime trials immediately preceded target trials (corresponding to what is referred to as immediate priming above), whereas priming effects diminished with an increasing number of intervening fillers (two or six) between prime and target trial. In addition, participants were more accurate in reproducing a prime sentence with the target syntactic structure when fewer filler sentences occurred between the prime and the attempt to reconstruct the prime verbatim, showing that explicit memory of the prime decreases with increasing intervening fillers. Furthermore, there were effects of cumulative priming, regardless of the number of intervening fillers. Together, these findings led Bernolet and colleagues to conclude that different processes, while all part of one larger mechanism, contribute to immediate compared to cumulative and long-term priming effects: Immediate priming effects are boosted by explicit memory of prime sentences, whereas implicit learning governs cumulative and long-term priming effects, with no effect of explicit memory of prime sentences (see also Hartsuiker & Bernolet, 2017, for a parallel account of grammar learning in L2 learners).

The fact that priming can persist over longer periods of time and shows inverse frequency effects underscores the claim that it may be part of a mechanism that contributes to systematic adaptation and therefore drives language learning (Branigan & Messenger, 2016; Chang et al., 2006; Dell & Chang, 2014). Similarly, it underscores the usefulness of priming paradigms as a tool to research error-based learning in language.

2.1.2 Errorful Generation

In addition to investigating the gradual adaptation of the linguistic system via priming paradigms, many lab-based experiments have shown that the generation of errors, i.e., incorrect predictions, leads to better memory for the correct response. This has been shown both for the learning of general knowledge (Brod & Breitwieser, 2019; Brod et al., 2018; Kang et al., 2009; Kornell et al., 2009, 2015) as well as in language, in particular for the learning of word pairs or novel lexical items (Kornell et al., 2009; Potts et al., 2019; Potts & Shanks, 2014). For instance, Kornell et al. (2009) investigated the learning of fictitious trivia knowledge in addition to weakly semantically related word pairs in a paired associate learning task. They compared the effectiveness of a condition in which participants simply saw the question-answer pair or word-pair, and one in which participants were asked to predict the answer or target word first before being given the correct answer as feedback. The latter condition nearly always resulted in the generation of an error, i.e., an incorrect prediction. Examples of a fictitious trivia question and of paired associates are given in (1).

(1) What is the last name of the person who proposed the theory of mal-adaptibility?
– Alfred
olive – branch; mouse – hole; train – caboose

The results of a cued recall posttest following the study phase showed that participants remembered the trivia answers and corresponding word pairs significantly more often when they had been asked to predict an answer and generated an error during the learning phase. These results were obtained at a posttest both 5 minutes and 38 hours following the study phase. One explanation for this effect offered by Kornell and colleagues is that predicting and generating a possible answer during study, even if the correct answer is not retrieved, involves the activation of semantic concepts related to the question or cue word. In turn, this may increase the semantic connection between these related words and the trivia answer or target word, allowing these related words to subsequently serve as mediators to the trivia answer or target word during retrieval at test (see also the "elaborative generation hypothesis" in Carpenter, 2009). Similar results were obtained by Kornell et al. (2015), who used a paired associate learning task and showed that it is not critical to correctly predict an answer as long as a prediction is attempted at all. Prior knowledge and elaborate semantic networks may thus contribute to the beneficial effect of incorrect predictions in learning fictitious knowledge and paired associates.

However, not all learning involves the activation of semantic networks, and the activation of related semantic concepts thus cannot be the sole explanation for the learning benefit of generating errors. For instance, errors have been shown to be beneficial in motor learning (e.g., Diedrichsen et al., 2010; van Vugt & Tillmann, 2015). Similarly, it is unlikely that semantic networks are involved in the learning of some syntactic or morphosyntactic structures in language, such as finite verb placement or grammatical gender marking in German, as these aspects of language knowledge are located outside of semantic networks. Alternative accounts posit that the generation of predictions, especially if errorful, benefits subsequent memory by enhancing the encoding of feedback provided to speakers (e.g., Brod et al., 2018; Brod & Breitwieser, 2019; Potts et al., 2019). Indeed, for benefits related to the generation of predictions to appear, positive evidence, i.e., feedback with the correct answer, needs to be presented shortly after the generation of a prediction (Pashler et al., 2005; Vaughn & Rawson, 2012). It is also crucial that learners attend to the feedback provided and understand it. Studies that do not ensure that feedback is processed often show no benefit of feedback at all (see Metcalfe, 2017).

2.1.3 Feedback

2.1.3.1 Surprisal

One factor that might contribute to the strength with which feedback is encoded is its surprisal value, i.e., the degree to which the feedback received is surprising to the receiver. Butterfield and Metcalfe (2001), for instance, had participants answer general knowledge questions as well as rate their confidence in their answer on a 7-point scale, ranging from *sure wrong* to *sure correct*. Participants then received feedback on their answers. While the authors had hypothesized that erroneous answers made with high confidence to be the most difficult for participants to correct at a recall posttest, they actually found the reverse: Highly confident errors were more likely to be changed to a correct response at posttest compared to erroneous answers made with low confidence. This applied to both answers of which participants thought they were *sure correct* but turned out to be incorrect, and for which participants thought they were *sure wrong* but turned out to be correct. The authors introduced the term 'hypercorrection effect' to describe this effect. Butterfield and Metcalfe concluded that feedback to an error that was committed with a high degree of confidence might be more surprising, and therefore arousing, recruiting more cognitive capacities to encode and store the contents of the feedback. In a follow-up experiment, Butterfield and Metcalfe (2006) showed that surprising or unexpected feedback indeed received more attention from participants than unsurprising or expected feedback. In this follow-up study, after answering a general knowledge question and rating their confidence in their answers, participants received feedback and, on some trials, heard an auditory tone played as well. The participants' task was to detect and indicate whether a tone had been played during the presentation of the feedback, with the idea being that participants' attentional resources would have to be split between feedback processing and tone detection. More surprising feedback should receive more attention, in turn decreasing the amount of attention that can be allocated to tone detection. Butterfield and Metcalfe (2006) found that participants performed worse in the tone detection task if they had indicated high confidence in an answer that then turned out incorrect, compared to instances in which they only indicated low confidence in their response. This indeed indicates that participants paid less attention to the tones during feedback presentation if the feedback was surprising. In addition, performance on the tone detection task was negatively related to final test performance: The more accurate participants performed on the tone detection task, the less accurate their performance on the recall post-test was, indicating that less attention had been devoted to the feedback presented (see also Fazio & Marsh, 2009).

Relatedly, Brod et al. (2018) showed that the generation of a prediction and the degree of surprise, as measured via pupil dilations in response to feedback, is related to memory benefits. In their study, participants generated a prediction either before being presented with facts about the number of inhabitants in European countries and the results of German soccer games, or after being presented with these facts. In this postdiction condition, participants would first see the outcome and were then asked to indicate which outcome they would have predicted. The reults showed that prediction led to better memory for both types of facts than postdiction. In addition, expectancy-violating feedback was accompanied by larger pupil dilations in the prediction condition compared to the postdiction condition. Finally, pupil size in the prediction condition was positively correlated with later memory, such that larger pupil dilations in response to expectancy-violating feedback were more likely to lead to learning. No such relationship was found in the postdiction condition. These findings suggest that generating a prediction before being presented with feedback allows a learner to be surprised about the mismatch between their expectation and the outcome, leading to better feedback encoding. In addition, it demonstrates that the degree of surprise is measurable using pupil dilations, a methodology that will also be exploited in the second experiment in this dissertation.

Overall, these findings (Brod et al., 2018; Butterfield & Metcalf, 2001, 2006) and others (e.g., Brod et al., 2022; Frömer et al., 2021; Theobald et al., 2022) support the idea that unexpected, and therefore surprising feedback, can lead to a larger expenditure of effort to encode feedback, leading to better memory for the feedback content as well as of other features related to the presentation of feedback.

2.1.3.2 Curiosity

Again, surprise cannot be the sole factor leading to enhanced encoding of feedback, however, as there are cases where participants are highly unlikely to be confident in their responses and therefore cannot experience a mismatch between the high confidence in their response and the correct answer (see, for instance, Seabrooke et al., 2019). The learning of novel knowledge constitutes one such case. The act of generating an expectation here is more akin to guessing, as prior knowledge is not available to guide a prediction. In fact, studies couched in the framework of generative learning, in which learners are actively engaged in the construction of new knowledge, have frequently found benefits of generating answers in the absence of a previously established relation to prior knowledge – i.e., guessing – across a variety of educational situations (e.g., Brod et al., 2018; Inagaki & Hatano, 1977; Marsh et al., 2012; Pressley et al., 1992). In such frameworks, guessing is seen as a learning strategy in its own right, alongside prediction – i.e., generating an answer based on an established relation of this answer with prior knowledge (Brod, 2021).

Related to language learning, Potts and Shanks (2014) found that guessing was indeed beneficial for the learning of rare English words as well as of translations of foreign vocabulary words. Participants either generated a guess followed by feedback, chose from two answer possibilities followed by feedback, or simply read the word pair before completing a multiple-choice post-test. Results from four separate experiments converged to suggest that errorful generation followed by feedback benefited participants' memory of the newly learned English and foreign vocabulary words more so than choosing from correct answer possibilities or simply studying word pairs. Importantly, these benefits appeared even in the absence of semantic co-activation of related concepts that might act as mediators for later retrieval. In their general discussion, Potts and Shanks hypothesized that generating a genuine guess may increase participants' curiosity to find out the correct answer, consequently leading to enhanced encoding of feedback presented (see also Kang et al., 2009; Exp. 3-6 in Kornell et al., 2009).

Indeed, in a study investigating the learning of numerically-based trivia facts – presumably unknown to participants – Brod and Breitwieser (2019) compared the effects of generating a guess versus generating an example on the levels of curiosity and subsequent retention. Participants saw a statement like "X of all animals are insects" and were asked either to guess the value of X or to come up with an example of an insect. They then rated their curiosity to find out the correct answer, followed by feedback presenting the correct value for X. The results of the curiosity ratings and a multiple-choice posttest showed that not only did guessing lead to higher levels of curiosity, but higher levels of curiosity also led to better memory for the correct answer. However, when curiosity ratings were disregarded, there was no direct effect of learning condition on posttest performance, i.e., guessing did not lead to more learning than giving an example. In terms of pupillary response, high curiosity ratings as compared to low curiosity ratings were related to larger pupil dilations during feedback anticipation. In addition, pupil dilation was enhanced when participants generated a numerical guess rather than an example, both during feedback anticipation as well as during feedback presentation. Brod and Breitwieser take these findings to indicate that high levels of curiosity generated via guessing increased participants' awareness of their knowledge gap. In turn, this heightened participants' anticipation of the feedback, leading to subsequent better encoding of the feedback presented, as demonstrated by the pupil data. These findings seem to indicate that there might be a role for curiosity above and beyond that of guessing by increasing learners' readiness to process upcoming feedback.

In a follow-up study to Potts and Shanks (2014), Potts and colleagues (2019) demonstrated similar findings regarding the role of curiosity in word learning as well. Not only did Potts and colleagues replicate the findings from Potts and Shanks (2014) regarding the benefits of guessing, but they also found a relationship with curiosity. Curiosity ratings were higher when participants generated a guess before rating their curiosity compared to the absence of generation. They hypothesized that these increased levels of curiosity are due to the knowledge gap becoming apparent to participants. Furthermore, curiosity ratings were higher after generating than before generating a guess, which indicates that the act of committing to a guess may make participants more curious to find out the correct answer, preparing them for enhanced feedback encoding (see also Brod & Breitwieser, 2019). However, curiosity ratings were not related to better performance on a final test, a finding that did not completely mirror Brod and Breitwieser (2019). It is thus an open question whether the act of attempted retrieval or guessing is therefore necessary for enhanced feedback encoding to occur, or whether high levels of curiosity are sufficient (see

Murayama, 2022, for a curiosity-driven learning account motivated by the intrinsic reward value of gaining knowledge).

Nevertheless, assigning benefits of generating erroneous predictions, whether based on prior knowledge or not, to an enhanced encoding of feedback is particularly appealing in the case of learning novel knowledge, as well as knowledge that is not semantic in nature. If prior knowledge and experience with the language were necessary prerequisites for error-based learning to occur, then arguably, children should show no or only attenuated learning from prediction error, as they also have less real-world experience and prior language knowledge. However, research has demonstrated that error-based learning mechanisms indeed support learning among children (e.g., Emberson et al., 2015; Fazekas et al., 2020; Gambi & Messenger, 2021; Havron et al., 2019; Romberg & Saffran, 2013; Zhang et al., 2018; but see Gambi et al., 2021). Emberson et al. (2015), for instance, used a cross-modal stimulus omission paradigm to investigate whether infants as young as 6 months were subject to prediction error. In an exposure phase, infants were first exposed to audio input that was predictive of a subsequent visual stimulus. Immediately after, infants were exposed to normal audio-visual trials, as well as trials that included the audio stimulus but omitted the visual stimulus. The results showed that these infants had similar cortical responses as adults to stimuli violating, as well as conforming to, previously established expectations. Similar findings were obtained by Zhang et al. (2018) using pupillometry. However, keep in mind that infants in both studies first went through an exposure phase, creating relevant associations before being subsequently tested on their response to violations of these associations.

Related to language, Gambi and Messenger (2021) investigated whether prediction error could help 4-year-old children with the acquisition of direct object datives. While the use of prepositional object datives is already established at that age, direct object datives are not, and thus constitute an unmastered, although potentially not entirely novel, syntactic structure. Following a pretest to establish the baseline use of direct object datives, children completed a training phase. Children saw a touchscreen with three objects, heard a sentence, and were asked to act out the sentence on the touchscreen pointing and moving the objects around. In a between-subjects design, participants either heard only prepositional object or direct object datives. Importantly, the animacy status of the three objects in the sentence was manipulated between-subjects as well, such that the display either featured three animate objects, or two animate and one inanimate object. In a display featuring two animate and one inanimate object, children hearing “Winnie the Pooh will give …” could reasonably expect the inanimate object to be mentioned next in the sentence as the theme, followed by the animate object as the recipient, adhering to a prepositional object dative structure. However, the direct object structure requires the animate object to appear before the inanimate object (“Winnie the Pooh will give the duck the frisbee” vs. “Winnie the Pooh will give the frisbee to the duck.”). Children in the direct object group presented with an inanimate and two animate objects were thus exposed to sentences that repeatedly led to prediction error by disconfirming children’s predictions based on their knowledge of the prepositional object dative — i.e., they might have predicted “Winnie the Pooh will give …” to be continued by “… the frisbee to the duck”, but only ever heard continuations with “… the duck the frisbee”. The results of an immediate posttest showed priming effects for children exposed to direct object sentences with inanimate themes, as well as larger increases in the comprehension of direct object datives compared to baseline compared to children who were exposed to direct object sentences with animate objects only. While these results were only found on test trials with two animate objects – not on trials with an inanimate and an animate object – they provide preliminary evidence that prediction error can indeed drive the learning of novel syntactic structures in children.

The fact that guessing, in the absence of prior knowledge, has been shown to be beneficial for learning, has consequences for frameworks of error-based learning. Such findings imply that effects of prediction and prediction error may extend to situations in which the relevant knowledge system has not yet been established and learners are therefore truly testing hypotheses about what the target prediction may be. Benefits of generating an expectation even in the absence of related prior knowledge have currently only been demonstrated in studies that investigated the learning of items related to episodic memory, i.e., fictitious/non-fictitious facts or vocabulary words (Brod & Breitwieser, 2019; Brod et al., 2018, 2022; Butterfield & Metcalfe, 2001, 2006; Kang et al., 2009; Kornell et al., 2009; Potts et al., 2019; Potts & Shanks, 2014; Theobald et al., 2022), or the learning of syntactic structures (Gambi & Messenger, 2021). The question remains whether the generation of expectations, regardless of whether prior knowledge exists or not, can also facilitate learning of other types of information. Accordingly, this dissertation focuses on the role of feedback in error-based learning for the learning of a morphosyntactic alternation in L2 learners of German with varying proficiency, some of whom are yet unfamiliar with the target structure.

2.2 Error-Based Learning in an L2

Error-based learning can account for such phenomena as structural priming and adaptation in the L1, but its role in L2 learning is still far from clear (see Kaan, 2015; Kaan & Grüter, 2021). L2 learners, particularly at low proficiency levels, engage in online predictive processing at a much lower level, if at all, compared to native speakers (Kaan, 2014; Kaan & Grüter, 2021). In particular, engagement in predictive processing among L2 learners is modulated by the target structure in question, especially when predictive cues presented to learners are morphosyntactic rather than semantic in nature (Dussias et al., 2013; Lew-Williams & Fernald, 2010; Grüter et al., 2017; Hopp, 2013, 2015, 2016; for a review, see Kaan, 2014; Kaan & Grüter, 2021). In addition to the type of target structure in question, L1-L2 structural similarity is a further factor that determines whether learners experience difficulties in using cues predictively (e.g., Alemán Bañón & Martin, 2021; Dussias et al., 2013; Hopp & Lemmerth, 2018; Van Bergen & Flecken, 2017; see also Kaan, 2014; Kaan & Grüter, 2021). Predictive processing in L2 learners may furthermore be modulated by variables such as learner proficiency (Dussias et al., 2013; Hopp, 2015a). Hopp (2015a), for instance, showed that neither low-intermediate nor high-intermediate or advanced English L2 learners of German were able to use case-marking in object-verb-subject sentences predictively, although advanced learners were able to recover from initial misanalyses faster than learners with lower proficiency. Effects of target structure or L1-L2 similarity may therefore be further exacerbated in lower proficiency learners as well (Kaan, 2014; Kaan & Grüter, 2021). Other individual difference variables, such as lexical variability, retrieval speed, or working memory (e.g., Curcic et al., 2019; Hopp, 2018; Ito et al., 2018) have also been shown to have an effect on anticipatory processing.

Although research has shown less engagement in predictive processing in L2 comprehension and open questions therefore remain regarding the extent to which error-based learning applies in L2 learning, studies have begun to address whether L2 learners show adaptation effects in the L2 (e.g., Chun, 2020; Hopp, 2020; Jackson & Hopp, 2020; Jackson & Ruf, 2017, 2018; Kaan & Chun, 2018; Kaan et al., 2019). In a self-paced reading study that investigated whether L2 learners adapt their processing routines as a result of frequent prediction error, Kaan et al. (2019) failed to find adaptation effects. Using massed exposure, they investigated whether Spanish L2 learners of English with a wide range of proficiency levels adapt their reading of constructions which have a frequent versus a less frequent continuation. The less frequent structures – and therefore the structures of interest – were embedded *what*-questions and coordinating constructions with *and* (see (2)). The more frequent counterparts to each structure of interest were sentences with embedded *whether*-questions as well as coordinating constructions with *but*.

(2)	a. Low frequency:	The builder wondered what the worker repaired the leak with before going home.
	High frequency:	The builder wondered whether the worker repaired the leak with some tape before going home.
	b. Low frequency:	The servant cleaned the table and the floor was cleaned by the maid.
	High frequency:	The servant cleaned the table but the floor was cleaned by the maid.

Participants were presented with 72 experimental sentences, 18 of each type presented in (2). Kaan and colleagues hypothesized that participants would show increased reading times for those regions in the less frequent structures that would alert learners to their prediction error in comparison to the *whether*- or *but*-conditions. They also anticipated learners would show adaptation effects, i.e., reading time differences between sentences with less frequent and more frequent structures in (2) should become smaller over the course of the experiment. However, Kaan and colleagues did not find any adaptation effects – reading time differences did not become smaller as exposure to the target sentences increased. L2 learners were thus unable to adapt their processing of the target sentences, possibly because the target parse was too difficult for L2 learners to identify. Indeed, if the target parse is not accessible to learners, this limits their ability to compare their initial parse with the target parse, preventing the computation of prediction error and, in turn, the updating of the linguistic system. In fact, even the L1 English speakers in the study only adapted their reading of one of the structures, the more easily retrievable and recoverable coordinating construction with *and*. This finding underscores the fact that mere exposure to an infrequent target structure may not be sufficient for adaptation to occur. Instead, learners may need scaffolding in order to achieve the appropriate parsing of the infrequent target structure. This way, they can become able to recognize discrepancies between their current linguistic system and the target, as they may be unable to access the target parse on their own.

Indeed, there is evidence that learning from prediction error can occur when access to the target parse is provided. In two experiments, Hopp (2020) investigated whether inducing prediction error would lead to the revision of processing routines for two structures that have been shown to be difficult for German L2 learners of English to process correctly. The first experiment targeted sentences with intransitively used transitive verbs, as in (3a). Such sentences have been shown to be difficult for L2 learners, as the noun phrase *he* following the verb is temporarily ambiguous between an object reading for the verb *play* and a subject reading for the main clause, showing that learners might not take case marking into account when processing online (Hopp, 2015b).

(3) a. While the boy was <u>playing</u> he pleased his parents with the music.
b. The boy played and the music pleased the parents.
c. The boy played the music and it pleased the parents.

After a pretest using eye-tracking, the intermediate to advanced German L2 learners of English in Hopp (2020) received massed exposure to sentences containing the target structure to induce prediction error and to provide positive evidence of the correct alternative sentence parse. For instance, learners in the experimental group received sentences as in (3b), where verbs such as *played* occurred with a direct object and were thus never used transitively. Learners in the control group received sentences that were not designed to induce prediction error (3c). At posttest, overall fixation times showed that learners in the experimental group adapted to the targeted processing patterns through the systematic and massed exposure to input inducing prediction error, whereas learners in the control group did not.

In the second experiment, Hopp targeted the processing of tense-marked verbs as in (4a) and (4b) using a self-paced reading paradigm. In English, temporal adverbials that mark events with reference and event time preceding speech time require the use of a past tense verb, whereas temporal adverbials that mark events with current relevance – event time and speech time are identical – require the use of present perfect. Roberts and Liszka (2013) found that intermediate to advanced German L2 learners of English were insensitive to mismatches between temporal adverbials and grammatical tense marking, due to German not encoding grammatical aspect and the use of present perfect being acceptable for both past and completed events or events that still have current relevance (Rothstein, 2008).

(4) a. Last week, James went swimming every day.
b. Since the summer, James has gone swimming every day.
c. Last year, Bruno liked the new classmate, but he has since then made other friends than him. Now they do not even talk to each other.
d. Last year, Bruno liked the new classmate, but he did back then not make a new friend in him. Now they do not even talk to each other.

Similar to the first experiment, German L2 learners of English completed a pretest and then received massed exposure to sentences containing the target structures to induce prediction error and to provide positive evidence of the correct alternative sentence parse. Here, the experimental group received sentences such as in (4c), which contrast the use of temporal adverbials and their appropriate accompanying grammatical tense marking. Learners in the control group read sentences that were not designed to induce prediction error (4d). The results were very similar to those of the first experiment, in that reading times showed that learners in the experimental group adapted their processing patterns after experiencing frequent prediction error through massed exposure to the target parse. In both experiments, however, it is important to point out that the mere experience of prediction error after predicting a certain sentence continuation (4a or 4b) in the pretest or posttest was insufficient to adapt; participants did not adapt just from being exposed to sentences in which they experienced prediction error, as there was no effect of trial number on reading behavior, mirroring effects found in Kaan et al. (2019). Instead, adaptation required exposure to the alternative sentence parse as offered to learners in the experimental groups during an exposure phase (4b and 4c).

The extent to which learners actually create and commit to predictions during language processing, however, is uncertain. For instance, Kaan and Chun (2018) tested whether Korean L2 learners of English exhibit priming effects with the production of direct object and prepositional object structures in English. Participants described pictures by completing sentence fragments provided to them, including all characters shown. In prime trials, sentence fragments forced either the use of a direct object ("The nurse gave the patient ...") or a prepositional object structure ("The nurse gave the medicine …"). Sentence fragments in target trials, which followed prime trials after one to three fillers, only provided the subject and verb ("The clown showed …"). Kaan and Chun measured both immediate priming and cumulative adaptation. Immediate priming referred to priming effects in response to a structure that was just encountered. Cumulative adaptation, on the other hand, described priming effects as a result of the sum of previous encounters to a structure. While the results showed cumulative priming effects for both structures, learners did not exhibit any effects of immediate priming on either structure. Indeed, one explanation that Kaan and Chun offer is that it is unclear to what extent participants created predictions for material succeeding the verb in each sentence, such that they may not have experienced prediction error and, in turn, been primed for immediate production following the prime. A second explanation is that the sentence completion task may not have forced deep processing of the target structures and therefore was not successful at drawing learners' attention to the targeted structure (see also Pickering & Branigan, 2008). Alternatively, filler sentences intervening between prime and target trials may have weakened any effects.

Nevertheless, there is evidence that adaptation can succeed when learners are forced to make predictions. Grüter et al. (2021) compared a group of Korean L2 English learners who participated in a guessing game to a control group of the same population on their written production of double object ("The girl feeds the squirrel a peanut.") and prepositional object sentences ("The girl feeds a peanut to the squirrel."). In prime trials in the guessing game, participants were encouraged to guess how Jessica, a fictitious native speaker of English, would describe a picture. Participants were given the specific verb and two noun phrases to use in their description, and in generating this description, participants in essence generated a prediction about Jessica's upcoming utterance. After typing their answer, participants then saw the prime sentence, i.e., how Jessica in fact described the picture, which was always a double object construction in experimental trials. Participants then had to describe a new picture without receiving feedback from Jessica. In contrast, participants in the control group saw Jessica's description of the sentence and had to re-type the sentence, before they described a picture on their own. The analyses compared the production of double object and prepositional object sentences between a baseline phase and the priming phase, as well as the baseline and an immediate posttest. While both groups increased in their production of double object sentences in both the priming phase and at immediate posttest compared to baseline, effects were larger for participants who completed the guessing game, indicating that forcing predictions indeed led to greater longer-term adaptation, or learning.

While forcing predictions can thus benefit L2 learning – at least when participants are already partially familiar with the target structure – it is not entirely clear what processes contribute to the development of prediction and predictive processing in an L2. One requirement seems to be that learners develop awareness of a cue's predictive use. In a separate line of research, Curcic et al. (2019) investigated the relationship between predictive looks to a target object in a visual world paradigm and three different degrees of awareness: whether participants were aware of the use of grammatical gender in the artificial language based on Fijian (gender awareness), whether they were aware of the gender agreement pattern (pattern awareness), and whether they were aware that grammatical gender helped them look at targets faster, i.e., that they could use it predictively (prediction awareness). The authors found a relation between type of awareness and predictive looks: only those learners who indicated prediction awareness also showed predictive looks to the target object. In contrast, gender awareness or even pattern awareness, were not sufficient to develop predictive processing. Similarly, Andringa (2020) reported that predictive looks to target objects emerged in the block immediately after the one in which participants indicated they had become aware of the target structure. However, it is unclear whether this type of awareness matches the one reported to have an effect on predictive processing in Curcic et al. (2019), as Andringa (2020) did not distinguish between different awareness types in their analysis. Together, these two studies highlight that awareness of a cue and its predictive value plays a role in the development of prediction and predictive processing in L2 learners.

Indeed, there is evidence that learners need to be able to detect – or become aware of – a novel target structure in the input in order for priming and adaptation to occur. For novel morphological features of a target language, McDonough and Trofimovich (2015) found that learners were mostly unable to detect and produce object-first sentences in Esperanto, which are formed with an unfamiliar syntactic structure in addition to morphological case marking, following a structural priming task. The authors suggest that structural priming in the absence of other types of input or feedback may not be a useful learning strategy for novel features of an L2. Rather, structural priming may enhance existing knowledge in L2 learners. McDonough and Fulga (2015) expanded on this finding and showed that only those learners who successfully detected the structure in a learning activity prior to the priming task were able to be primed, indicating that rudimentary knowledge may be necessary. This is also in line with Curcic et al.'s (2019) and Andringa's (2020) findings: If awareness contributes to the development of predictive processing — and priming involves predictive processes (Dell & Chang, 2014; Chang et al., 2006) — then awareness necessarily also plays a role in priming and adaptation in L2 learning. The studies reviewed above (Grüter et al., 2021; Hopp, 2020; Kaan & Chun, 2018; Kaan et al., 2019), however, all have in common that learners were already familiar with the target structure. It therefore remains unclear whether error-based learning is possible when the target structure is novel to L2 learners (Jackson & Ruf, 2017; McDonough & Trofimovich, 2015), especially in lower-proficiency learners, as some baseline knowledge of the target structure may be necessary for structural priming, and therefore learning, to occur (Hartsuiker & Bernolet, 2017; McDonough & Fulga, 2015). Furthermore, the processes by which learners detect target structures in the input, and thereby become able to detect discrepancies between their predictions or expectations and the target structures, in particular for novel and unfamiliar structures, remain unclear.

In a recent study, Bovolenta and Marsden (2021) used an artificial language learning paradigm to investigate whether the element of surprise, induced via prediction error, could lead to the acquisition of a novel structure, the passive in an artificial language they call Yorwegian (see 5).

(5) Active: *Lu meeb flugat lu prag.* "The girl calls the boy."
Passive: *Lu prag fluges ka lu meeb.* "The boy is called by the girl."

Participants were exposed to active sentences in Yorwegian on Day 1, while they completed simple forced-choice picture decisions. On Day 2, participants were exposed to six learning blocks with both active and passive sentences and also received feedback on their forced-choice picture decisions in 75% of trials; this feedback was congruent in blocks 1-2, meaning that after hearing a passive or active sentence and choosing a picture, participants would see the correct picture while hearing the same structure they had just been exposed to. Importantly, the experimental group received incongruent feedback in 25% of feedback trials in blocks 3-6: After being exposed to an active sentence and choosing a picture, feedback showing the correct picture was paired with a passive sentence. While learning of vocabulary words was equal across the control and experimental groups, the learning of the target structure was not: Participants in the experimental group were more accurate in a forced-choice test using untrained verbs than the control group on both active and passive sentences, while the main difference between groups on a grammaticality judgment task emerged on active rather than passive sentences. The results of this study thus indicate that surprise induced via unexpected feedback may indeed contribute to learning L2 target structures, although the exact mechanisms remain unclear since participants in the experimental group also showed better learning of the non-targeted active sentences. While the authors argue that the passive in Yorwegian constitutes a novel structure to participants, this may be too strong of a claim: it seems that the meaning of the target structure and that it can occur in an alternation with active sentences was already familiar to participants, as English also has an alternation between passive and active sentences that differs in form. However, the exact realization of the alternation in Yorwegian may have been novel to participants, in particular the differential verbal inflection used in the alternation. It thus remains unclear to what extent error-based learning applies in the context of learning a previously unfamiliar target structure, either with an unfamiliar meaning or an unfamiliar alternation.

While previous studies grounded in error-based learning frameworks have thus shown greater learning in errorful conditions for rare L1 words and L2 vocabulary words (Kornell, 2009; Potts et al., 2019; Potts & Shanks, 2014) or familiar syntactic or morphosyntactic structures (Bovolenta & Marsden, 2021; Hopp, 2020; Jackson & Hopp, 2020), there have not been any investigations of error-based learning for unfamiliar morphosyntactic alternations of a language (but see McDonough & Fulga, 2015 and McDonough & Trofimovich, 2015). Investigating whether error-based learning also applies in such cases can help clarify the role of error-based learning in L2 contexts in general.

In summary, error-based learning may apply in certain contexts of L2 learning, in particular when target structures are already familiar to learners. However, the role of proficiency and familiarity with the target structure is still unclear. This dissertation therefore contributes to closing these gaps in the literature and answering these open questions by investigating the role of proficiency and target structure familiarity for the learning of a morphosyntactic alternation in German L2 learners across the proficiency spectrum.

2.3 Corrective Feedback in L2 Learning

Largely separate from research on error-based learning mechanisms, investigations of corrective feedback and its effectiveness for L2 learning have played a major role in the field of Second Language Acquisition (SLA). Corrective feedback is of both theoretical as well as pedagogical interest to the field of SLA. On one hand, findings on the effectiveness of corrective feedback can contribute to our understanding of the cognitive mechanisms that underlie L2 acquisition. On the other hand, understanding which types of corrective feedback are most beneficial for L2 learners is also of importance to language instructors, as the goal in language pedagogy is to facilitate language learning. Since feedback also has the potential to trigger prediction error, as described in Section 2.1.3, an understanding of how corrective feedback facilitates learning will provide insights into error-based mechanisms and their applicability in L2 learning more generally.

Most L2 learners have experienced corrective feedback and will know that it is ubiquitous in the L2 classroom, as well as other contexts of L2 acquisition. The term ‘corrective feedback’ thereby describes the feedback that learners receive in response to linguistic errors, either in their written or oral production. In an early review of corrective feedback and its application in L2 classrooms, Lightbown and Spada (1999) define corrective feedback as “any indication to the learners that their use of the target language is incorrect” (p. 171). For example,

> when a language learner says, ‘He go to school everyday’, corrective feedback can be explicit, for example, ‘no, you should say goes, not go’ or implicit ‘yes he goes to school every day’, and may or may not include metalinguistic information, for example, ‘Don’t forget to make the verb agree with the subject’. (p. 171-172)

Corrective feedback thus describes instructors’ reactive moves to errors, which serve not only to alert learners of the error, but also to push them to attend to the accuracy of their productions in subsequent turns (Carroll & Swain, 1993; Sheen 2011).

2.3.1 Theoretical Underpinnings for the Effectiveness of Corrective Feedback

As it is of theoretical importance to SLA, it is critical to understand why corrective feedback may be effective. Perhaps unsurprisingly, most researchers agree that corrective feedback is beneficial for L2 acquisition, and several theories and hypotheses grounded in the field of SLA can account for its effectiveness. A commonality between all these theories and hypotheses is that learners need to recognize feedback as such. This means that for learning to occur, learners need to notice that there are discrepancies between their own output and the target form provided to them via corrective feedback (Gass, 1988; Robinson, 1995; Schmidt, 1990, 2001; Tomlin & Villa, 1994). Gass (1988) proposed a framework of five processing stages through which learners need to proceed in order to turn input into output. These include 1) apperception, 2) comprehended input, 3) intake, 4) integration, and 5) output. For the purpose of understanding why corrective feedback is an effective way to enhance learning, the first stage – apperception – is most relevant. Apperception describes attention afforded to and selection of input for further processing. Essentially, the process of apperception is “the process of understanding by which newly observed qualities of an object are related to past experiences … the selection of what we might call noticed (apperceived) material” (p. 201). It thereby involves the conscious relation of some form in the input to “some prior knowledge which has been stored in our experience” (p. 201). Gass thus argues that the extent of prior knowledge determines whether apperception will occur and whether certain aspects of the input will be selected for further processing. By immediately following on the heels of an erroneous form produced by a learner, corrective feedback is then a way to support apperception by increasing the probability of learners relating the target form in the input to the information stored in their linguistic system. Successfully apperceived input can then proceed through the remaining processing stages and trigger linguistic development.

In his widely cited Noticing Hypothesis, Schmidt (1990, 2001) similarly argues for a close relationship between that which learners “notice” and that which they might later produce themselves. Importantly, according to Schmidt, L2 learning is a conscious process, meaning that it cannot occur without attention to L2 forms in the input, in tandem with the subjective experience of “noticing”, i.e., becoming aware of these specific forms in the L2 input and being able to refer to them. Only once these conditions are met can input be translated to intake, with intake referring to language forms that are available for further processing to the learner, such as commitment to long-term memory and the concomitant development of the linguistic system. If a learner thus notices a mismatch between — or a gap in — the form of their current linguistic system and some incoming input — either from an interlocutor or while reading — this will then trigger an adjustment of the internal linguistic system, leading to L2 learning (e.g., Altmann & Mirković, 2009; Godfroid et al., 2013; Huettig, 2015). Indeed, L2 learners have been shown to learn better when they consciously register mismatches between what they say and what a more proficient interlocutor says (Philp, 2003).

The necessity of awareness for L2 learning, however, has been contested by other researchers, including Tomlin and Villa (1994). While Tomlin and Villa essentially agree with Schmidt (1990, 2001) that attention is necessary for L2 learning, they argue that awareness is not a necessary prerequisite for noticing and, in turn, learning to occur. Instead, they propose a tripartite model of interrelated networks, which serve the purpose of detecting stimuli in the environment (i.e., to cognitively register them), as well as to alert (i.e., prepare to attend) and orient (i.e., deploy attentional resources) the learner towards these stimuli. According to Tomlin and Villa, however, none of those functions require awareness, i.e., the subjective experience of noticing, as described in Schmidt (1990, 2001). To reconcile these differences, Robinson (1995) proposed a model that closely parallels Schmidt's Noticing Hypothesis but accommodates some of the more fine-grained distinctions made in Tomlin and Villa's tripartite model. Here, noticing describes the "detection plus rehearsal in short-term memory, prior to encoding in long-term memory" (Robinson, 1995, p. 296). Only detected and attended to input, which is a subset of the overall input, enters short-term memory. It is in that environment that noticing then may take place. Thus, in Robinson's model, detection refers to a stage that precedes noticing. While researchers have argued that feedback encourages focal attention to form, the debate of whether conscious awareness is required for L2 acquisition underscores that it is still unclear what processes exactly take place during feedback processing (but see Andringa, 2020; Curcic et al., 2019).

Several theories and hypotheses about the effectiveness of corrective feedback in L2 acquisition assign a fundamental role to attention and the "noticing" or "detection" of stimuli, regardless of the role afforded to awareness in the process of acquisition. On one hand, situations can arise in which learners ineffectively try to communicate their message to an interlocutor, as they may be using incorrect language forms. Such types of situations offer learners an opportunity to test and reformulate their hypotheses about the correct use of specific language forms. One type of support for the effectiveness of corrective feedback therefore comes from an interactionist perspective. Here, the assumption is that interaction creates opportunities for the noticing and detection of relevant language forms through negotiation. 'Negotiation' hereby refers to the act of modifying one's output to repair breakdowns in communication (Gass, 1997, 2003; Long, 1996). Breakdowns in the conversation can be signaled through interactional feedback, i.e., feedback that is implicitly or explicitly generated by the conversation partner about the comprehensibility of the learners' utterance. These breakdowns necessarily provoke the adjustment of linguistic form in tandem with conversational structure or message content, until understanding between interlocutors is achieved (Long, 1996). Under the interaction hypothesis (Long, 1980; 1996) then, conversational interaction, especially when it includes negotiation for meaning through interactional feedback, is seen as facilitatory for L2 acquisition, as it draws learners' attention to linguistic structures that are problematic in their current language output. According to the interaction hypothesis, the optimal environment for acquiring an L2 thus consists of meaning-based activities, during which learner attention is briefly directed towards more formal properties of the language via interactional feedback, such that learners may notice the discrepancies between output and target form. Such environments include activities and tasks that require learners to interact in order to successfully complete a task, such as information-gap tasks. As interactional feedback is provided in a moment when learners process both meaning and form while they are trying to convey a message to complete the task at hand, interactional feedback and subsequent negotiation may then provide opportunities to develop form-meaning mappings, i.e., to create a connection between a form to the referential meaning it encodes, like the English plural marking -*s* (VanPatten et al., 2004). The main tenet of the interaction approach (Gass & Mackey, 2014), which derives from the interaction hypothesis, is then that interaction facilitates L2 learning because it creates opportunities for negotiation for meaning through interactional feedback, which ideally draws learners' attention towards problems in the language output they produced and in turn encourages the production of modified output (Gass & Mackey, 2014). Modified output hereby describes when learners change their original, non-targetlike output in response to receiving feedback in an attempt to match the target form. Indeed, studies investigating the effects of interaction find learning benefits for interactions when negotiation is part of the interaction (Mackey, 1999; but see Muranoi, 2000), but not when interaction occurs without negotiation (Ellis et al., 1994; Pica, 1992). The benefits of interaction, and specifically of interactional feedback, are reminiscent of situations in which language learners are provided with corrective feedback: By providing corrective feedback, the feedback provider attempts to alert the learner of a discrepancy between their erroneous output and the target form, drawing their attention to this mismatch and, in some cases, providing additional information as to what the correct form should be.

Of importance here is also the role of output produced by learners, in particular of pushed and modified output in response to feedback (Swain, 1993, 2005). In her Output Hypothesis, Swain argues that learners

> need to be pushed to make use of their resources; they need to have their linguistic abilities stretched to the fullest; they need to reflect on their output and consider ways of modifying it to enhance comprehensibility, appropriateness, and accuracy. (Swain, 1993, p. 160)

Pushed output draws learners' attention towards possible gaps in their linguistic system and helps them notice discrepancies between their non-target output and the correct input. This realization can occur either internally, with the learner noticing that they are missing a specific form or that their output did not match the form they were trying to produce, or directed from the exterior, when feedback is provided to the learner. Swain thus stresses the need for L2 learners to produce output in order to test hypotheses about the L2, with feedback allowing the learner to produce modified output and enabling them to update their hypotheses, if necessary. However, modified output may remain only an attempt to produce the target form, as learners do not always achieve to match their output with the target. Nevertheless, these modifications, in particular in response to feedback that provides target forms, help the learner update their hypotheses for renewed testing, thereby facilitating the eventual internalization of target-like forms.

2.3.2 Types of Corrective Feedback

Corrective feedback is ubiquitous in the language classroom, and so are the types of corrective feedback that instructors can provide to learners. However, not all corrective feedback is created equal and addresses learner errors to similar extents. This review will provide an overview over the different types of corrective feedback that have commonly been investigated in both classroom- and lab-based studies, with a focus on oral corrective feedback, i.e., feedback provided in response to errors orally produced by learners (for a review of written corrective feedback, see R. Ellis (2009) and Li and Vuono (2019)).

Over the years, researchers have made different attempts at classifying oral corrective feedback. One of the most widely accepted classification schemes (Lyster & Ranta, 1997) identified six different types of corrective feedback based on observations of student-teacher interaction in a French immersion classroom, which were subsequently reclassified into two broader corrective feedback categories: reformulations and prompts (Lyster and Ranta, 2007). Reformulations include those kinds of feedback that provide the learner with a model utterance in response to their non-targetlike output. These include recasts and explicit corrections. Prompts, on the other hand, include those kinds of feedback that push learners to self-repair, i.e., to produce corrected versions of their original utterance, rather than providing the target reformulation. However, this doesn't mean that reformulations cannot elicit responses from learners as well.

Later, Sheen and Ellis (2011) refined this taxonomy by including a distinction between implicit and explicit kinds of feedback (see Table 2.1, with examples given in 6)). This distinction is again grounded in the processes that are hypothesized to take place within the learner when receiving feedback: Explicit feedback directly tries to draw learner attention towards problems in their output, whereas implicit feedback leaves it to the learner to notice the difference between their output and the target form. While recasts are generally considered implicit types of feedback (Long, 2007), research has shown that whether learners perceive them as implicit or explicit is dependent on a number of factors, such as the linguistic target, whether the recast is a full or partial reformulation of the original utterance, and the number of changes made to the original utterance (e.g., Ellis & Sheen, 2006; Sheen, 2004, 2006). However, with explicitness being defined in terms of perceptual salience or the overt marking of feedback as such, for instance via metalanguage (Ortega, 2009, p. 75), the degree of explicitness of a specific type of corrective feedback is difficult to hold constant across studies, as learner perceptions of salience may vary with individual learner variables, such as metalinguistic knowledge, as well as contextual variables, such as the instructional context.

In addition to distinguishing between conversational and didactic recasts, Sheen and Ellis (2011) make a further distinction between explicit reformulations that provide a model utterance, and explicit prompts that withhold the correct form from the learner and push for self-repair. Lyster et al. (2013) visualized this classification on a continuum from implicit to explicit types of corrective feedback while also making the distinction between input-providing and output-pushing types of corrective feedback (see Figure 2.1). Naturally, there are variations to these types of feedback, and, rather than always appearing as clear-cut iterations, feedback types can also appear in combination with others.

The distinction between corrective feedback types requiring self-repair on the part of the learner, i.e., prompts, and those providing target-like utterances in response to learner errors, i.e., reformulations, is also of theoretical interest to research in SLA due to the different types of linguistic evidence that they provide: positive or negative evidence. Positive evidence offers information about what is possible in a language by exposing learners to target utterances in the input. Negative evidence, on the other hand, provides the learner with information about what is not possible in a language (Gass, 1997). In terms of corrective feedback types, prompts, including clarification requests, repetitions, elicitations, and metalinguistic cues provide negative evidence; recasts provide positive evidence in addition to negative evidence, if the learner recognizes that the recast indicates an error in their utterance; and finally, explicit corrections always provide negative and positive evidence.

Table 2.1

Types of corrective feedback (adapted from Kartchava, 2019; Lyster et al., 2013)

	Implicit	**Explicit**
Reformulations – teacher corrects (input-providing)	**Conversational recasts** – a reformulation of a student utterance in an attempt to resolve a communication breakdown – often take the form of confirmation checks	**Didactic recasts** – a reformulation of a student utterance in the absence of a communication problem **Explicit correction** – a reformulation of a student utterance plus a clear indication of the error
Prompts – student self-corrects (output-prompting)	**Repetition** – a verbatim repetition of the student utterance, often with adjusted intonation to highlight the error **Clarification request** – a phrase such as 'Pardon?' and 'I don't understand' following a student utterance to indirectly signal an error	**Metalinguistic cue** – a brief metalinguistic statement aimed at eliciting a self-correction from a student **Elicitation** – directly elicits a self-correction from the student, often in the form of a *wh*-question **Paralinguistic signal** – an attempt to non-verbally elicit the correct form from the learner

6) Examples of each corrective feedback type in response to the following learner production: "He goed to the store yesterday."

a. Conversational recast:	"He went to the store yesterday to do what?"
b. Repetition:	"He goed to the store yesterday?"
c. Clarification request:	"I'm not sure I understand."
d. Didactic recast:	"He went to the store yesterday."
e. Explicit correction:	"He went to the store yesterday. The verb has an irregular past tense form."
f. Metalinguistic cue:	"Remember that some verbs have irregular past tense forms."
g. Elicitation:	"What did he do yesterday?"
h. Paralinguistic signal:	Non-verbal response

Figure 2.1

Types of corrective feedback on the implicit-explicit continuum (Lyster et al., 2013)

PROMPTS

	clarification request	repetition	paralinguistic signal	elicitation	metalinguistic cue
IMPLICIT ⟷ EXPLICIT					
	conversational recast	didactic recast	explicit correction		explicit correction + metalinguistic explanation

REFORMULATIONS

2.3.3 Effects of Corrective Feedback on L2 Learning

Due to the theorized effectiveness of corrective feedback and its importance both for theories of SLA as well as pedagogy, corrective feedback has been the focus of many classroom- and lab-based studies investigating its effectiveness and its relationship to L2 acquisition. Researchers generally agree that corrective feedback is beneficial in contexts of language instruction and helps learners acquire lexis, phonological and grammatical features, as well as pragmatics (e.g., Goo & Mackey 2013; Li, 2010; Lyster & Saito, 2010; Lyster, et al., 2013; Mackey & Goo 2007; Sheen, 2011). Among these features, grammatical errors have been shown to be addressed most frequently by language teachers, compared to other linguistic features (Brown, 2016; Mackey et al., 2000). Despite the fact that corrective feedback most often targets grammar errors, however, learners tend to recognize corrective feedback aimed at lexical errors more frequently (Lyster et al., 2013; Mackey et al., 2000; Mackey & Goo, 2007). Positive effects of corrective feedback have also been observed for L2 pronunciation development (e.g., Saito & Lyster, 2012a, 2012b). These larger benefits of corrective feedback for lexis and pronunciation have been argued to stem from the negative effects that choosing an inappropriate lexical item or severely mispronouncing a lexical item can have on the intended meaning of an utterance, possibly leading to communication breakdowns (Egi, 2007; Mackey et al., 2000). In addition, feedback targeting lexical errors may be more salient than feedback targeting grammar errors, in particular morphosyntactic errors. Consider, for instance, the following exchange, where the learner makes a lexical and morphosyntactic error at the same time:

(7) Learner: “He try to let the door open.”
Instructor: ”Okay, he tries to leave the door open.”

The instructor corrects both errors in the same feedback move. However, if detecting and/or noticing the error is a necessary requirement for feedback to be perceived as such, then feedback targeting lexical errors is simply easier to detect.

A large strand of corrective feedback research has focused on comparing the effectiveness of different types of feedback (e.g., Ammar, 2008; Ammar & Spada, 2006; Carroll & Swain, 1993; Doughty & Varela, 1998; Ellis, 2007; Ellis et al., 2006; Ishida, 2004; Leeman, 2003; Loewen & Nabei, 2007; Long et al., 1998; Lyster, 2004; Lyster & Izquierdo, 2009; Mackey & Philp, 1998;

McDonough, 2007; McDonough & Mackey, 2006; Nassaji, 2009; Sheen, 2007, 2008; Thomas, 2018; Zhao & Ellis, 2020; Yang & Lyster, 2010; see Li, 2010; Lyster & Saito, 2010 for meta-analyses). Generally speaking, implicit feedback in these studies is operationalized via recasts. Recasts are also the most frequent type of feedback found in language classrooms (Brown, 2016; Mackey & Philp, 1998), and its popularity may be due to its contingency, nonintrusiveness, and possibly affordance of both positive and negative evidence (Leeman, 2003), making it an ideal device for focusing on form during meaning-oriented communication. Regarding the effectiveness of recasts, results are mixed. Recasts have been shown to be significantly more beneficial than no feedback in lab settings (Iwashita, 2003; Leeman, 2003; Lyster & Izquierdo, 2009; McDonough, 2007; McDonough & Mackey 2006; Nassaji, 2009; Philp, 2003; Saito & Lyster, 2012a), where individual learners receive intensive recasts focused on one specific structure, potentially making its corrective nature quite salient. In classroom settings, on the other hand, recasts show small or no effects (Ellis et al., 2006; Lyster, 2004; Sheen, 2010; Yang & Lyster, 2010). Still, recasts in classroom settings can be very effective when they are intensive and particularly salient (Han, 2002; Ammar & Spada, 2006). In general, however, the effect of recasts in classrooms is smaller than in lab settings (Li, 2010). To my knowledge, no corrective feedback studies have investigated the relative effects of feedback ordering, i.e., providing two or more different types of feedback in different sequential orders to learners. However, the relative ordering may have a long-term impact on the type of information learners focus on when processing relevant feedback (see Ellis, 2006; Ellis & Sagarra, 2010; Sagarra & Ellis, 2013). This dissertation will address this gap in the literature in Experiment 1.

On the other end of the continuum, we find more explicit corrective strategies, such as elicitation and metalinguistic cues. Overall, elicitation strategies have shown to lead to more significant gains than no feedback (Ammar, 2008; Ammar & Spada, 2006; Lyster, 2004; Lyster & Izquierdo, 2009; McDonough, 2007; Nassaji, 2009; Yang & Lyster, 2010). Metalinguistic feedback has been found beneficial for classroom-based learning as well (e.g., Ellis et al., 2006; Goo, 2012). Many of these experiments contrasted different kinds of feedback to investigate the differential effects of feedback types on language learning, in particularly for morphosyntax, where the difference between the incorrect output and the correction may be less salient and therefore more difficult to perceive. Ammar and Spada (2006) investigated the learning of English possessive adjectives (*his* and *her*) in grade 6 learners with different proficiencies. The comparison of recasts and prompts that included metalinguistic cues revealed no differences on an oral picture description task at immediate posttest. However, the group that received prompts did show an advantage at delayed posttest. McDonough (2007) compared the effects of recasts, prompts (operationalized as clarification requests), and no feedback on the learning of English past tense forms elicited by information-gap tasks in a lab setting using oral production measures. While recasts and prompts led to significantly higher accuracy than no feedback, there were no significant differences between recasts and prompts. Yang and Lyster (2010), on the other hand, who also investigated English past tense forms, albeit in a classroom setting, reported that the group receiving prompts was significantly better at retrieving irregular past tense forms in an oral task at delayed posttest and regular forms at both testing times. Here, however, prompts took on one of four different forms: either metalinguistic clues, repetition, clarification requests, or elicitation. Similarly, Ellis (2007), Ellis et al., (2006), and Sheen (2007) all report that metalinguistic feedback led to more significant gains than either no feedback or recasts for English comparative forms, past tense *-ed*, and articles, respectively.

Taken together, findings regarding the effects of different kinds of feedback are mixed and seem to be modulated by a number of factors beyond just experimental setting. These factors include learner proficiency and other individual differences, as well as the targeted structure, which are discussed in the following sections.

2.3.3.1 The Role of Proficiency

Some studies have investigated whether proficiency impacts the effect of recasts and more explicit types of feedback on subsequent learning (Ammar, 2008; Ammar & Spada, 2006; Li, 2014; Mackey & Philp, 1998). This brief review focuses, in particular, on the effectiveness of recasts, as this corresponds loosely to the feedback type implemented in this dissertation. Mackey and Philp (1998), for instance, found that those L2 learners who were more developmentally advanced benefited from recasts while learning question formation in English, while those who were less advanced did not. Similarly, Ammar and Spada (2006) found that those who scored above 50% on a pretest benefited from both prompts and recasts to the same degree, while those who scored below 50% on the pretest benefited from prompts only (see also Iwashita, 2003). Ammar (2008) found similar results for the acquisition of English possessive determiners, with the additional difference that prompts also led to faster retrieval of the possessive determiner than recasts in a speeded multiple-choice test. In addition to developmental readiness, general proficiency also plays a role in the effectiveness of different types of feedback.

In a study that investigated both the role of proficiency as well as target structure, Li (2014) found that only learners at higher proficiency levels benefited from recasts for the learning of the Chinese perfective marker *-le*, whereas recasts were effective for the learning of a Chinese classifier for learners at both high-level and low-level proficiency. For both target structures, metalinguistic corrections led to larger gains than recasts for low-level learners, although recasts — in the absence of metalinguistic corrections — still increased accuracy for classifiers. For high-level learners, on the other hand, the effects across both feedback types and target structures were similar. Li argued that high-level proficiency learners may have had more cognitive resources available to process the corrective intent of recasts. Furthermore, the benefit of recasts for classifier learning in low-proficiency learners was ascribed to the relative saliency of classifiers compared to other studies that investigated structures of lower saliency (e.g., Ellis et al., 2006; Lyster, 2004; Yang & Lyster, 2010). These findings suggest that explicit feedback may be more effective for lower-proficiency learners because they lack a stable mental representation of the structure due to possibly limited exposure and experience. This uncertainty in the use of the perfective marker may have made it more difficult for lower-proficiency learners to recognize the corrective intent of recasts. By providing them with explicit feedback which includes information about the target structure, their noticing of the correct target form may therefore have been aided.

2.3.3.2 The Role of Target Structure

In addition to proficiency modulating the effectiveness of different kinds of feedback, the difficulty associated with learning the target feature may also have an impact on the effectiveness of different types of corrective feedback. In Long et al. (1998), the effectiveness of recasts was demonstrated for adverb placement but not topicalization in third-semester L2 learners of Spanish. In Iwashita (2003), recasts were effective in facilitating the acquisition of the *te*-verb form (similar to English progressive), but not that of the two locative-initial targets in L2 Japanese. While these studies investigated the effectiveness of recasts compared to a control group, other studies investigated the differential effects of different types of feedback on different target structures.

Ellis (2007), for instance, investigated whether recasts and metalinguistic feedback had differential effects for the learning of two morphosyntactic forms in L2 English, namely past tense *-ed* and comparative *-er*. Results showed that recasts did not facilitate the learning of either of these structures, while metalinguistic feedback did. There are theoretically grounded reasons for these findings: Implicit corrective feedback targeting structures that are complex in form or meaning or that are low in saliency to the learner may require greater cognitive resources, both to notice the intent of the corrective feedback but also to subsequently update the internal linguistic system (e.g., Goo, 2012; Révész, 2012). Alternatively, it is possible that some baseline knowledge of the target structure may be necessary to figure out the mismatch between output and target form through implicit corrective feedback. An appropriate level of baseline knowledge may be more readily achieved with less complex structures.

In addition to investigating the effects of proficiency as reviewed above, Li (2014) also investigated the effects of target structure on the effectiveness of feedback. The two target structures chosen for investigation were the Chinese perfective *-le*, which was deemed difficult to learn due to its lower perceptual saliency, its opaque form-meaning mapping, and the complexity of the rule governing the use of perfective *-le*, and Chinese classifiers, which were deemed easier to learn. The results indicate that for low-proficiency learners, metalinguistic feedback had larger effects than recasts for the more complex target feature, while the effects across the two feedback types were comparable for high-proficiency learners. Together, these findings suggest that recasts might be particularly beneficial for the learning of target features that are relatively more complex and therefore more difficult to acquire than others, although this again may be modulated by factors such as proficiency.

This is supported by Yang and Lyster (2010), who found that prompts worked better for learning regular English past tense forms, whereas recasts and prompts worked equally well for regular and irregular past tense forms. Varnosfadrani and Basturkmen (2009), on the other hand, showed that recasts were effective for increasing accuracy in the use of developmentally late linguistic features, i.e., such as indefinite articles, regular past tense forms, relative clauses, active and passive voice, and third person singular *-s* in upper-intermediate learners of English. Developmentally early features, i.e., definite articles, irregular past tense forms, and plural *-s*, however, increased more in accuracy when targeted by explicit corrections including metalinguistic information. The authors argue that recasts may be more effective for target features that are difficult to articulate as a set of rules, whereas explicit correction with metalinguistic feedback may help learners in formulating and testing hypotheses about rules existing in the language that apply to early-learned features. Arguably, the equation of complexity with late acquired features is questionable. However, the fact that all learners had similar proficiency levels points to the fact that some other variable, in this case the stage within a developmental sequence, may play a role as well.

2.3.4 Intersections between Error-Based Learning and Corrective Feedback

With a reduced engagement in prediction, learners' opportunities for prediction error, and consequently for priming, adaptation, and error-based learning, may be compromised (Phillips & Ehrenhofer, 2015; but see Kaan, 2015). Nevertheless, error-based learning has a parallel in the field of SLA through the concept of 'noticing the gap' (Schmidt, 1990; see also Section 2.3.1), which has also been invoked as one of the theoretical underpinnings for the effectiveness of corrective feedback in L2 learning: L2 learners learn better when they consciously notice a mismatch between what they say and what a more proficient interlocutor says, or vice versa (Philp, 2003). For such noticing to occur, however, a learner must have predicted either an output form they produced or upcoming input produced by the interlocutor before they actually produce it. If the learner notices a mismatch — or a gap — this will then trigger an adjustment of the internal linguistic system, leading to L2 learning (e.g., Altmann & Mirković, 2009; Godfroid et al., 2013; Huettig, 2015). Under such a view, the learner has to draw on production processes to predict an interlocutor's upcoming input during comprehension (see Pickering & Garrod, 2013), and it is this tightly interwoven system of production and comprehension that "underlies people's ability to predict themselves and each other" (p. 329). The processes of noticing a gap and the subsequent adaptation as a result of this noticing, drawing on production and comprehension processes, are thus very similar to how error-based learning accounts (e.g., Dell and Chang's, 2014, P-chain model) conceptualize adaptation and learning via prediction error. These parallels thus suggest that error-based learning may play a crucial role in L2 learning as well.

The connection between error-based learning mechanisms and research on corrective feedback should thus be apparent: In order to learn and adapt from prediction error in the realm of language, learners have to compare their prediction with other linguistic material available to them — either internally available in their linguistic system or provided to them from an external source — and notice the discrepancies. One way that linguistic material can be provided to learners from an external source is via corrective feedback, which minimally provides information about the accuracy of the prediction, and maximally provides the correct alternative form as well. Just as proficiency and the target structure play a role in whether L2 learners are able to generate predictions (see Section 2.2.2), these very same variables also seem to mediate the effectiveness of implicit corrective feedback (see Sections 2.3.3.1 and 2.3.3.2). Crucially, the effectiveness of error-based learning relies not only on the extent to which L2 learners are able to generate predictions and the extent to which they commit to these predictions, but also hinges on whether learners are able to recognize and process the feedback provided to them in order to compute a prediction error and to update their internal linguistic system.

The literature reviewed in Section 2.3.3 demonstrates that feedback is generally effective in aiding L2 acquisition when learners notice the feedback as such, with the impact of corrective feedback modulated, among other factors, by learner proficiency and the linguistic target. However, it is still unclear what exactly learners do with feedback in the moment they receive it and how it is processed by the cognitive and linguistic system. To date, many studies have made use of uptake to determine whether learners notice and process the feedback targeted at their error. Lyster and Ranta (1997) define uptake as "a student's utterance that immediately follows the teacher's feedback and that constitutes a reaction in some way to the teacher's intention to draw attention to some aspect of the student's initial utterance" (p. 49). For instance, learners might repair their utterance and produce modified output (e.g., Egi, 2010; Gurzynski-Weiss & Révész, 2012; McDonough & Mackey, 2006; Sheen, 2008), as illustrated in the example (8) below.

(8)	Learner:	"Yesterday, he goes to the store."
	Instructor:	"Okay. Remember that you have to use a past tense form of the verb when you talk about something that already happened."
	Learner:	"Yesterday, he went to the store."

Findings regarding utterance repair and modified output have been mixed, however, especially the relationship between modified output and learning as evidenced on posttests (e.g., Loewen, 2005; Loewen & Philp, 2006; Mackey & McDonough, 2006; Mackey & Philp, 1998;

McDonough, 2004, 2005, 2007). For instance, while Mackey and Philp (1998) found no difference in the frequency of producing modified output between developmentally more and less advanced learners, more advanced learners still produced the target structure more frequently at posttest compared to pretest than less advanced learners. McDonough (2004) on the other hand, showed that learners who produced more modified output during learning tasks had higher production accuracy of English conditionals at immediate and partially also at delayed posttests, compared to learners who produced less modified output (see also McDonough, 2005). Still, utilizing modified output as a measure of feedback processing is problematic, as the occurrence of uptake and modified output does not directly indicate that the feedback was processed and resulted in learning (Mackey, 2006). Similarly, the absence of uptake and modified output does not necessarily preclude that learning occurred: "the lack of evidence of noticing or attention is not equivalent to proof that attention or noticing is not present; in other words, absence of evidence is not the same thing as evidence of absence" (p. 409). Even in the absence of modified output, processing of the feedback may have taken place.

The fact that feedback may be processed without an overt response from the learner has been demonstrated with the use of electroencephalography (EEG) and event-related potentials (ERPs), specifically the feedback-related negativity (FRN; Falkenstein et al., 1991; Gehring et al., 1993). The FRN is argued to reflect the process of evaluating an expected outcome with the actual outcome (Ernst & Steinhauser, 2012). This ERP component is elicited when unexpected feedback is presented after participants made a decision in a learning task, i.e., they generated a prediction about an outcome and then experience a prediction error in response to the feedback presented. Importantly, the presence of the FRN indicates feedback processing even in the absence of modified output. Some studies have demonstrated that a stronger FRN effect — the experience of a greater prediction error — is associated with a larger proportion of subsequent corrections or behavioral adaptations (van der Helden et al., 2010). Other studies have shown that improvements in performance over the course of a learning task result in a smaller FRN in response to negative feedback — i.e., an outcome that is contrary to what was expected — and that steeper reductions in the FRN are associated with faster learning (Arbel & Wu, 2016). Crucially, these findings indicate that the way individuals process feedback has implications for the quality of learning, regardless of whether learners immediately produce and correct a problematic utterance.

However, to date no study has investigated the processes that unfold during online feedback processing, their development over the course of a learning task, and how online feedback processing relates to learning target features of an L2 (but see Bultena et al., 2017, 2020). These are important aspects, however, as studying these processes not only addresses the way in which learners notice and process feedback, but also may help explain why certain types of feedback are more effective with regards to certain target structures compared to others. In addition, it can help adjudicate to what extent error-based learning via prediction and prediction error indeed applies in L2 learning. The second experiment in this dissertation therefore addresses the question of how learners process feedback in real time and use it after generating predictions to adapt their linguistic system. Rather than using EEG, however, this dissertation employs pupillometry to measure feedback processing online in the moment learners receive it. Pupillometry has been deemed a valid measure to index different types of cognitive functions, including the computation of prediction error (Zhang & Emberson, 2020; see Section 2.4 Pupillometry). In addition, it can be used in conjunction with other eye-tracking measures to track the development of predictive eye-movements; this is not possible with EEG.

2.4 Pupillometry

Pupillometry, as the name suggests, is a methodology used to measure the change in size of pupil dilation in response to a stimulus. Although the methodology has been around for more than 60 years, researchers have only recently started to employ pupillometry in linguistic research (see Schmidtke, 2018, for a recent review). While the size of the pupil increases and decreases with changes in luminance, also known as the light reflex, another factor that influences pupil size is related to cognitive activity in response to a stimulus or task. These are also known as Task-Evoked Pupillary Responses (TERPs; Schmidtke, 2018). In fact, electrophysiological and neuroimaging studies have revealed a correlation between pupil dilation and activity in the locus coeruleus (LC; Aston-Jones et al., 1999; Murphy et al., 2014), a brainstem nucleus that is the source of the neuromodalutor norepinephrine (NE), which in turn has been implicated in cognitive functions such as memory consolidation and retrieval, working memory, attention, and expectancy violations (Brod & Breitwieser, 2019; Brod et al., 2018; Cohen Hoffing & Seitz, 2015; Hochmann & Papeo, 2014; Miller & Unsworth, 2019; 2021; Kafkas, 2021; Robison & Unsworth, 2019; Sara, 2009; Sara & Bouret, 2012; Scheepers et al., 2013; Zhang et al., 2018).

For instance, Hochmann and Papeo (2014) demonstrated that pupil dilation in infants is modulated by stimuli frequency — 3- and 6-moth old infants showed larger pupil dilations when presented with an infrequent string of syllables compared to a frequent syllable string. Cohen Hoffing & Seitz (2015) showed that pupil dilation size is related to stimulus memory — larger pupil dilations during the encoding phase of to-be-remembered target stimuli were related to better memory of and faster reaction times for both the target stimulus itself as well as of a secondary stimulus presented at the same time as the target. Finally, Robison and Unsworth (2019) investigated whether changes in pupil size are related to fluctuations in deployment of working memory. Indeed, they found that smaller pupil dilations coincided with fewer items being held in working memory, whereas pupil dilations were larger when more items were temporarily stored. As can be seen from the many different types of cognitive functions that have been investigated with pupillometry, it should be noted that TEPRs do not correspond to a singular cognitive process. Rather, this eye-tracking measure can function as an index of the relative cognitive activity associated with any of the functions described above, including memory consolidation and retrieval, working memory, attention, and expectancy violations (Fritzsche, 2021).

Of specific interest to this dissertation, pupillometry has previously been used to investigate the amount of attention deployed while encoding information during a learning task (e.g., Miller & Unsworth, 2019). Miller and Unsworth (2019) investigated the relationship between fixations, pupil dilation as an index of the amount of attention allocation, and learning in a paired associates task and found that larger pupil dilations during encoding correlated with learning: Participants with greater accuracy at recall showed larger pupil dilations during encoding, compared to learners who performed worse at the recall task. According to the authors, it is thus attentional allocation during encoding, measured via pupil dilations, that predict the degree of learning. Besides functioning as an index of attention allocation, pupillometry has also been used to investigate prediction error in linguistic (Sekicki & Staudte, 2018; Scheepers et al., 2013) and non-linguistic tasks (Braem et al., 2015; Brod & Breitwieser, 2019; Brod et al., 2018; Nassar et al., 2012; Raisig et al., 2010; Scheepers et al., 2013; Zhang et al., 2018). Zhang et al. (2018), for instance, trained both infants and adults with an implicit learning task, in which sounds and images were associated with one another, such that the occurrence of one sound predicted the appearance of a corresponding image. The authors found similarly increased pupillary responses in both infants and adults when an image failed to appear after a cueing sound, i.e., when a trial violated participants' predictions, compared to when the image reliably appeared in a testing phase subsequent to the exposure phase, and interpreted this as the manifestation of a prediction error. Similarly, Brod et al. (2018) as well as Brod and Breitwieser (2019) found larger pupil dilations when participants received surprising feedback, i.e., feedback that disconfirmed a strong expectation, compared to less surprising feedback.

Investigating prediction error during linguistic processing, Sekicki and Staudte (2018) used the visual world paradigm to investigate cognitive load in listeners processing incoming visual and linguistic material. In three experiments, participants' gaze was tracked during the unfolding of a felicitous auditory sentence (e.g., "The man now spills the water."). Visual scenes across the three experiments showed either two objects that were felicitous with the verb in the sentence but one had a better fit than the other (e.g., a glass of water or an ice cream cone); two objects, of which only one was felicitous with the verb in the sentence (e.g., a glass of water and a sausage); or two objects that both highly fit the verb in the sentence (e.g., a glass of juice and a cup of tea). In addition, participants received visual prompts that directed their gaze towards one of these objects in the visual scene. In the first two experiments, Sekicki and Staudte demonstrated larger pupil dilations — and thus higher cognitive load or prediction error — when participants had been prompted to gaze at the less expected (e.g., ice cream cone) or incongruent (e.g., sausage) object but then heard the expected continuation of the sentence, as opposed to when their gaze had been directed to the expected or congruent object (e.g., a glass of water). In the third experiment, however, pupil dilations were smaller when gaze prompts matched the referent in the sentence, but there were no differences between mismatching gaze prompts compared to no gaze prompts, indicating that processing of the referent was not complicated by the misleading gaze prompt. Still, together with Zhang et al. (2018), Brod et al. (2018), and Brod and Breitwieser (2019), these studies demonstrate that the pupillary dilation response is indeed affected by cognitive surprise and prediction error, in both non-linguistic as well as linguistic environments. Importantly, the measure of pupil dilation change is sensitive enough to indicate varying degrees of prediction error, i.e., the degree or magnitude of mismatch between a prediction and an outcome (Zhang & Emberson, 2020).

Despite being in its infancy in the field of L2 learning, pupillometry as a methodology thus has great potential in its application as a measure of cognitive activity. Previous studies have shown that it can reliably index attention and prediction error, both variables of interest to this dissertation. Pupillometry is thus a useful measure that can complement other eye-tracking measures, such as gaze fixations: Fixations indicate the locus of attention, while pupil dilations inform researchers about the amount of attention or other cognitive activity involved in the processing of incoming input. The second experiment in this dissertation thus uses both eye fixations and pupillometry with an adapted VWP to investigate the emergence of predictive processing in L2 learners (see also Schremm et al., 2017) and the amount of cognitive activity during feedback presentation — the moment in which learners would experience prediction errors.

2.5 The Present Study

2.5.1 Motivations and Goals

The literature reviewed in this chapter has illustrated the current state of research concerning error-based learning mechanisms and their applicability in contexts of L2 learning, as well as the role that corrective feedback plays in L2 acquisition. By doing so, the literature review brought together two strands of research that are directly relevant for one another — consider the role of feedback in error-based learning — but so far have only been considered separately. At the same time, the literature review also raises important questions regarding the role of proficiency and target structure, as well as the processes that unfold while feedback is being processed online. This dissertation addresses these issues and has the following main goals:

1. To explore the role and interactions of proficiency, target structure, and feedback type for the applicability and success of error-based learning mechanisms in L2 learners.
2. To investigate whether training involving error-based learning mechanisms leads to the emergence of online predictive behavior.
3. To explore changes in online feedback processing and how they relate to learning the targeted structure.

In order to achieve these goals, I present two studies that test error-based learning mechanisms and the role of corrective feedback for the learning of a morphosyntactic alternation among German L2 learners. Experiment 1 addresses the general applicability of error-based learning mechanisms in L2 learning, specifically across the proficiency spectrum, by creating contexts that invite prediction. In addition, Experiment 1 investigates the relationship between different feedback types in response to learners' predictions and their benefits for learning the targeted morphosyntactic alternation. Experiment 2, on the other hand, focuses on the emergence of online predictive processing in beginning to intermediate German L2 learners using the visual world paradigm. In addition, Experiment 2 investigates whether and how L2 learners process corrective feedback by measuring and tracking changes in pupil dilation in response to feedback over the course of the experiment. In both experiments, offline forced-choice picture matching tasks and picture description tasks measure learning in the modalities of comprehension and production.

Together, the experiments provide a detailed picture of error-based learning in contexts of an L2. As error-based learning mechanisms include the formulation of predictions (or guesses) as well as the processing of feedback, an investigation of error-based learning mechanisms in L2 learning does not only provide information about whether error-based learning applies in contexts of L2 learning in general, but also provides a window into the role of prediction and feedback processing as well. As such, this dissertation has both theoretical value as well as pedagogical implications, in particular for the structuring of learning tasks.

2.5.2 The Target Structure

To investigate the applicability of error-based learning mechanisms, i.e., the role of prediction, corrective feedback, and prediction error, for L2 grammar learning, a target structure is necessary that includes a predictive cue signaling to speakers possible continuations of the speech input. To this end, the target structure in the present study is the present perfect, or *Perfekt*, in German. In contrast to English, *Perfekt* in German does not encode perfective aspect (Schmuck, 2013). Instead, it is equivalent to the English simple past in that it refers to events that have been concluded in the past (although there are stylistic and regional differences between the two forms; see Schmuck, 2013). Crucially, there are additional differences between English present perfect and German *Perfekt*: In German, the *Perfekt* is formed with one of two auxiliary verbs in the second (V_2) position in a main clause, depending on the semantics of the verb phrase: the present indicative of *haben* 'to have' is used with transitive atelic verbs, such as 'to eat', while the present indicative of *sein* 'to be' is used with intransitive telic verbs that involve movement between two end points or a change in condition, such as 'to run' or 'to wake up' (Musan, 2002; Schmuck, 2013; see Example 9 below). The alternation thus has an underlying semantic component that is difficult for L2 learners to acquire, and learners across various proficiency levels often make errors when choosing the appropriate auxiliary, generally relying on *haben* as a default auxiliary for all main verbs (e.g., Bordag & Sieradz, 2014).

In addition to the auxiliary alternation, main verbs form the past participle in different ways, depending on whether they are strong verbs (*ge*- + main verb [with or without stem vowel change] + *-en*, e.g., *gelaufen* 'walked/ran'), weak verbs (*ge*- + main verb [without vowel change] + *-t*, e.g., *gekocht* 'cooked'), or mixed verbs (*ge*- + main verb [with stem vowel change] + *-t*, e.g., *gerannt* 'ran') (see Schmuck, 2013). The past participle is always placed at the end of a main clause.

9) a. Tina hat gestern Abend spät gekocht. 'Tina cooked late last night.'
b. Tina ist gestern Abend spät gelaufen. 'Tina walked/ran late last night.'
c. Tina ist gestern Abend spät gerannt. 'Tina ran late last night.'

The *Perfekt* was chosen for the purpose of this study for two main reasons: 1) The underlying semantic rule to the auxiliary alternation potentially creates a context in which speakers can predict the upcoming past participle of the main verb at the end of the sentence. If speakers hear or read the auxiliary *sein* in a context that indicates the use of *Perfekt,* they should be able to predict that the upcoming past participle denotes an action that involves movement or a change in condition. And 2), while German *Perfekt* is a grammatical structure, it includes both a semantic component, namely the relationship between the auxiliary and the past participle form of the main verb, as well as a morphological component, namely the form of the past participle. The target structure thus offers the unique opportunity to investigate the effects of both semantically meaningful prediction as well as the differential effects of feedback targeting either the semantic aspect of the structure only, or both the semantic aspect and the form of the past participle, during L2 learning.

CHAPTER 3: EXPERIMENT 1

3.1 Introduction and Research Questions

The previous chapter outlined several open questions regarding error-based learning and its applicability to L2 learning. Central to these questions is the role of feedback and learner proficiency. The first experiment in this dissertation therefore broadly addresses whether prediction — together with feedback, i.e., the experiencing of prediction error — can function as a learning mechanism in the L2. More specifically, it investigates whether encouraging learners to predict the verb that completes a sentence fragment — in conjunction with its past participle form — helps with the acquisition, primarily, of the auxiliary alternation in German *Perfekt,* and secondarily, of the past participle forms of each targeted verb. In addition to the role of prediction, Experiment 1 investigates whether the relative ordering of two different types of feedback, one targeting the auxiliary alternation only and the other targeting both the auxiliary alternation and the past participle form, has differential impacts on the learning of German *Perfekt.* Finally, Experiment 1 investigates the role of proficiency in the applicability of prediction as a learning mechanism at different stages of L2 learning by including German L2 learners across the whole proficiency spectrum. By combining a training task that encourages prediction and provides feedback with outcome measures that test receptive and productive knowledge of *Perfekt*, this experiment offers insights into whether and how error-based learning may apply in an L2 context. Experiment 1 thus has three primary research questions:

RQ1) Do learners acquire the target structure after receiving prediction-based training, as measured via accuracy on independent behavioral posttest measures?

RQ2) Does the order in which participants receive two different types of feedback have an impact on learning of the target structure, both during training as well as on independent behavioral posttest measures?

RQ3) Does proficiency have an effect on the learning of the target structure via prediction and feedback, as measured via accuracy on independent behavioral posttest measures?

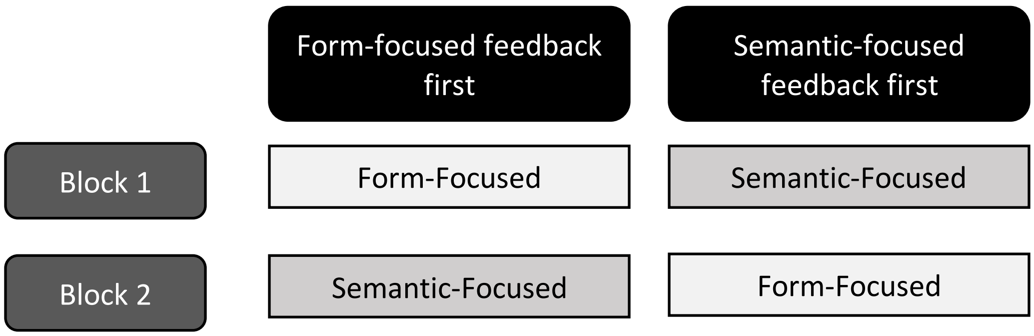
The following sections outline the experimental method applied to investigate these questions. First, I introduce the experimental approach, including the general design and time course, a description of participants in the study, the training task, as well as the outcome measures used to evaluate learning. Section 3.3 then describes the approach to data processing as well as results of the experiment. Chapter 3 concludes with a discussion of the results in light of the research questions posed above.

3.2 Overview of Experimental Approach

3.2.1 General Design and Time Course of Study

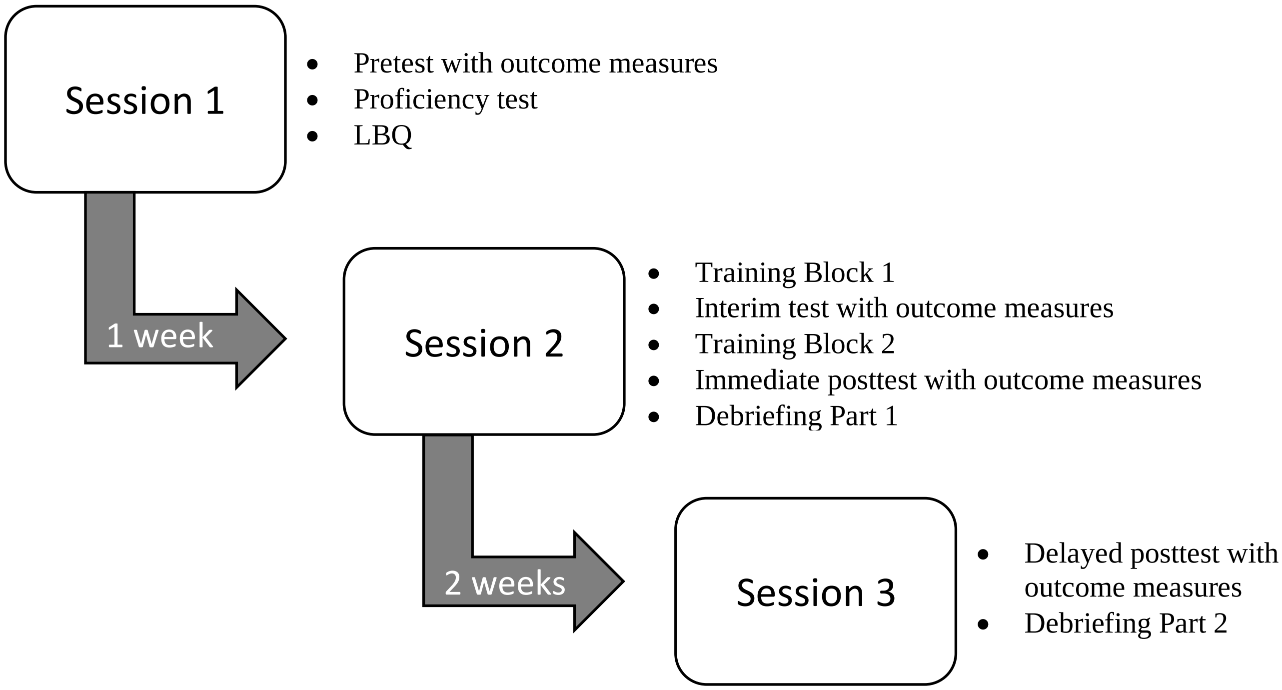
Due to the COVID-19 pandemic, all measures, including the training task, outcome measures, and measures of proficiency, were conducted online and remotely using Labvanced, a cloud-based platform for web-based experiments (Finger et al., 2017). Participants were split across two groups that differed only in the order of feedback type they received. Participants in semantic-focused feedback first group received semantic-focused feedback targeting the auxiliary alternation only in the first training block, followed by form-focused feedback targeting both the auxiliary alternation and the form of the past participle in the second training block. Participants in form-focused feedback first group received the same kinds of feedback, but the order of feedback conditions was reversed (see Figure 3.2.1).

Figure 3.2.1
Experimental design with feedback condition counterbalanced within groups and across blocks of training



Participants from Penn State completed the pretest in Week 9 of the semester. The training and interim test were then administered in Week 10, followed by the immediate posttest in the same session. The delayed posttest was then administered two weeks later in Week 12 of the semester (see Figure 3.2.2). This timing was chosen so that all elements of the experiment, including the delayed posttest, would be completed before the targeted structure was introduced to GER 1 students via their textbook and classroom materials in Week 13. Participants from the University of Texas at Austin completed the pretest in what would have been Week 13 of Penn State's semester, followed by the training, interim test, and immediate posttest in Week 14, and the delayed posttest in Week 16. Students enrolled in GER 1 were excluded in this round of data collection at UT-Austin, as they would have already been introduced to the target structure at the time of participating in the study. Both testing and training were conducted outside of participants' regularly scheduled class time. Total participation time across all sessions was around 2.5 hours.

Figure 3.2.2
Timeline of testing and training in Experiment 1



3.2.2 Participants

Participants in this experiment were students enrolled at various levels of German courses at Penn State's University Park campus, ranging from beginning-level GER 1 to intermediate and to advanced 400-level German language courses. Because of attrition in study participation, additional participants were recruited from the University of Texas at Austin. They were enrolled in equivalent German courses at this institution, excluding GER 1. Participants were recruited from these courses to participate in the experiment outside of their regular classroom instruction in Spring 2021.

In total, 71 participants were recruited across all targeted German courses. These 71 participants were randomly and evenly split across two groups that differed only in the order of feedback type they received: thirty-four participants in the form-focused feedback first group and 37 participants in the semantic-focused feedback first group. Of those 71 participants, all participated in Session 1, 58 participated in Session 2, and 48 participated in Session 3. Participants received monetary compensation in the form of electronic Amazon gift cards at the rate of \$10/hour. For participating in all study sessions, they thus received a total of \$25: \$6.67 for the first and third session and \$11.67 for the second session. One participant was excluded because they indicated being a native speaker of German on the language background questionnaire (LBQ), which included questions about participants' previous experience with German as well as their native language and other foreign languages. Most participants were native speakers of English, but other L1s included Mandarin Chinese (*N*= 1) and Spanish (*N*= 1). Two other participants were excluded for being at ceiling on both outcome measures at pretest already. Finally, only data from participants who competed all three sessions were included in the analyses. Data from 46 participants are therefore included in the analyses and results. Of these 46 participants, 22 were in the form-focused feedback first group and 24 were in the semantic-focused feedback first group.

3.2.2.1 Screening and Proficiency Measures

As mentioned above, participants completed two measures that were designed to assess participants' demographic and language learning history, as well as their proficiency in German. Together, these tasks ensured that both experimental groups were comparable with respect to their language background and proficiency. In addition, proficiency was also measured as it is one of the variables of interest in this experiment.

3.2.2.2 Materials and Data Processing

The language background questionnaire (Appendix A) consisted of multiple choice, Likert-scale, and open-ended questions designed to collect information about the languages participants knew as well as basic demographic information, such as age and gender. Most relevant for the present experiment, participants provided information about their native language and the language(s) spoken in the home, as well as detailed information about their language learning history, including the foreign languages they had learned or were currently learning, when they had started learning these languages, how much time they had spent learning these languages (measured in years), and whether they had lived or studied abroad in a German-speaking country for more than four consecutive weeks (measured in months). In addition, the language background questionnaire asked participants to rate their proficiency in English, German, and any other languages they know with regards to reading, spelling, writing, listening, and speaking the respective language. Participants rated their skills on a 10-point Likert scale, with a score of 1 indicating "poor" skills and a score of 10 indicating "very good" skills.

The independent proficiency test consisted of a combination of two German language placement tests, one used at the University of Wisconsin (University of Wisconsin Center for Placement Testing, 2006), the other used by the Goethe Institut (Goethe Institut, 2020). The combined test (see Appendix B) consisted of multiple-choice questions, five of which were embedded in a consecutive reading paragraph, and targeted a variety of grammatical structures, such as subject-verb agreement, word order, case, tense, and the use of prepositions. Each question had four answer possibilities, of which only one was correct. Two examples are illustrated in (1) and (2): Each item consisted of a sentence, from which a word or phrase had been deleted. Participants had to choose between four answer possibilities to complete the sentence. If they did not know the correct answer, they were encouraged to guess. For each item, there was only one correct answer that would render the sentence grammatical. Answers were thus scored with 1 point for a correct answer and 0 points for any incorrect answer. The combined proficiency test consisted of 42 questions in total; participants could thus score between 0 and 42 points. Scored points were then transformed to proportions for each participant, i.e., a participant who scored 30 points out of 42 received a proficiency score of 0.71.

(1) Ich habe mit ___ Freund getanzt.

a. mein
b. meinem
c. meinen
d. meiner

(2) ___ nach Qualität kosten die Jacken 100 bis 200 Euro.
a. Entlang
b. Jede
c. Mal
d. Je

3.2.2.3 Results

The means and standard deviations for variables of interest from the language background questionnaire and from the independent proficiency test are presented in Table 3.2.1.

Table 3.2.1
Participant demographics

Group	Form-focused feedback first			Semantic-focused feedback first		
Variable	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Current age (years)	20.1	2.1	18-27	21.1	5.9	18-48
L2 proficiency test (0-1)	0.49	0.21	0.12-0.95	0.49	0.23	0.12-0.90
L2 proficiency ratings (1 - 10)						
Reading	5.9	2.0	1-9	5.2	2.1	2-9
Spelling	6.3	2.4	1-10	5.9	2.7	1-10
Writing	5.6	1.9	1-8	5.1	2.0	1-8
Speaking	5.4	1.6	1-8	4.7	2.1	1-8
Listening	5.5	2.0	1-9	4.8	2.4	1-9
Overall	5.7	2.0	-	5.1	2.3	-

To ensure that both groups were comparable on these measures before administering the outcome measures and the training unit, a series of independent *t*-tests — or Kruskal-Wallis tests where distribution was non-normal — were conducted to identify any differences. There were no differences between groups for age at the time of the study (p = .785), nor for the independent proficiency test (p = .974) or any of the self-rated proficiency measures (all p > .172).

3.2.3 Training

3.2.3.1 Materials

Materials included in the training consisted of six sentence fragments in *Perfekt*, which left out the past participle form of the target verb (see Appendix C). Each sentence fragment was constructed with the following structure: 1) Proper name 'Tina', 2) *Perfekt* auxiliary, 3) a temporal adverbial, 4) a second temporal adverbial, and 5) a blank for the past participle. These six sentence fragments each appeared with the *Perfekt* auxiliaries: 1) either the auxiliary *haben* or 2) the auxiliary *sein,* conjugated for third person singular (see Examples 3 a and b).

3) a. Tina hat gestern Abend spät _____________.
b. Tina ist gestern Abend spät _____________.

Tina has yesterday night late ____________.
'Tina ____________ late yesterday night.'

Target verbs in the experiment included both strong verbs that take the *-en* inflection for their past participle form (e.g., *gelaufen* 'walked) and weak verbs that take the *-t* inflection for their past participle form (e.g., *gekocht* 'cooked'). To focus participants' attention on the inflections, strong verbs that have a stem vowel alternation in the past participle form (e.g., *getrunken* 'drunk') were excluded and were not trained. For the same reason, mixed verbs (e.g., *bringen-gebracht*) were also excluded. A total of 12 verbs were targeted in the training, of which four take the auxiliary *sein* and the other eight take the auxiliary *haben*. Although this is not perfectly counterbalanced, it more accurately represents the frequency differences with which the *Perfekt* auxiliary *sein* is encountered in natural language use (Bordag & Sieradz, 2014). Of these 12 verbs, five were strong verbs (three with *haben*, two with *sein*) and seven were weak verbs (five with *haben* and two with *sein*; see Appendix C). These 12 verbs were chosen from *Sag Mal 2nd Edition* (Anton et al., 2017), the textbook used in Penn State's introductory German course GER 1. Only verbs that fit the requirements described above and that had been introduced to participants enrolled in GER 1 by the time the study began were chose for inclusion. This was done to ensure that all participants, including those at the lowest levels of proficiency, were familiar with the meaning of all trained verbs prior to the experiment in order to prevent cognitive resources needing to be deployed to learn word meanings. Target verbs were not controlled for their cognate status with English, as this would have reduced the already limited pool of possible target verbs by another significant amount — however, I acknowledge that this factor may have an impact on the processing and learning of both the *Perfekt* auxiliary alternation as well as past participles (see Lemhöfer et al., 2010).

Due to the inclusion of four main verbs that take the auxiliary *sein* and eight main verbs that take the auxiliary *haben* in the experiment, traditional counterbalancing through the use of a Latin Square design was not possible. Instead, six lists were created to cover the two training blocks and test measures at each testing time (pretest, interim test, immediate posttest, and delayed posttest). Within each list, each of the 12 main verbs appeared once as the target verb, each main verb with *sein* appeared twice as a foil verb for target main verbs with *haben*, and four out of eight main verbs with *haben* appeared as foil verb for target main verbs with *sein*. Across lists, each main verb was paired with each of the six sentence fragments only once. Due to the uneven distribution of main verbs across the two auxiliaries, each sentence fragment appeared with the auxiliary *sein* four times and with the auxiliary *haben* eight times over the course of the experiment. Participants were assigned a different list at each testing time and in each round of training to ensure that they only saw each main verb paired with each sentence fragment once throughout the experiment.

3.2.3.2 Training Procedure

3.2.3.2.1 Training Task

Participants completed two rounds of training, separated by an interim test, the contents of which are described in Section 3.2.4. First, participants were presented with a context that was designed to encourage prediction while making the act of predicting as natural as possible. To ensure that all participants, including the beginning-level L2 German learners, understood the context, it was presented in English, along with all other task-specific instructions. The context is provided below in Figure 3.2.3. After reading the context in which the training task was embedded, participants completed a practice trial to familiarize themselves with the task. Then, participants completed the first training block. For more specific task instructions, see Appendix D.

Figure 3.2.3

Context encouraging prediction provided to participants at the beginning of the training session

> Imagine you are living in Germany with your roommate Tina. You should remember Tina from the first session.
>
> You receive a phone call from a detective to schedule an interview. They would like to ask you about your roommate, who is suspected of having caused a car accident. The detective would like to know if you can confirm your Tina's whereabouts and activities leading up to the night of the accident. You know that Tina is innocent, but you can't quite recall what your roommate did these past couple of days.
>
> On the day of the interview, the detective shows you incomplete sentences of what Tina said she was doing, along with two pictures. Only one of these pictures fits with what she said she did. Your task is to guess everything Tina said she did by choosing a picture and completing the incomplete sentences. The closer your guesses are to what she said she did, the more convinced the detective will be that Tina is innocent!
>
> Each time you choose a picture and complete an incomplete sentence, the detective will give you feedback whether it fits with what Tina said she did.

Each training block consisted of 12 trials. In each trial, participants were presented with a sentence fragment in *Perfekt* (see Examples 4 a and b).

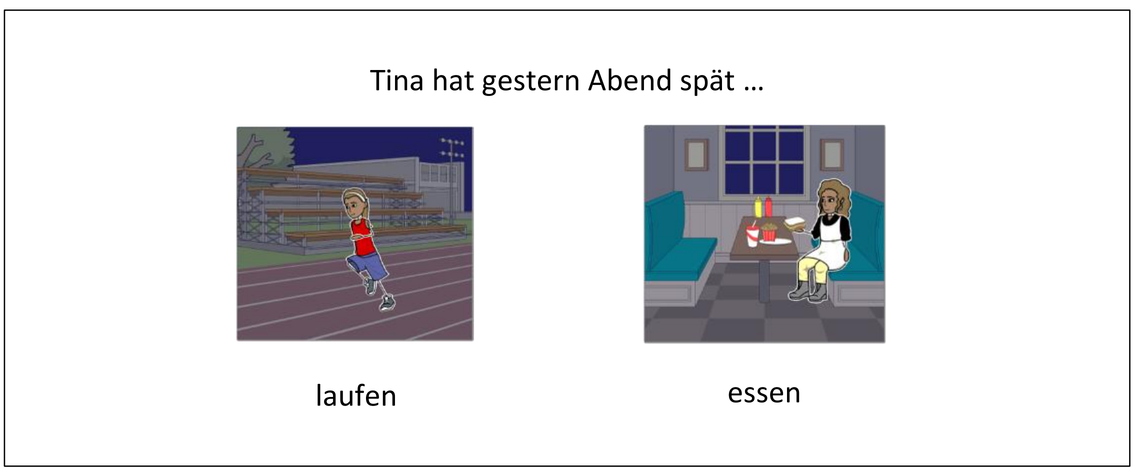
4) a. Tina hat gestern Abend spät _____________.
   b. Tina ist gestern Abend spät _____________.

Participants were presented with the same 12 main verbs in each training block, meaning that in each training block, participants were presented with eight sentence fragments with target verbs that take the auxiliary *haben* and four that target the auxiliary *sein*. However, each main verb was embedded in a different sentence fragment across training blocks. In addition to the sentence fragments, a set of 36 pictures was created with Storyboardthat.com, three that correspond to each target verb. These pictures depicted the action denoted by each main verb being performed by the main character Tina, with slight changes in Tina's clothes, position, or the background image (see Figure 3.2.4 for an example). Three pictures were created for each verb, such that participants would not be able to solely rely on their visual memory to accurately complete the task, but would have to make grammatical associations between the main verb and the auxiliary instead. Each picture from the set of three appeared twice across the six lists.

In each trial, participants were presented with two pictures on the screen along with a sentence fragment: a target picture, whose corresponding verb matched the auxiliary in the sentence fragment presented, as well as a foil picture whose corresponding verb mismatched the auxiliary in the sentence fragment. In addition, the infinitive form of the verb corresponding to each picture was provided underneath each picture on the screen. For instance, participants were presented with the sentence fragment in 4a) along with the two pictures below (see Figure 3.2.4). In this case, *essen* ‘to eat’ would be the matching picture, as it takes the auxiliary *haben*. Target and foil pictures were balanced for which side of the screen they appeared on. In addition, each target picture appeared with a different foil picture on each of the six different lists, as far as possible given the unbalanced number of verbs taking *haben* versus *sein* overall.

Participants were instructed to choose the picture that they believed accurately completed the sentence fragment to describe what their roommate did the day prior by pressing one of two previously assigned keys on their keyboard: the *F* key for the picture on the left and the *J* key for the picture on the right. They were encouraged to do so as quickly and accurately as possible. After choosing the picture, they were then prompted to complete the sentence fragment by typing the appropriate past participle form of the verb associated with the picture they chose into a text input field. Immediately following that, they received one of two types of feedback. These are described in more detail in Section 3.2.3.2.2. After the first training block, participants completed the interim test followed by the second training block. The second training block was identical to the first one, meaning that participants completed the same tasks with the same targeted structure and main verbs but with new sentence fragment-verb pairings and previously unseen picture pairings. However, the type of feedback was now switched between groups, such that after completing both training blocks, both groups had received both types of feedback. Which type of feedback they received in Block 1 versus Block 2 was counterbalanced across groups.

Figure 3.2.4
Sample trial for the picture choice task



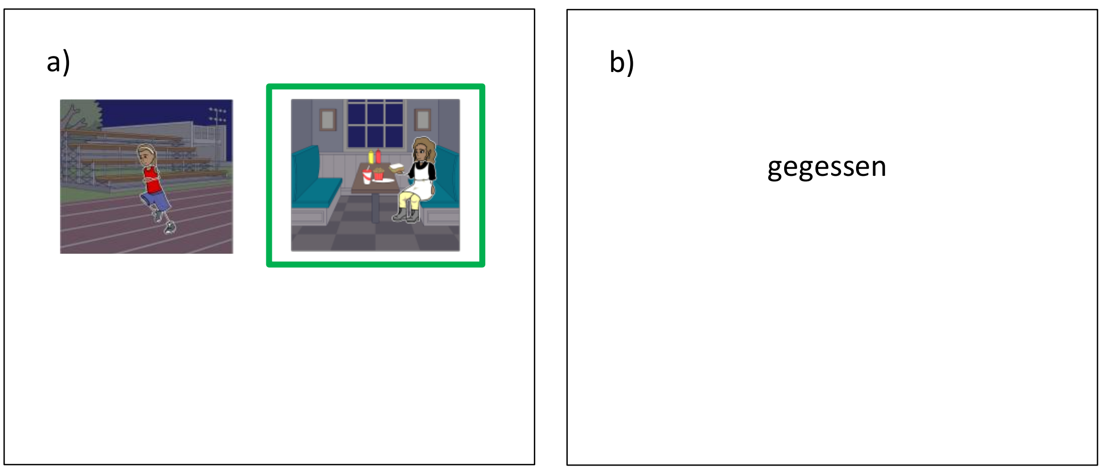
3.2.3.2.2 Feedback Conditions

There were two types of feedback that participants received during the training session, with the order in which they received each type counterbalanced by participant group. The two types of feedback differed in scope. The first type of feedback only informed participants as to whether they chose the correct picture — and therefore whether they had chosen the verb that matched the auxiliary. After choosing a picture and completing the sentence fragment with what they believed to be the correct past participle form of the target verb, participants saw both pictures again on a separate screen with a green box around the correct picture (see Figure 3.2.5a).

Feedback was presented for three seconds, after which the next trial began automatically. Participants received this feedback regardless of whether they chose the correct picture or not. Since this type of feedback exclusively targets the semantic-based relationship between the auxiliary and the main verb, it will be referred to as 'semantic-focused feedback'.

The second type of feedback informed participants whether they chose the correct picture and, therefore, verb that matched the auxiliary, in addition to whether they correctly formed the past participle form of the target verb. Participants chose a picture and completed the sentence fragment with the form of the past participle. After submitting their response, participants saw the correct form of the participle on a separate screen for three seconds, after which the next trial ensued automatically (see Figure 3.2.5b). Again, participants received this feedback regardless of whether they chose the correct picture and/or produced the correct form of the past participle or not. Since this type of feedback targets both the semantic-based relationship between the auxiliary and the main verb as well as the form of the past participle, it will be referred to as 'form-focused feedback'.

Figure 3.2.5
Sample of the a) semantic-focused feedback presentation and b) form-focused feedback presentation



The semantic-focused feedback does not have a direct equivalent in research on corrective feedback, as it relies on the non-verbal indication of what the correct picture choice would have been. It therefore does not have a direct equivalent in the language learning classroom, and, as such, may have been an unfamiliar way for participants to receive feedback on their performance. However, as it only provides positive evidence of what the correct picture choice is, it may be an effective way to provide feedback about the auxiliary alternation. Form-focused feedback, on the other hand, closely aligns with what is called a partial recast in research on corrective feedback (Ellis et al., 2006) and thus represents a type of implicit feedback that participants also typically receive in their classroom language learning experience. By giving participants the correct form of the participle, this type of feedback provides information to the learner about both the semantic-based relationship between the auxiliary and the main verb as well as the form of the past participle. This feedback type is thus richer in information than the semantic-focused feedback, as it provides two types of information about participants' performance. However, this also means that participants have to process feedback in which the saliency of the word form of the past participle may mask the semantic information conveyed about the auxiliary alternation. This setup allows us to investigate whether the effectiveness of implicit corrective feedback is related to the saliency and amount of information encoded in the feedback.

Importantly, the training units were identical across both groups and only differed within the relative ordering of which type of feedback participants received in the first training block compared to the second, i.e., ‘semantic-focused feedback’ or ‘form-focused feedback’.

3.2.4 Outcome Measures and Materials

Participants completed a written picture description task to measure their productive knowledge and a picture matching task to measure their receptive knowledge of the target structure. Picture description tasks have been frequently used in studies investigating structural priming (e.g., Grüter et al., 2021; McDonough & Kim, 2016; Schoonbaert et al., 2007; Shin & Christianson, 2012) and to measure productive knowledge of a targeted structure (e.g., Nassaji, 2017). In this study, the picture description task provided a measure of participants’ productive knowledge of the auxiliary alternation and the past participle. Similarly, picture matching tasks have also found significant use in studies investigating error-based learning (e.g., Bovolenta & Marsden, 2021; Weber et al., 2019) and receptive knowledge (e.g., Hopman & MacDonald, 2018; Keppenne et al., 2021; McDonough & Fulga, 2015), and it is here used to capture participants’ receptive knowledge of the auxiliary alternation. In the experiment, the picture description task always preceded the picture matching task to prevent priming effects from exposure to the auxiliaries in the picture matching task.

3.2.4.1 Picture Description Task

In this task, participants were presented with pictures similar to the ones they saw during training, accompanied by the corresponding infinitive verb of the depicted action. Each picture appeared on a separate screen. The participants’ task was to describe the picture using the targeted *Perfekt* structure (temporal adverbials were not required for this task). Participants typed their description into a text box underneath the picture and the infinitive verb. At each testing time, participants described pictures that corresponded to all targeted verbs in the training task. The order of picture presentation was randomized across participants. The picture descriptions were then coded for target use of the auxiliary and of the past participle inflection and prefix (see Section 3.3.1.2). Participants completed 12 target trials at each round of testing.

In addition to testing the trained target verbs, six additional trials were included to test for generalization effects, three with main verbs that take *haben* and three with main verbs that take *sein* as its auxiliary. These verbs were not trained and thus allow us to measure whether participants were able to generalize the underlying semantic regularity of the auxiliary alternation to untrained verbs. These generalization trials used the same six sentence fragments as the trials targeting trained verbs, just with the six untrained verbs. Here, a set of 12 pictures was created, with two pictures corresponding to each verb, such that participants would not see the same picture across all tasks. Again, each picture appeared twice throughout the experiment. Generalization trials in the picture description task were again coded for accuracy of the auxiliary as well as accuracy of the past participle as described in Section 3.3.1.2.

3.2.4.2 Picture Matching Task

This task was similar to the training task. Participants were presented with pairs of pictures on the screen, together with the infinitive forms of the targeted verb associated with each picture. Participants then read a sentence fragment, e.g., "Tina ist gestern Abend spät ___________." 'Tina ___________ late last night.', and were asked to match the sentence fragment to one of the pictures as fast as possible by pressing one of two pre-assigned keys: the *F* key for the picture on the left and the *J* key for the picture on the right. In each picture pair, only one picture corresponded to the sentence fragment, based on the form of the auxiliary, i.e., a picture that corresponds to a verb that takes the auxiliary *sein* was always paired with a picture that corresponds to a verb that takes the auxiliary *haben*. Thus, the correct picture choice solely depended on the auxiliary in the sentence fragment and the corresponding verb, but not on other differences in the two pictures. Feedback was not provided after any trial. While the task is similar to the training task at hand, the fact that no feedback was given, as well as the fact that the targeted main verbs were paired with different sentence fragments and pictures than during training, should prevent this measure from providing additional training. Participants completed 12 target trials at each round of testing. Both accuracy and reaction times (RTs) were recorded for subsequent analyses.[3]

In addition to testing the trained target verbs, six trials were included to test for generalization effects, with the same six untrained verbs as described above for the picture description task. These trials used the same six sentence fragments as the trials targeting trained verbs. Similarly, only one picture corresponded to the sentence fragment, based on the form of the auxiliary, i.e., a picture corresponding to an untrained verb that takes the auxiliary *sein* was always paired with a picture corresponding to an untrained verb that takes the auxiliary *haben*. The same pictures were used as in the generalization picture description task. At each testing time, verbs targeting generalization effects appeared with a different sentence fragment and a different foil picture. Both accuracy and RTs were measured.[1]

3.2.4.3 Debriefing

Participants completed two debriefing questionnaires, which were split across the second and third session. After the immediate posttest, the questionnaire included questions about (1) how much effort they put into the training tasks, (2) how difficult they perceived the training task to be, (3) whether they used any types of strategies to successfully complete the tasks, (4) whether they enjoyed completing the training tasks, (5) how curious they were to find out whether their picture choices were correct, (6) how they used the feedback they received to revise their strategies, and (7) which type of feedback they found more useful. After the delayed posttest, the questionnaire included questions about (1) what participants believed to be the goal of the training session, (2) whether participants noticed anything about the auxiliary alternation and whether and how this impacted their picture choice, (3) whether they noticed anything about the targeted main verbs, and (4) how important participants think it is to be grammatically correct in German, both on a personal level as well as to be generally understood. In both questionnaires, questions were ordered in such a way that they became increasingly more specific in asking about the targeted structure and the use of feedback (see Appendix E). Participants' answers to these questions supplement the outcome measures and provide insights into participants' internal thought processes during the

[3] Due to a software malfunction, RTs were only collected for some participants during Session 1, and no RTs were collected during any of the other sessions. Comparisons between groups and within participants are therefore not possible.

training, including how they arrived at their predictions and perceived the feedback provided to them.

3.2.5 Norming Study

Participants in the main study were encouraged to generate predictions about the past participle left out of the sentence fragment based on the auxiliary. To ensure that native speakers of German can indeed pick out the correct main verb based on the auxiliary included in the sentence fragment, as regional variation exists with regard to auxiliary choice in German, I first conducted a norming study with twenty-two German native speakers from German-speaking countries to test the materials created for the main study. Participants were asked to complete two versions of the picture matching task described above (Section 3.2.4.2): One without auxiliaries, where the goal was to ensure that the adverbials used in each sentence fragment did not bias participants to choose one picture over the other i.e., that the sentence fragments used were semantically neutral. When averaged across all participants, if the picture choice in the absence of the auxiliary approaches chance, then this indicates that the adverbials used in the sentence fragment did not bias participants to choose one picture over the other. The second task included auxiliaries to ensure that German native speakers would choose the appropriate picture based on the auxiliary provided to them in the sentence fragment. The picture matching task without auxiliaries always preceded the picture matching task with auxiliaries to prevent priming. Participants completed the norming study in approximately 15 minutes and were not compensated for their participation.

Results showed that participants were indeed able to accurately choose the appropriate verb picture based on the auxiliary in the sentence fragment (M = 0.88, SD = 0.32), and that, overall, the sentence fragments did not bias participants to choose one verb picture over the other (M = 0.45, SD = 0.50)[4]. The materials were thus deemed acceptable for use in the main study.

3.3 Analysis and Results

Two different sets of analyses were run to investigate the research questions, which are repeated below.

RQ1) Do learners acquire the target structure after receiving prediction-based training, as measured via accuracy on independent behavioral posttest measures?

RQ2) Does the order in which participants receive two different types of feedback have an impact on learning of the target structure, both during training as well as on independent behavioral posttest measures?

RQ3) Does proficiency have an effect on the learning of the target structure via prediction and feedback, as measured via accuracy on independent behavioral posttest measures?

In the first set of analyses, which addresses research questions 1 and 2, the development of auxiliary and participle accuracy over the four testing times on the picture matching and picture

[4] As the large standard deviation indicates, there were some items for which there seemed to be bias for one specific picture. However, discussion with native speakers of German who did not participate in the norming study indicated that this bias may have come from semantic constraints of the depicted verbs rather than the adverbials used in the sentence fragments. The same native speakers also asserted it was highly unlikely for L2 learners of German to be sensitive to these semantic constraints.

description task were analyzed as a function of group, i.e., of the relative ordering of the different feedback types participants received (semantic-focused feedback first vs. form-focused feedback first). In the second set of analyses, which addresses research question 3, the development of auxiliary and participle accuracy over the four testing times on the picture matching and picture description task were analyzed as a function of proficiency, independent of the relative ordering of the different feedback types. Trained and untrained generalization items were separately subjected to both sets of analyses, although only the results of the analyses for trained item are reported in this chapter, as the research questions did not directly address generalization. Results of the analyses for untrained generalization items can however be found in Appendix F. Finally, post-hoc analyses were run on the accuracy data from the training tasks, and exploratory analyses were conducted to investigate the role of prior knowledge and target structure awareness for learning.

3.3.1 Data Scoring

3.3.1.1 Picture Matching

The picture matching task consisted of 18 items — 12 trained items that were included in the training, and 6 untrained items that were used to test for generalization effects. For each of these items, the correct choice of a picture was scored with a 1, while incorrect choices were scored with a 0. Scores were kept separate for trained vs. untrained items, such that the maximum score for trained items analyzed here was 12.

3.3.1.2 Picture Description Task

The picture description task consisted of the same number of items — 12 trained items that were included in the training, and 6 untrained items that were used to test for generalization effects. The accuracy of the auxiliary as well as that of the participle were scored separately, although participle accuracy was only of secondary interest. Auxiliaries were scored based on whether participants produced the appropriate auxiliary — *haben* or *sein* — for a given target item, receiving a score of 1 for the correct auxiliary and score of 0 for the incorrect auxiliary or if no auxiliary was supplied. Although I also coded whether participants conjugated the verb correctly and used the present tense, this was not considered when scoring for overall accuracy. For participles, participants received a score of 1 if they included the *ge*- prefix, spelled the stem correctly, and used the lexically appropriate irregular *-en* or the regular *-t* suffix. If any of these components was incorrect or missing, they scored a 0 for that item. Again, scores were kept separate for trained vs. untrained items, such that the maximum score for trained items analyzed here was 12 for auxiliaries and participles each. For the auxiliary accuracy analyses in the picture description task described in the following section, however, only the four *sein* items were included, as *haben* is the default *Perfekt* auxiliary in German, is more frequent than *Perfekt* formed with *sein*, and allows for direct transfer from the English Perfect auxiliary ‘to have’.

3.3.1.3 Training

Items that were included in the training were scored similarly. For the picture choice in each of the 12 trials of each block, the correct choice of a picture was scored with a 1, while incorrect choices were scored with a 0. Within each training block, the maximum possible score was thus 12. For typing the participle, scoring mirrored that of the participle in the picture description task. Within each training block, the maximum possible score was again 12.

3.3.2 Learning and the Role of Feedback Ordering

3.3.2.1 Analyses

To address overall learning (research question 1) and learning differentiated by experimental group (research question 2), auxiliary accuracy data from the picture matching task and auxiliary and participle accuracy data from the picture description task were analyzed in separate models using mixed-effects logistic regression analyses with the *lme4* package version 1.1-21 (Bates et al., 2015) in R version 4.1.0 (R Core Team, 2021) using the *bobyqa* optimizer and Laplace approximation. For all analyses, the initial models included an interaction term between experimental group (FORM vs. SEM), sum-coded −0.5 and 0.5, and testing time (PRE, INT, IMM, DEL). Testing time was coded via a backward-difference contrast matrix, such that each testing time was compared only to the testing time immediately prior to it. For instance, accuracy at immediate posttest was only compared to accuracy at the interim test. In addition, proficiency was added as a covariate to the models to control for differences in accuracy at pretest, given that the study included participants with a large range of L2 German proficiency. Raw proficiency scores were transformed to *z*-scores and centered around 0. The final random effect structure for each model was determined by starting with the maximum structure justified by the experimental design (Barr et al., 2013), which included random intercepts for participants and items and correlated random by-participant slopes for testing time and by-item slopes for group. If a model failed to converge without warnings or showed a singular fit, models were gradually simplified first by removing the appropriate correlated random slopes, then by removing the appropriate random slope or intercept altogether. Model assumptions (collinearity, overdispersion, and autocorrelation) were checked using the *performance* package version 0.9.0 (Lüdecke et al., 2022) in R. Ninety-five percent Wald's confidence intervals for fixed effects in each model were computed using the *broom.mixed* package version 0.2.9.3 (Bolker et al., 2022) in R. Marginal and conditional R^2 are reported for each final model to assess model fit. For all models, the alpha-level was set to .05.

3.3.2.2 Results

3.3.2.2.1 Picture Matching Task

Descriptive results for the picture matching task can be found in Table 3.3.1. The *BootES* package version 1.2.1 (Gerlanc & Kirby, 2021) was used in R to generate bootstrapped 95% confidence intervals with 2000 resamples for the mean of each outcome variable. The results are also visually represented in Figures 3.3.1 and 3.3.2 using violin plots. A violin plot, which is a combination of a box plot and a kernel density plot, is used to visualize not only summary statistics like medians or means, but also the density and distribution of individual datapoints contributing to those summary statistics. The clustering of many individual datapoints around a similar value on the y-axis leads to a wider peak in the plot compared to when datapoints are distributed across a wide range of values on the y-axis.

In all tables and graphs, accuracy is calculated as the ratio of correct items to all items on any given test. For instance, if a participant chose the correct picture on 8 out of 12 items at immediate posttest in the picture matching task, their accuracy score corresponds to 0.75 at immediate posttest.

Table 3.3.1

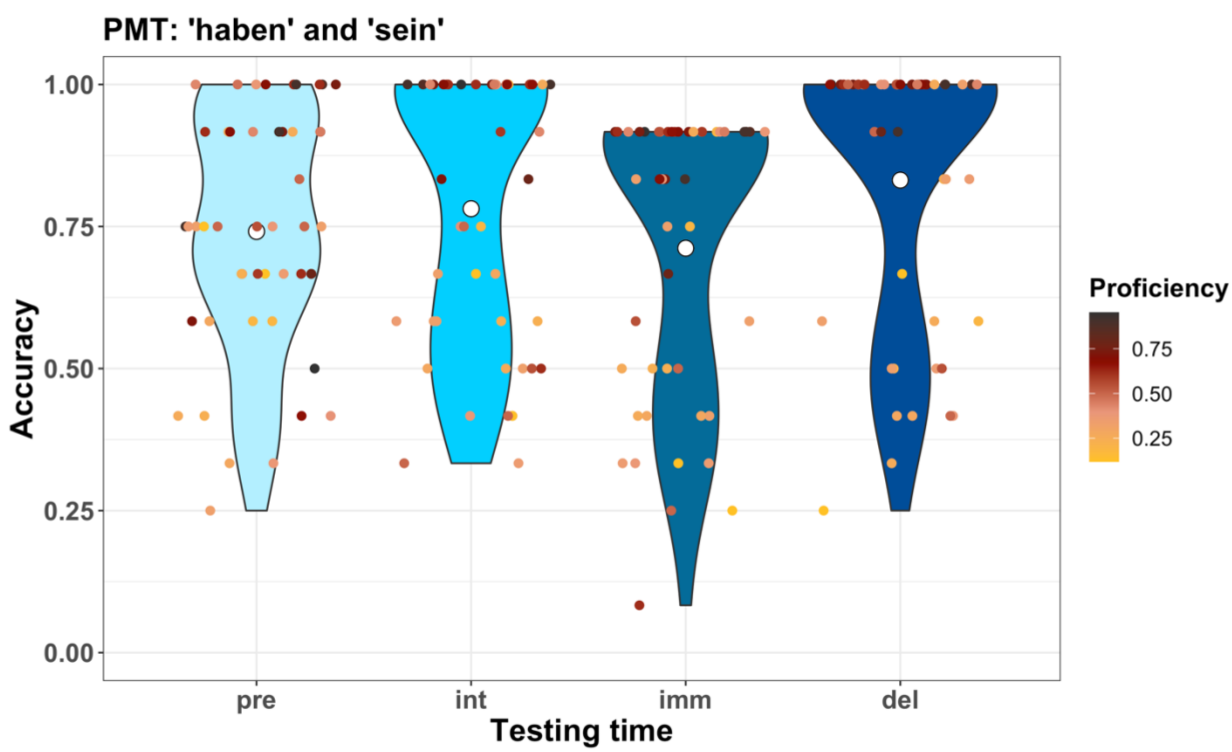
Means for auxiliary accuracy on the Picture Matching Task with bootstrapped 95% CIs

	Form-focused feedback first		Semantic-focused feedback first		Across groups	
	M (*SD*)	95% CI	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Pretest	0.69 (0.46)	[0.59,0.78]	0.77 (0.42)	[0.69,0.83]	0.73 (0.44)	[0.67,0.79]
Int. test	0.76 (0.42)	[0.67,0.85]	0.77 (0.42)	[0.66,0.87]	0.77 (0.42)	[0.69,0.83]
Imm. posttest	0.74 (0.44)	[0.63,0.82]	0.67 (0.47)	[0.55,0.77]	0.70 (0.46)	[0.62,0.77]
Del. posttest	0.81 (0.40)	[0.67,0.90]	0.84 (0.37)	[0.73,0.91]	0.82 (0.38)	[0.75,0.88]

The final generalized linear mixed effects model included random intercepts for subjects and items and a random slope for testing time by subject. The model (see Table 3.3.2) revealed significant main effects of testing time, such that accuracy across groups increased from pretest to interim test, decreased from interim test to immediate posttest, and increased again from immediate to delayed posttest. These effects can also be traced in Figure 3.3.1, where the white dots represent mean accuracy at each testing time across groups.

Figure 3.3.1

Accuracy for the Picture Matching Task by testing time

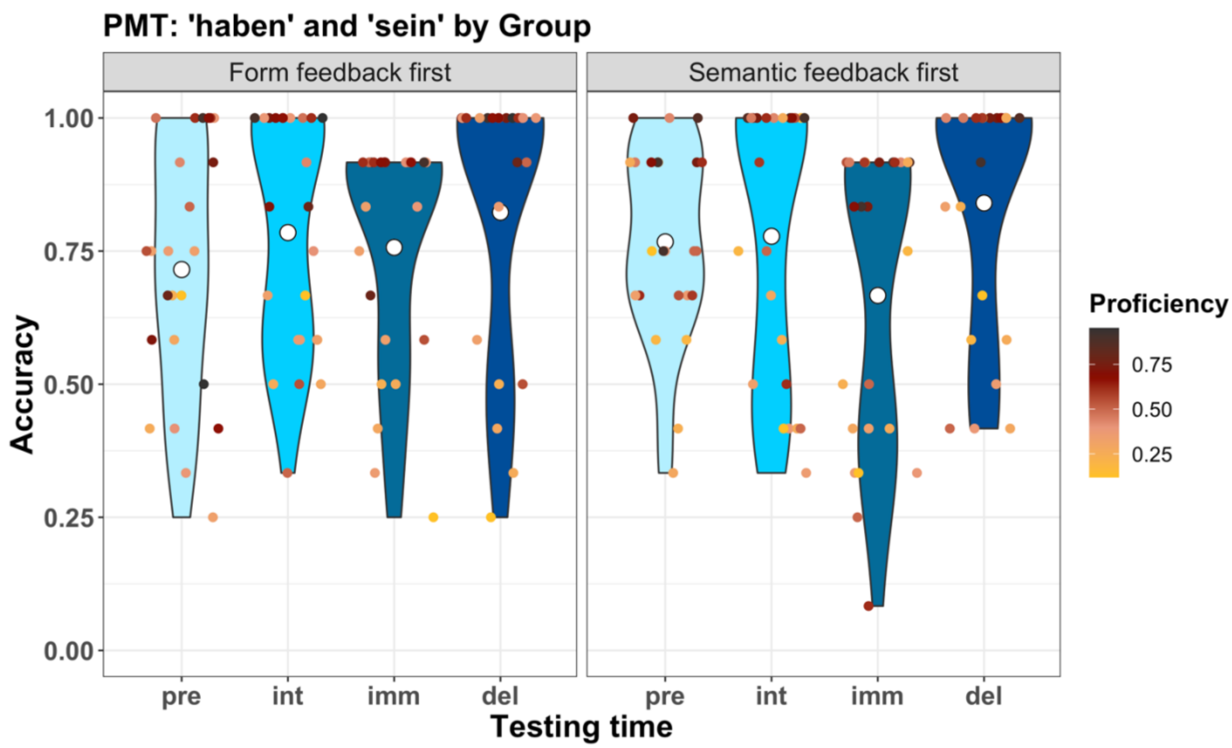


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

There were no main effects of group, but there was a marginally significant interaction between group and testing time, such that the semantic-focused feedback first group decreased more in accuracy from interim test to immediate posttest than the form-focused feedback first group (see Figure 3.3.2). However, the confidence interval for this interaction crosses zero and therefore needs to be interpreted with caution. Nevertheless, to inspect this marginal interaction in more detail, Figure 3.3.3 plots the accuracy at interim test and immediate posttest by individuals in each experimental group. It shows that while accuracy remained stable or decreased from interim test to immediate posttest for all participants in the semantic-focused feedback first group, participants in the form-focused feedback first group either decreased in accuracy, remained stable, or also increased in accuracy. Finally, the analysis also revealed a main effect of proficiency, showing that more proficient participants were more accurate than less proficient participants across all testing times.

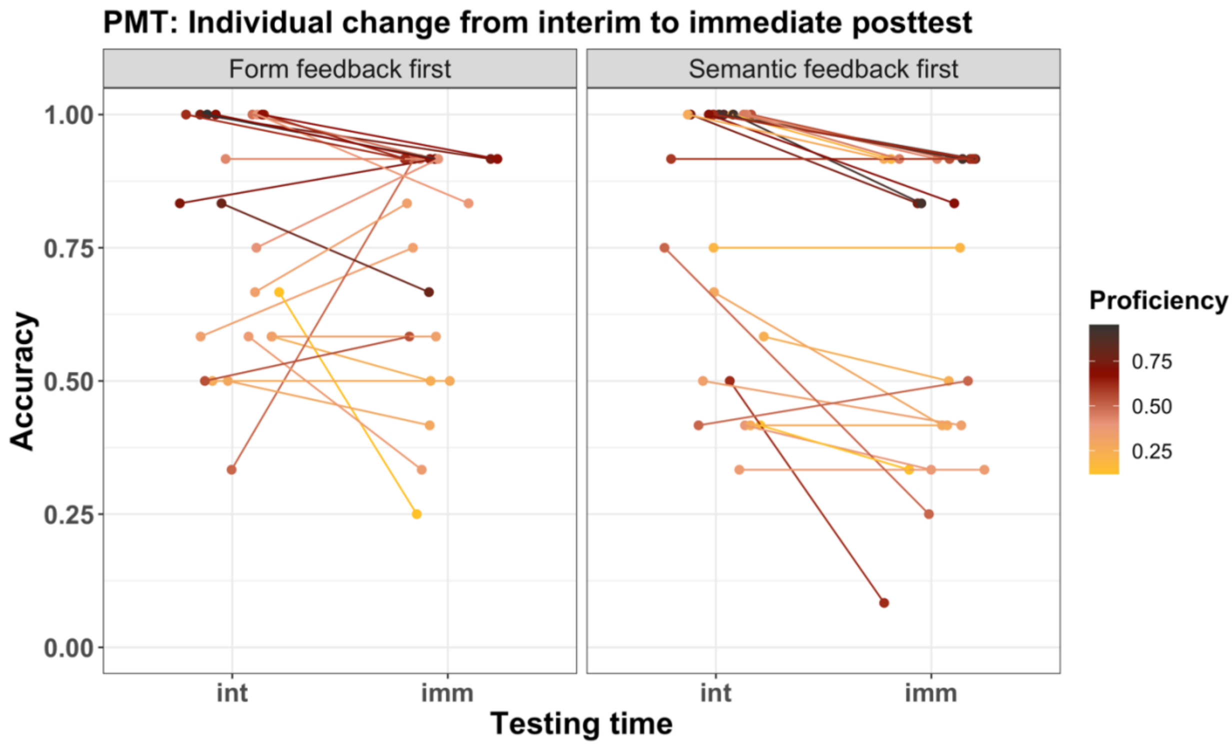
Figure 3.3.2

Accuracy for the Picture Matching Task by group and testing time



Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Figure 3.3.3
Graphic representation of the marginal interaction effect between testing time and group



Note. Colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Table 3.3.2

Model output for auxiliary accuracy in the Picture Matching Task as a function of group

						95% Confidence Interval	
Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		Lower Bound	Upper Bound
(Intercept)	1.948	0.267	7.301	2.85E-13	***	1.425	2.471
Pre-Int	0.773	0.327	2.365	0.018	*	0.132	1.414
Int–Imm	-0.813	0.255	-3.195	0.001	**	-1.312	-0.314
Imm–Del	2.054	0.482	4.266	1.99E-05	***	1.110	2.998
SemFirst	0.248	0.450	0.552	0.581		-0.633	1.130
Proficiency	0.441	0.182	2.431	0.015	*	0.086	0.797
Pre–Int:Sem	-0.134	0.579	-0.231	0.817		-1.270	1.002
Int–Imm:Sem	-0.749	0.394	-1.898	0.058	†	-1.522	0.025
Imm–Del:Sem	0.989	0.658	1.504	0.133		-0.300	2.278

Random Effects

Groups	Names	Variance	Std. Deviation	Correlation		
Subject	Intercept	1.9711	1.4061			
	Pre-Int	2.3463	1.5318	0.68		
	Int-Imm	0.4362	0.6604	-0.77	-0.78	
	Imm-Del	2.4550	1.5668	0.88	0.58	-0.53
Item	Intercept	0.1524	0.3903			

Model fit

	Marginal	Conditional
R^2	0.131	0.544

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

Model equation: Accuracy ~ TestingTime*Group + Proficiency + (1+TestingTime|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

3.3.2.2.2 Picture Description Task (Auxiliaries)

Descriptive results for auxiliary accuracy on the picture description task can be found in Table 3.3.3, and the results are graphically represented in Figures 3.3.4 and 3.3.5.

Table 3.3.3

Means for sein auxiliary on the Picture Description Task with bootstrapped 95% CIs

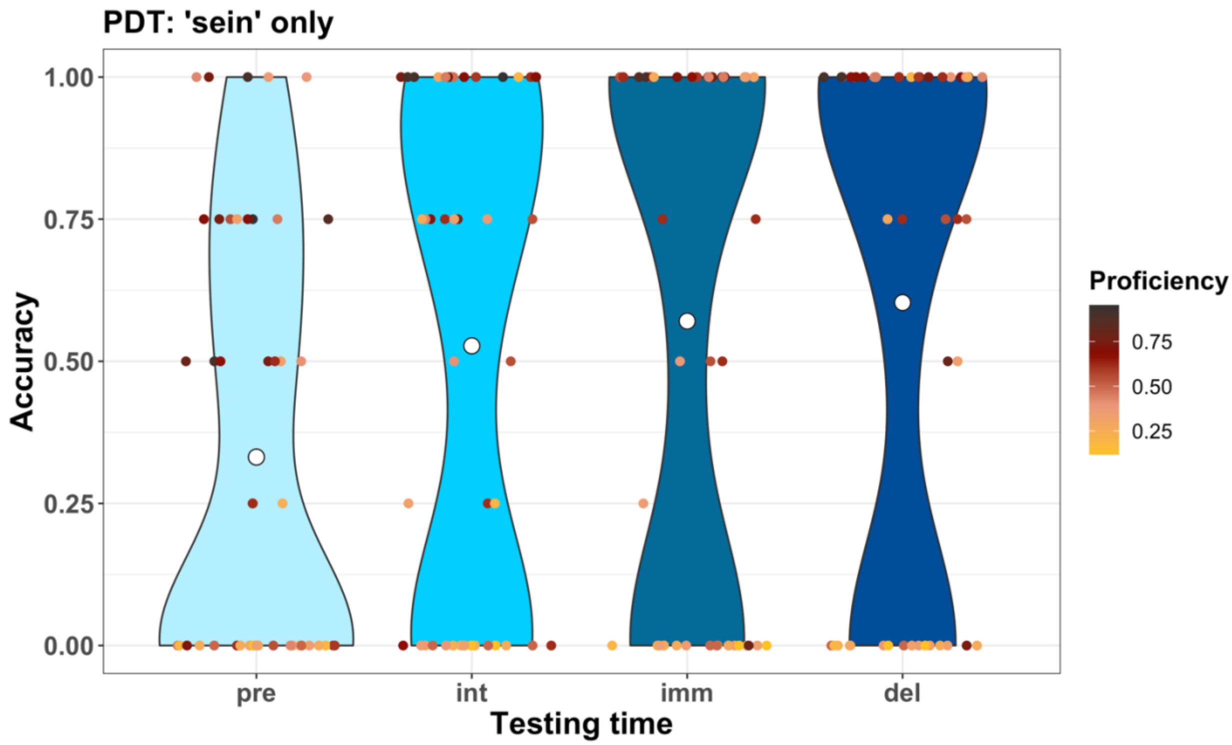
	Form-focused feedback first		Semantic-focused feedback first		Across groups	
	M (*SD*)	95% CI	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Pretest	0.41 (0.49)	[0.25,0.56]	0.26 (0.44)	[0.13,0.42]	0.33 (0.47)	[0.23,0.44]
Int. test	0.48 (0.50)	[0.30,0.63]	0.57 (0.50)	[0.38,0.74]	0.53 (0.50)	[0.39,0.64]
Imm. posttest	0.58 (0.50)	[0.36,0.74]	0.56 (0.50)	[0.33,0.73]	0.57 (0.50)	[0.42,0.69]
Del. posttest	0.66 (0.48)	[0.46,0.81]	0.55 (0.50)	[0.34,0.73]	0.60 (0.49)	[0.46,0.72]

The final generalized linear mixed effects model (see Table 3.3.4) included random intercepts for subjects and items and a random slope for testing time by subject. The model revealed one significant main effect of testing time, such that accuracy across groups increased from pretest to interim test. However, there were no other effects of testing. Again, these effects can be observed in Figure 3.3.4, with the white dots representing mean auxiliary accuracy across groups[5]. While mean accuracy increased even after interim test, the violin plots also show that the relative distribution of participants' accuracy, as indicated by the width of the violin plots at interim, immediate, and delayed posttest, remained similar. The statistical non-significance of what descriptively seems to be a constant increase after the interim test is bolstered by confidence intervals that span zero. However, this also means that accuracy remained statistically stable after the interim test and, importantly, did not decrease across testing times, even though decreases from immediate to delayed posttest are quite common.

[5] Separate visualizations showed that the increase in accuracy on *sein* items across testing times did not lead to a decrease in accuracy on *haben* items — in fact, accuracy also descriptively increased over time. This means that the increase in accuracy on *sein* items is not a result of overgeneralization.

Figure 3.3.4

Accuracy of sein *auxiliaries for the Picture Description Task by testing time*

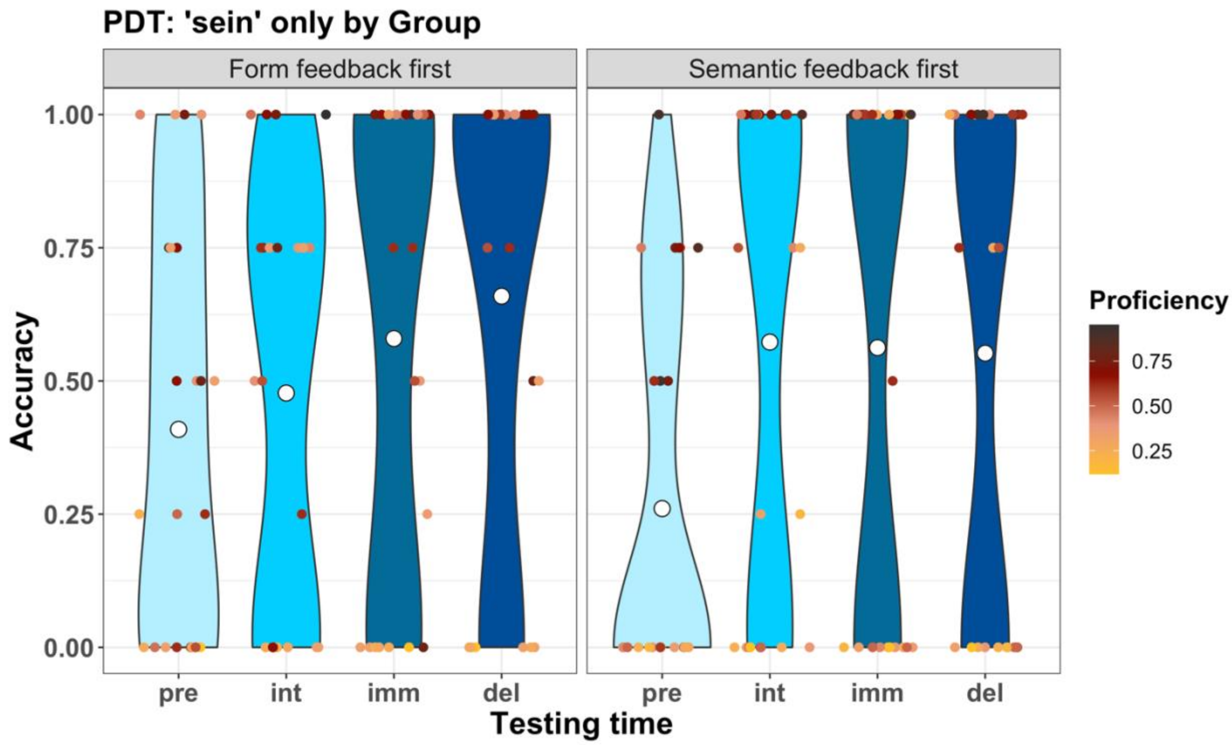


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Just like the picture matching task, there was no main effect of group. However, there was a significant interaction between group and testing time, such that the semantic-focused feedback first group increased more in accuracy from pretest to interim test than the form-focused feedback first group. While the group means in Figure 3.3.5 look very similar across groups, Figure 3.3.5 also shows a stronger change in distribution in the semantic-focused feedback first group: In the semantic-focused feedback first group, the violin plot at interim test is widest at the top, whereas the widest point in the form-focused feedback first group is slightly lower. In addition, confidence intervals for this effect do not cross zero and therefore represent a stable effect. To inspect the significant interaction in more detail, Figure 3.3.6 plots the accuracy at pretest and interim test by individuals in each experimental group. It shows that while accuracy remained stable or increased from pretest to interim test for all participants in the semantic-focused feedback first group, participants in the form-focused feedback first group either increased in accuracy, remained stable, or also decreased in accuracy. Finally, the analysis again revealed a main effect of proficiency, showing that more proficient participants were more accurate than less proficient participants across all testing times (see Figure 3.3.4).

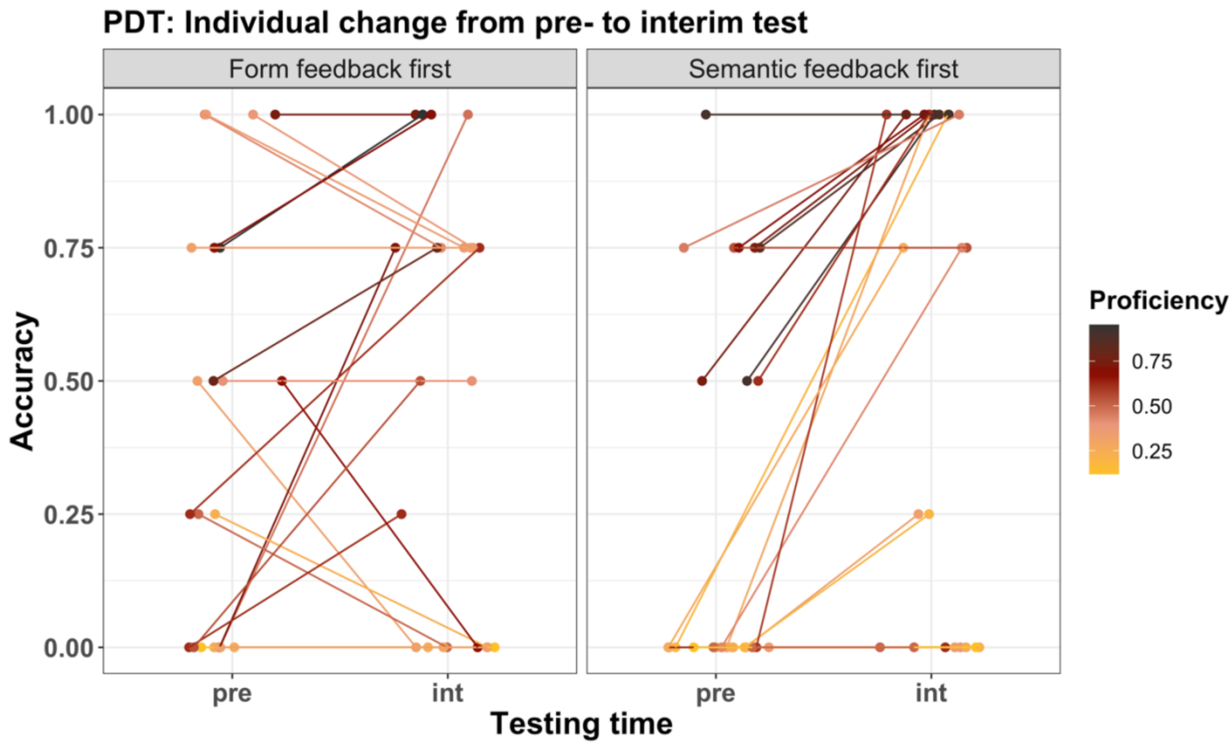
Figure 3.3.5

Accuracy of sein *auxiliaries for the Picture Description Task by group and testing time*



Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Figure 3.3.6
Graphic representation of the significant interaction effect between testing time and group



Note. Colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Table 3.3.4

*Model output for auxiliary accuracy (*sein *only) on the Picture Description Task as a function of group*

Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		95% Confidence Interval Lower Bound	Upper Bound
(Intercept)	-0.067	0.754	-0.089	0.929		-1.546	1.411
Pre-Int	2.681	0.873	3.070	0.002	**	0.970	4.393
Int–Imm	0.617	0.819	0.753	0.451		-0.988	2.223
Imm–Del	0.419	0.804	0.521	0.603		-1.158	1.995
SemFirst	-0.254	1.313	-0.193	0.847		-2.827	2.319
Proficiency	2.789	1.072	2.600	0.009	**	0.687	4.891
Pre–Int:Sem	4.704	1.610	2.921	0.003	**	1.548	7.861
Int–Imm:Sem	-1.352	1.602	-0.844	0.399		-4.491	1.788
Imm–Del:Sem	-1.875	1.598	-1.174	0.241		-5.007	1.257

Random Effects

Groups	Names	Variance	Std. Deviation	Correlation		
Subject	Intercept	15.0468	3.8790			
	Pre-Int	7.4556	2.7305	0.32		
	Int-Imm	12.1163	3.4808	0.65	0.33	
	Imm-Del	7.1739	2.6784	-0.16	-0.02	-0.46
Item	Intercept	0.2994	0.5472			

Model fit

	Marginal	Conditional
R^2	0.309	0.906

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

Model equation: Accuracy ~ TestingTime*Group + Proficiency + (1+TestingTime|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

3.3.2.2.3 Picture Description Task (Participles)

Descriptive results for participle accuracy on the picture description task can be found in Table 3.3.5, and the results are graphically represented in Figures 3.3.7 and 3.3.8.

Table 3.3.5

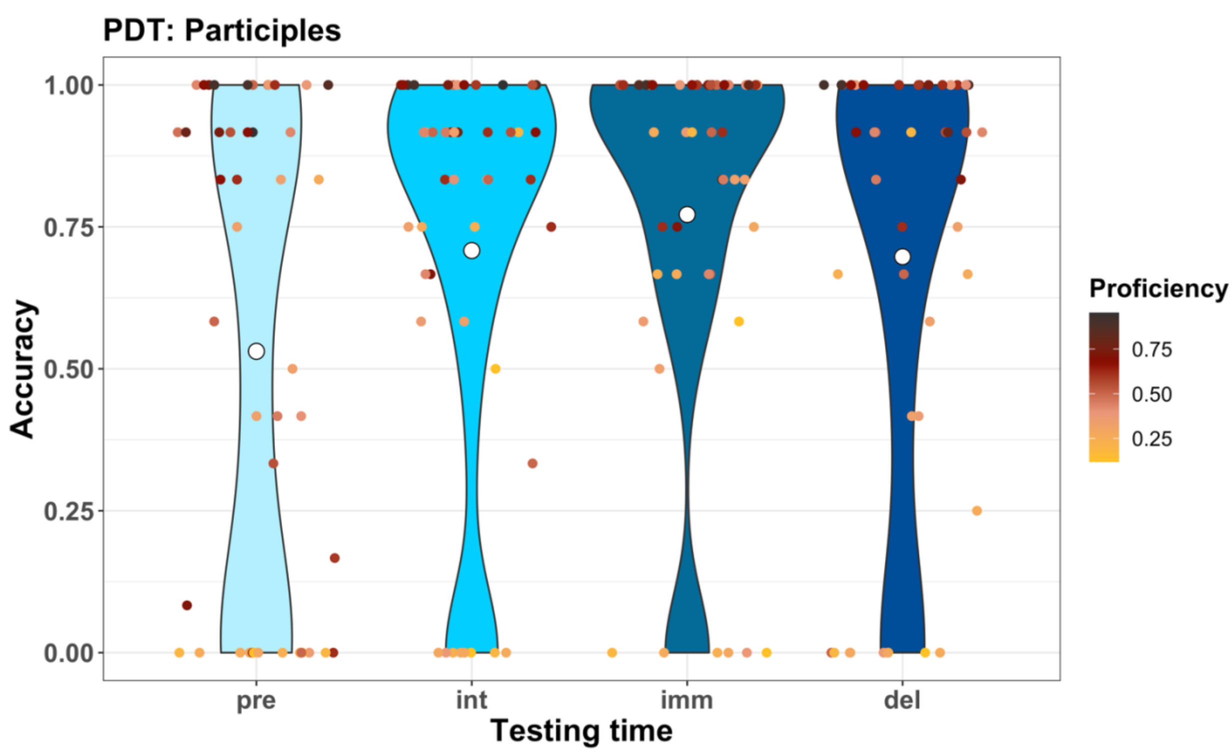
Means for participles on the Picture Description Task with bootstrapped 95% CIs

	Form-focused feedback first		Semantic-focused feedback first		Across groups	
	M (*SD*)	95% CI	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Pretest	0.56 (0.50)	[0.33,0.70]	0.50 (0.50)	[0.27,0.62]	0.53 (0.50)	[0.35,0.61]
Int. test	0.78 (0.41)	[0.54,0.84]	0.64 (0.48)	[0.41,0.74]	0.71 (0.45)	[0.53,0.76]
Imm. posttest	0.78 (0.41)	[0.57,0.85]	0.76 (0.43)	[0.56,0.88]	0.77 (0.42)	[0.63,0.84]
Del. posttest	0.78 (0.41)	[0.55,0.85]	0.62 (0.49)	[0.41,0.75]	0.70 (0.46)	[0.53,0.75]

The final generalized linear mixed effects model included random intercepts for subjects and items, but no random slopes (see Table 3.3.6). The model revealed significant main effects of testing time, such that accuracy across groups increased from pretest to interim test and from interim test to immediate posttest, but decreased again from immediate to delayed posttest. These effects are visible in Figure 3.3.7, where the white dots represent mean accuracy across groups.

Figure 3.3.7

Accuracy of participles on the Picture Description Task by testing time



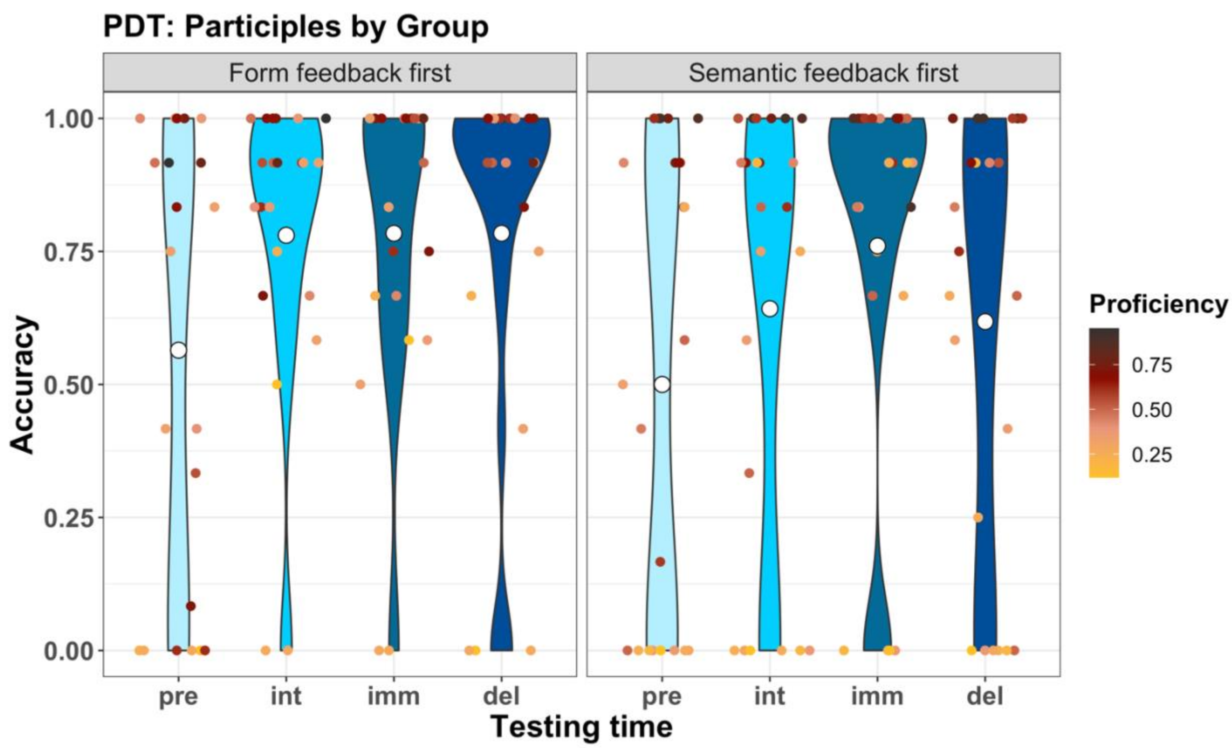
Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

There were no main effects of group, however there were significant interactions between group and testing time, such that the semantic-focused feedback first group increased more in accuracy from interim test to immediate posttest than the form-focused feedback first group, but also decreased more from immediate to delayed posttest. These effects are clearly visible in Figure 3.3.8, where white dots again represent group means. Whereas the form-focused feedback first group remains stable in accuracy after pretest, there is more variability across testing times in the semantic-focused feedback first group.

To inspect these significant interactions in more detail, Figure 3.3.9 plots the accuracy at interim test, immediate, and delayed posttest by individuals in each experimental group. Comparing individuals' change in performance from interim to immediate posttest across groups it becomes apparent that more participants in the semantic-focused feedback first group exhibited a drastic increase in accuracy, whereas any individual changes in the form-focused between these two testing times were more moderate. This trend is similar for the change from immediate to delayed posttest, although this time the effect is negative, with most participants in the semantic-focused feedback first group experiencing significant losses in accuracy, whereas most individual differences in performance in the form-focused feedback first group seem more moderate. Again, the analysis also revealed a main effect of proficiency, showing that more proficient participants were more accurate than less proficient participants across all testing times (see Figure 3.3.7).

Figure 3.3.8

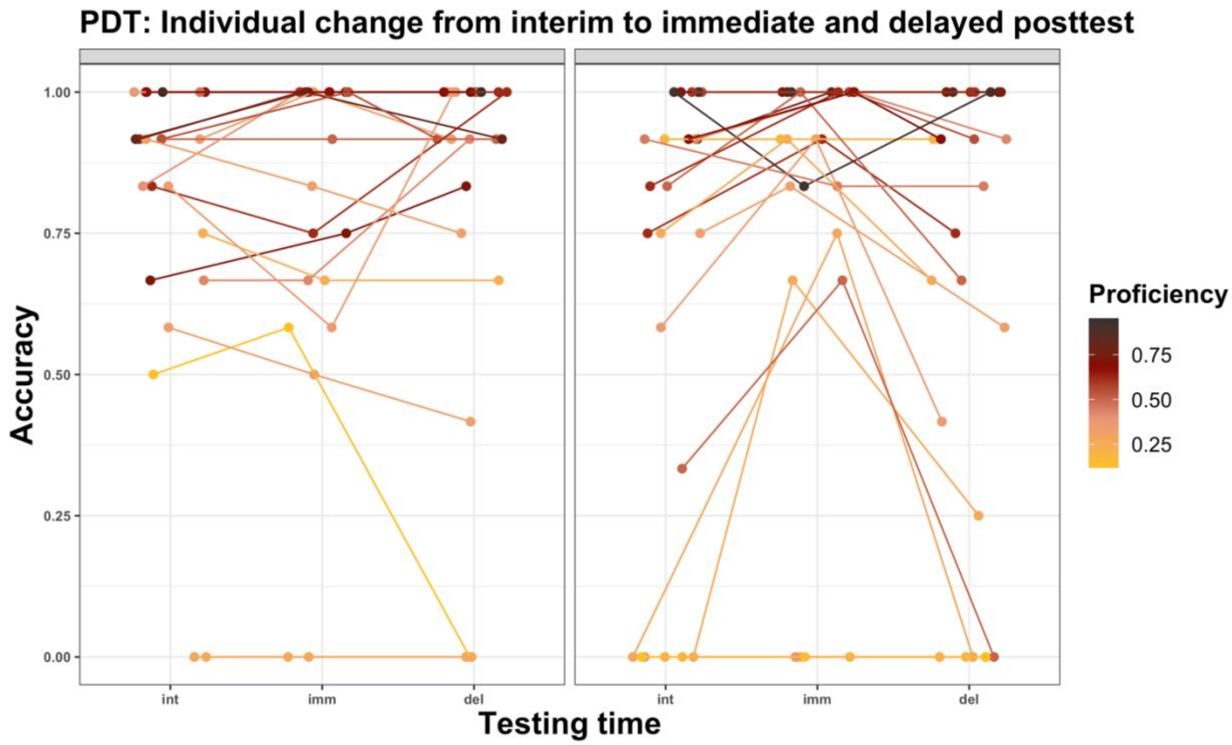
Accuracy of participles on the Picture Description Task by group and testing time



Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Figure 3.3.9

Graphic representation of the significant interaction effects between testing time and group



Note. Colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Table 3.3.6

Model output for participle accuracy on the Picture Description Task as a function of group

						95% Confidence Interval	
Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		Lower Bound	Upper Bound
(Intercept)	1.432	0.398	3.599	3.19E-04	***	0.652	2.212
Pre-Int	1.730	0.202	8.546	1.27E-17	***	1.333	2.127
Int–Imm	0.908	0.232	3.913	9.13E-05	***	0.453	1.363
Imm–Del	-1.030	0.233	-4.414	1.01E-05	***	-1.487	-0.573
SemFirst	-0.820	0.770	-1.064	0.287		-2.330	0.690
Proficiency	2.538	0.420	6.050	1.45E-09	***	1.716	3.360
Pre–Int:Sem	-0.463	0.402	-1.151	0.250		-1.252	0.325
Int–Imm:Sem	1.724	0.464	3.715	2.03E-04	***	0.815	2.634
Imm–Del:Sem	-2.060	0.467	-4.415	1.01E-05	***	-2.974	-1.145

Random Effects

Groups	Names	Variance	Std. Deviation	Correlation
Subject	Intercept	5.8827	2.4254	
Item	Intercept	0.1185	0.3443	

Model fit

	Marginal	Conditional
R^2	0.453	0.806

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

Model equation: Accuracy ~ TestingTime*Group + Proficiency + (1|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

3.3.2.3 Post-hoc Training Analysis

3.3.2.3.1 Analysis

In order to investigate the unexpected decrease in auxiliary accuracy in the picture matching task from interim test to immediate posttest across groups, but also with a stronger effect in the semantic-focused feedback first group, I conducted a post-hoc analysis of auxiliary accuracy during training. Mean accuracy on each trial for each group was visualized split by training block in order to explore possible causes for the decrease from interim to immediate posttest that may have emerged during training, in particular in Block 2. A mixed-effects logistic regression model was then created following the same procedures as described for investigating the effects of feedback ordering on the test data. For these analyses, the model included trial number instead of testing time, coded as a continuous variable.

For completeness' sake, a similar post-hoc analysis was conducted on participle accuracy during training. Only items where participants had correctly chosen the picture before typing the corresponding participle were included in the analyses, leading to the exclusion of 19.6% of all data points, i.e., 80.4% of all training data were submitted to the analysis.

3.3.2.3.2 Results

3.3.2.3.2.1 Auxiliaries

Descriptive results for auxiliary accuracy during training can be found in Table 3.3.7, and the results are graphically represented in Figure 3.3.10.

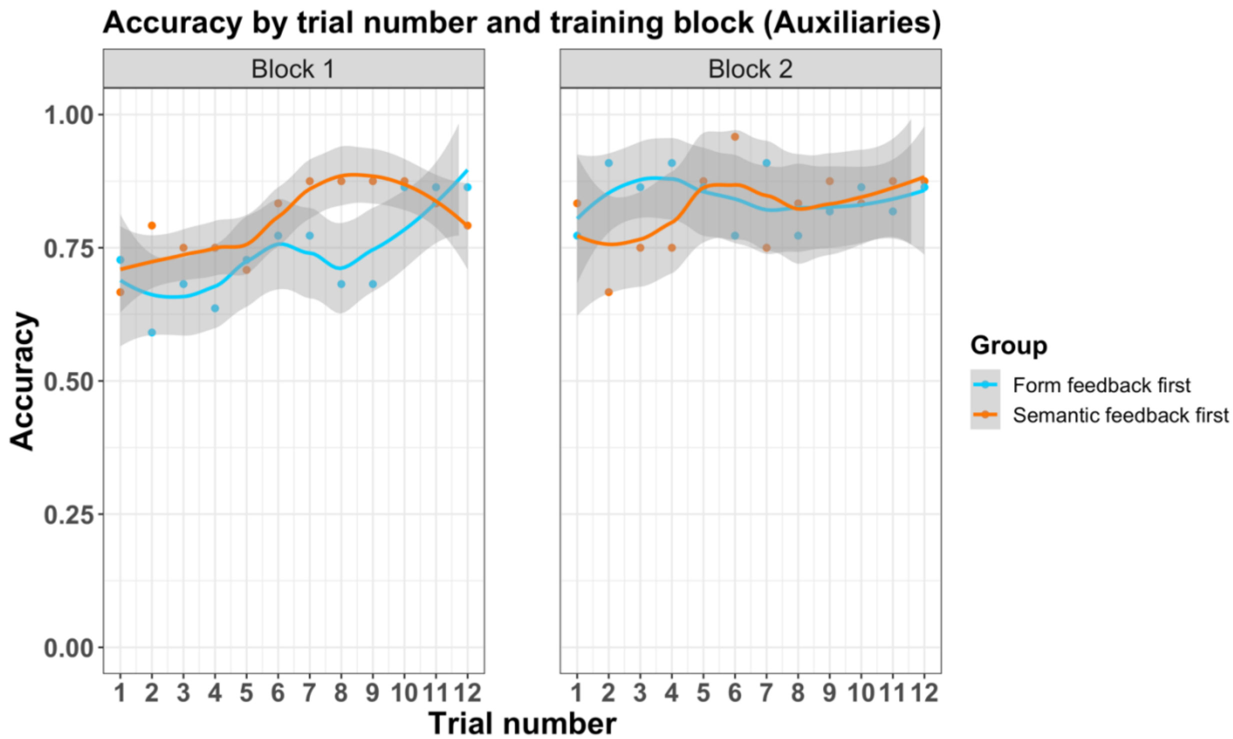
Table 3.3.7

Means for auxiliary accuracy during training with bootstrapped 95% CIs

	Form-focused feedback first		Semantic-focused feedback first		Across groups	
	M (*SD*)	95% CI	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Block 1	0.74 (0.44)	[0.64,0.83]	0.80 (0.40)	[0.70,0.87]	0.77 (0.42)	[0.70,0.83]
Block 2	0.84 (0.36)	[0.76,0.91]	0.82 (0.38)	[0.74,0.89]	0.83 (0.37)	[0.77,0.88]

Results from the analysis by trial and group are presented below in Figure 3.3.10, which shows a steady increase in accuracy in each group across trials and in each training block. In Block 1, it seems like accuracy in the semantic-focused feedback first group is consistently higher across trials than in the form-focused feedback first group, matching the differential focus of the feedback in each group. In Block 2, the pattern is reversed for the first trials, but by trial 5, both groups perform similarly.

Figure 3.3.10
Accuracy of auxiliaries during training by group and trial



Note. Lines represent smoothed conditional means, gray areas represent 95% CIs around smoothed conditional means.

The final generalized linear mixed effects model included random intercepts for subjects and items as well as a random by-subject slope for trial. The model (Table 3.3.8) revealed a significant main effect of trial, such that accuracy across groups increased as participants progressed in the training, although the effect size is very small. There was no main effect of group and the interaction between group and trial was also non-significant (see Table 3.3.8). Finally, there was a main effect of proficiency in the expected direction, showing that more proficient participants were more accurate than less proficient participants overall.

Table 3.3.8

Model output for auxiliary accuracy during training as a function of group (by trial)

						95% Confidence Interval	
Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		Lower Bound	Upper Bound
(Intercept)	1.411	0.309	4.563	5.04E-06	***	0.805	2.017
Trial	0.057	0.018	3.195	0.001	**	0.022	0.092
SemFirst	0.582	0.490	1.187	0.235		-0.379	1.542
Proficiency	1.078	0.253	4.260	2.04E-05	***	0.582	1.574
Trial:Sem	-0.032	0.026	-1.234	0.217		-0.083	0.019

Random Effects				
Groups	Names	Variance	Std. Deviation	Correlation
Subject	Intercept	1.2401	1.1136	
	Trial	0.0001	0.0089	0.79
Item	Intercept	0.1534	0.3917	

Model fit		
	Marginal	Conditional
R^2	0.214	0.472

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

Model equation: Accuracy ~ Trial*Group + Proficiency + (1+Trial|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

3.3.2.3.2.2 Participles

Descriptive results for participle accuracy during training can be found in Table 3.3.9, and the results are graphically represented in Figure 3.3.11.

Table 3.3.9

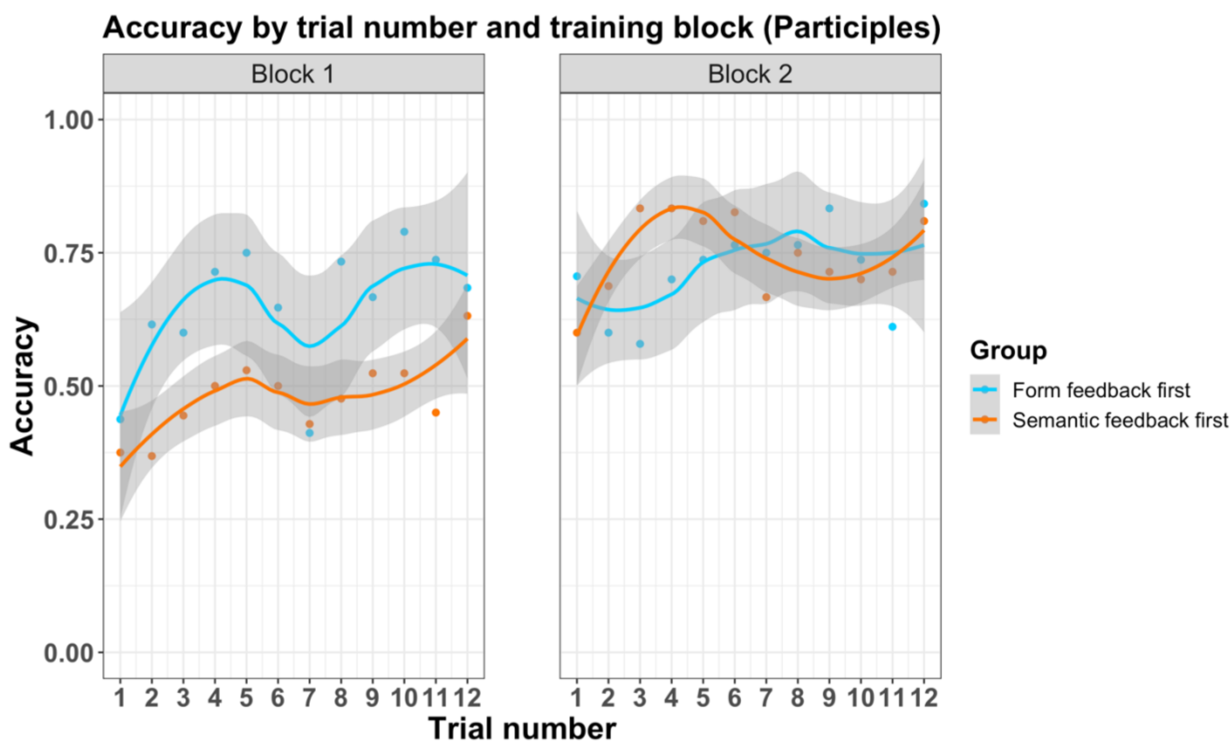
Means for participle accuracy during training with bootstrapped 95% CIs

	Form-focused feedback first		Semantic-focused feedback first		Across groups	
	M (*SD*)	95% CI	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Block 1	0.65 (0.48)	[0.45,0.70]	0.48 (0.50)	[0.23,0.61]	0.56 (0.50)	[0.38,0.61]
Block 2	0.72 (0.45)	[0.51,0.81]	0.75 (0.44)	[0.54,0.81]	0.73 (0.44)	[0.58,0.78]

Results from the analysis by trial and group are presented below in Figure 3.3.11, which again shows mostly increases in accuracy across trials in both groups and training blocks. In this graph, the form-focused feedback first group exhibits higher accuracy across all trials in the first training block, matching the expected performance based on the focus of the feedback for this group in the first training block. Again, the pattern is somewhat reversed in the second block, although accuracy is similar for both groups by the second half of the block.

Figure 3.3.11

Accuracy of participles during training by group and trial



Note. Lines represent smoothed conditional means, gray areas represent 95% CIs around smoothed conditional means.

The final generalized linear mixed effects model included random intercepts for subjects and items as well as a random by-subject slope for trial. The model (see Table 3.3.10) revealed a significant main effect of trial, such that accuracy across groups increased across trials. There was also a main effect of group, as well as a significant interaction between group and trial, such that the semantic-focused feedback first group showed a steeper increase across trials than the form-focused feedback first group. Again, the analysis revealed a main effect of proficiency, showing that more proficient participants were more accurate than less proficient participants across all trials.

Table 3.3.10

Model output for participle accuracy during training as a function of group (by trial)

Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		95% Confidence Interval Lower Bound	Upper Bound
(Intercept)	-0.594	0.544	-1.091	0.275		-1.659	0.472
Trial	0.152	0.032	4.801	1.58E-06	***	0.089	0.214
SemFirst	-2.293	1.058	-2.167	0.030	*	-4.367	-0.219
Proficiency	2.538	0.448	5.669	1.43E-08	***	1.661	3.416
Trial:Sem	0.156	0.062	2.520	0.012	*	0.034	0.278

Random Effects

Groups	Names	Variance	Std. Deviation	Correlation
Subject	Intercept	8.011	2.8302	
	Trial	0.0202	0.1423	-0.64
Item	Intercept	0.2052	0.4530	

Model fit

	Marginal	Conditional
R^2	0.563	0.807

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

Model equation: Accuracy ~ Trial*Group + Proficiency + (1+Trial|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa)

3.3.2.4 Summary: Learning and the Role of Feedback Ordering

The results from the picture matching task suggest that learning occurred across both groups, with accuracy at delayed posttest being about 9% higher than at pretest (see Table 3.3.1). More proficient learners were more accurate at all testing times, which is unsurprising, seeing that they have more experience with the target structure. While there was no overall difference between groups in how accurately they performed, there seem to have been differences between groups when looking at the decreases from the interim test to immediate posttest, as indicated by the negative coefficient of the marginal interaction between group and the change from interim test to immediate posttest. In particular, the semantic-focused feedback first group decreased more in accuracy than the form-focused feedback first group.

The results for auxiliary accuracy on the picture description task suggest that both groups increased in accuracy immediately after the first training block and remained stable thereafter, with accuracy at delayed posttest being about 27% higher than at pretest (see Table 3.3.3). Again, and unsurprisingly, more proficient learners were more accurate at all testing times. While there was no overall difference between groups in how accurately they performed, the significant interaction between group and the change from pretest to interim test indicates that the semantic-focused feedback first group increased more in accuracy than the form-focused feedback first group. These results differ from the auxiliary accuracy results of the picture matching task in important ways: (1) There was no decrease in accuracy from interim test to immediate posttest, suggesting that potential difficulties in a comprehension-based task do not translate to a different task, namely the production of relevant auxiliaries; and (2) there was a stronger increase in auxiliary accuracy from pretest to interim test in the semantic-focused compared to the form-focused feedback first group, an interaction that is not present in the picture matching task.

Finally, for participle accuracy in the picture description task, results again indicate that more proficient learners were more accurate at all testing times. Additionally, they showed learning across groups up until immediately after training, i.e., at immediate posttest, but also a decrease thereafter i.e., at delayed posttest. This overall decrease in participle accuracy at delayed posttest was driven by the semantic-focused feedback first group, as accuracy in the form-feedback focused first group overall remained stable from interim test to immediate and delayed posttest. Importantly, however, accuracy at delayed posttest was still about 17% higher than at pretest (see Table 3.3.5). The results for participle accuracy on the picture description task differ from those of auxiliary accuracy on the same task in two ways: (1) the first significant group interaction occurred between the interim test and immediate posttest rather than going from pretest to interim test; and (2), whereas the semantic-focused feedback first group remained stable in auxiliary accuracy from immediate to delayed posttest, they significantly decreased in participle accuracy during the same time period.

These findings and their implications for the role prediction and feedback ordering in L2 learning are discussed in Section 3.4 Discussion.

3.3.3 Learning and the Role of Proficiency

3.3.3.1 Analyses

To address the role of proficiency in learning (research question 3), data from the picture matching and picture description task were again analyzed using mixed-effects logistic regression analyses, mirroring the model building procedures described in Section 3.3.2.1. For both picture matching and picture description tasks, the initial models for accuracy included an interaction term between testing time (PRE, INT, IMM, DEL), using backward-difference coding as described in

Section 3.3.2.1, and proficiency, which was transformed to *z*-scores and centered around 0. Importantly, however, group was not included as a fixed effect, as this set of analyses investigates the effects of proficiency on the learning trajectory independent of feedback ordering.[6]

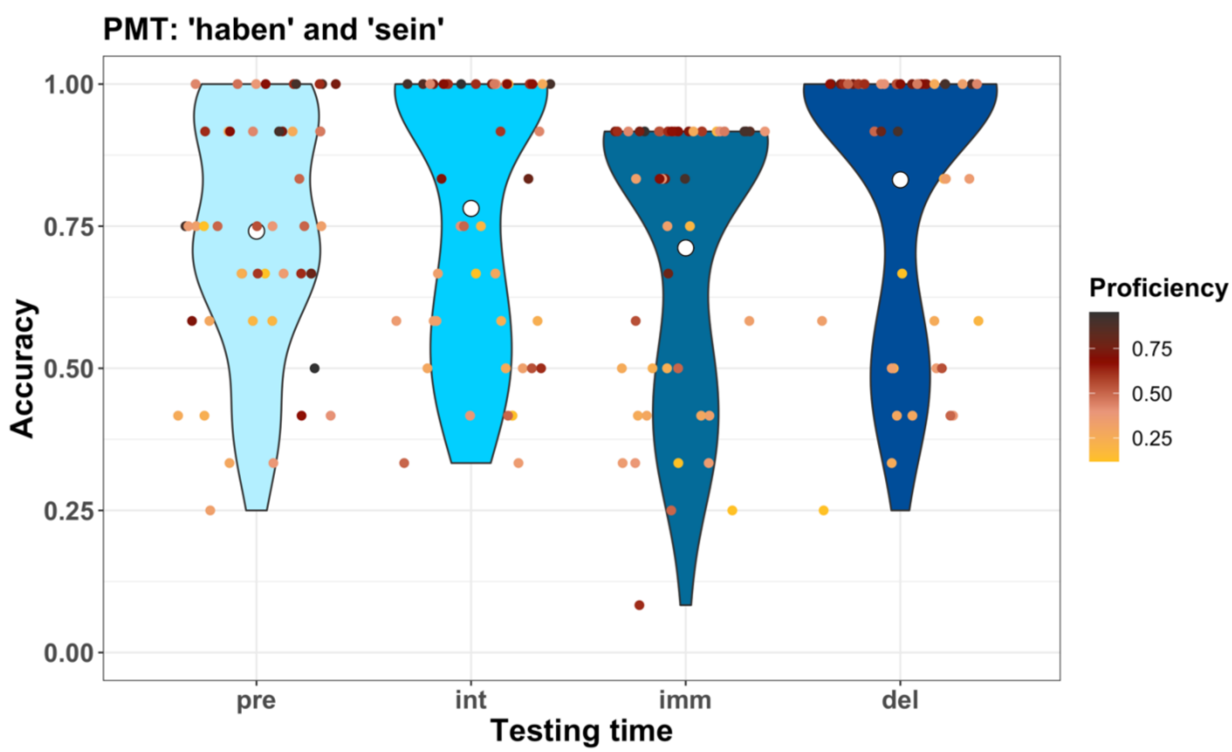
3.3.3.2 Results

3.3.3.2.1 Picture Matching Task

As proficiency was coded as a continuous variable, descriptive results are only graphically represented in Figure 3.3.12, which corresponds to Figure 3.3.1 in Section 3.3.2.2.1. Figure 3.3.12 shows increases in accuracy from pretest to interim test, with the largest density of data points moving towards ceiling, as well as from immediate to delayed posttest, where the distribution of data points seems to stay identical but collectively move upwards. However, it also shows a decrease in accuracy from interim test to immediate posttest, with the distribution of data points becoming wider than it was at interim test—ranging almost across the whole range of accuracy.

Figure 3.3.12

Accuracy for the Picture Matching Task by testing time



Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

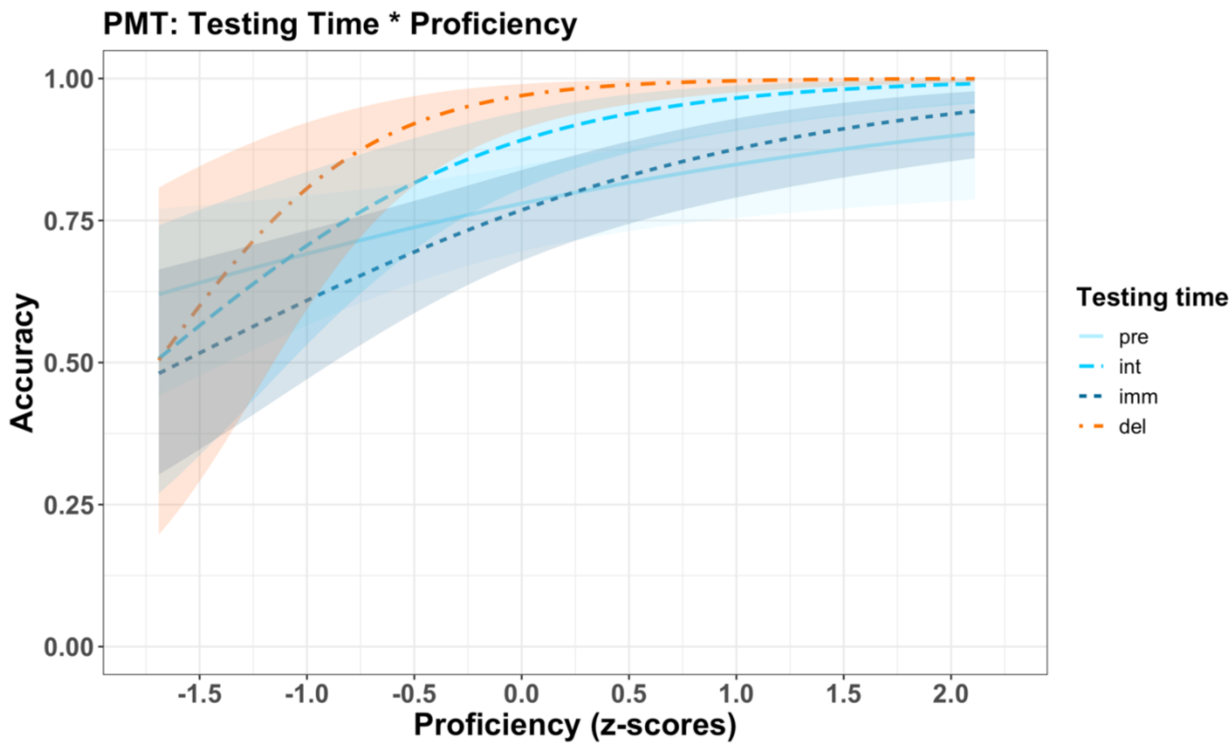
[6] In the previous section, proficiency was included as a covariate to control for between-group differences that emerged at pretest as a function of proficiency. Including group as a covariate in the analyses in the current section does not make sense, however, as I was interested exactly in the role that proficiency plays in learning, irrespective of feedback type. Furthermore, including a three-way interaction between testing time, proficiency, and group was not possible due to issues of statistical power.

The final generalized linear mixed effects model included random intercepts for subjects and items and a random slope for testing time by subject. The model (see Table 3.3.11) revealed significant main effects of testing time, such that accuracy across all participants independent of proficiency increased from pretest to interim test, decreased from interim test to immediate posttest, and increased again from immediate to delayed posttest. This mirrors the effects of testing time found in the primary analyses described in Section 3.3.2.2.1, where the decrease from interim to immediate posttest was largely driven by the semantic-focused feedback first group. Furthermore, like in the primary analyses described in Section 3.3.2.2.1, there was a main effect of proficiency, revealing that more proficient participants were more accurate than less proficient participants across all testing times. Critically, these main effects of testing time and proficiency were qualified by interactions between the two: There was a significant interaction between proficiency and the change in accuracy from pretest to interim test. There was also a marginally significant interaction between proficiency and the change in accuracy from interim test to immediate posttest, with the effect being in the opposite direction to the previous one. However, the confidence interval for this interaction crosses zero, meaning that caution in interpreting this effect is warranted. Finally, there was a significant interaction between proficiency and the change in accuracy from immediate to delayed posttest.

Interactions with continuous variables, such as proficiency, are difficult to interpret. For instance, it is unclear from the significant interaction terms in the model which participants — those who are less proficient or those who are more proficient — increased or decreased most in accuracy at any testing time. The interaction terms between proficiency and testing time were therefore visualized using the interact_plot function of the i*nteractions* package version 1.1.5 (Long, 2021) in Figure 3.3.13, with proficiency as the predictor variable and testing time as the moderator variable. For any given *z*-score of proficiency, model estimates of accuracy are plotted on the y-axis as accuracy between 0 and 1, with each line representing a different testing time. Where the difference between lines is largest is where increases/decreases are strongest. In viewing this figure, the interaction terms described above become clearer. The significant interaction between proficiency and the change in accuracy from pretest to interim test corresponds to a larger increase in accuracy from pretest to interim test for more proficient learners, as the distance between the solid light blue line and the dotted light blue line in Figure 3.3.13 becomes larger as a function of proficiency. The marginally significant interaction between proficiency and the change in accuracy from interim test to immediate posttest shows that more proficient participants decreased more in accuracy from interim test to immediate posttest than less proficient participants, as the dotted dark blue line is lower than the dotted light blue line in Figure 3.3.13, and the distance between the two is largest in the middle range of proficiency. Finally, the significant interaction between the change in accuracy from immediate to delayed posttest becomes visible from the dotted dark blue and orange lines in Figure 3.3.13, again revealing a larger increase in accuracy from immediate to delayed posttest for learners across the middle of the proficiency spectrum. Overall, Figure 3.3.13 shows that participants roughly around proficiency *z*-scores ranging from -0.5 to 1.0 exhibited the largest changes from pretest to any following tests.

Figure 3.3.13

Visualization of interaction terms between proficiency and testing time on the Picture Matching Task



Note. Error bands represent 95% confidence intervals.

Table 3.3.11

Model output for auxiliary accuracy on the Picture Matching Task as a function of proficiency

						95% Confidence Interval	
Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		Lower Bound	Upper Bound
(Intercept)	2.011	0.263	7.633	2.30E-14	***	1.495	2.528
Pre-Int	0.840	0.337	2.492	0.013	*	0.179	1.500
Int–Imm	-0.906	0.287	-3.157	0.002	**	-1.468	-0.343
Imm–Del	2.275	0.526	4.324	1.53E-05	***	1.244	3.306
Proficiency	1.122	0.226	4.973	6.59E-07	***	0.680	1.565
Pre–Int:Prof	0.770	0.311	2.474	0.013	*	0.160	1.380
Int–Imm:Prof	-0.474	0.252	-1.881	0.060	†	-0.968	0.020
Imm–Del:Prof	1.291	0.423	3.056	0.002	**	0.463	2.120

Random Effects						
Groups	Names	Variance	Std. Deviation	Correlation		
Subject	Intercept	1.6144	1.2706			
	Pre-Int	1.8852	1.3730	0.58		
	Int-Imm	0.4581	0.6768	-0.64	-0.57	
	Imm-Del	1.4533	1.2055	0.86	0.35	-0.45
Item	Intercept	0.1513	0.3890			

Model fit		
	Marginal	Conditional
R^2	0.304	0.594

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

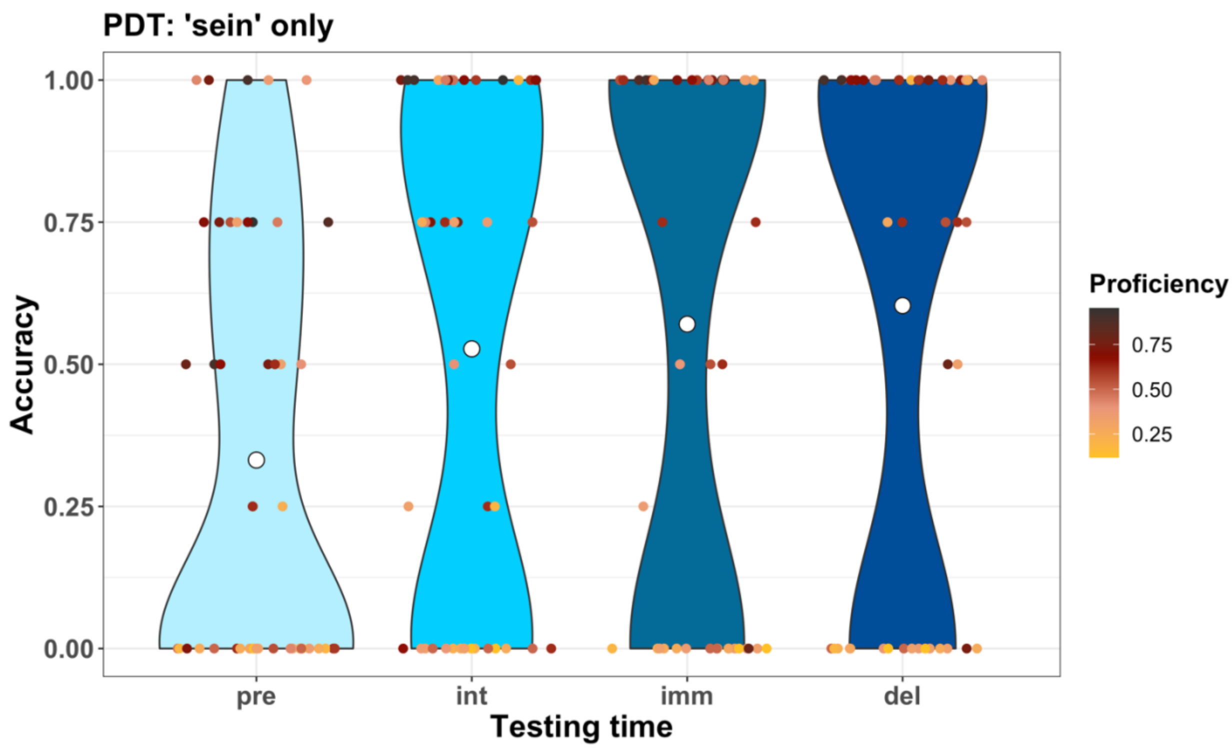
Model equation: Accuracy ~ Proficiency*Group + (1+TestingTime|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

3.3.3.2.2 Picture Description Task (Auxiliaries)

Again, descriptive results are graphically represented in Figure 3.3.14, which corresponds to Figure 3.3.4 in Section 3.3.2.2.2. Figure 3.3.14 shows a steady increase from each testing time to the next. While the density of data points at pretest is highest at floor, the largest density consistently moves upward towards ceiling, such that the distribution of data points becomes almost bimodal.

Figure 3.3.14

Accuracy of sein *auxiliaries for the Picture Description Task by testing time*



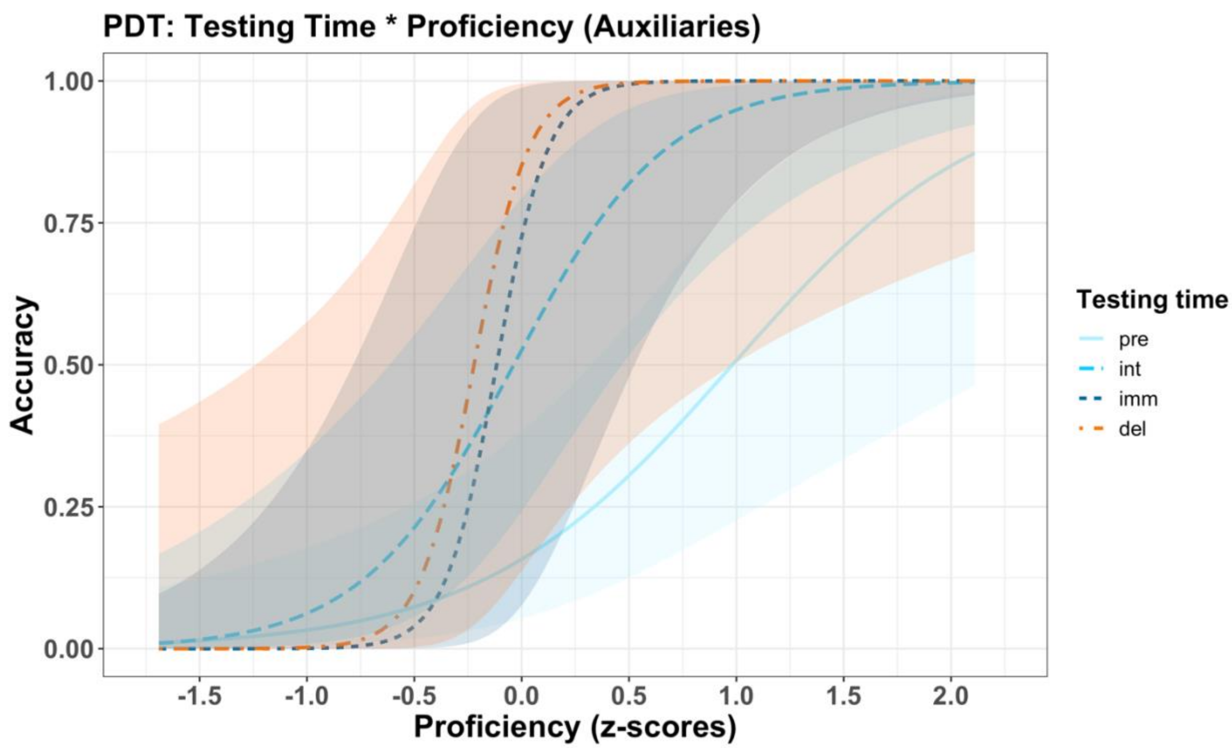
Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

The final generalized linear mixed effects model included random intercepts for subjects and items and a random slope for testing time by subject. The model (see Table 3.3.12) revealed one significant main effect of testing time, such that accuracy across all participants independent of proficiency increased from pretest to interim test. However, there were no main effects for the change in accuracy from interim test to immediate posttest or from immediate to delayed posttest. This mirrors the effects of testing time found in the primary analyses described in Section 3.3.2.2.2. Furthermore, like in the primary analyses described in Section 3.3.2.2.2, there was a main effect of proficiency, revealing that more proficient participants were more accurate than less proficient participants across all testing times. In terms of interactions between the two predictor variables, there was only one marginally significant interaction between proficiency and the change in accuracy from interim test to immediate test. However, the confidence interval for this interaction cross zero and exhibit a very large range, meaning that the effect is most likely spurious. The interactions between proficiency and the change from pretest to interim test and between proficiency and the change in accuracy from immediate to delayed posttest were not significant.

Despite only one marginally significant interaction between testing time and proficiency — with confidence intervals crossing zero — the interaction terms between proficiency and testing time were again visualized in Figure 3.3.15. The marginally significant interaction between proficiency and the change in accuracy from interim test to immediate posttest can be interpreted by looking at the light blue and dark blue dotted lines, which cross each other at a proficiency score of about -0.2. This means that more proficient participants increased in accuracy from interim test to immediate posttest, while less proficient participants actually decreased. Overall, Figure 3.3.15 below then shows that participants with proficiency *z*-scores of -0.2 or higher showed increases from pretest to all other tests. However, confidence intervals for each testing time span a wide range of auxiliary accuracy — together with the large ranges and zero-crossing confidence intervals for the interactions in the model, this suggests that there is a lot of variation between participants, even at similar levels of proficiency.

Figure 3.3.15

Visualization of interaction terms between proficiency and testing time for auxiliaries on the Picture Description Task



Note. Error bands represent 95% confidence intervals.

Table 3.3.12

Model output for auxiliary accuracy on the Picture Description Task as a function of proficiency

						95% Confidence Interval	
Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		Lower Bound	Upper Bound
(Intercept)	0.286	0.971	0.295	0.768		-1.617	2.189
Pre-Int	1.780	0.683	2.607	0.009	**	0.442	3.119
Int–Imm	0.868	1.461	0.594	0.553		-1.996	3.731
Imm–Del	0.775	1.680	0.461	0.645		-2.517	4.067
Proficiency	5.182	1.450	3.573	3.53E-04	***	2.339	8.025
Pre–Int:Prof	1.110	0.872	1.273	0.203		-0.599	2.819
Int–Imm:Prof	5.518	3.134	1.761	0.078	†	-0.625	11.661
Imm–Del:Prof	-0.469	4.656	-0.101	0.920		-9.595	8.658

Random Effects

Groups	Names	Variance	Std. Deviation	Correlation		
Subject	Intercept	29.0560	5.3904			
	Pre-Int	7.3461	2.7104	0.53		
	Int-Imm	42.7262	6.5365	0.91	0.51	
	Imm-Del	17.5864	4.1936	-0.01	-0.23	-0.22
Item	Intercept	0.2664	0.5162			

Model fit

	Marginal	Conditional
R^2	0.435	0.962

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

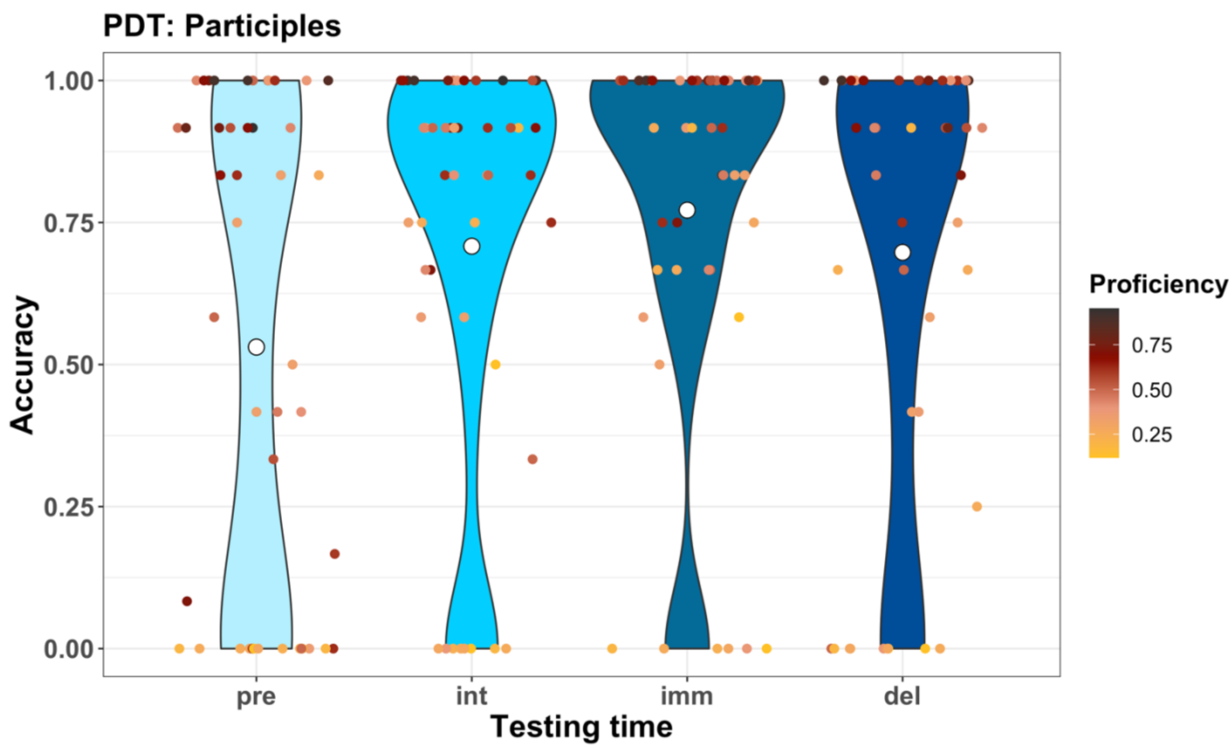
Model equation: Accuracy ~ Proficiency*Group + (1+TestingTime|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

3.3.3.2.3 Picture Description Task (Participles)

Descriptive results are once again graphically represented in Figure 3.3.16, which corresponds to Figure 3.3.7 in Section 3.3.2.2.3. Figure 3.3.16 shows at-chance performance at pretest, with the density of data points being largest right below ceiling as well as at floor. There is a steady increase from pretest to interim test, with the largest density of data points being in the highest quarter of the accuracy range, as well as from interim test to immediate posttest, with the largest density now located at ceiling. From immediate to delayed posttest, however, there is a decrease back to the levels of the interim test, with data points being distributed more variably across the whole range of accuracy.

Figure 3.3.16

Accuracy of participles for the Picture Description Task by testing time



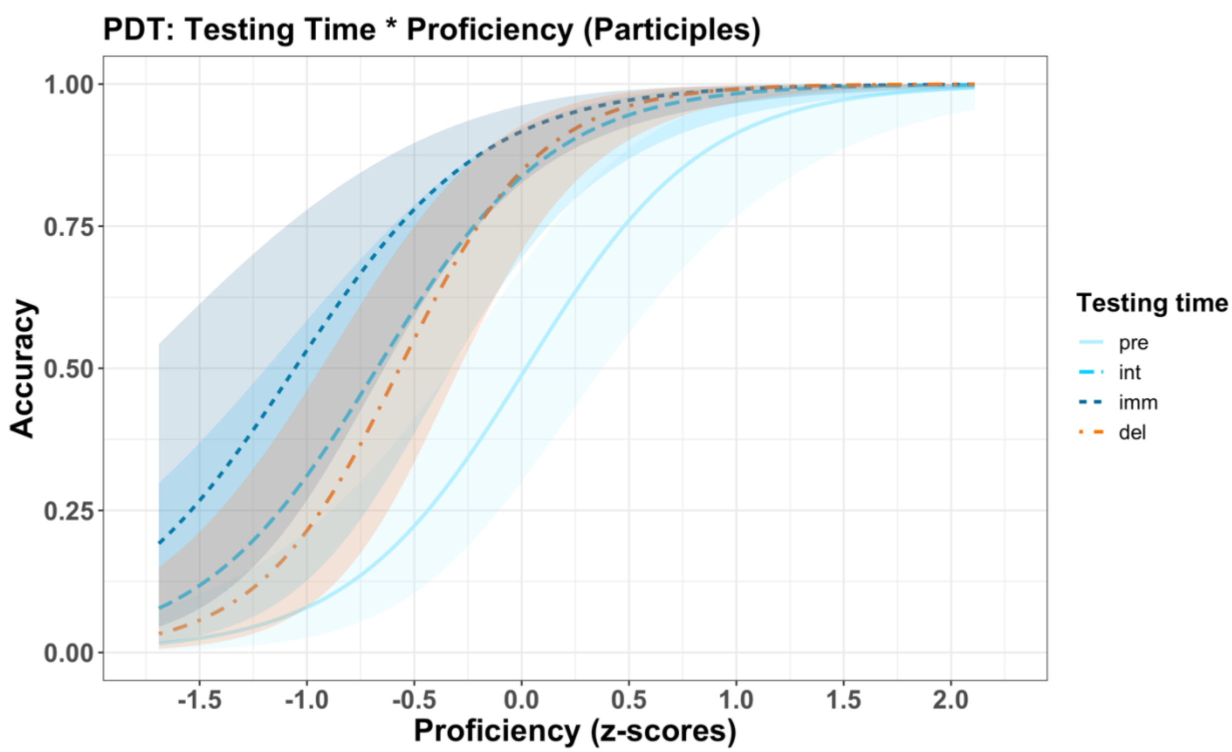
Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

The final generalized linear mixed effects model included random intercepts for subjects and items, but no random slopes. The model (Table 3.3.13) revealed significant main effects of testing time, such that accuracy across all participants independent of proficiency increased from pretest to interim test and from interim to immediate posttest, but decreased from immediate to delayed posttest, mirroring those effects found in the primary analyses described in Section 3.3.2.2.3. Similarly, there was a main effect of proficiency just like in the primary analyses described in Section 3.3.2.2.3, revealing that more proficient participants were more accurate than less proficient participants across all testing times. In terms of interactions between the two predictor variables, there was a significant interaction between proficiency and the change in accuracy from immediate to delayed posttest. The interactions between proficiency and the change from pretest to interim test and between proficiency and the change in accuracy from interim test to immediate posttest were not significant.

Again, the interaction terms between proficiency and testing time were visualized in Figure 3.3.17. The graph shows increases of similar magnitude and with large degrees of variation from pretest to interim test and immediate posttest for almost the whole proficiency band, as shown by the distance between the solid light blue, dotted light blue, and dotted dark blue lines. Figure 3.3.17 also allows the interpretation of the significant interaction between the change in accuracy from immediate to delayed posttest, as accuracy on the delayed posttest decreased from the immediate posttest mostly for participants with proficiency *z*-scores between -1.5 and 0.5, but not above 0.5. Even at proficiency scores of around -0.25, the decrease from immediate to delayed posttest starts to become much smaller, as the distance between the dark blue and the orange dotted lines becomes smaller.

Figure 3.3.17

Visualization of interaction terms between proficiency and testing time for participles on the Picture Description Task



Note. Error bands represent 95% confidence intervals.

Table 3.3.13

Model output for participle accuracy on the Picture Description Task as a function of proficiency

						95% Confidence Interval	
Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		Lower Bound	Upper Bound
(Intercept)	1.426	0.395	3.613	3.02E-04	***	0.653	2.200
Pre-Int	1.686	0.205	8.210	2.21E-16	***	1.284	2.089
Int–Imm	0.762	0.242	3.141	0.002	**	0.286	1.237
Imm–Del	-0.687	0.253	-2.718	0.007	**	-1.182	-0.192
Proficiency	2.531	0.415	6.091	1.12E-09	***	1.716	3.345
Pre–Int:Prof	0.030	0.250	0.119	0.906		-0.461	0.520
Int–Imm:Prof	-0.161	0.274	-0.587	0.557		-0.699	0.377
Imm–Del:Prof	0.741	0.294	2.519	0.012	*	0.164	1.318

Random Effects				
Groups	Names	Variance	Std. Deviation	Correlation
Subject	Intercept	5.736	2.3949	
Item	Intercept	0.115	0.3392	

Model fit		
	Marginal	Conditional
R^2	0.444	0.800

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

Model equation: Accuracy ~ Proficiency*Group + (1|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

3.3.3.3 Summary: Learning and the Role of Proficiency

The results of the picture matching task indicate that proficiency plays a role in the effectiveness of prediction-/error-based training, at least for comprehending the role of the auxiliary in *Perfekt s*entences. In particular, more proficient learners were overall more accurate than less proficient learners. This effect mirrors results described in the analysis of group effects in Section 3.3.2 and is unsurprising, as more proficient learners not only have prior knowledge of but also experience with the target structure and would therefore be expected to exhibit higher accuracy. The visualization of the interaction terms between proficiency and testing time in Figure 3.3.13 suggests that participants with proficiency *z*-scores ranging from about -0.75 to 0.75 exhibited the greatest increase across the three experimental sessions. Interestingly, while Figure 3.3.13 also shows a strong decrease from interim test to immediate test for the same proficiency range between -0.75 and 0.75, the confidence intervals produced by the model, which cross zero, suggest that the effect is spurious. The visualization also suggests that participants at the higher end of the proficiency scale — with *z*-scores of 1.5 and higher — were almost at ceiling from the very beginning, whereas there did not seem to be much learning or any learning for participants at the lower end of the proficiency scale—with *z*-scores of -1.0 and lower.

The results for auxiliary production in the picture description task differ from those of auxiliary comprehension in the picture matching task. While more proficient participants were, again, more accurate overall, the only other significant effect was a main effect of testing time, with a significant increase in accuracy from pretest to interim test. Figure 3.3.15 suggests that participants with a proficiency *z*-score higher than -0.5 increased from pretest to delayed posttest, while those with a proficiency *z*-score lower than -0.5 actually decreased. The results of the statistical analyses, however, do not confirm this trend. In fact, the results differ from the auxiliary accuracy results of the picture matching task, in that there were no significant interactions between proficiency and testing time, the exception being a marginally significant interaction with the change from interim test to immediate posttest, but with confidence intervals crossing zero, indicating a spurious effect.

Finally, for participle accuracy in the picture description task, results again indicate that more proficient learners were more accurate at all testing times, mirroring the results of the analyses conducted in Section 3.3.2. The significant effects of testing time suggest consistent increases across the whole proficiency scale up until immediately after training, i.e., at immediate posttest, followed by decreases from immediate to delayed posttest. This latter effect is qualified by an interaction, with Figure 3.3.17 showing that accuracy on the delayed posttest decreased from the immediate posttest mostly for participants with proficiency *z*-scores between -1.5 and 0.5, but not above 0.5.

These findings and their implications for the role of prediction and proficiency in L2 learning are discussed in Section 3.4 Discussion.

3.3.4 Exploratory Analysis: The Role of Prior Knowledge

Another, equally important question that cannot be answered by investigating the role of proficiency scores alone is whether the amount of prior knowledge of the target structure plays a role for error-based learning mechanisms. As outlined in the literature review, it is possible to differentiate between two different processes for creating expectations: 1) guessing, which is the creation of (weak) expectations based on no existing relevant prior knowledge, and 2) predicting, which in turn is the creation of expectations based on some degree of relevant prior knowledge (see Brod, 2021).

3.3.4.1 Analyses

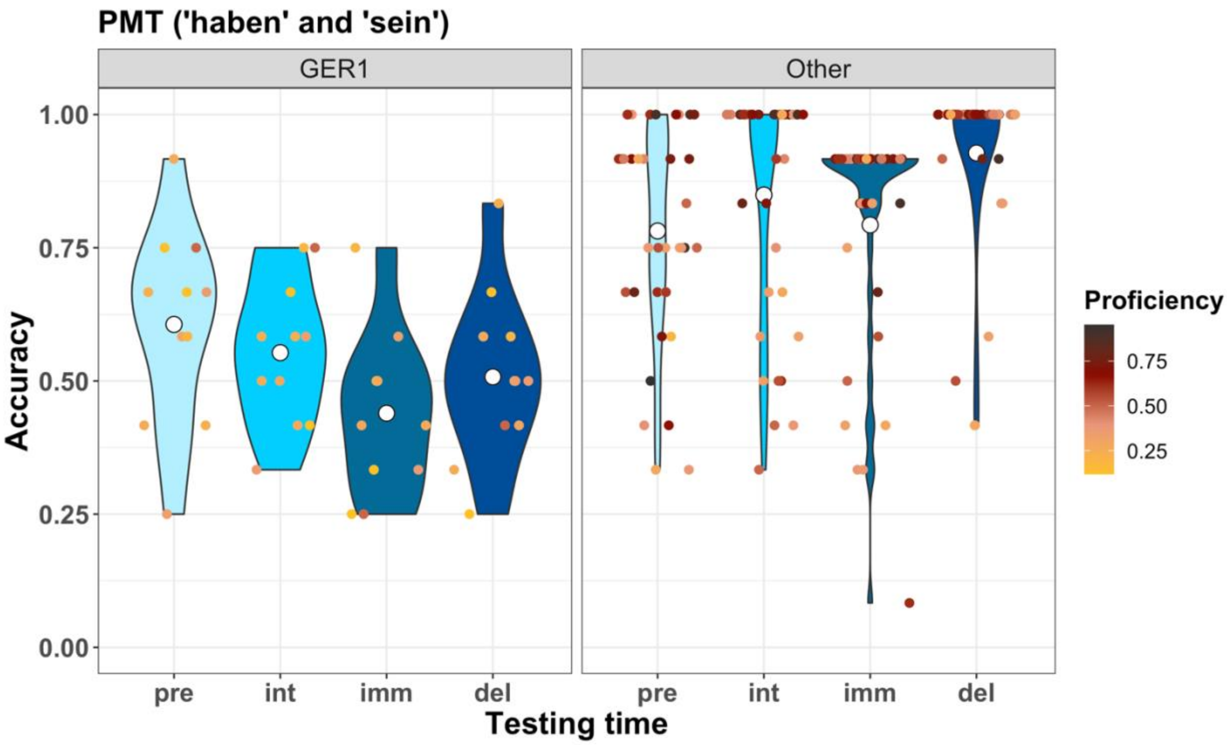
In this experiment, I approximated the distinction between guessing and predicting by looking at participants' course level: Participants in GER1 (N = 11) had not been introduced to *Perfekt* and its auxiliary alternation at the time they participated in the experiment, meaning they had no formal prior knowledge of the primary target feature. Participants in course levels above GER1 (N = 35), on the other hand, had already been formally introduced to the target feature and therefore had prior experience with it at the time they participated in the experiment. Participants were therefore split into two groups according to course level in order to investigate the role that prior knowledge — as operationalized in this experiment — plays for error-based learning.

3.3.4.2 Results

Figures 3.3.18 and 3.3.19 show auxiliary accuracy on the picture matching (Figure 3.3.18) and picture description task (Figure 3.3.19) across testing times and split by level of prior knowledge. For the picture matching task, the results for participants with prior knowledge ("Other") mirrored those found across all participants (see Figure 3.3.1). The results for participants without prior knowledge ("GER1), on the other hand, followed a different pattern: Rather than increasing in accuracy from pretest to any of the following tests, auxiliary accuracy decreased up to the immediate posttest before slightly increasing again at delayed posttest. For the picture description task, the results for participants with prior knowledge ("Other") again mirrored those found across all participants (see Figure 3.3.4). Participants without prior knowledge ("GER1"), on the other hand, remained at floor across all testing times. The results for both the picture matching and picture description task are sufficiently different between groups that no statistical analyses were conducted.

Figure 3.3.18

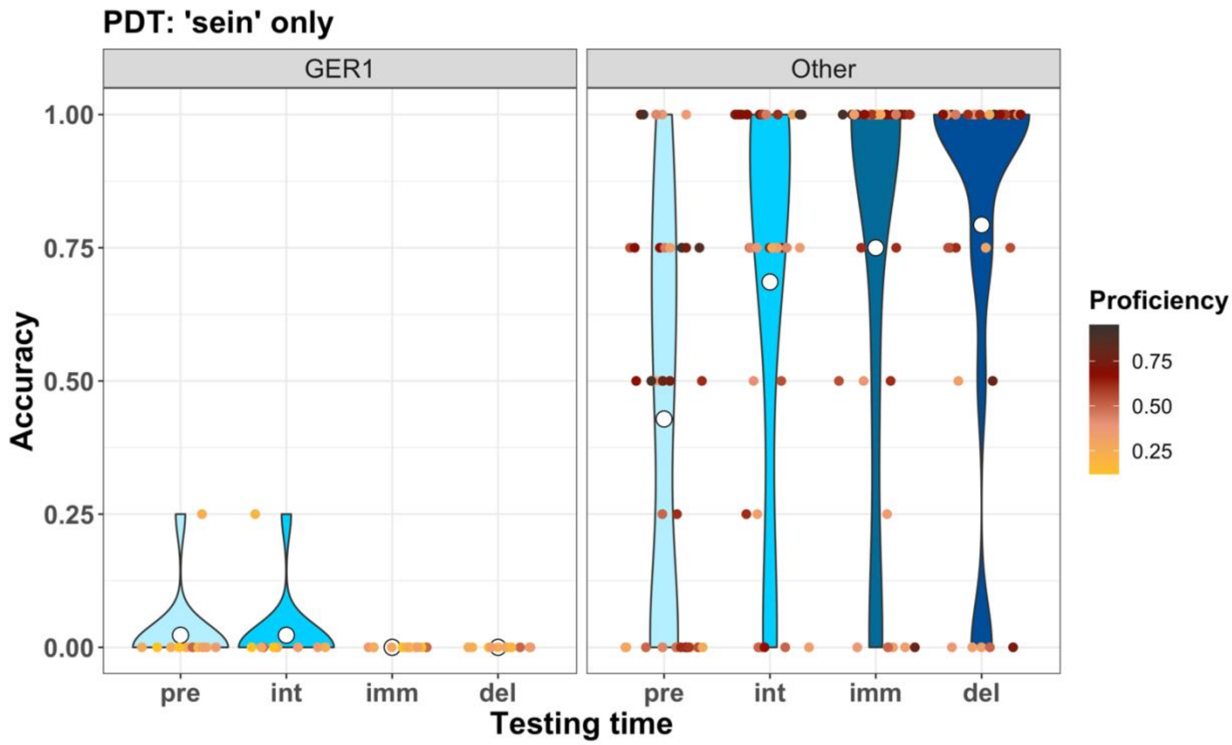
Auxiliary accuracy for the Picture Matching Task by course level and testing time



Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Figure 3.3.19

*Auxiliary accuracy (*sein *only) for the Picture Description Task by course level and testing time*



Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

3.3.4.3 Summary: The Role of Prior Knowledge

The descriptive results and the graphic representations of both the picture matching and the picture description task split by course level—GER1 versus Other—indicate that prior knowledge plays a role in the effectiveness of error-based learning in conjunction with implicit feedback. In the picture matching task, participants in GER1 hovered around chance after a slight decrease from pretest, while participants in higher course levels mirrored the results across all participants more closely. In the picture description task, participants in GER1 remained at floor across all testing times, while participants in higher course levels again mirrored the results across all participants. These findings and their implications for the role of prediction-/error-based learning in L2 learning are discussed in Section 3.4 Discussion.

3.3.5 Exploratory Analysis: The Role of Awareness

Although some debate still exists as to whether awareness of a target structure is a necessary prerequisite for learning (Robinson, 2005; Schmidt, 1990, 1995, 2001; Tomlin & Villa, 1994), it is undebated that it can at least facilitate and potentially also accelerate the learning of a target structure (e.g., Andringa, 2020; Curcic et al., 2018; Leow, 1997, 2001). While the role of awareness in the learning process was not of primary interest in the experiment, the debriefing questionnaires nevertheless collected information about participants' awareness of the auxiliary alternation during the training tasks and during testing. Data from these two debriefing questionnaires were compiled and scored to investigate if participants were aware or became aware of the auxiliary alternation and its underlying semantic aspect during training. Participants were classified as unaware (1), emergent aware (2), or pattern aware (3) based on the following criteria:

1. Unaware: Participants did not mention the auxiliary alternation as something that helped them make decisions. They also answered “no” to the question whether they noticed that some sentences had a form of *haben* while others had a form of *sein*.
2. Emerging aware: Participants mentioned the auxiliary alternation, but either didn’t know the underlying regularity behind the alternation or didn’t explicitly mention it.
3. Pattern aware: Participants mentioned the auxiliary alternation and mentioned the underlying regularity behind the alternation.

Two research assistants familiar with German and the general goals of the experiment coded the debriefing data after being briefed by myself. Interrater reliability after the first round of coding was 60.4%. Discrepancies arose mainly because one of the two research assistants consistently applied the coding scheme more conservatively than the other research assistant. Both research assistants and I therefore met a second time to discuss and resolve any remaining coding discrepancies.

3.3.5.1 Analyses

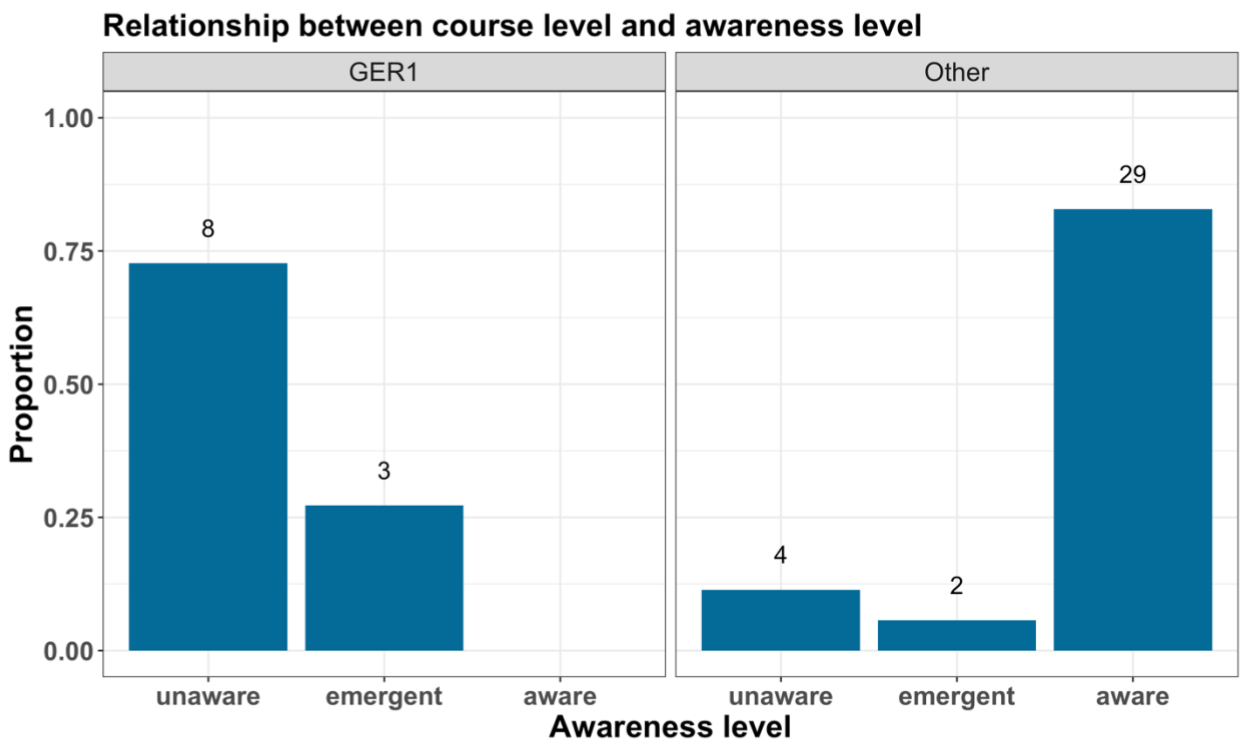
An exploratory analysis was conducted using the coded data from the debriefing questionnaires. Specifically, participants’ awareness status was analyzed in its relationship to participants’ course level using correlation analysis. Spearman correlations were used due to non-normal distribution of the dependent variable. A second set of analyses explored whether awareness status was related to auxiliary accuracy on the picture matching and picture description tasks. Since the debriefing questionnaire was only administered at the end of the second session after the immediate posttest and the third session after the delayed posttest and therefore would not have captured any group-differences that may have emerged as a function of feedback ordering, awareness status and its relationship with auxiliary accuracy was only analyzed at pretest, immediate, and delayed posttest. Furthermore, as the debriefing questionnaire was not fine-grained enough to distinguish between potentially very subtle differences in awareness and their emergence, the relationship between awareness and accuracy was investigated by looking at whether participants performed below, at, or above chance on a given test at any given testing time. To determine performance above chance, at chance, and below chance on a given test, I used a binomial probability calculator, with the probability of success on a trial set to 0.5 and the number of trials set to 12 and 4 for the picture matching and picture description task respectively. Chance was defined as the cumulative probability that a given number of successful trials had an alpha-level smaller or larger than 0.05. On the picture matching task, participants were consequently classified as performing above chance if their accuracy was above 0.66, at chance if it was between 0.33 and 0.66, and below chance if their accuracy was below 0.33. On the picture description task, participants were classified as performing above chance if their accuracy was above 0.75, at chance if it was between 0.25 and 0.75, and below chance if their accuracy was below 0.25. To visualize this in a simple manner, graphs include a gray area, representing chance. Chi-Square Tests of Independence were then performed separately for each task and testing time to assess the relationship between awareness status and performance, using the *chisq.test* function from the *stats* package version 4.1.0 (R Core Team, 2021) in R. For these tests, participants below and at chance were collapsed into a single category.

3.3.5.2 Results

Figure 3.3.20 and shows the proportion and absolute numbers of participants at each awareness level within the two categories of course level, i.e., "GER1" or "Other". As can be seen in Figure 3.3.20, the large majority of participants in GER1 were unaware of the auxiliary alternation and only a few were classified as emergent aware. The large majority of participants in course levels higher than GER1, however, were classified as aware. Unsurprisingly, the results of the Spearman correlation analyses confirm that there is a clear relationship between course level and awareness status, as there was a moderately strong positive correlation between the two variables, $r(44)=0.73$, $p < .001$.

Figure 3.3.20

Relationship between course level and awareness status



3.3.5.2.1 Picture Matching Task

For the second set of analyses, descriptive results for the picture matching task can be found in Table 3.3.14 and Figure 3.3.21.

Table 3.3.14

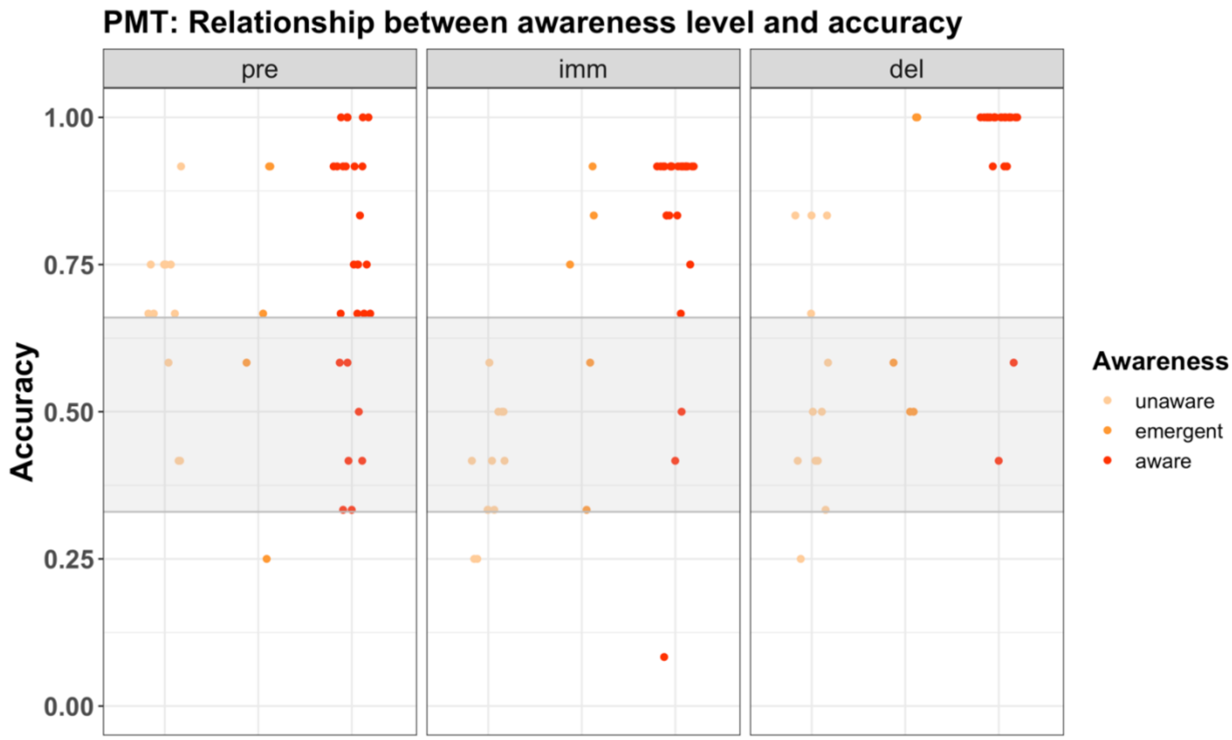
Counts of participants below, at, or above chance on the Picture Matching Task by awareness level and testing time

Pretest	Unaware	Emergent	Aware
Below chance	0	1	0
At chance	3	1	7
Above chance	9	3	22
Immediate posttest			
Below chance	2	0	1
At chance	10	2	2
Above chance	0	3	26
Delayed posttest			
Below chance	1	0	0
At chance	7	3	2
Above chance	4	2	27

As can be seen in Figure 3.3.21 — and keeping in mind the different numbers of participants at each level of awareness — most participants at pretest, regardless of awareness level, were performing above chance. The results of the Chi-Square Test indicate that there were no differences in chance level based on awareness level at pretest, $X^2(2, N = 46) = 0.57$, $p = .753$. At immediate and delayed posttest, the distribution by awareness level looks slightly different. At immediate posttest, participants coded as unaware were closer to or below chance, whereas participants coded as aware were mostly above chance, with only three participants at or below chance. Participants coded as emergent followed a similar pattern as those coded as aware. The results of the Chi-Square Test confirm that there were differences in chance level based on awareness level at immediate posttest, $X^2(2, N = 46) = 29.31$, $p < .001$. At delayed posttest, participants coded as unaware were almost evenly distributed above and below chance, whereas participants coded as aware were mostly performing above chance and participants coded as emergent were mostly at chance. Again, results of a Chi-Square Test confirm that there were differences in chance level based on awareness level at delayed posttest, $X^2(2, N = 46) = 17.74$, $p < .001$.

Figure 3.3.21

Auxiliary accuracy on the Picture Matching Task by testing time and awareness level



Note. The gray area represents chance level. Dots represent individual participants.

3.3.5.2.2 Picture Description Task

For the second set of analyses, descriptive results for the picture description task can be found in Table 3.3.15 and Figure 3.3.22.

Table 3.3.15

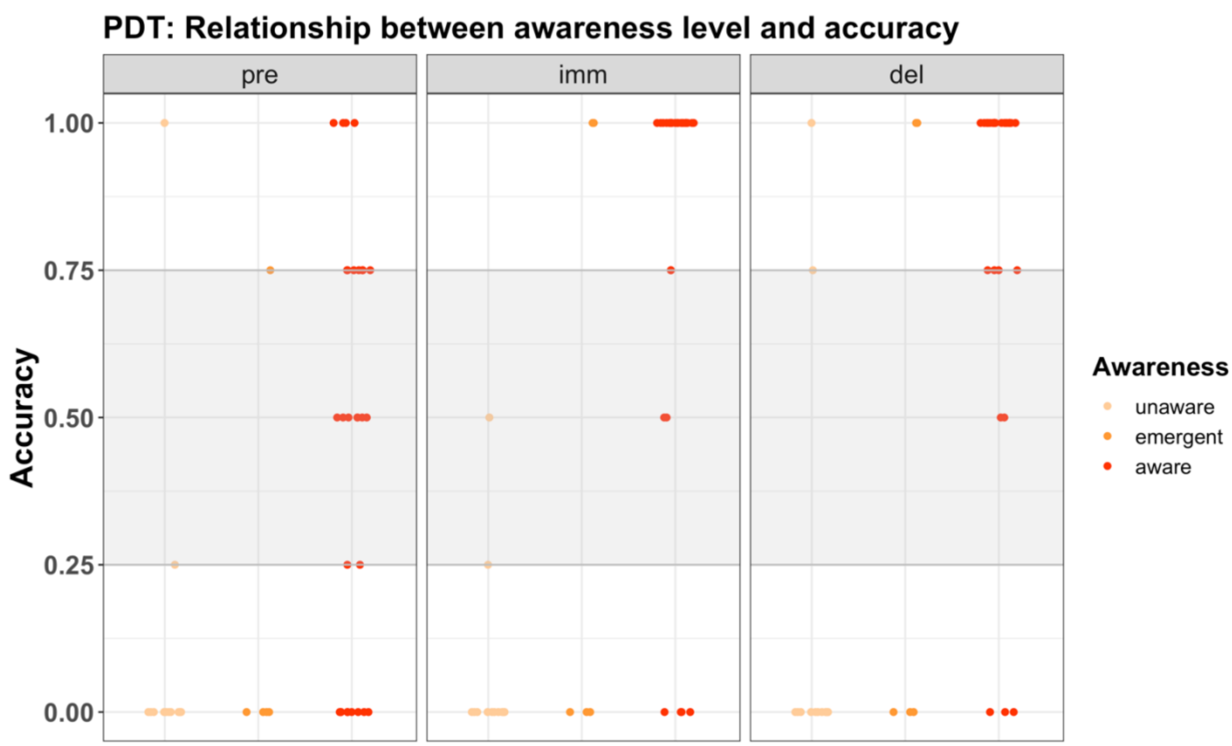
Counts of participants below, at, or above chance on the Picture Description Task by awareness level

Pretest	Unaware	Emergent	Aware
Below chance	11	4	11
At chance	0	0	7
Above chance	1	1	11
Immediate posttest			
Below chance	11	3	4
At chance	1	0	2
Above chance	0	2	23
Delayed posttest			
Below chance	10	3	3
At chance	0	0	2
Above chance	2	2	24

As can be seen in Figure 3.3.22 — and again keeping in mind the different numbers of participants at each level of awareness — different patterns emerged for the different awareness levels at each testing time. At pretest, participants coded as unaware and emergent were mostly performing below chance — at floor actually — with the exception of one participant in each group who was above chance. The picture is more varied for participants coded as aware with performance below, at, or above chance being approximately evenly distributed. The results of the Chi-Square Test indicate that there were no differences in chance level based on awareness level at pretest, $X^2(2, N = 46) = 3.86$, $p = .145$. At immediate and delayed posttest, the picture looks almost unchanged for participants coded as unaware and emergent, with one or two participants moving up above or down below chance. For participants coded as aware, the majority was above chance at both immediate and delayed posttest, although a few participants remained below chance as well. The results of the Chi-Square Test confirm that there were differences in chance level based on awareness level at immediate posttest, $X^2(2, N = 46) = 21.98$, $p < .001$, as well as at delayed posttest, $X^2(2, N = 46) = 16.59$, $p < .001$.

Figure 3.3.22

*Auxiliary accuracy (*sein *only) on the Picture Description Task by testing time and awareness level*



Note. The grey area represents chance level. Dots represent individual participants.

3.3.5.3 Summary: The Role of Awareness

The results of the first analysis, investigating course level and awareness level, show a clear relationship between the two. Of those participants in GER1, 72.7% were unaware of the auxiliary alternation and 27.3% showed emergent awareness. On the other hand, of those participants in course levels higher than GER1, only 11.4% were unaware of the alternation, 5.7% showed emergent awareness, and 82.9% were fully aware of the alternation. These results indicate that course level, which is also an approximation to prior knowledge, is indicative of awareness of the targeted auxiliary alternation.

The results of the second set of analyses, investigating the relationship between awareness level and accuracy at pretest, immediate, and delayed posttest, also reveal a relationship between the two. While at pretest the results of the Chi-Square tests indicate that participants who were coded as unaware or emergent and those who were coded as aware showed similar accuracy and had similar probabilities of performing above or below chance, the relationship between awareness level and accuracy emerged at the immediate and delayed posttests, i.e., after training. Participants who were aware of the auxiliary alternation were more likely to perform above chance at immediate and delayed posttest on both the picture matching and picture description task than participants coded as unaware or emergent aware (see Figures 3.3.21 and 3.3.22).

These findings and their implications for the role of prediction-/error-based learning in L2 learning are discussed in Section 3.4 Discussion.

3.4 Discussion

The first experiment in this dissertation investigated the role of two implicit feedback types and their relative ordering, as well as the role of proficiency for error-based learning mechanisms in L2 learning. In addition, the role of prior knowledge and awareness of the auxiliary alternation were investigated as well. Findings for each of these topics of interest and their relevance for understanding the applicability of error-based learning in contexts of L2 learning are discussed below.

3.4.1 The Role of Feedback Ordering in Error-Based L2 Learning

When considering the effects of feedback ordering — i.e., the first set of analyses described in Section 3.3.2 — the results of the picture matching and picture description tasks showed main effects of testing time for the auxiliary alternation, as well as for participle form in the picture description task, indicating that overall, participants improved over the course of the experiment. One surprising finding, however, was a significant decrease from interim to immediate posttest on the picture matching task for both groups. On the other hand, there were no main effects of group on either test, indicating that the different ways feedback was ordered within each group did not lead to any differences in overall learning, i.e., no group was consistently more accurate than the other. Nevertheless, different interaction patterns between testing time and feedback group emerged for auxiliary comprehension and production, as well as for participle production. On the picture matching task, there was only a marginal interaction with CIs crossing zero between the change from interim to immediate posttest and feedback group, indicating that the semantic-focused feedback first group may have decreased in accuracy more than the form-focused feedback first group. However, this may be a spurious effect. For auxiliary accuracy on the picture description task, there was a significant interaction between group and the change from pretest to interim test, indicating that the semantic-focused feedback first group initially increased more in accuracy than the form-focused feedback. For participles on the picture description task, there was a stronger increase from interim test to immediate posttest in the semantic-focused feedback first group compared to the form-focused feedback first group, but also a stronger decrease going from immediate to delayed posttest.

Taking a look at the significant interactions first, the interactions between testing time and feedback group found on the picture description task can be ascribed to the type and content of the implicit feedback provided in each block of training (see e.g., de Haas et al., 2020 and Hervais-

Adelman et al., 2008, for relevant studies in other learning domains). For instance, the semantic-focused feedback provided to the semantic-focused feedback first group in the first block of training focused their attention and learning on the auxiliary alternation, boosting their auxiliary accuracy from pretest to interim test more so than the feedback focusing on participle form provided to the form-focused feedback first group in the first block of training. On the other hand, however, there were also significant differences between groups at pretest on this task, with the semantic-focused feedback first group scoring lower than the form-focused feedback first group. While this difference turned out to be an effect of proficiency differences between groups — which is why proficiency was included as a covariate variable in all feedback ordering analyses — the stronger increase in the semantic-focused feedback first group going from pretest to interim test could also be an effect of starting out with lower accuracy. Still, this would indicate a boost provided to this group based on the content of the semantic-focused feedback, which focuses on the relationship between the auxiliary and the participle rather than on participle form. A similar explanation applies to participle accuracy: the form-focused feedback that participants in the semantic-focused feedback first group received in the second block of training increased their participle accuracy more than the semantic-focused feedback provided to the form-focused feedback first group, which is unsurprising. It is unclear, however, why there was no stronger increase in the form-focused feedback first group going from pretest to interim test than in the semantic-focused feedback first group, mirroring the effects found for auxiliary production. Similarly, it is unclear why there is a stronger decrease from immediate to delayed posttest in the semantic-focused feedback first group for participle accuracy, as both groups were exposed to both types of feedback by this point.

From an error-based learning perspective, the absence of main effects of group on any test indicate that the relative ordering of implicit feedback does not play a role in error-based learning when considering overall gains: No feedback group had an overall advantage over the other for learning the targeted auxiliary alternation and, of secondary interest, the participle form, especially by the delayed posttest. However, the findings regarding the interactions do point to the importance of matching the feedback provided to language learners to the content of the learning task and errors that occur (see for instance Hopp, 2020). If learners make an inaccurate prediction of possible language forms, the feedback that alerts participants to this error needs to match the locus of the error to allow for the computation of a prediction error. Similarly, if an accurate prediction is formed, the feedback needs to enable participants to confirm their prediction (see Pashler et al., 2005). Irrelevant information, such as information about form when the target is a semantically conditioned morphosyntactic alternation, diverts attention and cognitive resources away from the critical information encoded in the feedback. These effects become visible in the interactions on the picture description task, but also in the post-hoc training analysis, where descriptively, the group that received the most relevant feedback performed more accurately than the other across trials: The semantic-focused feedback first group showed higher accuracy for the auxiliary alternation in the first block of training, but this was reversed in the second block; the form-focused feedback first group showed higher accuracy for participle form in the first block of training, but this is again reversed in the second block.

Turning to the main effects of testing time for auxiliary accuracy, participants descriptively increased in their accuracy on the picture matching and picture description task going from pretest to delayed posttest. However, the significant decrease in accuracy from interim to immediate posttest on the picture matching task is an unexpected finding. One potential explanation for why the decrease occurs across <u>both</u> groups is an inability among participants to adapt to a changed task. In particular, the switch to a different feedback type in the second block may have confused participants about the focus of the task and which linguistic feature to pay attention to during training. This may have led to a shift in attention, caused by the different amount of information encoded and focused on by the feedback. The change in feedback type may have thus confused participants as to what the goals of the learning task were. Indeed, Sabah and colleagues (2019, 2021) found similar difficulties in adapting to changed task demands in studies investigating the effects of variable task content. The post-hoc analysis of the training data in the current study, however, shows that task difficulty cannot fully explain the decrease in accuracy among both feedback groups at immediate posttest. Rather than showing a decrease in the second block of training, participants in both feedback groups continually increased in accuracy as they progressed through the training, as evidenced by a main effect of trial number. It therefore remains unclear why participants' accuracy decreased on the immediate posttest but not in the second block of training. Relatedly, there was no decrease in auxiliary accuracy for either group on the picture description task going from interim to immediate posttest. However, the picture description task differed from the training task and was arguably more difficult. This suggests that compared to receptive (i.e., comprehension-based) knowledge, productive knowledge of the auxiliary alternation might be less susceptible to modulations in training task demands, and/or that such short-term difficulties may be specific to tasks that most closely resemble training.

Another important finding concerns the variable effects of short-term learning, with individuals within each group increasing as well as decreasing in accuracy during the span of the training task. For the picture matching task, however, the delayed posttest shows about 9% higher accuracy on average than the pretest, indicating long-term improvement from pretest to delayed posttest. These long-terms gains for auxiliary accuracy are even higher in the picture description task, averaging at 27%, despite considerable short-term variability across individuals in each feedback group. These findings mirror corrective feedback research showing short-term variability in learning as a result of implicit feedback, but more stable long-term performance (Li, 2010; see also Shin & Christianson, 2012). Similarly, studies investigating adaptation to changed task demands have invoked the notion of "desirable difficulty" (Schmidt & Bjork, 1992; Bjork & Bjork, 2011), arguing that "creating challenging practice conditions can facilitate deeper learning and retention yet withot [*sic*] any observable improvement during training" (Sabah et al., 2021). This underscores the importance of including delayed posttests in any investigations of error-based learning, as immediate learning effects may be attenuated or vary widely between individual participants, as seen in the current study.

The explanations and hypotheses regarding the role of feedback ordering and feedback type presented above remain necessarily speculative, as I am unaware of other research that has specifically investigated ordering effects of different types of implicit feedback. However, it is important to understand the potentially differential short-term effects of feedback orders — despite the fact that long-term effects of providing different feedback types may be comparable — as feedback provided in language classrooms often switches between alternate types and is not restricted to a single type of feedback at any given time or during any given task. It remains unclear where the current findings fall compared to other contexts of L2 learning where a manipulation of feedback ordering could be investigated. However, a replication study with a slightly different learner population is currently under way, and its results may shed more light on the robustness of the findings related to feedback ordering in the current experiment. Future studies in L2 learning, as well as in other domains in which prediction and feedback drive learning, could also investigate the processes involved in choosing between relevant and irrelevant information encoded in the feedback for the task at hand (cf. 'attentional learning'; for an overview, see Kruschke, 2011).

3.4.2 The Role of Proficiency in Error-Based L2 Learning

When considering the effects of proficiency as a full variable of interest in the absence of the feedback group variable — i.e., the second set of analyses described in Section 3.3.3 — the effects of testing time for the auxiliary alternation on the picture matching and the picture description task, as well as for the participle on the picture description task, mirrored the feedback group analyses discussed in the previous section. Furthermore, the results again showed that more proficient learners were more accurate in comprehending and producing auxiliaries and producing participles compared to less proficient learners, as evidenced by the main effects of proficiency on all tests and measures. However, different interaction patterns between testing time and proficiency emerged for auxiliary comprehension and production, as well as for participle production. First, significant interactions emerged from pretest to interim test and from immediate to delayed posttest on the picture matching task, indicating that learners around the center of the proficiency scale increased more than lower- or higher proficiency learners. In addition, a marginal negative interaction emerged from interim to immediate posttest on the same task, indicating that these same learners also decreased more than lower- or higher-proficiency learners. However, this might be a spurious effect, as the confidence interval for the effect crosses zero. Second, a marginally positive interaction emerged from interim test to immediate posttest test for auxiliary accuracy on the picture description task, this time indicating a larger increase from interim test to immediate posttest for learners around the center of the proficiency scale compared to lower-proficiency or higher-proficiency learners. Finally, a significant interaction emerged for participle accuracy on the picture description task, indicating that accuracy decreased for participants with low and intermediate proficiency going from immediate to delayed posttest, but not for those with high proficiency levels. Overall, accuracy on the picture description task remained lower than on the picture matching task across all testing rounds.

The finding that more proficient learners were generally more accurate at all testing times for auxiliaries and participles on both the comprehension and production tests is unsurprising and expected, as more proficient participants are not only familiar with the target structure, but likely have encountered and used it more often as well. This greater knowledge at outset then facilitated and further strengthened their learning of the auxiliary alternation and the participle forms over the course of the training task. Similarly, it is unsurprising that auxiliary accuracy in the production task lagged behind that of the comprehension task, not just in mean accuracy, but also in the amount of variation between learners of similarly high proficiency (see Figure 3.3.15). Research has repeatedly shown that accuracy in production tasks compared to comprehension tasks is generally lower, e.g., for L2 vocabulary (e.g., Laufer, 1998; Webb, 2008) and morphosyntax (e.g., Grüter et al., 2012; Keppenne et al., 2021).

As seen in the interactions between testing time and proficiency and their visualizations, the largest gains (and losses) across testing times emerged for learners around the center of the *z*-scored proficiency scale, roughly between *z*-scores of -0.5 and 1.0. Participants at the lower end of the scale either showed smaller gains or even decreases across testing times, whereas participants at the higher end of the scale were close to ceiling already at pretest and, as a result, only showed smaller increases over time. Why might that be the case?

There are complementary explanations for these findings, which include (1) the ability to create expectations, (2) the effectiveness of implicit feedback, and (3) prior knowledge of *Perfekt*.

Regarding the ability to create expectations and/or to predict upcoming sentence material, previous research has shown that engagement in predictive processing among L2 learners is modulated by the target structure in question (Dussias et al., 2013; Lew-Williams & Fernald, 2010; Grüter et al., 2017; Hopp, 2013, 2015, 2016; for a review, see Kaan, 2014; Kaan & Grüter, 2021), as well as by proficiency (Dussias et al., 2013; Hopp, 2015). For instance, Dussias et al. (2013) showed that only highly proficient English L2 learners of Spanish used grammatical gender information to predict upcoming sentence referents, whereas less proficient learners did not. These findings extend to the realm of priming research, which shows larger within-L2 priming effects for learners with higher L2 proficiency compared to learners of lower proficiency (Bernolet et al., 2013). Kaan and Chun (2018) and Kaan et al. (2019) also discuss the possibility that learners do not in fact create predictions that could facilitate learning the targeted structures. For example, Kaan et al. (2019) demonstrated that learners may be unable to correctly parse — and in turn, predict the structure of — infrequent embedded *what*-questions and coordinating constructions with *and*, precluding the occurrence of prediction error and adaptation. Similarly, Kaan and Chun (2018) hypothesized that learners may not have made strong predictions about sentence fragment continuations in a priming task, eliminating priming effects and learning. It is thus possible that learners at the lower end of the proficiency scale in the current study did not create expectations, which necessarily mitigates the possibility for prediction errors to occur and, in turn, error-based learning to take place. Additionally, even if participants at the lower end of the proficiency scale created expectations as to which picture matched the auxiliary presented in sentence fragments in the current study, they might not have committed to these predictions with sufficient confidence for prediction errors to be surprising and to lead to better memory of these newly learned items (e.g., Brod et al., 2022; Theobald et al., 2022).

Similarly, the ability to efficiently process and use implicit feedback may be attenuated in learners at the lower end of the proficiency scale (e.g., Ammar & Spada 2006; Li, 2014). For instance, Li (2014), who investigated the effects of different oral feedback types for learners of different proficiencies, showed that participants with lower proficiency scores only benefited from prompts but not recasts, while participants at higher proficiency equally benefited from both. Li specifically argues that learners with low proficiency may lack the cognitive resources necessary to recognize the corrective force of recasts and other implicit feedback types, and that, as a result, these learners are not able to utilize the feedback effectively. On the other hand, studies that investigate different feedback types in computer-mediated environments generally show recasts to be at least somewhat effective (e.g., Sagarra & Abbuhl, 2013; Sauro, 2009; Baralt, 2013; but see Loewen & Erlam, 2007), as learners' errors and the corrective feedback are visually juxtaposed, allowing for an effective comparison between the two. While participants in the current study received computer-mediated rather than oral feedback, the provision of feedback was still visually and temporally separated from potential learner errors in choosing the appropriate picture or completing the sentence fragment during training. A direct comparison between learner error (or non-error) and correct form, as it is generally implemented in other studies that use computer-mediated feedback (e.g., Loewen & Erlam, 2007; Sagarra & Abbuhl, 2013), was thus not possible for learners. Lower-proficiency learners who lack the necessary cognitive resources to make this comparison on their own may therefore not have been able to compute prediction errors (or to confirm their predictions) and use the implicit feedback to update their hypotheses and language model.

Finally, participants at the lower end of the proficiency scale may not have had sufficient basic knowledge of the target structure to create informed expectations and to leverage the information provided with the corrective feedback in order to hone in on the correct language forms (see McDonough and Fulga, 2015 for similar findings regarding structural priming). Importantly, this explanation is connected to the ability to create predictions and successfully process implicit feedback. In other words, these three possible explanations are not mutually exclusive and may interact with one another.

These three complementary explanations can account for differences in learning gains between learners at the lower end versus learners at the center of the proficiency scale. What about participants at the other end of the scale, who have high levels of proficiency? If anything, participants at the higher end of the proficiency scale should be better able to create predictions, process implicit feedback, and update their prior knowledge as a result. However, one has to keep in mind the main effect of proficiency across all measures. Participants with high levels of proficiency were more accurate across all testing times than participants with low levels of proficiency—including at pretest. Participants with high levels of proficiency only showed small performance improvements at each testing time because they started out with higher accuracy. This does not mean that error-based learning does not apply in learners with higher proficiency, only that there might be little room for improvement to begin with. Participants around the center of the *z*-scored proficiency scale, on the other hand, may have found themselves in a potential "sweet spot" for error-based learning to take place, as they (1) possessed only basic knowledge of the target structure — leaving sufficient room for improvement — and (2) most likely had the cognitive resources available to make (3) more informed predictions with (4) a sufficient degree of confidence, such that they would be able to efficiently process and use (5) implicit feedback to confirm or disconfirm their predictions and to (6) update their language model. In fact, similar findings were obtained by Grüter et al. (2021), who showed stronger long-term priming effects for English double object constructions in Korean L2 English learners around the center of their *z*-scored proficiency scale —but keep in mind that this scale does not necessarily correspond to the one in the current study.

3.4.3 The Role of Prior Knowledge and Awareness

In the current experiment, participants with prior knowledge — operationalized as participants' course level being higher than GER1— mirrored the overall group effects, with accuracy between 75% and ceiling on the picture matching task and around 75% after pretest on the picture description task. Participants without prior knowledge, on the other hand, showed essentially no learning at all, i.e., they hovered around chance on the picture matching task and remained at floor in the picture description task.

Previous studies found benefits of guessing — i.e., creating expectations in the absence of connections to prior knowledge — for the learning of items related to episodic memory, for instance fictitious/non-fictitious facts or vocabulary words (Brod & Breitwieser, 2019; Brod et al., 2018, 2022; Butterfield & Metcalfe, 2001, 2006; Kang et al., 2009; Kornell et al., 2009; Potts et al., 2019; Potts & Shanks, 2014; Theobald et al., 2022. Items stored in episodic memory, such as facts or word-pair associates, seem to be acquirable via one-shot declarative learning — a form of error-based learning in which learning occurs rapidly via prediction error caused by a single pairing of a stimulus with an outcome, even in the absence of prior knowledge about this stimulus-outcome pair (Lee et al., 2015), essentially constituting a form of guessing. One explanation for why guessing was successful in these studies is that the generation of expectations, especially if errorful, benefits subsequent memory by enhancing the encoding of feedback provided to speakers (e.g., Brod et al., 2018; Brod & Breitwieser, 2019; Potts et al., 2019). However, the current experiment did not reveal similar learning in participants without prior knowledge. In fact, learners without prior knowledge showed no learning at all, a finding that was somewhat unexpected based on the previous studies on guessing and error-based learning cited above. Nevertheless, the findings from the current experiment are reminiscent of previous studies that investigated the differential effects of corrective feedback for learners of different proficiencies and/or different levels of prior knowledge (Ammar, 2008; Ammar & Spada, 2006; Li, 2014; see also Iwashita, 2003). For instance, Ammar and Spada (2006) showed that only participants who scored above 50% at pretest benefited from recasts, i.e., implicit feedback, whereas participants who scored below 50% on pretest did not. Similarly, Li (2014) showed that recasts were not effective for lower proficiency learners attempting to learn the complex Chinese perfective marker *-le.*

The finding that learners without prior knowledge did not show any learning in response to the prediction-based training task has important implications for the applicability of error-based learning in an L2. While clearly effective in situations where prior knowledge of the target feature exists (e.g., Grüter et al., 2021, McDonough & Trofimovich, 2015; Coumel et al., 2022) or where items related to episodic memory are targeted (e.g., Brod & Breitwieser, 2019; Brod et al., 2018, 2022; Butterfield & Metcalfe, 2001, 2006; Kang et al., 2009; Kornell et al., 2009; Potts et al., 2019; Potts & Shanks, 2014; Theobald et al., 2022), the benefits of error-based learning mechanisms might not extend to the learning of novel morphosyntactic features in the L2, like the German *Perfekt* auxiliary alternation targeted in this study (but see Bovolenta and Marsden, 2021). It seems like the learning of morphosyntactic features might be fundamentally different than the learning of items that are stored in episodic memory, such as facts or word-pair associations (akin to L2 vocabulary learning), in the absence of prior knowledge. This could be due to the nature and organization of morphosyntactic cues in learners' language systems, in particular if said cues do not exist or have no predictive value in a learner's L1. More specifically, learners' inability to overcome learned (in)attention to the auxiliary as a predictive cue without prior knowledge of this alternation may stem from the absence of this cue in a participants' L1 (see also Ellis, 2006; Ellis & Sagarra 2010; Sagarra & Ellis, 2013). In turn, it may be more difficult for learners to abstract morphosyntactic regularities using cycles of hypotheses and feedback if a cue has no initial value. Under such a view, for error-based learning of morphosyntactic features to take place, learners may need to become aware of the underlying morphosyntactic regularities first before being able to successfully explore and reject hypotheses about the possible applications of that regularity. Rather than enabling the learning of novel L2 morphosyntactic features, prediction-based learning tasks that involve implicit feedback may be most effective when they reactivate prior knowledge in those learners who possess it, making it a useful learning paradigm to reinforce previously learned L2 features or structures (see McDonough & Trofimovich, 2015). Within an error-based learning framework, guessing may thus be a learning mechanism in its own right, but only for specific types of item-based knowledge that can be captured within episodic memory. In fact, episodic knowledge, such as facts or word-pair associations, may be learnable via one-shot declarative learning, whereas exploring morphosyntactic regularities may lead to greater uncertainty about possible and impossible language forms, slowing down the learning rate (see Lee et al., 2015) and requiring longer exposure to more items featuring the morphosyntactic feature, in order for learning in the absence of prior knowledge to take place.

In the context of L2 learning, awareness may also play a key role in error-based learning tasks, especially when feedback is more implicit in nature. In the current experiment, participants showed different levels of awareness based on course level: Participants in GER1, i.e., beginning learners, showed no awareness of the auxiliary alternation based on the debriefing questions, whereas participants in higher course levels were largely aware of this feature. In addition, awareness had an impact on accuracy in both the picture matching and picture description task. More specifically, while there were no differences in accuracy as a function of awareness levels at pretest, only those participants who were aware of the auxiliary alternation — with some exceptions — performed above chance at immediate and delayed posttest. These findings are in line with McDonough and Fulga (2015), who showed that only those learners who successfully detected — or became aware of — a novel morphological feature of the target language in a learning activity prior to a priming task were able to be primed (see also Coumel et al., 2022; McDonough & Trofimovich, 2015).

Importantly, in this experiment course level also coincided with how I operationalized prior knowledge: Participants in GER1, who hadn't been introduced to the *Perfekt* in the German classroom yet were classified as having no prior knowledge whereas participants in all other course levels were assumed to have prior knowledge, as they had been explicitly introduced to the target structure in their previous German courses. This then means that participants who had no prior knowledge were more likely to be unaware of the auxiliary alternation in the input whereas participants with prior knowledge were mostly aware of it. However, this also means that it is impossible to tease apart the individual contributions that prior knowledge vs. awareness make in error-based learning. Indeed, some participants who were classed as unaware — and either had or did not have prior knowledge — performed above chance at delayed posttest on the picture matching task. Nevertheless, the absence of learning effects in most participants without prior knowledge in the present study underscores not just the role of prior knowledge for error-based learning, but also the role that awareness plays in kickstarting and accelerating the learning process (see Andringa, 2020; Curcic et al., 2018) — although it may not be an absolute requirement for learning to occur.

Admittedly, the role of awareness was not a primary focus in the current study, which explains the coarseness of the awareness measures used: Participants completed debriefing surveys related to their overall experience and their awareness of the targeted auxiliary alternation after the immediate and the delayed posttest. They therefore had to make retrospective judgments about their awareness of the targeted feature during the training, a type of measure prone to interference (Gass & Mackey, 2017). In order to explore the role of awareness —in conjunction with the role of attention and surprisal — at a more fine-grained level, Experiment 2 uses pupillometry to explore the processing and use of implicit corrective feedback as it is provided to learners (Godfroid, 2019).

Overall, the findings related to the role of prior knowledge indicate that error-based learning — at least in conjunction with implicit feedback— may not facilitate learning a morphosyntactic alternation for which the relevant knowledge system has not been established yet, because learners may not be able to make efficient cyclical use of errors and implicit feedback to narrow in on the correct language forms. Future research should investigate whether this inability can be overcome with increased exposure to the target structure.

3.4.4 Implications for Error-Based L2 Learning

The present findings have important implications for whether error-based learning is possible in contexts of L2 learning, mainly related to the role of proficiency, prior knowledge and awareness. For one, the findings show that participant proficiency and experience with the L2 target feature plays a role. Learners around the center of the proficiency scale in the current study may have found themselves in a potential "sweet spot" for error-based learning to take place. This suggests that some basic prior knowledge and experience with the L2 target feature as well as sufficient cognitive resources to process implicit feedback and use it to update learners' internal language model are prerequisites for efficient cycles of prediction-feedback-prediction error-update to occur.

Secondly, creating expectations without any prior knowledge of the target structure or feature — i.e., guessing — may not contribute to error-based learning compared to creating expectations with prior knowledge. Thus, in contexts of L2 learning, guessing and predicting might be two different components of a larger learning mechanism that apply to the learning of different types of knowledge: Items stored in episodic memory, such as L2 vocabulary words vs. other linguistic knowledge, such as morphosyntactic features. Whereas guessing — and one-shot declarative learning — benefits L2 vocabulary acquisition (e.g., Potts & Shanks, 2014; Potts et al., 2019; Kornell et al., 2009), the learning of morphosyntactic features may require a longer sequence of trial and error, leading to the gradual learning of the underlying morphosyntactic regularities by slowly establishing a system of relevant knowledge. This rudimentary system of relevant knowledge regarding the morphosyntactic target structures may then allow for predictions and a more targeted exploration of hypotheses. It remains unclear when exactly the transition between less-beneficial guessing and more beneficial predicting occurs during the learning of L2 morphosyntax. Similarly, it remains unclear if guessing and predicting are two points on a continuum or whether they represent qualitatively different learning mechanisms. These would be valuable questions for future research.

Relatedly, the type of feedback provided to learners within an error-based learning task may play an important role. While both types of implicit feedback provided in the current study seemed to benefit learners with prior knowledge, implicit feedback did not help learners without prior knowledge become aware of the target auxiliary alternation and facilitate learning. It is thus possible that learners without prior knowledge of specific L2 target feature require different types of corrective feedback to kickstart the learning process. While error-based learning models generally consider this type of learning to occur implicitly and without learners' awareness, it is certainly possible that this is not the case for inexperienced L2 learners. Instead, corrective feedback may need to include explicit information to draw learners' attention to the error that they produced, in order to facilitate the computation of prediction error and the development of new hypotheses about possible language forms. This speculation is corroborated by previous studies in the realm of corrective feedback research, which have shown that explicit feedback is often more beneficial to learners of lower proficiency (e.g., Ammar & Spada, 2006; Li, 2014).

The finding that error-based learning may not be possible for novel morphosyntactic features in an L2 when it has been shown to be beneficial for vocabulary learning parallels findings from other fields of research that are all relevant to investigations of error-based learning, namely predictive processing, corrective feedback, and priming research. Research on predictive processing has shown that while L2 learners of different proficiency can use lexical cues predictively (e.g., Chambers & Cooke, 2009; Chun et al., 2021; Dijkgraaf et al., 2017, 2019; Hopp, 2015; Ito et al., 2018), they use predictive morphosyntactic cues much less consistently (Dussias et al., 2013; Frenck-Mestre et al., 2019; Lew-Williams & Fernald, 2010; Grüter et al., 2017; Hopp, 2013, 2015, 2016; see also Mitsugi & MacWhinney, 2016; for a review, see Kaan, 2014; Kaan & Grüter, 2021). Similarly, research on corrective feedback has shown that while feedback targeted at grammatical — including morphosyntactic — errors is most frequent (Brown, 2016; Mackey et al., 2000), learners recognize corrective feedback aimed at lexical errors more frequently as actual feedback (Lyster et al., 2013; Mackey et al., 2000; Mackey & Goo, 2007). Finally, research into

L2 priming has shown reliable semantic priming effects independent of participants' proficiency levels (e.g., Altarriba & Knickerbocker, 2011; Williams & Cheung, 2011; for a review, see McDonough & Trofimovich, 2008), whereas morphosyntactic priming in production in the absence of prior knowledge or awareness is generally not observed (e.g., McDonough & Fulga, 2015; McDonough & Trofimovich, 2015; see also Thomas, 2018). These parallels strengthen the conclusion that errors in the absence of prior knowledge and explicit instruction may not drive the learning of novel morphosyntactic features of an L2, since the crucial processes involved in and representative of error-based learning — prediction, feedback, and priming — also showcase difficulties in using or processing L2 morphosyntactic cues.

3.4.5 Limitations and Future Studies

While the current study has some limitations that impact the conclusions that can be drawn from the results and analyses, these also offer exciting avenues for future studies. In the current study, 46 L2 German learners participated and were split into two fairly even experimental groups. While this sample size allowed to run models with two-way interactions to investigate the roles of feedback ordering and proficiency separately, the sample size was too small to include the critical three-way interaction between testing time, group, and proficiency. Future studies should therefore attempt to replicate the current study with a larger sample size or increase the number of target items included, to the extent possible considering the limited availability of participants and known vocabulary items, to increase statistical power and investigate the joint effects of feedback ordering and proficiency in error-based learning tasks.

Secondly, participants in the current study completed two blocks of training, each block containing the same 12 target items (8 with *haben*, 4 with *sein*). This means that participants had one exposure to each target item with each type of feedback, and thus only two exposures to each target item overall. This limited exposure may not have been sufficient for participants, especially those without prior knowledge of the auxiliary alternation, to become aware of it and to exhibit behavioral signs of learning, in particular because implicit learning may take longer than other types of learning (Ellis, 2005). Exposures could therefore be increased in two ways, either by increasing the total amount of targeted verb types (e.g., 16 items with *haben*, 8 items with *sein*, each repeated twice) or by increasing the number of tokens per verb type (e.g., 8 items with *haben*, 4 items with *sein*, each repeated 4 times). In studies that have investigated the role of type vs. token frequency for the acquisition of syntactic structures in an L2 (e.g., Fulga & McDonough, 2016; McDonough & Nekrasova-Becker, 2014; McDonough & Trofimovich, 2013; see also Ellis, 2002), the number of types has ranged between three types with nine tokens of each (McDonough & Trofimovich, 2013) and five types with four tokens of each (McDonough & Nekrasova-Becker, 2014). In studies that have investigated error-based learning mechanisms and the role of prediction error, the number of total trials included in the learning tasks has ranged between 8 items with two exposures each (Fazekas et al., 2020 for syntax) and 30 items with one exposure to each (Brod et al., 2018 for episodic items). For the present study, it remains unclear if a greater number of total exposures — either in terms of the quantity of types or the quantity of tokens per type (or both) — might ultimately lead to learning in learners without prior knowledge of the auxiliary alternation. Future studies could therefore attempt a partial replication of the current study by increasing the number of training blocks for participants without prior knowledge, exploring whether prediction-based learning with implicit feedback simply necessitates larger amounts of input — either in terms of types or tokens — and, in turn, errors, for learning to occur when prior knowledge is absent.

CHAPTER 4: EXPERIMENT 2

4.1 Introduction and Research Questions

Experiment 1 revealed both a role for prior knowledge and general L2 proficiency as well as a role for target structure awareness for learning the German *Perfekt* auxiliary alternation. However, Experiment 1 was not an online experiment, meaning that neither the creation of expectations nor the use of feedback, nor participants' awareness of the target auxiliary alternation were investigated in real time. Furthermore, the literature review presented in Chapter 2 outlined several open questions regarding what learners do with and how they process feedback in the moment it is provided to them. The second experiment in this dissertation addresses whether beginning- to intermediate- level German learners can use the German *Perfekt* auxiliary alternation predictively, how feedback is processed, and whether feedback processing is indicative of learning. More specifically, this experiment investigates whether learners start to show behavioral signs of predictive processing by tracking their eye movements as they listen to sentence fragments and engage with pictures on the computer screen. In conjunction with the emergence of predictive processing, Experiment 2 investigates whether and how learners process feedback in the exact moment they receive it, using pupil size as a proxy measure for cognitive functions such as attention, surprise, and awareness. In addition, the evolution of pupil size in response to feedback over the course of the training task will be analyzed to track changes in feedback processing. Finally, Experiment 2 investigates whether feedback processing, as measured via pupillary response, correlates with the learning of the auxiliary alternation — and secondarily of the past participles — in participants. By investigating predictive processing as well as feedback processing in L2 learners with beginning to intermediate proficiency, this experiment sets out to create a better understanding of the role of prediction and feedback processing in error-based learning in an L2 context. In Experiment 2, I therefore asked the following research questions:

- *RQ1) Over the course of the training task, do learners start to engage in predictive processing, as measured by the proportion of looks to a target picture?*
- *RQ2) After listening to target sentence fragments, do learners process feedback they receive about the accuracy of their prediction, as measured via pupil dilation? Does feedback processing evolve over the course of the training task?*
- *RQ3) Do pupillary responses correlate with the amount of learning, as measured during training and on independent posttest measures?*

The following sections outline the experimental method applied to investigate these questions. First, I introduce the experimental approach, including the general design and time course, a description of participants in the study, the training task, as well as the outcome measures used to evaluate learning. Section 4.3 then describes the approach to data processing as well as results of the experiment. Chapter 4 concludes with a discussion of the results in light of the research questions posed above.

4.2 Overview of Experimental Approach

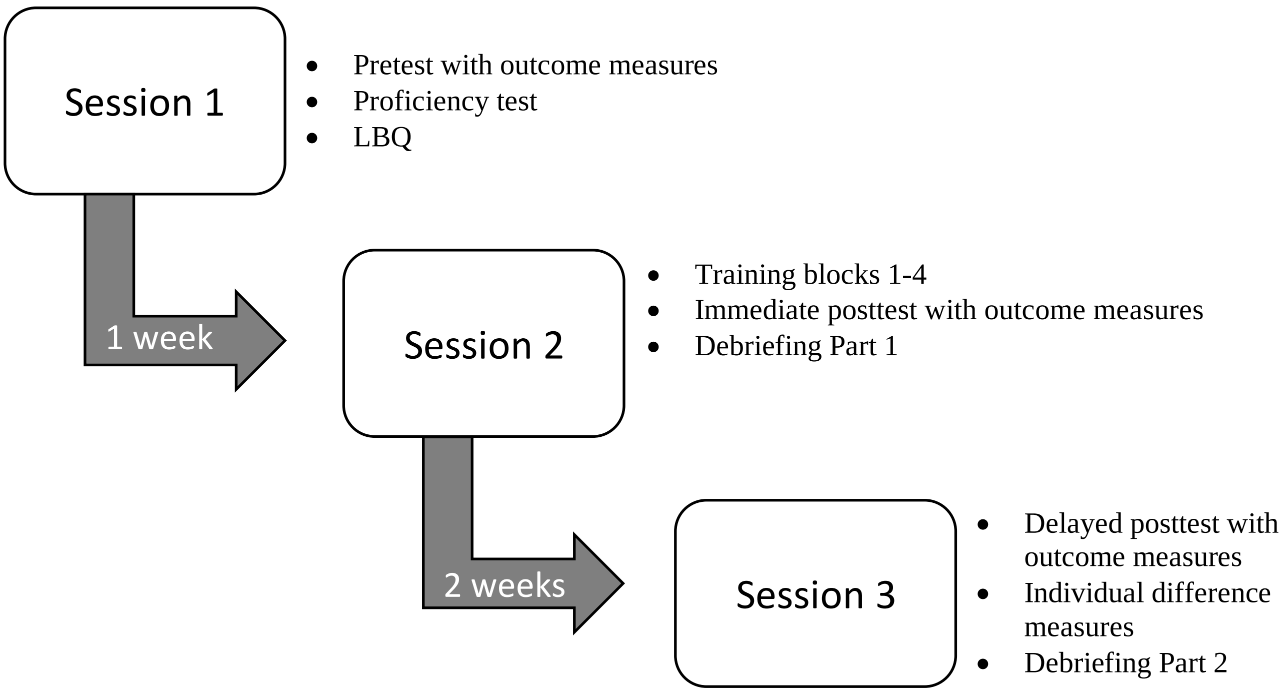
4.2.1 General Design and Time Course of Study

All measures were conducted in person in a research lab on Penn State's University Park campus. Participants either completed tasks using a desktop computer or using the eye-tracker-mounted laptop, depending on the experimental session and task.

Participants completed the pretest in Weeks 4 or 5 of the semester. The training and immediate posttest were then administered in Week 5 or 6, depending on the timing of the first session. Finally, the delayed posttest was then administered two weeks later in Weeks 7 or 8 of the semester, again depending on when Sessions 1 and 2 were completed (see Figure 4.2.1). All sessions were conducted outside of participants' regularly scheduled class time. Total participation time across all sessions was around 2.75 hours.

Figure 4.2.1

Timeline of testing and training in Experiment 2



4.2.2 Participants

Participants in this experiment were students enrolled at various levels of German courses at Penn State's University Park campus, ranging from beginning-level GER 2 and GER 3 to intermediate GER 201 language courses. Participants were recruited from these courses to participate in the experiment outside of their regular classroom instruction in Spring 2022.

In total, 21 participants were recruited across the targeted German courses. Of those 21 participants, 16 participated in all three sessions. Participants either received extra credit — the amount of which was determined by their respective course instructor — or monetary compensation at the rate of $10/hour. For participating in all study sessions, monetary compensation amounted to a total of $32.50: $7.50 for the first and third session, $12.50 for the second session, and a $5 bonus for completing all three sessions. Three participants were excluded from the analyses, as they did not complete the training task due to calibration issues with the eye-tracker. This means that data from only 13 participants were entered into the analyses.

4.2.2.1 Screening and Proficiency Measures

Participants completed two measures that were designed to assess their demographic and language learning history, as well as their proficiency in German. Together, these tasks ensured that participants were comparable with respect to their language background and proficiency. In addition, measuring proficiency allowed to enter proficiency as a control variable in any statistical analyses, should differences based on proficiency arise at pretest.

4.2.2.2 Materials and Data Processing

The language background questionnaire and the proficiency measure were identical to the ones used in Experiment 1 and described in Section 3.2.2.2.

4.2.2.3 Results

The means and standard deviations for variables of interest from the language background questionnaire and from the independent proficiency test are presented in Table 4.2.1.

Table 4.2.1
Participant demographics

Variable	*M*	*SD*	*Range*
Current age (years)	19.6	2.1	18-25
L2 proficiency test (0-1)	0.36	0.48	0.16-0.64
L2 proficiency ratings (1 - 10)			
Reading	4.6	1.9	1-8
Spelling	6.0	1.9	2-8
Writing	4.2	2.1	2-8
Speaking	4.2	2.6	1-10
Listening	4.4	1.9	1-8
Overall	4.7	2.1	-

No participants were excluded because of their language background, and L1s other than English included Hindi (*N*=1) and Turkish (*N*=1).

4.2.3 Training

Training was similar to Experiment 1 with three important differences: the modality of stimulus presentation, the number of training blocks, and the feedback provided. In Experiment 2, participants completed four training blocks without any interim tests. Each block of training consisted of 24 trials, targeting each verb in a different sentence fragment once per round of training. In each trial, participants heard the sentence fragment up until the offset of the second temporal adverbial. While participants heard the sentence fragment, their eye movements were tracked to see whether they began to fixate on the target picture more after hearing the predictive auxiliary verb in comparison to the foil picture which did not match with the auxiliary — similar to the training design in Experiment 1. After hearing the sentence fragment, participants chose the picture that they believed accurately completed the sentence fragment by pressing one of two previously assigned keys on their keyboard. Immediately after choosing the picture, they also completed the sentence fragment by typing into a textbox what they believed to be the past participle form of the verb associated with the picture they chose. Once participants completed their production of the past participle, they received feedback in the form of the correct past participle, regardless of whether participants had committed an error in their choice or when typing the participle form or not.

4.2.3.1 Materials

Materials were similar to the ones used in Experiment 1, with some differences regarding the total number of target verbs, the number and structure of sentence fragments used, and the modality of sentence fragment presentation.

Materials included in the training consisted of twelve sentence fragments in *Perfekt*, which left out the past participle form of the target verb (see Appendix C). Each sentence fragment was constructed with the following structure: 1) one of four noun phrases, 2) *Perfekt* auxiliary, 3) temporal adverbial, 4) a second temporal adverbial, and 5) a blank for the past participle. These twelve sentence fragments each appeared with the *Perfekt* auxiliaries, either the auxiliary *haben* or the auxiliary *sein*, conjugated for third person singular (see Examples 1 a and b).

1) a. Der Mann/Die Frau/Der Junge/Das Mädchen hat gestern Abend spät ________.
   b. Der Mann/Die Frau/Der Junge/Das Mädchen hat gestern Abend spät ________.

   The man/The woman/The boy/The girl has yesterday night late ________.
   'The man/The woman/The boy/The girl ________ late yesterday night.'

Target verbs in this experiment again included both strong verbs that take the *-en* inflection for their past participle form (e.g., *gelaufen* 'walked) and weak verbs that take the *-t* inflection for their past participle form (e.g., *gekocht* 'cooked'). Unlike in Experiment 1, strong verbs and mixed verbs that have a stem vowel alternation in the past participle form (e.g., *getrunken* 'drunk' or *gerannt* 'run') were not excluded in order to increase the total number of possible target verbs that could be incldued. A total of 24 verbs were included in the training, of which eight take the auxiliary *sein* and 16 take the auxiliary *haben*. Again, this balance represents the frequency differences with which the *Perfekt* auxiliary *sein* is encountered in natural language use (Bordag & Sieradz, 2014). Of these 24 verbs, ten were strong verbs (five with *haben*, five with *sein*), seven were weak verbs (eleven with *haben* and two with *sein*), and one was a mixed verb (one with *sein*; see Appendix C). These 24 verbs were chosen from *Sag Mal 2nd Edition* (Anton et al., 2017) and *Stationen* (Augustyn & Euba, 2015), the textbooks used in Penn State's GER 2 and GER 3 & GER 201 respectively. Similar to Experiment 1, this was done to ensure that all participants were familiar with the meaning of all trained verbs prior to the experiment to avoid processing difficulties due to unfamiliarity with word meanings. Again, target verbs were not controlled for their cognate status with English, as this would have reduced the already limited pool of possible target verbs by another significant amount.

Like in Experiment 1, traditional counterbalancing through the use of a Latin Square design was not possible. Instead, seven lists were created to cover the four training blocks and test measures at each testing time (pretest, immediate posttest, and delayed posttest). Within each list, each of the 24 main verbs appeared once as the target verb, each main verb with *sein* appeared twice as a foil verb for target main verbs with *haben*, and eight out of 16 main verbs with *haben* appeared as foil verb for target main verbs with *sein*. Across lists, each main verb was paired with each of the twelve sentence fragments only once. Participants completed a different list at each testing time and in each round of training to ensure that they saw each main verb paired with a different sentence fragment throughout the experiment.

4.2.3.2 Training Procedure

4.2.3.2.1 Training Blocks and Trial Setup

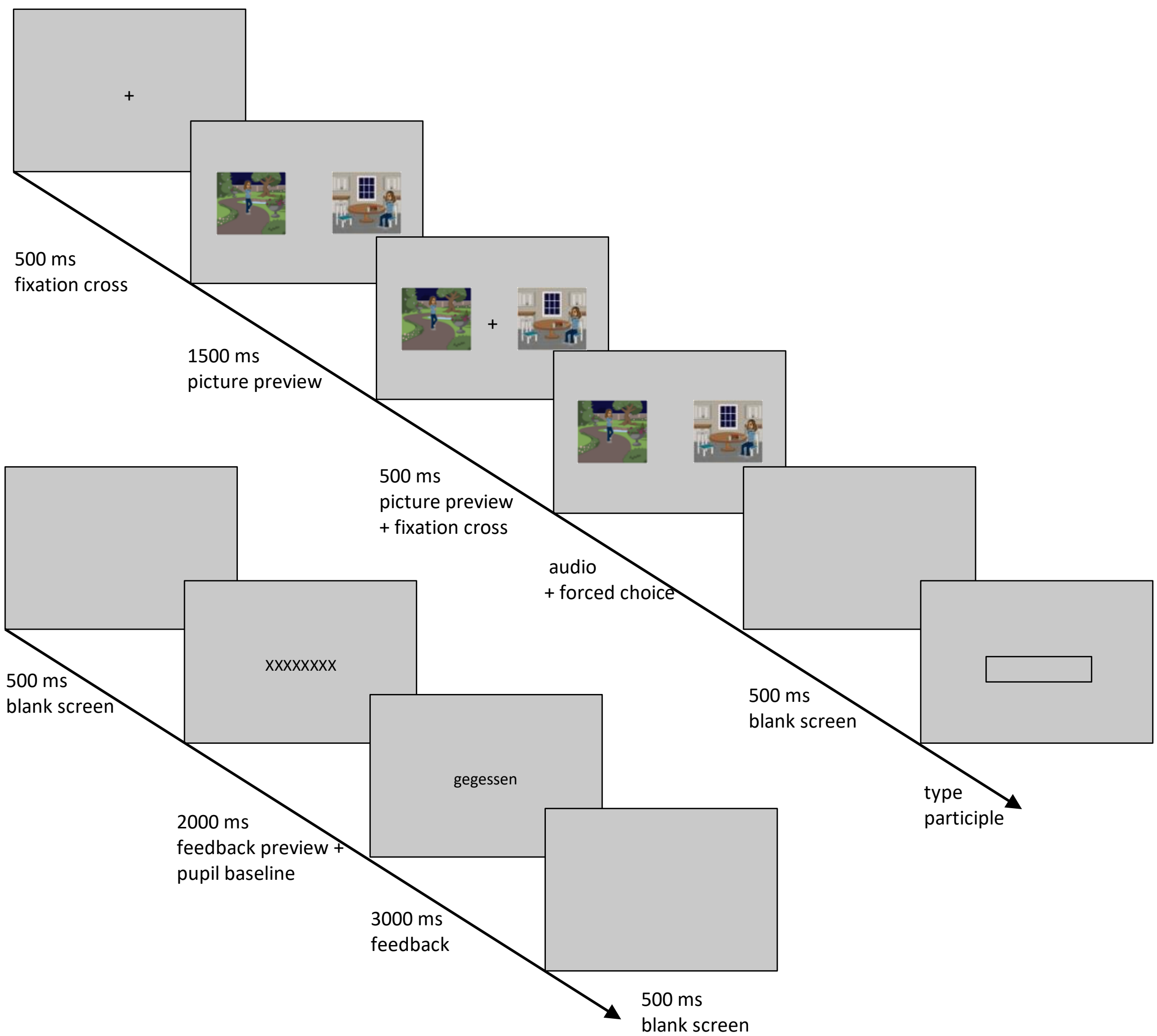
Participants completed four training blocks, each consisting of 24 trials. Because there were 24 target verbs but only twelve sentence fragments, each training block contained two subsets of trials. Each subset contained all twelve sentence fragments, with eight trials with *haben* as the auxiliary and four trials with *sein* as the auxiliary. Sentence fragments therefore occurred twice per training block, but only once per subset of trials. Trials were then randomized within but not across each of the two subsets of twelve sentence fragments in each block.

On each trial, participants first saw a black fixation cross in the center of the screen for 500 ms. The fixation cross was followed by a picture preview for 1500 ms: Both the target and foil picture appeared, either on the left or right side of the screen. The location of the target picture was counterbalanced across the experiment. To prevent participants from already looking at one of the pictures before the onset of the audio recording, a fixation cross appeared between the two images for another 500 ms and participants were instructed to redirect their gaze towards the fixation cross. The experiment was programmed such that the audio would only start playing once participants had looked at the fixation cross between the two pictures for 500 ms. While the audio played, participants were free to look anywhere on the screen while their eye movements were being tracked. After the audio file finished playing, participants had to choose the picture which they believed accurately completed the sentence fragment by pressing one of two previously assigned keys on their keyboard, the *F* key for the left picture and the *J* key for the right picture. After pressing one of these two keys, participants then saw a blank screen for 500 ms before being prompted to type the past participle form of the verb corresponding to the picture they chose into a textbox. This was then followed by another blank screen for 500 ms, before feedback was presented to participants.

Similar to the form-focused feedback in Experiment 1, participants saw the correct past participle on a separate screen, regardless if their picture choice and the past participle form they produced were correct or incorrect (see Figure 3.2.5b). This feedback contained two kinds of information about the participants' prediction: 1) whether their choice of picture was correct, and, if it was correct, 2) whether the form of the past participle participants produced was correct as well. The feedback therefore resulted in the computation of three different types of possible prediction errors. For instance, the feedback could confirm a participant's predictions in both picture choice and, thus, the semantic relationship between the auxiliary and the main verb, as well as the past participle form, e.g., if the participant produced the form *gelaufen* based on the auxiliary *sein* and a picture of a person running (SEM+FORM+). The feedback could also confirm a participant's prediction based on the semantics underlying the relationship between the auxiliary and the main verb but disconfirm their prediction about the past participle form, e.g., if the participant produced the form *gelauft* based on the auxiliary *sein* and a picture of a person running (SEM+FORM-). Finally, the feedback could disconfirm a participant's semantic prediction, e.g., if the participant chose the picture corresponding to *essen* when presented with a sentence fragment with the auxiliary *sein* (SEM-).

To avoid any effects of pupil dilation in response to feedback presentation to be contaminated by lingering effects from processing the sentence fragment and typing the past participle, a baselining screen was presented prior to feedback presentation. During this period, a blank screen was presented to participants for 2000 ms, with a string of capital Xs of fixed length in the center of the screen to keep luminosity between the baselining period and feedback presentation constant. Feedback then appeared in black font in the center of the screen for 3000 ms. Following feedback presentation, the screen went back to a blank screen for 500 ms. This allowed the pupil size to return to baseline before the next trial ensued. Participants were instructed to try not to blink from the onset of the Xs on the screen until the screen was blank again. A sample trial is presented below in Figure 4.2.2.

Figure 4.2.2
Sample trial in Experiment 2



4.2.3.2.2 Audio Files

The sentence fragments were recorded at a sampling rate of 44.1 kHz in a sound-proof recording studio by a female native speaker of German. The native speaker recorded each sentence with a participle at the end to ensure that sentence rhythm and intonation were as natural as possible. To create sentence fragments, the participle was then manually removed from each sentence fragment using Praat, ensuring that the remaining sentence fragment remained intact.

Once all sentence fragments were created, audio files were normalized to 60 dB using a Praat script. The mean duration of all audio files was 3.445 s, with no difference by auxiliary type (*haben* 3.526 s, *sein* 3.364 s; $p = .202$). The mean duration of the critical word, i.e., the auxiliary, was 230 ms and differed slightly by auxiliary type (*haben* 261 ms, *sein* 199 ms; $p < .01$). Finally, the mean onset time of the critical word, i.e., the auxiliary, was 813 ms. Mean onset time did not differ by auxiliary type (*haben* 789 ms, *sein* 837 ms; $p = .187$).

During the training task, the audio files associated with each trial were passed through a Steinberg UR22 mkII audio interface (24-bit/192 kHz) to ensure accurate timing and measuring of when the critical word was played. Audio volume on the display laptop and on the audio interface were kept constant.

4.2.3.2.3 Procedure

During training, participants were seated in front of the display laptop (Asus GL703GM-DS74) at a distance of approximately 20.5 in (52 cm) and rested their head on a chin rest to minimize head movements during trials. The eye-tracker, an SR Research Eye Link® Portable Duo remote laptop-mounted eye-tracker, was set to track participants' right eye at a sampling rate of 500 Hz. Pupil size was recorded as area. Prior to the first block of training, the eye-tracker was calibrated to participants' eyes using a 9-point calibration grid. Participants then received instructions for the task and completed three practice trials, after which they were given a chance to ask questions if they had any. Participants then began the main task with the first block of training. After each training block, participants were shown a message telling them that they could take a short break and relax. When participants were ready to resume with the next training block, the experimenter instructed participants to place their head back on the chin rest and initiated re-calibration, if necessary. After re-calibration, participants completed the next training block. If necessary, the eye-tracker was also re-calibrated during training blocks if the eye-tracker failed to detect a participant's look to the fixation cross prior to the start of a trial.

Instructions and trials were presented on the laptop monitor at a resolution of 1920 x 1080. The background color of each instruction and trials screen was set to rgb(169,169,169) with an opacity of 1.0. Light settings in the lab and on the laptop screen were kept constant.

4.2.4 Outcome Measures and Materials

4.2.4.1 Picture Matching and Picture Description Tasks

Participants completed similar outcome measures as those in Experiment 1, including a written picture description task to measure their productive knowledge and a picture matching task to measure their receptive knowledge of the target structure. For a description of these tasks and of how results were coded, see Sections 3.2.4.1 and 3.2.4.2. Participants completed 24 trials in the written production task and 24 trials in the picture matching task. These tasks were completed on a desktop computer in the research lab using EPrime version 3.0 (Psychology Software Tools, Pittsburgh).

4.2.4.2 Debriefing

Participants completed two debriefing questionnaires using Qualtrics, which were split across the second and third session. After the immediate posttest, the questionnaire included questions about (1) how difficult they perceived the training task to be, (2) whether they knew all the words used in the task and if not, which ones they did not know, and (3) how much caffeine they had consumed within the 12 hours prior to their participation. After the delayed posttest, the questionnaire included questions about (1) whether participants noticed anything about the sentence fragments, (2) whether there was anything about the sentence fragments that impacted their picture choice, (3) whether they noticed that some sentences had a form of *haben* and others a form of *sein*, (4) whether they felt like they waited until the end of the fragment to choose a picture, and (5) whether they used the feedback provided to them to become more accurate later in the task. Questions were ordered in such a way that they became increasingly more specific in asking about the targeted structure and the use of feedback (see Appendix E). Participants' answers to these questions supplement the outcome measures and provide insights into participants' internal thought processes during the training, including whether they consciously predicted, how they arrived at their predictions, and perceived the feedback provided to them.

4.2.4.3 Measures of Learning Aptitude

In addition to the comprehension and production outcome measures as well as the debriefing questionnaires, participants completed two measures of language learning aptitude: A Serial Reaction Time Task (SRT; Linck et al., 2013) to evaluate their implicit learning aptitude, and the LLAMA F test, which is available online (Lognostics; Meara & Rogers, 2021), to measure participants explicit learning aptitude. Implicit and explicit learning aptitude have been shown to be related to the benefits of corrective feedback (e.g., Granena & Yilmaz, 2019; Yilmaz & Granena, 2016, 2019, 2021) to varying degrees and were therefore measured to be included as control variables in any statistical analyses.

4.2.4.3.1 Serial Reaction Time (SRT) Task

Participants completed a Serial Reaction Time (SRT) Task (Linck et al., 2013) to evaluate their implicit learning aptitude. The task was programmed in EPrime version 3.0 (Psychology Software Tools, Pittsburgh). In this task, participants saw four boxes arranged horizontally on the screen. In one of these four positions, an asterisk would appear on each trial. The participants' task was to press a pre-assigned key on a keyboard that corresponds to the location of the asterisk on the screen. The test included six blocks of 96 trials each, with an intertrial interval of 500 ms. In the first and last blocks, asterisks appeared in a pseudo-random sequence. In the other blocks, asterisks consistently followed a repeating sequence of length 12.

Scores were calculated as the difference in median reaction times between the last sequential block (block five) and the last pseudo-random block (block six). The longer the times on the random block (block six) and the shorter the times on the last sequential block (block 5), the greater the implicit learning effect.

4.2.4.3.2 LLAMA F

Participants completed the LLAMA F (Meara & Rogers, 2021) to measure their explicit learning aptitude. The measure includes both a learning phase and a testing phase. In the learning phase, participants were instructed to learn the grammatical rules of an artificial language based on 20 picture-sentence pairs in just four minutes. Participants could freely click on the red buttons on the screen, and each button would show a different picture with the corresponding picture description underneath it.

After the learning phase, participants were tested on the extent to which they had learned the underlying rules of the artificial language by requiring them to select the grammatically correct sentence based on a picture presented to them. Participants saw a picture and could choose from the artificial language vocabulary to form a grammatically valid sentence. Similarly to the learning phase, the testing phase also included 20 items. Scores were calculated as number of correct items. The test phase was untimed but did not take any participant longer than 10 minutes.

4.3 Analysis and Results

Four sets of analyses were conducted to investigate the research questions in Experiment 2, which are repeated below.

RQ1) *Over the course of the training task, do learners start to engage in predictive processing, as measured by the proportion of looks to a target picture?*
RQ2) *After listening to target sentence fragments, do learners process feedback they receive about the accuracy of their predictions, as measured via pupil dilation? Does feedback processing evolve over the course of the training task?*
RQ3) *Do pupillary responses correlate with the amount of learning, as measured during training and on independent posttest measures?*

The first set of analyses addressed the development of auxiliary and participle accuracy over the three testing times on the picture matching and picture description task using robust statistics and non-parametric tests to investigate whether participants became more accurate on these measures during the time course of the experiment. The second set of analyses similarly addressed the development of auxiliary and participle accuracy, but over the course of the training task in the second experimental session. This included both average accuracy within each training block, as well as changes in accuracy across trials in each training block. In the third and fourth set of analyses, which address research questions 1 and 2 respectively, the time course of eye movement data from the picture choice portion of the training, as well as pupil size data from the feedback presentation, were visualized and analyzed descriptively. Due to the small sample size — a result of the COVID-19 pandemic that was still ongoing at the time of data collection — research question 3 could unfortunately not be answered within the scope of this dissertation. However, Section 4.3.6 addresses the kinds of analyses that may be conducted if more data is collected in the future.

4.3.1 Data Scoring

The data were scored similarly to Experiment 1, described in Section 3.3.1, with the only difference being that there were 24 items (16 *haben* and 8 *sein*) on the picture matching and picture description task, as well as in each round of training. This means that the maximal score for auxiliary accuracy and participle accuracy on a given task was 24. For the auxiliary accuracy analyses on the picture description task, however, only the eight *sein* items were included, similar to the analyses performed in Experiment 1.

4.3.2 Improvement on Outcome Measures

4.3.2.1 Analyses

For the outcome measures, i.e., auxiliary accuracy on the picture matching task and picture description task, as well as participle accuracy in the picture description task, I used robust descriptive statistics to analyze participant performance due to the small sample size. The statistics include the mean and standard deviation, the median, the interquartile range, as well as bootstrapped 95% confidence intervals. In order to calculate bootstrapped 95% confidence intervals, I used the BootES package (Kirby & Gerklanc, 2013) in R version 4.1.0 (R Core Team, 2021), with 2000 resamples for the mean of each outcome variable on each task.

In addition to the robust descriptive statistics, I also conducted non-parametric tests to analyze the change in accuracy from one testing time to another, given the non-normal distribution of the data, as indicated by Shapiro-Wilkes tests and visual inspection of QQ plots. Statistical analyses were conducted in R using the *stats* package version 4.1.0 (R Core Team, 2021). For each outcome measure and variable, I first conducted a Friedman test to compare the mean accuracy at pretest, immediate posttest, and delayed posttest. If the results of the Friedman test were significant at $p < 0.05$, I then conducted follow-up tests using Wilcoxon signed-rank tests to compare each testing time against one another (PRE – IMM, PRE – DEL, IMM – DEL). For these tests, the alpha level was Bonferroni-adjusted to .0167 (0.05/3 comparisons). Due to an EPrime malfunction during the picture description task at immediate posttest, data from only 12 participants are included in the statistical analyses. All other statistics, i.e., descriptive statistics and visualizations, include data from the 13 participants described in Table 4.2.1 in Section 4.2.2.3, unless specified otherwise.

Finally, Cohen's d effect sizes were calculated for all paired comparisons. These effect sizes were interpreted based on benchmarks for L2 research established by Plonsky and Oswald (2014). For within-group comparisons, the benchmarks are the following: d = .60 (small), 1.00 (medium), and 1.40 (large).

4.3.2.2 Results

4.3.2.2.1 Picture Matching Task

Descriptive results for the picture matching task can be found in Table 4.3.1. The results are also visually represented in Figures 4.3.1 and 4.3.2. In all tables and graphs, accuracy is calculated as the ratio of correct items to all items on any given test. Whether an item is correct or not depends on the auxiliary in a given sentence fragment and participants' picture choice for this item. If a participant chose the correct picture on 18 out of 24 items at immediate posttest in the picture matching task, their accuracy score corresponds to 0.75 at immediate posttest.

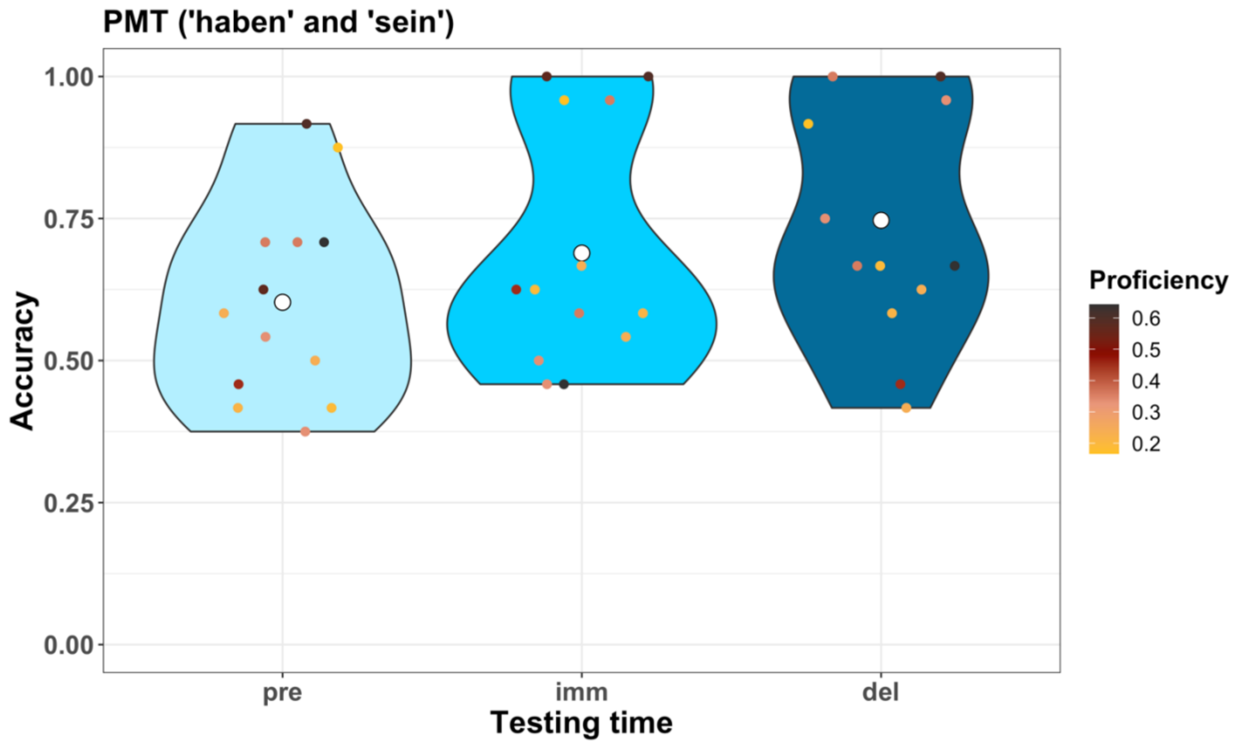
Table 4.3.1

Descriptive statistics for auxiliary accuracy on the Picture Matching Task

	M (SD)	Bootstrapped 95% CI	Mdn	IQR
Pretest	0.60 (0.49)	[0.52, 0.71]	0.58	0.25
Immediate posttest	0.69 (0.46)	[0.59, 0.82]	0.63	0.42
Delayed posttest	0.75 (0.44)	[0.64, 0.85]	0.67	0.33

Descriptively, participants increased 9% from pretest to immediate posttest, as well as 6% from immediate posttest to delayed posttest, as indicated by the mean accuracy in Table 4.3.1. These increases can also be traced in Figure 4.3.1, where the white dots represent mean accuracy at each testing time. However, these increases are not statistically significant: 95% CIs for accuracy at pretest overlap with those at the immediate posttest and those at the delayed posttest. Similarly, the CIs at the immediate posttest overlap with those of the delayed posttest. The non-significance of these differences is confirmed by the Friedman test (see Table 4.3.4).

Figure 4.3.1
Accuracy for the Picture Matching Task by testing time

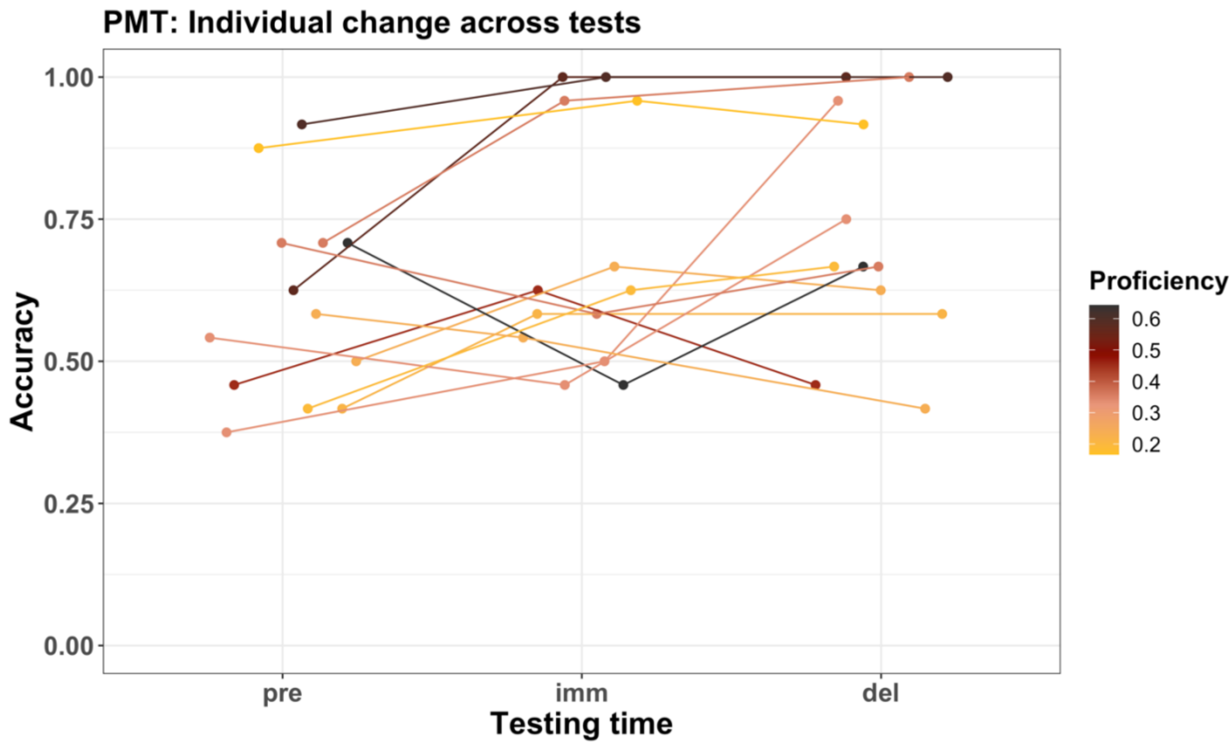


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

In terms of individual performance, Figure 4.3.2 shows that participants who started out with higher accuracy on the picture matching task at pretest are also those participants who increased to immediate posttest and maintained their accuracy to delayed posttest. For participants who started out with accuracy scores between 0.25 and 0.75, accuracy mostly stayed within that range at immediate posttest and delayed posttest with the exception of two participants, although a slight upward trend in noticeable for most others as well.

Figure 4.3.2

Graphic representation of individual participant's accuracy across testing times



Note. Colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

4.3.2.2.2 Picture Description Task (Auxiliaries)

Descriptive results for auxiliary accuracy on the picture description task can be found in Table 4.3.2. The results are also visually represented in Figures 4.3.3 and 4.3.4.

Table 4.3.2

*Descriptive statistics for auxiliary accuracy (*sein *only) on the Picture Description Task*

	M (SD)	Bootstrapped 95% CI	Mdn	IQR
Pretest	0.12 (0.32)	[0.0, 0.31]	0.0	0.0
Immediate posttest[a]	0.38 (0.48)	[0.14, 0.68]	0.0	0.91
Delayed posttest	0.76 (0.43)	[0.51, 0.90]	1.0	0.50

[a]For the immediate posttest, $N = 12$

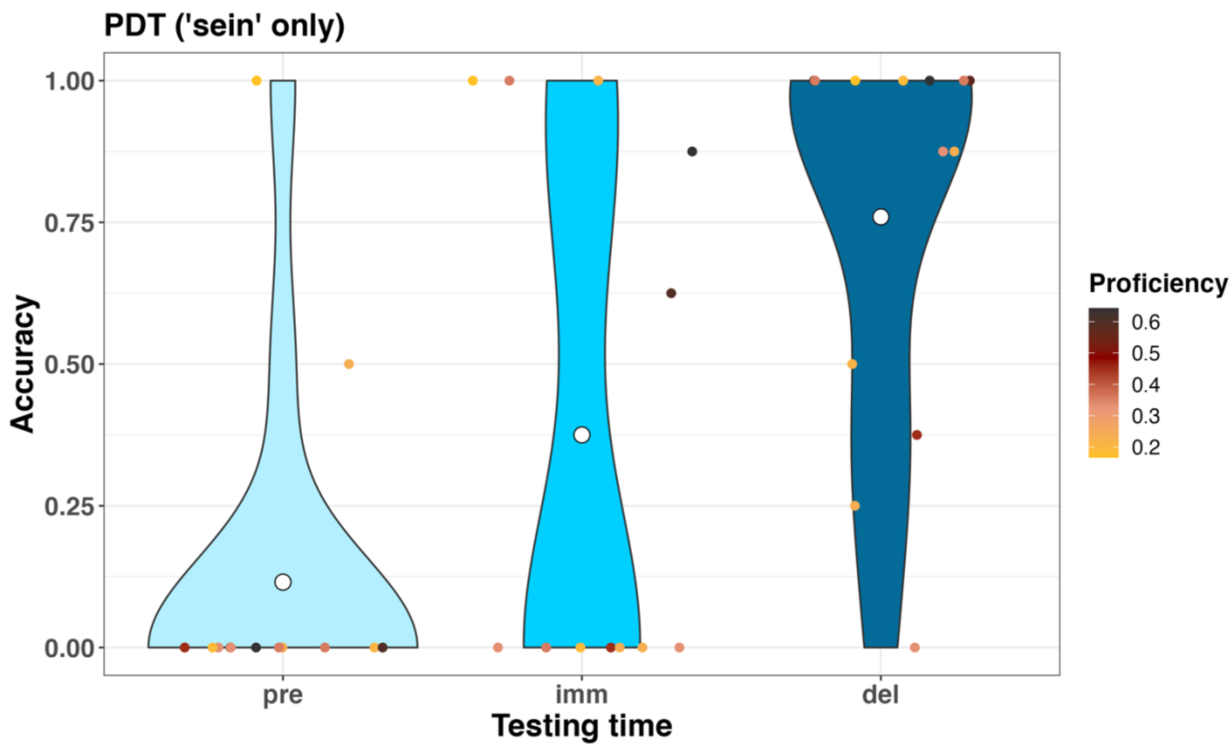
Descriptively, participants increased 26% from pretest to immediate posttest as well as 38% from immediate posttest to delayed posttest, as indicated by the mean accuracy in Table 4.3.2. These increases can also be traced in Figure 4.3.3, where the white dots represent mean accuracy at each testing time[7]. Figure 4.3.3 also indicates that the distribution of the data points changed,

[7] As in Experiment 1, separate visualizations showed that the increase in accuracy on *sein* items across testing times did not lead to a decrease in accuracy on *haben* items — in fact, accuracy also descriptively increased over time. This means that the increase in accuracy on *sein* items is not a result of overgeneralization.

with the violin shape at pretest being widest at floor, but widest at ceiling at delayed posttest. The Friedman test was significant, and follow-up Wilcoxon signed-rank tests indicated that the change in accuracy from pretest to delayed posttest was significant, whereas the changes from pretest to immediate posttest and from immediate to delayed posttest were not (see Table 4.3.4). This is confirmed by the non-overlapping CIs between pretest and delayed posttest, whereas all other CIs overlap (see Table 4.3.2). The effect size for the significant change was large.

Figure 4.3.3

Accuracy of sein *auxiliaries for the Picture Description Task by testing time*

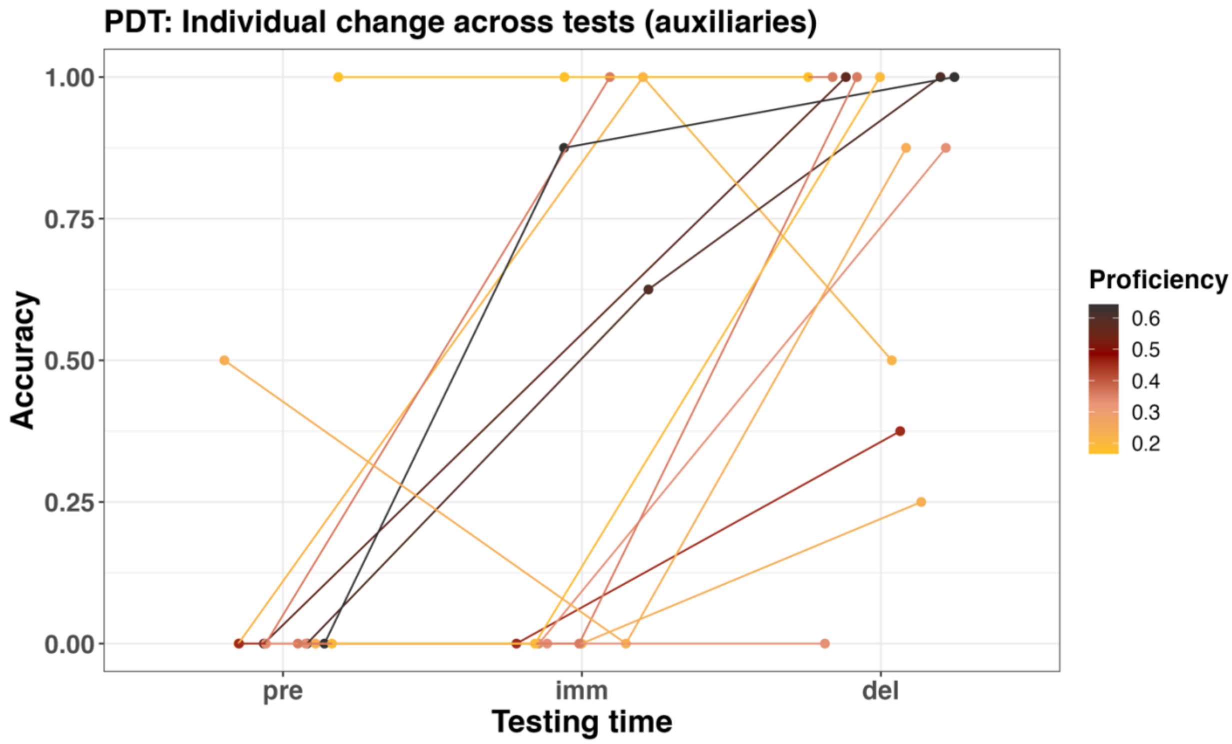


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

In terms of individual performance, Figure 4.3.4 shows that while some participants increased from pretest to immediate posttest and scored above 0.5 accuracy, most participants were still at floor at immediate posttest. This changes from immediate posttest to delayed posttest: Most of those who increased from pretest to immediate posttest increased even more, and most of those who were still at floor at immediate pretest then increased in accuracy on the delayed posttest.

Figure 4.3.4

Graphic representation of individual participant's accuracy across testing times



Note. Colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

4.3.2.2.3 Picture Description Task (Participles)

Descriptive results for participle accuracy on the picture description task can be found in Table 4.3.3. The results are also visually represented in Figures 4.3.5 and 4.3.6.

Table 4.3.3

Descriptive statistics for participle accuracy on the Picture Description Task

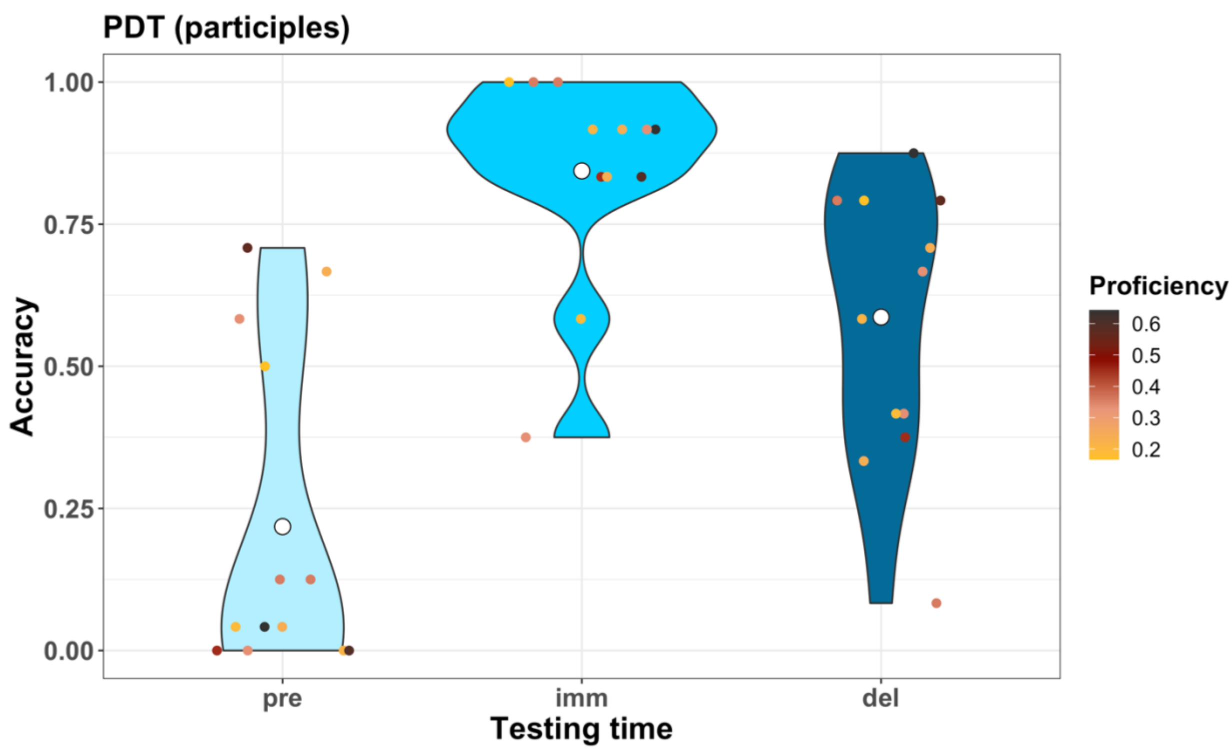
	M (SD)	Bootstrapped 95% CI	Mdn	IQR
Pretest	0.22 (0.41)	[0.08, 0.39]	0.04	0.50
Immediate posttest[a]	0.84 (0.36)	[0.68, 0.92]	0.92	0.10
Delayed posttest	0.59 (0.49)	[0.45, 0.69]	0.67	0.38

[a]For the immediate posttest, $N = 12$

Descriptively, participants increased 62% from pretest to immediate posttest, but also decreased in accuracy by 25% from immediate posttest to delayed posttest, as indicated by the mean accuracy in Table 4.3.3. These changes can also be traced in Figure 4.3.5, where the white dots represent mean accuracy at each testing time. Figure 4.3.5 also indicates that the distribution of the data points drastically changed from pretest to immediate posttest, with most participants performing close to ceiling at immediate posttest with only a couple of outliers. The distribution is much wider again at delayed posttest. The Friedman test was significant, and follow-up Wilcoxon signed-rank tests indicated that the increases in accuracy from pretest to immediate and delayed posttest were significant (see Table 4.3.4). The effect size for the change from pretest to immediate posttest was large, while the change from pretest to delayed posttest produced a small effect size. The decrease from immediate to delayed posttest was significant as well (see Table 4.3.4), although the effect size was negligible. These results are confirmed by the (mostly) non-overlapping CIs across testing times (see Table 4.3.3).

Figure 4.3.5

Accuracy of participles on the Picture Description Task by testing time

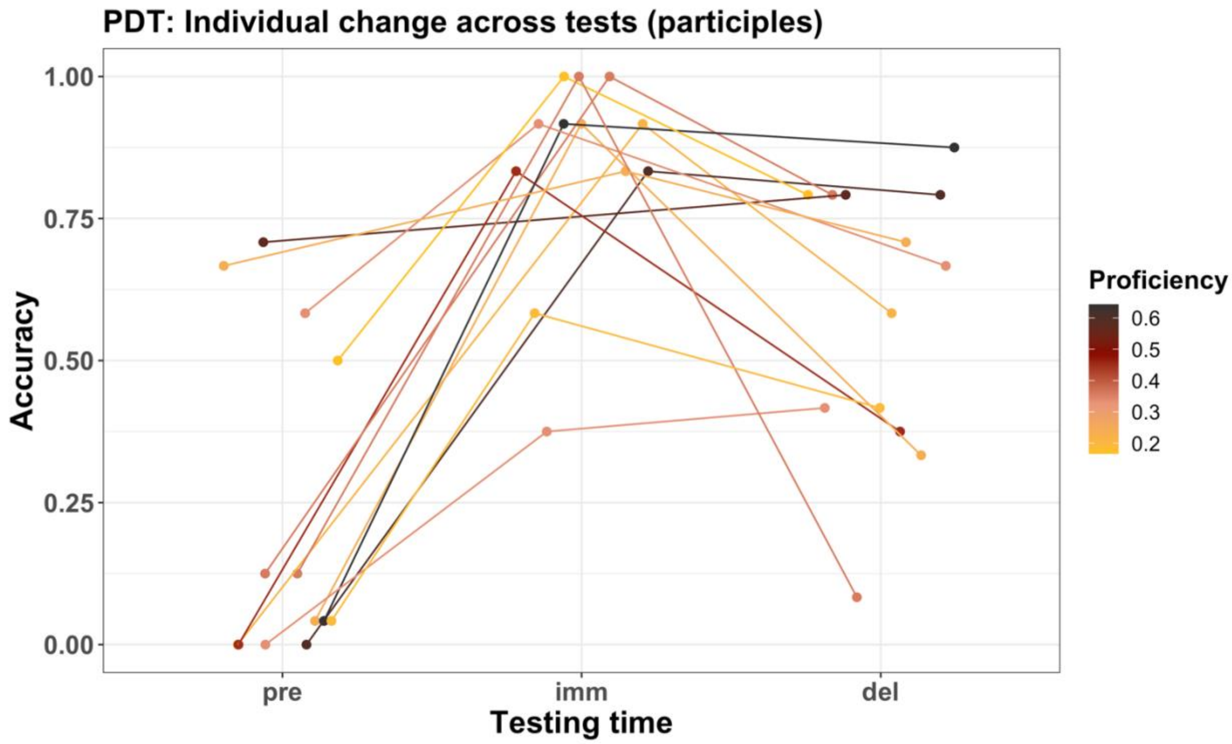


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

In terms of individual performance, Figure 4.3.6 shows that there were large increases from pretest to immediate posttest for all participants, including those who started out close to floor. These gains, however are not maintained by the delayed posttest, with accuracy decreasing for most participants from immediate to delayed posttest. Nevertheless, most participants were still more accurate at delayed posttest than at pretest.

Figure 4.3.6

Graphic representation of individual participant's accuracy across testing times



Note. Colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Table 4.3.4

Results from the Friedman and Wilcoxon signed-rank tests for the Picture Matching and Picture Description Tasks

	Friedman Test		Pretest - Immediate posttest			Pretest – Delayed posttest			Immediate-Delayed posttest		
	χ^2	*p*	*z*	*p*	*d*	*z*	*p*	*d*	*z*	*p*	*d*
PMT	4.04	.132	-	-	-	-	-	-	-	-	-
PDT (auxiliaries)	13.94	< .001*	-1.63	.104	0.64#	-2.79	.005*	1.69###	-2.02	.043	0.84#
PDT (participles)	20.17	< .001*	-3.03	.002*	1.60###	-2.94	.003*	0.82#	-2.86	.004*	-0.58#

* $p < .0167$. # = small effect size; ## = medium effect size; ### = large effect size

4.3.2.3 Summary: Improvement on Outcome Measures

For the picture matching task, while there were descriptive increases from pretest to immediate and delayed posttest, these differences did not reach statistical significance. The visualization of individual participants' performance shows that participants who started out with higher accuracy at pretest increased to immediate posttest and maintained this accuracy at delayed posttest. For participants who started out with slightly lower accuracy, a slight upward trend can be noted from pretest to immediate and delayed posttest.

For auxiliary accuracy on the picture description task, participants significantly increased in accuracy from pretest to delayed posttest (64% increase). Most participants started out at floor and many were still at floor at immediate posttest. However, most were at ceiling by the delayed posttest, indicating a delay in knowledge consolidation. These results are slightly different from those of the picture matching task, were there were descriptive increases but also larger within-group variation, which prevented these increases from reaching statistical significance.

For participle accuracy on the picture description task, results show a significant increase from pretest to immediate posttest, as well as from pretest to delayed posttest. However, accuracy also significantly decreases between immediate and delayed posttest. In terms of individual participants' performances, the behavior is more uniform, with all participants increasing from pretest to immediate posttest, and all participants decreasing from immediate to delayed posttest. These findings differ slightly from those of auxiliary production, where accuracy increases post-training, and the statistically significant increase is noted from pretest to delayed posttest only.

4.3.3 Improvement during Training: Behavioral Data

4.3.3.1. Analyses

Analyses for the behavioral training data — i.e., picture choice and participle typing — mirror those of the outcome measures, using robust descriptive statistics and non-parametric tests for statistical comparison. For the participle analysis, only those items where participants had correctly chosen the picture before typing the corresponding participle were included in the analyses, leading to the exclusion of 28.4% of all data points, i.e., 71.6% of all training data were submitted to the participle analysis. Changes in auxiliary accuracy and participle accuracy were then compared between each of the four training blocks using non-parametric tests. For these tests, the alpha level was Bonferroni-adjusted to .008 (0.05/6 comparisons).

4.3.3.2 Results

4.3.3.2.1 Auxiliaries

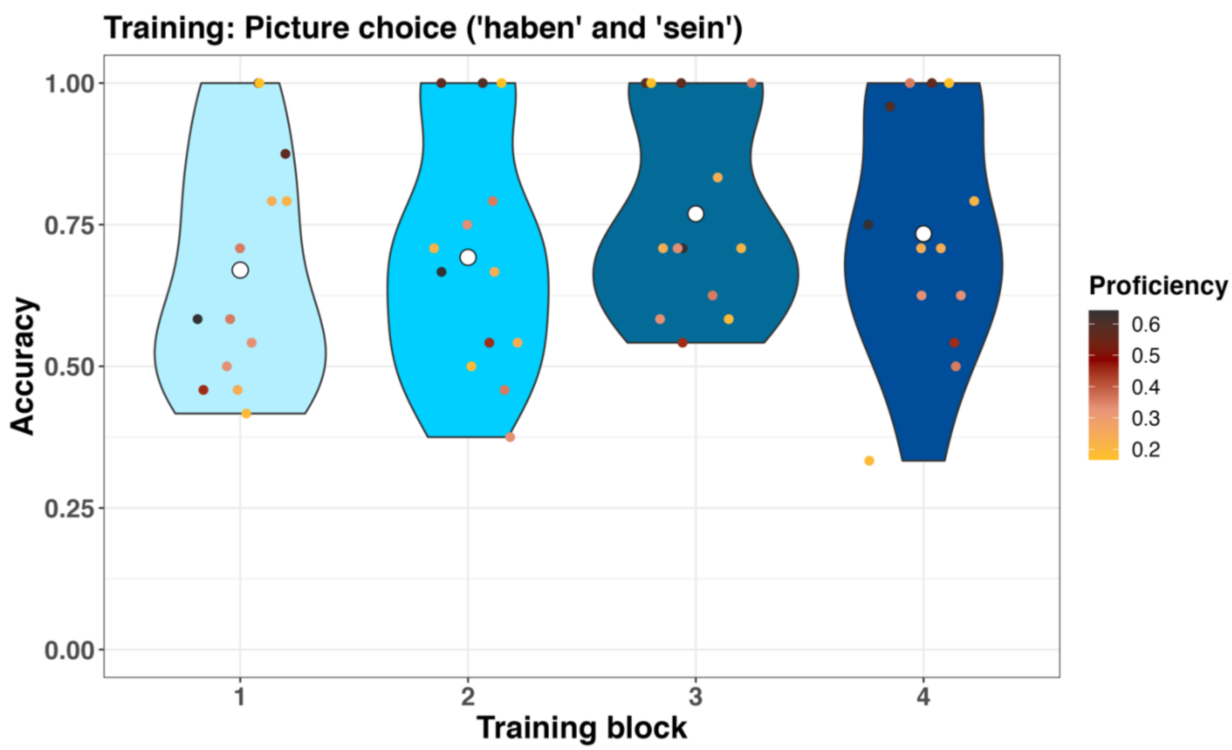
Descriptive results for auxiliary accuracy during training can be found in Table 4.3.5. The results are also visually represented in Figures 4.3.7 and 4.3.8. In all tables and graphs, accuracy is calculated as the ratio of correct items to all items on any given test. For instance, if a participant chose the correct picture on 18 out of 24 items during the first block of training, their accuracy score corresponds to 0.75 for Block 1.

Descriptively, participants increased 6% from Block 1 to Block 4, although the increase was as high as 10% from Block 1 to Block 3, as indicated by the mean accuracy in Table 4.3.5. These increases can also be traced in Figure 4.3.7, where the white dots represent mean accuracy at each testing time. However, these increases are not statistically significant: 95% CIs for auxiliary accuracy overlap across all training blocks. The non-significance of these differences is confirmed by the Friedman test (see Table 4.3.7).

Table 4.3.5
Descriptive statistics for auxiliary accuracy during training

	M (SD)	Bootstrapped 95% CI	Mdn	IQR
Block 1	0.67 (0.47)	[0.57, 0.78]	0.58	0.29
Block 2	0.69 (0.46)	[0.58, 0.81]	0.67	0.25
Block 3	0.77 (0.42)	[0.68, 0.87]	0.71	0.38
Block 4	0.73 (0.44)	[0.62, 0.85]	0.71	0.33

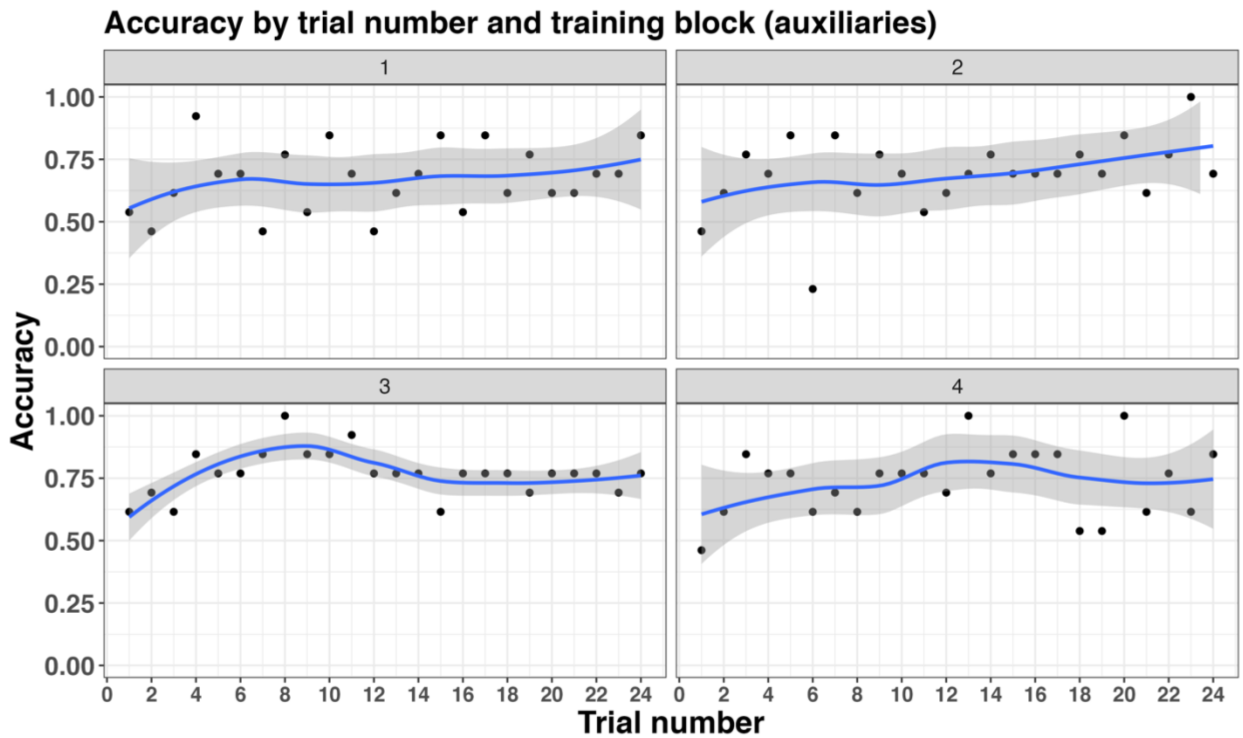
Figure 4.3.7
Accuracy of auxiliaries during training, by training block



Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

When inspecting performance within each training block across individual trials (Figure 4.3.8), it becomes apparent that accuracy seems to increase across trials within each training block. However, one can also see that the starting and end-point in accuracy is similar across all blocks: Accuracy on the first trial is somewhere around 50%, and accuracy on the last trial is somewhere around 75%. This visual inspection again confirms the statistically non-significant differences in auxiliary accuracy between training blocks.

Figure 4.3.8
Accuracy of auxiliaries during training by trial number and training block



Note. Lines represent smoothed conditional means, gray areas represent 95% CIs around smoothed conditional means.

4.3.3.2.2 Participles

Descriptive results for participle accuracy during training can be found in Table 4.3.6. The results are also visually represented in Figures 4.3.9 and 4.3.10.

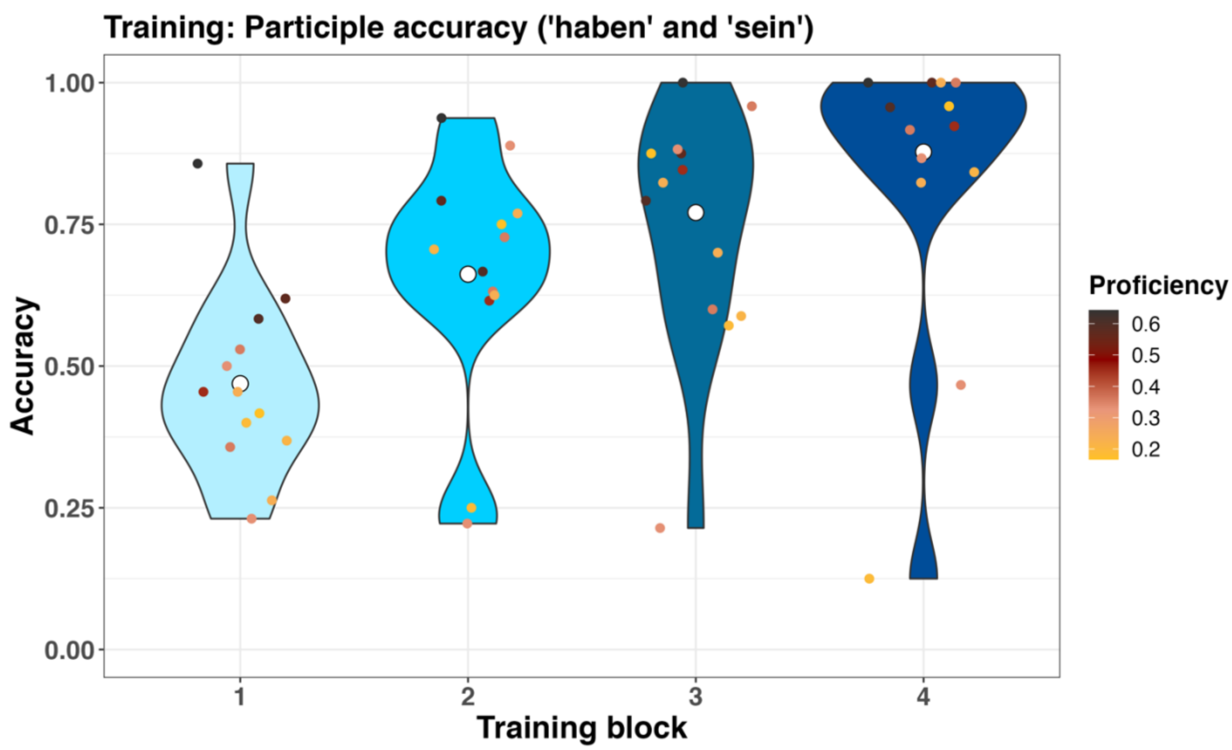
Table 4.3.6
Descriptive statistics for participle accuracy during training

	M (SD)	Bootstrapped 95% CI	Mdn	IQR
Block 1	0.47 (0.50)	0.39, 0.57	0.46	0.16
Block 2	0.66 (0.47)	0.53, 0.75	0.71	0.14
Block 3	0.77 (0.42)	0.61, 0.84	0.82	0.28
Block 4	0.88 (0.33)	0.62, 0.93	0.92	0.16

Descriptively, participants increased by 41% from Block 1 to Block 4, as indicated by the mean accuracy in Table 4.3.6. These increases can also be traced in Figure 4.3.9, where the white dots represent mean accuracy at each testing time. In contrast to the changes in auxiliary accuracy during training, some of these increases for participle accuracy are also statistically significant, as indicated by the follow-up Wilcoxon signed-rank test to the significant Friedman test: Accuracy significantly increased from Block 1 to all subsequent blocks, as well as from Block 2 to Block 4 (see Table 4.3.7). These significant differences are confirmed by the absence of overlapping CIs between these training blocks (with the exception of a small overlap between Blocks 1 and 2; see Table 4.3.6).

Figure 4.3.9

Accuracy of participles during training, by training block

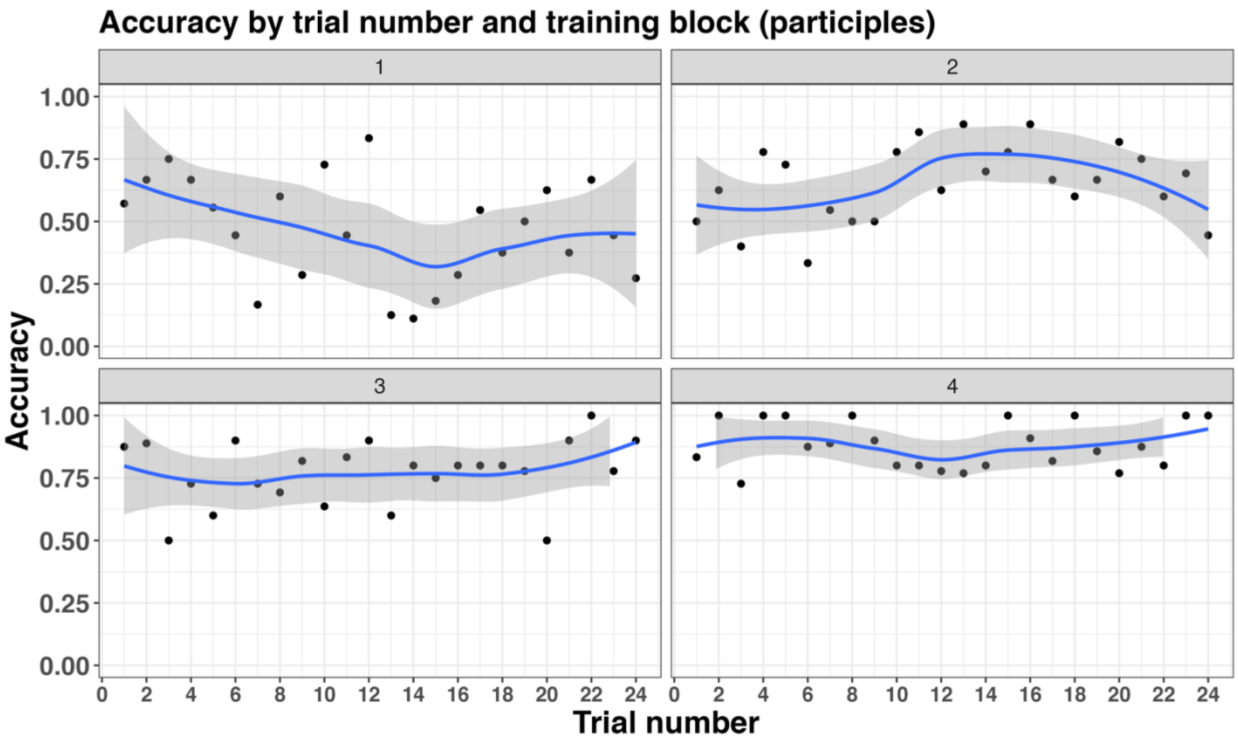


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

When inspecting performance within each training block across individual trials (Figure 4.3.10), it becomes apparent that accuracy wavered around chance — and lower than chance — in the first training block. While accuracy still hovered around chance at the beginning and end of the second block, there was also a slight increase around the items in the middle of the block, increasing to as high as 82.5%. However, Blocks 3 and 4 is where you see a stronger difference compared to Blocks 1 and 2, where accuracy was more uniform across trials, as indicated by the more narrow 95% confidence intervals. In addition, accuracy seems higher in these blocks than in Block 1 or 2, with accuracy ranging between 50% and 100% in Block 3 and between 71% and 100% in Block 4. This visual inspection again confirms the statistically significant differences in participle accuracy between training blocks.

Figure 4.3.10

Accuracy of participles during training by trial number and training block



Note. Lines represent smoothed conditional means, gray areas represent 95% CIs around smoothed conditional means.

Table 4.3.7

Results from the Friedman and Wilcoxon signed-rank tests for the picture choice and participle writing during training

	Friedman test		Block 1 – Block 2			Block 1 – Block 3			Block 1 – Block 4		
	χ^2	p	z	p	d	z	p	d	z	p	d
Picture choice	6.75	.080	-	-	-	-	-	-	-	-	-
Participle writing	20.49	<.001*	-2.73	.006*	0.39	-3.07	.002*	0.65#	-2.94	.003*	0.97##
			Block 2 – Block 3			Block 2 – Block 4					
			z	p	d	z	p	d			
			-1.75	.008†	0.25	-2.87	.004*	0.54#			
			Block 3 – Block 4								
			z	p	d						
			-1.84	.065	0.29						

* $p < .008$. # = small effect size; ## = medium effect size; ### = large effect size

4.3.3.3 Summary: Improvement during Training: Behavioral Data

For the picture choice during training, participants descriptively increased in accuracy by about 6% from Block 1 to Block 4. However, neither this increase nor the other comparisons between training block were statistically significant, and a lot of variation remained across participants. The visualization of the accuracy of individual trials within each block confirms this non-significance: Accuracy increased slightly from the first trial of a training block to the final trial, but the starting points and end points, as well as the slope of the (mostly) linear increase, are comparable across all blocks. These findings parallel the findings for auxiliary accuracy on the picture choice task from pretest to immediate and delayed posttest, where descriptive increases also remained statistically non-significant.

For producing the participle following the picture choice during training, participants descriptively increased in accuracy by about 41% from Block 1 to Block 4. In fact, increases from Block 1 to all other training blocks were significant, as well as from Block 2 to Block 4. Again, the visualization of the accuracy of individual trials within each block confirms these findings: Accuracy in Blocks 2, 3, and 4 is higher than in Block 1, while Blocks 3 and 4 look very similar in starting points, end points, and slope of the line. In addition, 95% confidence bands become narrower in later training blocks. While this finding contrasts with that for auxiliary accuracy during training, it mirrors the findings for participle accuracy on the picture description task from pretest to immediate and delayed posttest, where descriptive increases also reached statistical significance.

4.3.4 Eye Movements during Picture Choice

4.3.4.1 Analyses

The sample data were exported using the EyeLink Data Viewer software package version 3.2.1 (SR Research Ltd., 2018), relative to the onset of the audio file played during the picture choice task in each trial. The sample data were then processed using the *VWPre* package version 1.2.4 (Porretta, 2020) in R version 4.1.0 (R Core Team, 2021). First, all samples containing blinks and offscreen looks were removed from the data, leading to the exclusion of 5.2% of data points (blinks: 4.6 %, offscreen looks: 0.6 %). Following Barr (2008) and Godfroid (2020) looks that did not fall into any of the interest areas were not excluded from the data — instead, they were coded as looks to the "background". As a second step, any trials for which more than 50 % of the samples were removed in the first step were considered unreliable and were therefore also removed, leading to the exclusion of 12 trials, or another 0.3% of data points. Finally, I inspected the data to check if there were any participants for whom more than 70 % of all trials had been removed and who should thus also be excluded. However, no participant fit this category. Overall, 94.5% of all data points therefore remained for analysis and visualization. Following sample removal, the remaining data were binned into 50 ms time bins — each bin containing 25 samples — effectively down-sampling the data to a sampling rate of 20 Hz. The data in each bin were then converted to proportions indicating whether looks within each bin fell onto the target image, the foil image, or the background for each given trial. Proportions were calculated as looks to an interest area divided by all onscreen looks — i.e., looks to the target, foil, and the background — within a given time bin (Barr, 2008; Godfroid, 2020). The data were then plotted as time course from the onset of the target auxiliary in the audio file (which was aligned to the zero time point in Figure 4.3.11) to 3000 ms post-onset, showing only the proportion of looks to the target or foil image. I did not plot looks that fell outside of these regions of interest. Proportions therefore do not necessarily sum to 1.0. Upon inspection of plots for individual participants, one participant was excluded because they showed extremely low proportions of looks to the target and the foil picture during the relevant time window.

Due to the small number of participants and the low statistical power, data were not analyzed using statistical modeling. If more data is collected in the future, these will be analyzed using growth-curve modeling (see Mirman, 2014).

4.3.4.2 Results

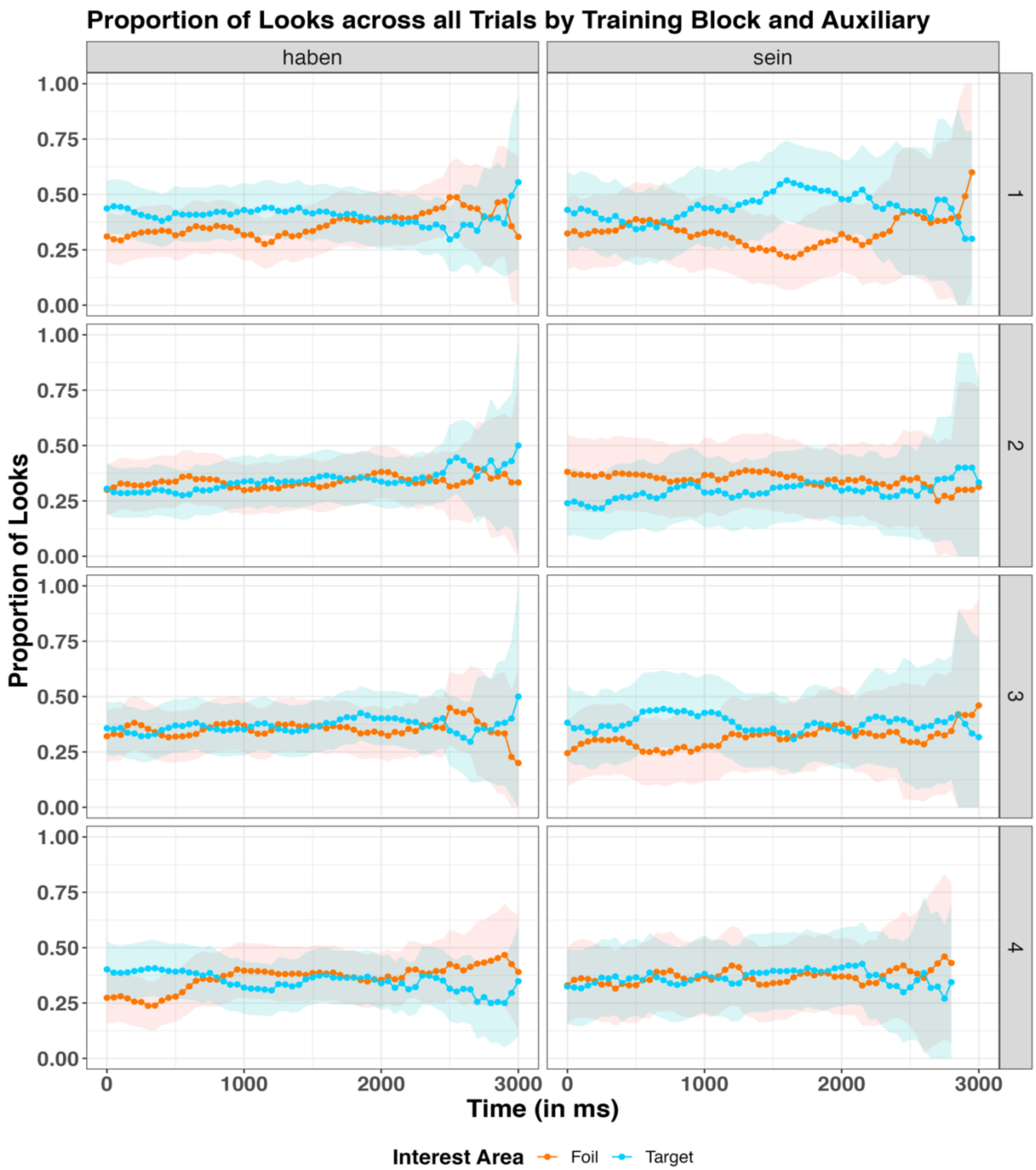
Visual representations of the time course of the proportions of looks to the target picture and the foil picture as a function of training block and auxiliary type can be found in Figure 4.3.11. As a reminder, the plot shows the time period from the onset of the auxiliary to 3000 ms post-onset. In this time period (in fact, even at the end of the auditory stimulus in each trial), participants never heard the participle of the target verb in the auditory stimulus, as only sentence fragments with the participle missing were presented to participants.

For *haben* auxiliaries, proportions of looks to the target and the foil picture are roughly equivalent in each training block, hovering between 0.25 and 0.5 across the different time bins, with some variability within and between blocks. 95% confidence intervals around the lines plotted to indicate the proportion of looks to the target and foil image in each training block overlap during the entirety of the time courses plotted, indicating that statistically significant differences between the proportion of looks to the target and foil images most likely do not exist. This means that for sentences with *haben* as their auxiliary, participants were equally likely to look to the target and foil pictures throughout the entirety of the auditory sentence presentation across the training task.

For *sein* auxiliaries, proportions of looks to the target and the foil picture also hover between 0.25 and 0.5 across the time bins. In Block 1, looks to the target picture are more frequent from 750 to around 2500 ms. In fact, the 95% confidence intervals around the lines do not overlap at around 1600 ms, indicating that there might be statistically significant differences in the proportion of looks to target compared to foil pictures around this time point (which roughly corresponds to the end of the first adverbial phrase included in the sentence fragments). However, this difference is reversed in Block 2, where looks to the foil are more likely from 0 to 1750 ms post-onset, and disappears in Block 4, where the plotted lines virtually overlap during the whole time course. Block 3, on the other hand, shows a temporary increase in looks to the target image from around 500 ms to 1250 ms post-onset, before the proportion of looks to each picture converges again.

Figure 4.3.11

Time course of looks to the target and foil pictures as a function of training block and auxiliary (from onset of auxiliary, i.e., 0 ms, to 3000 ms post-onset[8]). Proportions do not sum to 1.



[8] For *sein* items in Block 1 and Block 4, the time course is only plotted up to 2950ms and 2800ms post-onset, respectively. This is due to the fact that some auditory stimuli for *sein* items were only around 3000ms long overall. With Figure 1 plotting the proportion of looks from the onset of the auxiliary rather than the onset of the sentence fragment, the relevant time frame to be plotted is therefore shorter for some *sein* items in Block 1 and Block 4.

4.3.4.3 Summary: Eye Movements during Picture Choice

During the auditory stimulus presentation, in which the auxiliary functions as a predictive cue to which of the two pictures presented on screen is the target, looks to the target and foil picture were roughly equal for *haben* items in all blocks. For *sein* items, the proportion of looks to the target picture was temporarily higher in Block 1 than for looks to the foil picture, with the largest difference occurring at around 1600ms. In subsequent blocks, the proportions of looks to target and foil pictures were more similar again. However, overall, the proportion of looks to the target picture — and the foil picture — remained relatively low, and the width and overlap of the error bands representing 95% confidence intervals show that participants were equally likely to look at either the target or the foil picture in most blocks.

4.3.5 Pupil Size during Feedback Presentation

4.3.5.1 Analyses

The pupil size data were exported using the EyeLink Data Viewer software package version 3.2.1 (SR Research Ltd., 2018), relative to the onset of the feedback presentation in each trial. The data corresponding to 3000 ms of feedback presentation were exported in 150 time bins of 20ms (10 samples per bin). The binned data were then further processed using self-developed analysis scripts in R version 4.1.0 (R Core Team, 2021). Data from each participant's right eye were used for analysis. Missing data points resulting from off-screen looks, as well as looks that fell outside of the interest area corresponding to the feedback presented on screen were excluded from averaging in order to prevent changes in pupil size due to eye movements to impact the analysis. This resulted in the exclusion of 4.1% of all data points. In addition, any trials in which more than 33% of all data points were in a blink were excluded as well. This resulted in the further exclusion of 5.0% of data points, such that 91.9% of all data remained for analysis and visualization. For these remaining data, missing data points caused by blinks were linearly interpolated using the *na_approx* function from the *zoo* package version 1.8-10 in R (Zeileis et al., 2022). The binned and cleaned pupil data were then baseline-corrected on a trial-by-trial basis. A time period of 500ms immediately prior to the feedback onset was chosen as the baselining period (data in the baselining period were cleaned with a similar procedure as described above). Pupil size data within these 500ms were then averaged for each trial to create a baseline pupil size value. Then, the averaged baseline pupil size was subtracted from the average pupil size in each time bin of the feedback presentation period (see Mathôt et al., 2018). Finally, the difference value was divided by the averaged baseline pupil size to obtain a measure of percent change from baseline that is unaffected by changes in pupil size that last longer than an individual trial. This procedure is a standard approach for pupillometry data (Preuschoff et al., 2011; Lavin et al., 2014; Mill et al., 2016). A positive value for the percent change then corresponds to an increase in pupil size compared to baseline, whereas a negative value indicates a decrease in pupil size compared to baseline. These were the main pupil response measures reported in this experiment.[9]

Similar to the eye movement data, the pupil size timeseries were not analyzed using statistical modeling. Instead, the time course data and differences between conditions are described in prose. If more data is collected in the future, the time course data of pupil size change will be subjected to growth curve analysis (see Mirman, 2014).

---

[9] Mean pupil size as well as peak amplitude and latency can also be used to summarize pupil size differences between training blocks, auxiliaries, and error type. However, their utility is limited, as they do not allow to analyze the shape of these changes over the time-course of feedback presentation. Nevertheless, these measures are reported in Appendix G.

4.3.5.2 Results

Visual representations of the pupil size change and trajectory during feedback presentation can be found in Figure 4.3.12, which shows pupil size changes from feedback onset to 3000ms post-onset by training block, auxiliary, and error type. Error type hereby refers to the type of error that a participant may have committed in a given trial when completing the training task: "all correct" indicates that participants chose the correct picture and produced the correct form of the target participle; "all incorrect" indicates that participants chose the incorrect picture and therefore also produced an incorrect participle; "choice correct, participle incorrect" indicates that participants chose the correct picture but produced an incorrect form of the target participle. It is important to keep in mind that the number of data points contributing to each error type changes over the course of the four training blocks, as participants become more accurate over time (see Table 4.3.8).

Table 4.3.8

Proportion of error type per training block

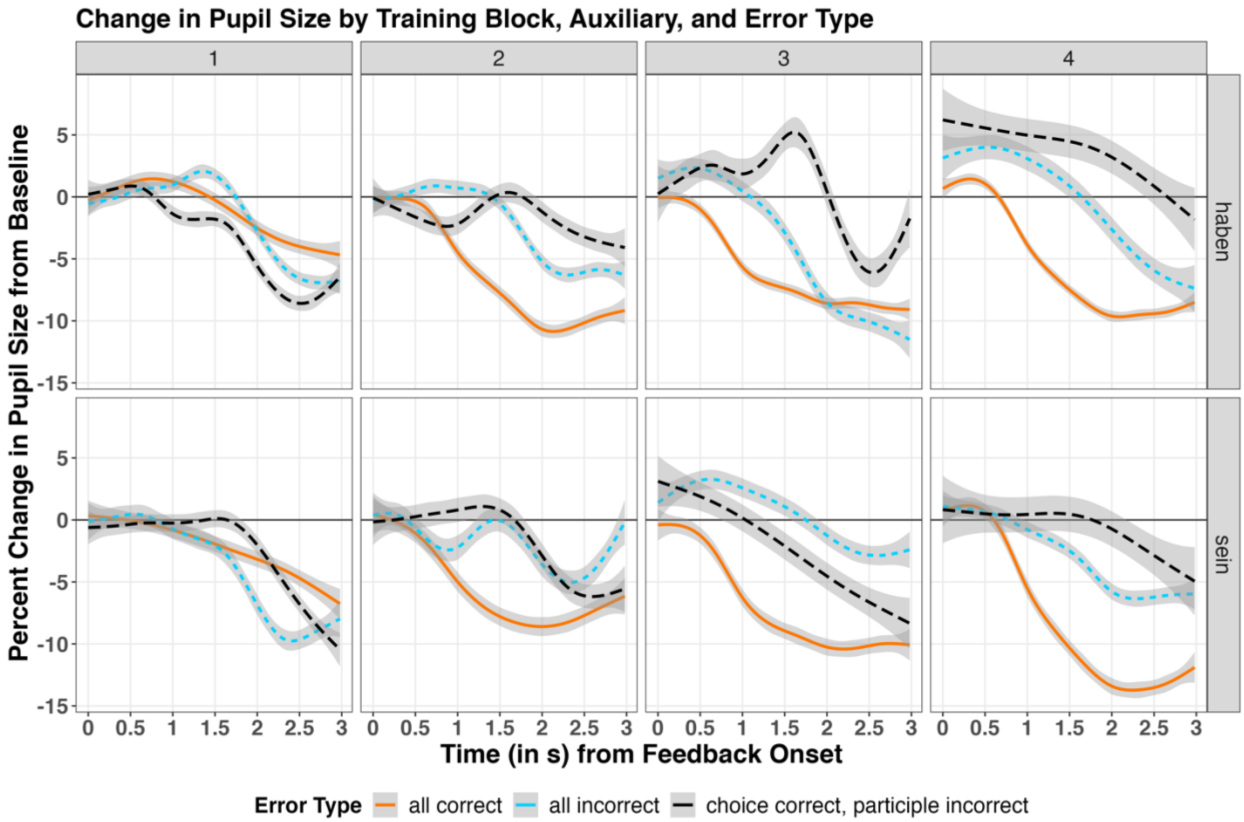
	all correct	choice correct, participle incorrect	all incorrect
Block 1	0.31	0.36	0.33
Block 2	0.48	0.22	0.30
Block 3	0.62	0.17	0.21
Block 4	0.68	0.08	0.24

The comparison of primary interest is between pupil responses evoked by feedback which indicated that participants had been correct in their picture choice and in the participle they typed ("all correct"; solid orange line in Figure 4.3.12) and pupil responses evoked by feedback which indicated that participants had been incorrect in both their picture choice and in the participle they typed ("all incorrect"; dashed blue line in Figure 4.3.12).[10] As can be seen in Figure 4.3.12, the pupillary response is similar for these two error types in the first block of training for both *haben* and *sein* items. However, the difference between these two lines increases as the training progresses, with pupil responses to errors of the type "all incorrect" almost always being larger than pupil responses to "all correct" items. In addition, this difference seems to be larger for *sein* items compared to *haben* items, especially in Blocks 3 and 4.

[10] Errors of the type "choice correct, participle incorrect" are of secondary interest, as it is unclear to what extent participants can or do make a distinction in degree of error compared to errors of the type "all incorrect" when being presented with the feedback.

Figure 4.3.12

Percent change in pupil size from baseline by training block, auxiliary, and error type



Note. Lines represent smoothed conditional means, gray areas represent 95% CIs around smoothed conditional means.

While only of secondary interest, errors of the type "choice correct, participle incorrect" (dashed black line in Figure 4.3.12) pattern mostly with errors of the type "all incorrect", in that pupil size in response to "choice correct, participle incorrect" errors are larger than for "all correct" items, with the exception of Block 1. In Block 3, differences emerge between *haben* and *sein* items with regard to errors of this type: For *haben* items, the dashed black line is consistently above the dashed blue and solid orange lines, whereas for *sein* items, it is between the two. In Block 4, on the other hand, "choice correct, participle incorrect" errors evoke the largest pupil responses for both *haben* and *sein* items again. Keep in mind, however, that only few data points contribute to this error type in Blocks 3 and 4, as participants become increasingly more accurate during training (see Table 4.3.8) — this is also visible in the relatively large error bands around the dashed black lines in Blocks 2, 3, and 4 compared to the other lines.

4.3.5.3 Summary: Pupil Size during Feedback Presentation

The results of the analyses for pupil size changes during feedback processing show that participants processed the feedback they received, but that the pupillary signature of this processing is dependent on whether participants committed an error prior to receiving said feedback. Different pupillary responses emerged in trials where participants were "all correct" compared to when they committed errors. While pupil size decreased overall during feedback presentation, the decrease was less pronounced for trials in which participants committed an error compared to when they made a correct picture choice and produced the correct target participle form — in other words, pupil sizes were larger during feedback presentation after errors than when participants had been "all correct". These differences became more pronounced over the course of the four training blocks. In addition, the difference in pupil size between "all correct" trials and "all incorrect" trials in the second half of the training (Blocks 3 and 4) were larger for *sein* items than for *haben* items.

4.3.6 Relationship between Pupil Size Changes and Outcome Measures

As pupil size during the encoding of feedback has been found to correspond to the degree of learning as exhibited by the increase in knowledge from pretest to posttest (e.g., Brod et al., 2018), the relationship between pupil size during feedback presentation and learning from pretest to posttests on the picture matching and picture description task as well as during training were of interest to this dissertation as well. Such a relationship would appear as a positive correlation between the difference scores for pupil size during feedback presentation for different error types and difference scores on the outcome measures going from pretest to immediate and delayed posttest or between the different training blocks (see Brod et al., 2018). However, due to the low number of participants in the current experiment, no such analyses were conducted. If more data is collected in the future, correlational analyses could be conducted to investigate this question. More specifically, these correlational analyses would investigate the relationship between the difference in pupil size for items in the fourth training block where participants were all correct and those where participants were all incorrect, and the difference between pretest and immediate posttest scores on a given outcome measure, i.e., auxiliary accuracy on the picture matching and description tasks as well as participle accuracy on the picture description task.[11]

4.3.6.1 Relationship between Learning Aptitude and Outcome Measures

Participants also completed the SRT task and the LLAMA F, as implicit and explicit learning aptitude may have an impact on participants' ability to learn from an error-based learning task without explicit instruction (see Granena, 2016). Such an impact could also be detected using correlational analyses between learning aptitude scores and improvements from pre- to posttests on both picture matching and picture description tasks as well as pupil size changes during feedback processing. As described above, given the small participant sample and the resulting low statistical power, no such analysis could be performed in this dissertation. Nevertheless, the results of both the SRT task and the LLAMA F are reported in Appendix H. If more data is collected in the future, two types of correlational analyses could be carried out: 1) between learning aptitude scores and pupil size differences as described above and 2) between learning aptitude scores and outcome measure difference scores (pretest to immediate posttest). These analyses would further our understanding of the roles of implicit and explicit learning aptitude in the context of error-based learning and feedback processing.

[11] A secondary analysis might investigate the correlation between the difference in pupil size for items in the fourth training block where participants were all correct and the averaged pupil size change for all items where an error was committed prior to feedback presentation, i.e., items that were all incorrect and items where participants made a correct picture choice but then produced an incorrect form of the participle.

4.4 Discussion

The second experiment in this dissertation investigated whether participants begin to launch predictive looks to target pictures using a visual world paradigm that incorporates an error-based learning task, whether feedback processing changes over the course of the training task as measured by pupil dilations, and whether feedback processing correlates with the amount of learning exhibited on the learning task and the outcome measures. Findings for each of these topics of interest and their relevance for understanding error-based learning, in particular of feedback processing, in contexts of L2 learning are discussed below. For this discussion, it is important to keep in mind that the participant sample was fairly small as a result of the COVID-19 pandemic, so all interpretations should be made with caution.

4.4.1 Improvement during Training and on Outcome Measures

The first and second set of analyses described in Sections 4.3.2 and 4.3.3 analyzed whether participants improved on auxiliary and participle accuracy, both in the training task and in the picture matching and picture description tasks, administered during pretest, immediate posttest, and delayed posttest. During training, accuracy in choosing the target picture — i.e., using the auxiliary as a cue — did not significantly improve over the four blocks of training, and participants' accuracy descriptively only increased by about 6%. On the other hand, participle accuracy increased significantly over the four blocks of training, in particular from Block 1 to all other blocks. Descriptively, the increase from Block 1 to Block 4 was around 41%. This opposing pattern of findings was mirrored in the outcome measures. The results of the picture matching task showed no significant increase in accuracy from pretest to any posttest, and the descriptive increase from pretest to delayed posttest was around 15%. The results of the picture description task, on the other hand, follow the trends from the training task: Auxiliary accuracy for *sein* items increased significantly from pretest to delayed by about 64%. Participle accuracy increased significantly from pretest to immediate (62%) and delayed posttest (37%), but there was also a significant decrease going from immediate to delayed posttest (23%).

The results of both training and outcome measures show that participants did not increase in accuracy when it comes to using the auxiliary as a cue to choosing the correct target verb, i.e., in comprehending the relationship between the auxiliary and the target verb. However, when it comes to producing the auxiliary in the picture description task, as well as producing the participle during training and in the picture description task, participants did improve. It therefore seems that effects of creating expectations during training do not necessarily manifest themselves in tasks that are similar in mode (i.e., comprehension vs. production) to the training task, but rather that learning is transferred to a new task, in this case the production of the *Perfect* auxiliary alternation and of the corresponding participles. This is reminiscent of the findings in Experiment 1, where larger increases were found in the production task compared to the comprehension task over the course of the experiment.

For the picture description task, the significant increase in auxiliary accuracy did not appear immediately after training, but was delayed to two weeks post-training. This matches findings from Experiment 1, where the largest descriptive increase was from pretest to delayed posttest as well (although the increase to immediate posttest was significant in Experiment 1 as well). It therefore seems like the learning of the underlying rule to the auxiliary alternation, while kick-started by the training provided to learners, takes time to be consolidated and to become available for productive language use.

The pattern is slightly different for participle accuracy on the picture description task, where the strongest increase in accuracy was found on the immediate posttest. This increase, both during training and on the outcome measures, can be traced back to the feedback presented to participants, which in Experiment 2 only took the form of the target participle, presented to participants at the end of each trial. This feedback encoded information immediately relevant for the participle form, which likely led participants to focus on producing correct participle forms as they progressed through the training. Similar findings were obtained in Experiment 1 during training: When participants received form-focused feedback, i.e., the correct participle form, participle accuracy was higher in that particular training block than when they received semantic-focused feedback. In contrast, in Experiment 1, between-group differences never appeared on the immediate posttest, as both participant groups had received both types of feedback at that point. The decrease in participle accuracy from immediate to delayed posttest on the picture description task in the current experiment, on the other hand, is most likely due to the number of different participle forms corresponding to the target verbs, and the fact that the form of a participle does not follow a simple underlying rule in the context of this experiment. Rather, participants needed to retrieve grammatical information from the lexical entry for each individual verb, encoding whether it is a weak, strong, or mixed verb. While this knowledge may be highly activated and therefore easily retrievable during and immediately following the training — due to participle forms being presented as feedback— this activation likely decayed over the two-week period intervening between immediate and delayed posttest.

These findings are relevant to investigations of language learning through error-based paradigms, as they demonstrate the importance of the type of information encoded in the feedback that is given in response to an error (see also Hopp, 2020). Participants significantly increased in their participle accuracy during training and on the picture description task on posttest measures. Importantly, the feedback given to participants in this task focused on the participle form. Indirectly, the feedback also provided information whether participants had made a correct picture choice, i.e., whether they had correctly used the auxiliary to predict the upcoming verbal material. This information about participants' picture choice, however, is clearly secondary to the form of the participle — still participants also increased in auxiliary accuracy, but only in the productive mode. These findings then point to the fact that the type of information encoded in the feedback plays an important role in how learners can and do use the feedback provided to them to improve and update their language model.

4.4.2 Predictive Eye Movements in an Error-Based Learning Task

The third set of analyses, described in Section 4.3.4, investigated whether participants launched predictive looks to the target image during the picture choice portion of the training task. Due to the small sample size, no statistical analyses were performed. However, Figure 4.3.11, which visualizes the proportion of looks to the target and foil pictures by training block and auxiliary type, indicates that the proportion of looks to the target picture generally did not increase in either training block after the onset of the predictive cue, the auxiliary, in particular for *haben* items. For *sein* items, on the other hand, Figure 4.3.11 shows a somewhat higher proportion of looks to the target picture compared to the foil picture in Block 1 of the training task, at around 1600 ms post-auxiliary onset. This difference in proportion of looks to the target and foil picture, however, diminished in subsequent blocks, and the likelihood of participants looking at the target or the foil picture for *sein* items became more equal as the training progressed. In fact, only 5 out of 12 participants (41.6%) who completed the second debriefing questionnaire indicated that they decided prior to the end of the sentence fragment which picture best fit the incomplete sentence. Even for these 5 participants, however, they would not have necessarily launched predictive looks, as this decision could also be made without looking at the pictures at all.

Nevertheless, it is somewhat surprising that a difference in proportion of looks to target and foil pictures emerged in Block 1 for *sein* items, at the very beginning of the training task, and disappeared as participants advanced through the training. Why might that be the case? Remember that *Perfekt* in German uses *haben* as the default auxiliary and that participants overwhelmingly had English as their L1, where the Perfect is always formed with the auxiliary *to have*. Participants may therefore not have expected *sein* to appear in the auditory sentence fragments. In turn, this means that participants may have experienced a certain level of surprise upon noticing the auxiliary alternation, in particular the *sein* forms, leading to an increased proportion of looks to the target pictures (see Henry, 2022; Jackson et al., 2012). Remember also that participants in this experiment, as opposed to Experiment 1, were already familiar with the auxiliary alternation, as all of them had been formally introduced to it in their German courses prior to participating in the experiment. It is therefore not surprising that participants were able to identify many of the target pictures in trials that included *sein* as the auxiliary. As participants then progressed through the training, the occurrence of sentence fragments with *sein* auxiliaries could have become less surprising or attention-grabbing, and as a result, participants may have abandoned the redirection of looks to the target *sein* items. Equal proportions of looks to the target and foil picture after the first block of training — for both *sein* and *haben* items — may then simply reflect an effect of adaptation to the task, where *sein* items become less surprising as participants advance in the training task, rather than an indication that participants became increasingly unable to launch predictive looks to the target picture. However, it cannot be completely ruled out that there are no other factors at play that impact the proportion of looks to the target and foil pictures, and that participants may indeed regress in their ability to launch predictive looks as the training progresses. Future research therefore needs to tease apart simple task effects from effects that may be of theoretical interest.

It is important to point out that it remains unclear if the auxiliary alternation for the German *Perfekt* has a predictive value to it at all — no data currently exist from native speakers of German investigating whether the *Perfekt* auxiliary alternation would lead to the creation of strong expectations, and, in turn, increased proportions of looks to the target image. However, Bovolenta and Husband (2022) used ERPs to show that Italian Perfect auxiliaries impact native Italian speakers' predictions of upcoming noun material, providing evidence that at least in another language, such predictive processing based on auxiliary selection is possible. Nevertheless, future research could investigate the predictive value of the auxiliary alternation for native speakers of German using a similar task.

Regardless of whether the auxiliary alternation has predictive value for native speakers, an important finding of the current experiment is that online predictive processing — i.e., creating expectations based on predictive cues within a temporally restricted time frame — may not be necessary for learning to occur (Huettig, 2015; Kaan, 2015). Despite the absence of predictive looks during the picture choice portion of the training task, participants still increased in auxiliary accuracy on the picture description task on immediate and delayed posttest. Given the nature of the training task, however — participants had to make picture choices without ever hearing the missing target participle at the end of the sentence fragment — participants were still engaged in the creation of expectations regarding which verb was missing from the fragment, albeit not necessarily in a temporally restricted time frame. It may therefore be the case that the offline creation of expectations, in conjunction with feedback, is sufficient for learning to occur. In such a case, error-based learning may well apply in contexts of learning a wide range of target L2 features, even those for which L2 learners are less likely to show online predictive behavior.

4.4.3 Pupil Dilations during Feedback Processing

The fourth set of analyses described in Section 4.3.5 investigated the time course of pupil size data in response to feedback presentation following the picture choice and participle production during training, to provide an online measure of feedback processing. The results showed that participants processed the feedback they received, but that the pupillary signature of this processing is dependent on whether participants committed an error prior to receiving said feedback. Different pupillary responses emerged in trials where participants were “all correct” compared to when they committed errors. While pupil size decreased overall during feedback presentation, the decrease was less pronounced for trials in which participants committed an error compared to when they made a correct picture choice and produced the correct target participle form — in other words, pupil sizes were larger during feedback presentation after errors than when participants had been “all correct”. These differences became more pronounced over the course of the four training blocks. In addition, the difference in pupil size between “all correct” trials and “all incorrect” trials in the second half of the training (Blocks 3 and 4) were descriptively larger for *sein* items than for *haben* items.

At the beginning of training in Block 1, all error types showed similar signatures, and trials in which participants did not commit any errors did not show larger pupil sizes during feedback presentation than trials in which participants had committed an error. This is not entirely unexpected: In the first block of training, participants were likely trying to figure out the task at hand and may therefore not have committed to their predictions with sufficient confidence for either errors or correct choices to lead to any type of surprise, and in turn larger pupil sizes, in the first block of training (Brod et al., 2018, 2022; Theobald et al., 2022; see also Kaan & Chun, 2018). This explains why peak pupil size and the time course of pupil size change is very similar for all three types of errors (“all correct”, “choice correct, participle incorrect”, or “all incorrect”) in this block.

Over the course of the training, however, the time course of pupil size change becomes different for the different error types. As participants became increasingly more accurate over the course of the training, their predictions were likely made with higher confidence than at the beginning of training. It is therefore highly likely that errors in either picture choice and/or participle production became more unexpected over the course of the training. The pupil signature during feedback presentation confirms this hypothesis: In later training blocks, pupil dilations for errors (“choice correct, participle incorrect”, or “all incorrect”) were much larger than for trials in which participants were correct. This finding is similar to findings from previous studies that have used pupil size changes to investigate prediction error and surprise, showing that unexpected outcomes lead to larger pupil sizes than outcomes that conform to expectations (Brod & Breitwieser, 2019; Brod et al., 2018; 2022; Hochmann & Papeo, 2014; Sekicki & Staudte, 2018; Scheepers et al., 2013; Zhang et al., 2018). Similarly, this finding also mirrors other studies which have shown adaptation in the size of pupillary response over the course of an experimental session (Brown et al., 2020; Sibley et al., 2019). However, it is important to keep in mind that a different number of data points contributed to each error type over the four training blocks, as participants became increasingly accurate in choosing the correct picture and producing the correct form of the corresponding participle. It is therefore important to collect data from more participants in order to increase the statistical power of any type of analysis in order to confirm the validity of the current findings.

As pupil size measures function as an index of the relative cognitive activity associated with functions such as memory consolidation and retrieval, working memory, attention, and expectancy violations (Brod & Breitwieser, 2019; Brod et al., 2018; Cohen Hoffing & Seitz, 2015; Hochmann & Papeo, 2014; Miller & Unsworth, 2019; 2021; Kafkas, 2021; Robison & Unsworth, 2019; Sara, 2009; Sara & Bouret, 2012; Scheepers et al., 2013; Zhang et al., 2018; see also Fritzsche, 2021), it remains unclear what exactly the difference in pupil size between errors and correct trials indicates in the current experiment. On one hand, pupil size could index the unexpectedness of an error in later training blocks compared to trials in which participants are all correct. On the other hand, it is also possible that pupil size in the current paradigm is related to the degree of attention participants allocated to processing any given feedback: feedback indicating an error may require more attention and cognitive effort to be processed, as participants are likely trying to make sense of the feedback in response to their prediction. It is also possible that the pupillary response corresponds to any number of combinations of the factors listed above, such as expectancy violation, attention, and memory consolidation. Within the current experiment, it is impossible to tease apart the exact source of the pupil size difference.

To my knowledge, this dissertation is one of the first — if not the only — to investigate online feedback processing and its evolution over the course of a training task using pupillometry to measure cognitive activity in the moment participants receive said feedback. The findings therefore contribute to our understanding of online feedback processing specifically, as well as to our understanding of feedback efficiency more broadly. By providing a window into participants' cognitive activity during feedback processing, the current experiment uncovers what exactly learners do with feedback in the moment they receive it and how it is processed by the cognitive and linguistic system. It thereby supplements years of research on corrective feedback which has used utterance repair and modified output to determine whether learners notice and process the feedback targeted at their error (e.g., Egi, 2010; Gurzynski-Weiss & Révész, 2012; Loewen, 2005; Loewen & Philp, 2006; Mackey & McDonough, 2006; Mackey & Philp, 1998; McDonough, 2004, 2005, 2007; McDonough & Mackey, 2006; Sheen, 2008). In addition, the findings also supplement attempts in other fields of inquiry that have investigated the correlates of error monitoring during feedback processing, for instance using EEG (e.g., Bujok, 2019; Bultena et al., 2017, 2020; Davidson & Indefrey, 2011; Opitz et al., 2011; Sebastian-Gallés et al., 2006). Finally, the rather large differentiation of pupil size in response to feedback confirming or disconfirming participants' offline predictions over the course of the experiment — pupil signatures were virtually identical in Block 1 but very different by Block 4 — occurred in the absence of the emergence of predictive looks to the target picture prior to feedback presentation. This indicates that error-based learning can occur in contexts of L2 learning, but that it does not necessarily go hand in hand with prediction as currently conceptualized in psycholinguistic research, namely as the creation of expectations in a temporally-restricted frame. The question then emerges whether using methodologies that investigate such online predictive processing are useful to investigations of error-based learning in L2 contexts.

4.4.4 The Relationship between Feedback Processing and Learning

As described in Section 4.3.6, the necessary analyses to answer this research question could not be conducted due to the small sample size. If more data is collected in the future, these data will shed light on the relationship between feedback processing and learning to the extent that such a relationship exist in the context of an error-based L2 learning task.

4.4.5 Implications for Error-Based L2 Learning

The present findings have important implications for our understanding of the role of prediction and feedback processing in error-based learning in an L2 context. It is important, however, to reiterate that these findings come from a small sample of German L2 learners with beginning to intermediate proficiency levels.

One major aspect of error-based language learning models is that prediction errors — which occur when a speaker makes a prediction about upcoming language material that is not met — are what guide the internal learning mechanisms to adjust linguistic representations and their weights, in particular to avoid similar prediction errors in the future (e.g., Chang et al., 2006; Dell & Chang, 2014; Jaeger & Snider, 2013; Kaschak, 2007). This process then underlies the implicit adaptation to and learning of specific language structures, especially in contexts of an L1. However, one major finding of the current experiment is that prediction as it is usually discussed in frameworks of error-based learning, as well as in psycholinguistics more generally (see Bovolenta & Marsden, 2021), may not be necessary for prediction error, and in turn adaptation through error-based learning, to occur. As the results showed, predictive looks during the auditory stimulus did not emerge throughout the entirety of the training, yet the pupil signature during feedback processing in Block 4, i.e., at the end of training, clearly showed that some type of adaptation must have occurred. Indeed, it is possible that learners created expectations outside of the temporally limited timeframe that is the auditory presentation of the stimulus. Since sentence fragments never actually contained the target participle, participants necessarily must have formed an expectation about the appropriate continuation of the sentence, even if creating this prediction did not occur within the bounds of the auditory stimulus. This finding thus expands our understanding of what prediction can look like in frameworks of error-based learning, and that prediction need not occur online for learning to be observed. For contexts of L2 learning more specifically, this means that structures for which L2 learners typically have shown a reduced engagement in online prediction — such as morphosyntactic cues (Dussias et al., 2013; Frenck-Mestre et al., 2019; Lew-Williams & Fernald, 2010; Grüter et al., 2017; Hopp, 2013, 2015, 2016; see also Mitsugi & MacWhinney, 2016; for a review, see Kaan, 2014; Kaan & Grüter, 2021) — may still be learnable via an error-based learning task. An alternative interpretation of these results is that participants may well have created predictions within the temporally limited timeframe of the auditory stimulus, but that these predictions did not translate into looks to the target picture in the setup of the visual world paradigm in the current experiment. It is thus possible that traditional tools of measuring prediction in L2 learners — such as eye-tracking in a visual world paradigm — may not be sufficient to capture all manifestations of prediction in L2 learners.

Related to the crucial role of prediction in error-based learning frameworks is that of feedback, which is necessary for speakers to compute prediction errors. Indeed, for benefits related to the generation of predictions to appear, positive evidence, i.e., feedback with the correct answer, needs to be presented shortly after the generation of a prediction (Pashler et al., 2005; Vaughn & Rawson, 2012). In the current experiment, the time course of the pupillary response to feedback evolved over the course of the training task, such that by the end of the training, pupillometric signatures of feedback processing differed between items for which participants made an incorrect prediction — i.e., they committed some type of error — compared to those where they made a correct prediction. By the end of the training, the pupillometric signatures suggest that errors had become more unexpected than correct responses (see e.g., Brod et al., 2018). Crucially, participants also significantly increased in their participle accuracy during training, showing that indeed participants used the feedback presented to them — which took the form of the target participle — in order to adapt their language system and to prevent prediction errors in subsequent trials. In addition, the current findings underscore the fact that uptake or modified output following the provision of feedback is not necessary for benefits of the feedback to appear: The evolution of the pupillary response to feedback over the course of the training task demonstrates that learners processed the feedback even in the absence of an overt response in the form of modified output (see Mackey, 2006; see also Bultena et al., 2020).

4.4.6 Limitations and Future Studies

Just like Experiment 1, the current study has some limitations that impact the conclusions that can be drawn from the results and analyses. However, these same limitations offer exciting avenues for future studies. For instance, the COVID-19 pandemic impacted the sample size in the current study: Data from only 13 participants were included, which severely limited the analyses that could be conducted, in particular regarding the question addressing the relationship between pupil size and learning. Future studies should therefore attempt to replicate the current study with a larger sample size in order to allow for more fine-grained analyses.

Secondly, the current study did not find that predictive looks emerged during the presentation of the auditory stimuli in the training task, with the potential exception of *sein* items in Block 1. However, the setup of the current experiment did not allow me to locate the exact source of this null result: The lack of predictive looks to the target pictures in the task could either be a task effect — participants were initially surprised to hear *Perfekt* sentences with *sein*, which lead to more looks to the target pictures but this surprise waned over the course of the experiment — or indicate that indeed L2 learners are less likely to engage in predictive processing when the predictive cues are syntactic or morphosyntactic in nature (Kaan, 2014; ; Kaan & Grüter, 2021; Kaan et al., 2010). Furthermore, it remains unclear whether the auxiliary embedded in *Perfekt* sentences has any level of predictive value in German, even for native speakers, in the sense that the auxiliary choice would trigger predictive processing of upcoming verbal material. Future studies should therefore attempt to tease apart the exact reasons for the absence of predictive looks in the current training setup by replicating the experiment and by potentially implementing a similar task design with native speakers of German.

Similarly, while different pupillometric signatures emerged during feedback processing at the end of training, it remains unclear what type of cognitive activity exactly is measured as participants process the feedback presented to them. As discussed in Section 4.4.3, cognitive activity associated with changes in pupil dilation range from memory consolidation and retrieval, working memory, attention, to expectancy violations (Brod & Breitwieser, 2019; Brod et al., 2018; Cohen Hoffing & Seitz, 2015; Hochmann & Papeo, 2014; Miller & Unsworth, 2019; 2021; Kafkas, 2021; Robison & Unsworth, 2019; Sara, 2009; Sara & Bouret, 2012; Scheepers et al., 2013; Zhang et al., 2018; see also Fritzsche, 2021), as well as any combination of these. Future research should tease apart the relative contributions of different types of cognitive activity to the processing of feedback following different error types in the context of an error-based learning task in an L2.

A fourth limitation to the experiment is the relative length of the training task. Because there were four training blocks, and due to the combination of a visual world paradigm with subsequent typing of words and feedback presentation, participants needed around 40 minutes to complete the training task. This may have had an impact on participants' ability to pay attention to and attentively process the feedback presented to them. Although measures were taken to mitigate any effects in pupil size change over the course of the training task that would stem from tonic pupil size changes, for instance from fatigue (McGarrigle et al., 2016; see also Schmidtke, 2018), it cannot be guaranteed that no such effects remain in the pupil data. While effects such as fatigue would have necessarily impacted the pupillometric signature in response to feedback for correct and incorrect items equally and therefore would not be able to completely account for the differential evolution of pupil responses over the course of the training task, it is still imperative to collect more data and attempt to replicate the current findings in order to be able to draw stronger conclusions about the pupillometric signatures associated with the processing of feedback in response to different error types.

Furthermore, it remains unclear how explicit and implicit learning aptitudes may impact learning in an error-based L2 learning context, as these learning aptitudes have been shown to impact the effectiveness of various types of corrective feedback for L2 learners (e.g., Granena & Yilmaz, 2019; Yilmaz & Granena, 2016, 2019, 2021). As described in the first paragraph of this section, the small sample size prevented me from running any type of correlational analysis between learning aptitudes and the amount of learning demonstrated on the outcome measures. However, understanding the relationship between explicit and implicit learning aptitudes could have important implications for our understanding of error-based learning, in contexts of L2 learning more specifically: Generally, error-based learning is seen as an implicit learning mechanism (Dell et al., 2006; Dell & Chang, 2014), but as the findings from Experiment 1 in this dissertation and other studies have shown, it is possible that some type of explicit mechanism contributes to learning aspects of an L2 in an error-based framework (e.g., Coumel et al., 2022; Grüter et al., 2021; McDonough & Fulga, 2015; McDonough & Trofimovich, 2015). Future studies should therefore investigate the relationship between learning aptitudes and learning L2 features in an error-based learning task.

Finally, future research should investigate the effects of learner proficiency in the current learning paradigm, both on the possible emergence of predictive looks as well as on the pupillary responses during feedback presentation. Since proficiency has been shown to have an impact on both predictive processing (Dussias et al., 2013; Hopp, 2015a) and the benefits of implicit corrective feedback (Ammar, 2008; Ammar & Spada, 2006; Li, 2014; Mackey & Philp, 1998), investigating its effects in this framework might reveal more nuanced differences between participants as well. Similarly, future research should investigate whether learning paths and pupillary responses are identical or differ if other types of feedback are presented to participants. For instance, such feedback could take the form of the semantic-focused feedback as in Experiment 1, or it could simply indicate to participants the correctness of their response (*correct* or *incorrect*). Investigating whether other types of feedback lead to different learning outcomes and/or pupillary responses would illuminate the role of information conveyed by the feedback and whether participants can or cannot make use of different types of information in response to prediction errors.

CHAPTER 5: CONCLUSION

The goal of this dissertation was to investigate in how far error-based learning mechanisms apply in contexts of L2 learning. To do so, two experiments were conducted, which (1) investigated the role of feedback type and ordering, proficiency, prior knowledge, and awareness, and (2) the emergence of predictive looks as well as the signature of cognitive activity during feedback processing using eye-tracking and pupillometry. In the sections that follow, I discuss the major findings from Experiments 1 and 2 and how they relate to the overall applicability of error-based learning in an L2. I also discuss the relevance of these findings for pedagogical applications, and discuss open questions that remain and provide directions for future research.

5.1 Overview of Major Findings

The dissertation investigated the following research questions:

Experiment 1:

1. Do learners acquire the target structure after receiving prediction-based training, as measured via accuracy on independent behavioral posttest measures?
2. Does the order in which participants receive two different types of feedback have an impact on learning of the target structure, both during training as well as on independent behavioral posttest measures?
3. Does proficiency have an effect on the learning of the target structure via prediction and feedback, as measured via accuracy on independent behavioral posttest measures?

Experiment 2:

1. Over the course of the training task, do learners start to engage in predictive processing, as measured by the proportion of looks to a target picture?
2. After listening to target sentence fragments, do learners process feedback they receive about the accuracy of their prediction, as measured via pupil dilation? Does feedback processing evolve over the course of the training task?
3. Do pupillary responses correlate with the amount of learning, as measured during training and on independent posttest measures?

The major findings addressing each research question (as well as exploratory analyses) are summarized below.

Experiment 1:

1. Overall, results showed significant improvements from pretest to subsequent posttests on both the picture matching and picture description task for both auxiliary and participle accuracy.
2. Feedback order played a role for auxiliary and participle accuracy during training, such that descriptively, auxiliary accuracy was higher in the training block in which participants received semantic-focused feedback, whereas participle accuracy was higher in the training block in which participants received form-focused feedback. There were no main effects of feedback ordering on any of the outcome measures, indicating that groups did not differ in their overall learning.

3. There were significant interactions between testing time and proficiency, indicating that participants around the center of the proficiency scale benefited most from the error-based learning task as implemented in this experiment. These effects appeared in particular for auxiliary accuracy on the picture matching task.
4. Exploratory analyses showed that not only general L2 proficiency played a role, but that whether participants possessed prior knowledge of the auxiliary alternation played a role as well. Participants without any prior knowledge of the auxiliary alternation did not show any improvements on either outcome measure.
5. Exploratory analyses showed a role for awareness, such that participants who were aware of the auxiliary alternation were more likely to show higher auxiliary accuracy at immediate and delayed posttests than participants who were unaware.

Experiment 2:

1. Visual inspections of looks to images showed that participants did not start to engage in predictive processing as measured via eye-tracking in a visual world paradigm. Small differences in proportions of looks to the target picture for *sein* items in the first training block may be attributed to surprisal effects.
2. Visual inspection of the pupillary time course during feedback presentation showed that participants processed the feedback they received, and that participants' pupillary signature in response to prediction errors compared to correct predictions evolved over the course of the training. More specifically, at the end of training, pupil sizes were larger during feedback presentation following errors than when participants had made a correct prediction.
3. The third research question could not be addressed with the current sample size (due to the COVID-19 pandemic).
4. The outcome measures showed different learning outcomes for each task: On the picture matching task, participants did not improve across testing times, whereas on the picture description task, participants significantly improved on both auxiliary and participle accuracy (from pretest to both immediate and delayed posttest). A similar distinction applies to the increase in accuracy during the four training blocks.

Because the two experiments did not investigate identical research questions and targeted slightly different populations in terms of proficiency ranges, the findings from the two experiments can, for obvious reasons, not directly be compared to one another. However, Experiment 2 still produced similar findings on the outcome measures as Experiment 1: Whereas participants significantly improved in accuracy across testing times on the picture matching task in Experiment 1, they did not do so in Experiment 2, although there appear to be descriptive trends in the same direction — keep in mind that the sample in Experiment 2 is almost four times smaller than in Experiment 1. Regarding auxiliary and participle accuracy on the picture description task, findings from Experiment 2 more closely mirrors Experiment 1, in that participants significantly increased across testing times in both experiments, and also showed a decrease for participle accuracy going from immediate to delayed posttest.

5.2 Implications for Error-Based L2 Learning and Corrective Feedback

The findings from the two experiments conducted as part of this dissertation have important theoretical implications for error-based learning models and their application to L2 learning, as well as for the role and importance of corrective feedback in contexts of L2 learning. These implications are discussed in turn below.

First, the findings of both experiments show that error-based learning in conjunction with implicit feedback in an L2 is possible, albeit with some limitations. For one, there seem to be differences between the benefits of error-based learning mechanisms when it comes to the learning of episodic items, such as vocabulary words and general knowledge facts, compared to syntactic or morphosyntactic knowledge. While previous research has shown benefits of guessing — creating an expectation in the absence of relevant prior knowledge — for the learning of episodic items (Brod & Breitwieser, 2019; Brod et al., 2018, 2022; Potts & Shanks, 2014; Potts et al., 2019; Theobald et al., 2022), the current experiments show that similar benefits of guessing might not exist when L2 morphosyntactic alternations constitute the linguistic target of the error-based learning task. Instead, error-based learning tasks that involve prediction and feedback might be most beneficial when there is a prior knowledge base that learners can exploit to inform their predictions and to more rapidly process and integrate the feedback they receive to update their language model, alongside future predictions (see McDonough & Trofimovich, 2015). Within an error-based learning framework in the context of L2 learning, guessing may thus be a learning mechanism in its own right, but only for specific types of item-based knowledge that can be captured within episodic memory. These may be learnable via one-shot declarative learning, whereas exploring morphosyntactic regularities may lead to greater uncertainty about possible and impossible language forms, slowing down the learning rate (see Lee et al., 2015) and requiring longer exposure to more items that include the morphosyntactic feature, in order for learning in the absence of prior knowledge to take place. In other words, prior knowledge potentially speeds up the exploration of patterns found in the language input, decreasing the randomness of predictions and making such predictions more targeted and therefore more useful. It remains unclear when exactly the transition between less-beneficial guessing and more beneficial predicting occurs during the learning of L2 morphosyntax. Similarly, it remains unclear if guessing and predicting are two points on a continuum or whether they represent qualitatively different learning mechanisms.

A second and important implication of the findings from Experiments 1 and 2 is that prediction within error-based learning may not need to occur online, i.e., in a temporally restricted timeframe, in order to contribute to the learning mechanism. For instance, in Experiment 1, participants read sentence fragments that included the auxiliary as the predictive cue as they were choosing a picture from the two pictures presented on screen. Participants could read the sentence as many times as they wanted before making a prediction, and there were no time restrictions on the task. Creating expectations in the absence of any sort of time pressure still proved beneficial as long as participants possessed prior knowledge of the target structure, as there were improvements during training and across testing times. This is corroborated by findings from Experiment 2, where one of the goals was specifically to track the emergence of predictive looks to target pictures on screen during the presentation of an auditory stimulus. While the auditory stimulus presentation forced participants to attend to the predictive cue within a temporally restricted timeframe, results did not show an increase of predictive looks to the target picture over the course of the training task, implying that participants did not engage in online prediction here either. However, it remains unclear if participants just did not show the expected signs of predictive behavior, i.e., predictive looks, or if the methodology used to measure predictive processing — eye-tracking in conjunction with the visual world paradigm — was simply unable to capture participants' predictive behavior. Nevertheless, participants again increased in accuracy across testing times. Together, Experiments 1 and 2 indicate that creating predictions offline, in conjunction with feedback, is sufficient for learning to occur. Indeed, prior research has shown benefits of creating expectations (Brod & Breitwieser, 2019; Brod et al., 2018, 2022; Potts & Shanks, 2014; Potts et al., 2019; Theobald et al., 2022), even in contexts of L2 learning that did not force participants to create these expectations within a temporally restricted timeframe (Bovolenta & Marsden, 2021; Grüter et al., 2021; Potts et al., 2019; Potts & Shanks, 2014). Conversely, error-based learning may well apply to a wide range of L2 features, even those for which L2 learners are less likely to show online predictive behavior. At the same time, proficiency may also play a role, both for the creation of expectations as well as regarding the processing of implicit corrective feedback, as shown in Experiment 1.

In fact, findings from both experiments also have implications that concern the effectiveness of corrective feedback in contexts of L2 learning. For one, findings show that the amount and type of information encoded in the feedback provided to learners has a direct short-term impact on what information participants extract from the feedback and use to adapt their language model. In Experiment 1, for instance, descriptive statistics during training showed that learners were more accurate on auxiliaries when they received semantic-focused feedback compared to form-focused feedback — which focused on the form of the participle and only indirectly provided information about auxiliary accuracy — whereas the reverse was true for participle accuracy. On the other hand, there were no main effects of group, showing that overall, the order in which the two feedback types were provided to participants did not matter in the long-term. Similarly, Experiment 2 only included form-focused feedback and the findings were similar: Participants increased significantly in participle accuracy during training (41% increase from Block 1 to Block 4), but only minimally in auxiliary accuracy (6% increase from Block 1 to Block 4). Similar trends are reflected in the outcome measures at immediate posttest. Together, the findings from these two experiments indicate that it is not only important to consider learner profiles and their compatibility with different corrective feedback types to maximize learning — e.g., learner proficiency, their learning aptitudes, working memory capacity, motivation, beliefs, etc. and their impact on the effectiveness of different types of explicit or implicit corrective feedback (e.g., Goo, 2012; Granena & Yilmaz, 2019; Kartchava & Ammar, 2014; Li, 2014) — but also to consider the error that was committed, as well as what information crucially needs to be encoded in the feedback to optimally target the error and to foster learning.

Finally, another finding that converges across both Experiments 1 and 2 is that participants saw the largest increases in auxiliary accuracy in the picture description task, i.e., testing productive knowledge, compared to the picture matching task, testing receptive knowledge of the auxiliary alternation. There are multiple potential explanations for this finding. One of these might be that we are dealing with ceiling effects in the picture matching task, in particular for learners with prior knowledge of the auxiliary alternation and higher levels of proficiency, leaving little room for improvement from one testing time to the next. On the picture description task, on the other hand, participants started out with much lower accuracy, allowing for larger increases across testing times. While ceiling effects may thus partially be responsible for the discrepancy in auxiliary accuracy increases between the picture matching and the picture description task, they cannot be the sole explanation: Even for participants around the center of the proficiency scale — i.e., those learners who benefited from the error-based training task the most — accuracy gains were larger in production than comprehension. Instead, researchers have posited that language users draw on production processes to predict upcoming sentence material during comprehension (see Pickering & Garrod, 2013; Martin et al., 2018). As the error-based learning task in the current experiment necessitated participants to make predictions based on their comprehension of the auxiliary provided in the sentence fragment, the engagement of the language production system during prediction may translate into increased production accuracy of the auxiliary as compared to its comprehension. While it is conceivable that retrieval practice — i.e., recalling and retrieving linguistic material from memory — may play a role as well (e.g., Hopman & MacDonald, 2018; Keppenne et al., 2021), this cannot be the case in the current experiment, as participants never actually produced the auxiliary while completing the error-based learning task. Future research could attempt more targeted investigations of mechanisms such as prediction-by-production as well as their differential contributions to the production and comprehension of target language features.

5.3 Pedagogical Implications

Besides contributing to theoretical discussions of L2 learning mechanisms, another goal of this dissertation was to provide insight into the effectiveness of error-based learning tasks for pedagogical applications in contexts of L2 learning. Together, the findings from Experiments 1 and 2 have pedagogical implications relating to the application of error-based learning tasks, specifically when targeting L2 morphosyntax.

First, findings from the first experiment indicate that error-based learning tasks targeting L2 morphosyntax — at least as implemented in the current experiment — may only benefit L2 learners who have prior knowledge of the target structure, in this case the German *Perfekt* auxiliary alternation. This finding resembles previous priming research, which found that rudimentary knowledge of the target structure may be necessary for priming and therefore learning to occur (McDonough & Fulga, 2015; McDonough & Trofimovich, 2015). While error-based learning tasks have thus been shown to be well suited for the acquisition of new episodic knowledge, i.e., vocabulary words or facts related to general knowledge (Brod & Breitwieser, 2019; Brod et al., 2018; Kang et al., 2009; Kornell, 2009; Potts et al., 2019; Potts & Shanks, 2014), when the target structure is not episodic in nature, such tasks may best be reserved for review activities among L2 learners who have been formally introduced to the target structure before.

In addition, these findings show that error-/prediction-based learning tasks constitute a viable alternative for review activities compared to how the *Perfekt* auxiliary alternation is traditionally introduced for review in German language classrooms. Activities in the German language classroom typically focus on the participle and how learners can infer from the participle which is the correct auxiliary to use in a sentence. However, this error-based learning task focused on the auxiliary and how the upcoming participle, or at least the category of the participle (i.e., is it a verb that indicates a change in place or state?), can be inferred from the auxiliary, reversing the directionality of the predictive relationship (cf. 'transitional probability', e.g., Thompson & Newport, 2007). Although more research is needed to validate the claim that error-based learning tasks are at least equally suited as traditional pedagogical methods to strengthen prior morphosyntactic knowledge, it is possible that the task in the current experiment strengthens learners' knowledge of the regularities that underly the auxiliary alternation more so than traditional methodologies, specifically which kinds of verb necessitate *sein* as the appropriate *Perfekt* auxiliary.

Finally, findings from Experiment 2, in particular, indicate that learners use feedback that is provided to them in response to their output, even in the absence of an overt response to the feedback on the learner's part. This underscores the importance of providing feedback to learners, not just in response to errors, but also in response to correct output in order to confirm or disconfirm learners' predictions about possible language forms (e.g., Altmann & Mirković, 2009; Godfroid et al., 2013; Huettig, 2015; Philp, 2003). While in the current experiment it remains unclear whether the benefits of feedback are equal in the absence versus presence of uptake or modified output (see e.g., Egi, 2010), findings from the second experiment provide evidence that uptake or modified output is not a requirement for feedback processing and learning to occur.

5.4 Open Questions and Directions for Future Research

For limitations and open questions relating to each of the two experiments included in this dissertation, please refer to Sections 3.4.5 and 4.4.6, respectively. Below, I discuss open questions and directions for future research more generally.

This dissertation provides and discusses important findings relating to the roles of proficiency and prior knowledge (Experiment 1), as well as the roles of predictive processing and feedback processing (Experiment 2) for error-based L2 learning. However, due to the COVID-19 pandemic, both experiments had a small sample size, which prevented me from conducting all the analyses that I ideally would have wanted to. Relatedly, it remains unclear if the current findings are limited to contexts of learning the German *Perfekt* auxiliary alternation and therefore are only relevant for L2 German as taught in US institutions of higher education or whether they would transfer to other target structures, languages, and populations as well. Currently, there are no theoretical grounds for assuming that the findings would not translate to other contexts, as long as a similar error-based learning task can be created and implemented. Future research therefore needs to not only attempt to replicate the current findings with a larger sample size of a similar population with the same target structure, but also to extend the investigations to other populations and target structures in order to provide stronger validity of the findings and to draw stronger conclusions about the applicability of error-based learning in L2 learning contexts.

Second, one of the limitations of Experiment 2 concerns the use of eye-tracking within the visual world paradigm to track the emergence of predictive processing in L2 learners as they complete the error-based learning task. As discussed in Section 4.4.5, it is possible that participants may have created predictions within the temporally limited time frame of the auditory stimulus, but that these predictions did not translate to behavior that is traceable with an eye-tracker. An open question is therefore whether participants really did not engage in online prediction, or whether the methodology used — eye-tracking and the visual world paradigm — was simply unable to capture participants' predictive behavior. Future research should therefore not only attempt to replicate the null-finding regarding predictive looks in the error-based learning task, but also investigate and explore whether this null-finding is duplicated using other methodologies that track predictive processing in language users, such as EEG.

Finally, another open question for future research relates to the role of curiosity as it is relevant to feedback processing in models of error-based learning. Previously, curiosity has been shown to have an impact on the degree of learning that occurs within error-based learning tasks (e.g., Brod & Breitwieser, 2019; Potts et al., 2019). In this dissertation, however, neither Experiment 1 nor Experiment 2 included measures of curiosity asking participants how curious they were to find out whether their prediction turned out correct or incorrect on each trial. Future research should investigate whether curiosity relates to feedback processing and the amount of learning as demonstrated on posttests when the target structure is morphosyntactic in nature. If indeed there is a relationship between curiosity and learning, future research should also investigate how to increase curiosity in order to maximize learning in an error-based learning task.

5.5 Final Remarks and Conclusion

This dissertation is one of the first — if not currently the only — to unite research and literature on error-based learning, L2 predictive processing, and corrective feedback in L2 learning. Furthermore, it is also one of the first — if not currently the only — to use pupillometry in order to investigate the processing of corrective feedback over the course of an L2 learning task. By conducting two experiments, this dissertation provides evidence that error-based learning mechanisms indeed apply in contexts of L2 learning, but that they may be more beneficial to re-activate and reinforce prior knowledge in L2 learners, particularly if the target is a morphosyntactic structure. Furthermore, it demonstrates the critical role that feedback plays in these mechanisms, as the information encoded in the feedback has a direct impact on what is learned from it. By linking psycholinguistic research to L2 learning, this dissertation also creates opportunities to translate its findings into pedagogical approaches that can positively benefit L2 learners' learning experiences. Finally, the findings of this dissertation, as well as the remaining limitations and open questions, point to exciting avenues for further investigations into error-based L2 learning, in particular the role of prediction and feedback processing.

References

Alemán Bañón, J., & Martin, C. (2021). The role of crosslinguistic differences in second language anticipatory processing: An event-related potentials study. *Neuropsychologia*, *155*, 107797. https://doi.org/10.1016/j.neuropsychologia.2021.107797

Altarriba, J., & Knickerbocker, H. (2011). Acquiring second language vocabulary through the use of images and words. In P. Trofimovich & K. McDonough (Eds.), *Applying priming methods to L2 learning, teaching and research: Insights from psycholinguistics* (pp. 21 – 48). John Benjamins.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264.

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Ammar, A. (2008). Prompts and recasts: Differential effects on second language morphosyntax. *Language Teaching Research*, *12*(2), 183–210.

Ammar, A., & Spada, N. (2006). One size fits all?: Recasts, prompts, and L2 learning. *Studies in Second Language Acquisition*, *28*(4), 543–574.

Andringa, S. (2020). The emergence of awareness in uninstructed L2 learning: A visual world eye tracking study. *Second Language Research*, *36*(3), 335–357.

Anton, C., Barkse, T., McKinstry, M. (2017). *Sag mal!. An introduction to German language and Culture* (2nd edition). Vista Higher Learning.

Arai, M., Van Gompel, R. P., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*(3), 218–250.

Arbel, Y., & Wu, H. (2016). A Neurophysiological examination of quality of learning in a feedback-based learning task. *Neuropsychologia*, *93*(A), 13–20.

Aston-Jones, G., Rajkowski, J., & Cohen, J. (1999). Role of locus coeruleus in attention and behavioral flexibility. *Biological Psychiatry*, *46*(9), 1309–1320. https://doi.org/10.1016/S0006-3223(99)00140-7

Augustyn, P., & Euba, N. (2015). *Stationen*. Cengage Learning.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*(3), EL174–EL180. https://doi.org/10.1121/1.4789864

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition*, *35*(4), 689-725.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457–474. doi:10.1016/j.jml.2007.09.002

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bernolet, S., Collina, S., & Hartsuiker, R. J. (2016). The persistence of syntactic priming revisited. *Journal of Memory and Language*, *91*, 99–116. https://doi.org/10.1016/j.jml.2016.01.002

Bernolet, S., & Hartsuiker, R. J. (2010). Does verb bias modulate syntactic priming? *Cognition*, *114*(3), 455–461. https://doi.org/10.1016/j.cognition.2009.11.005

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387. https://doi.org/10.1016/0010-0285(86)90004-6

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition*, *104*(3), 437–458. https://doi.org/10.1016/j.cognition.2006.07.003

Bolker, B. & Robinson, D. (2022). *broom.mixed: Tidying Methods for Mixed Models*. R package version 0.2.9.3. https://CRAN.R-project.org/package=broom.mixed

Bordag, D., & Sieradz, M. (2014). Ausgewählte Aspekte des Perfekt-und Passiverwerbs in DaF anhand einer Korpusuntersuchung. *Wege für Bildung, Beruf und Gesellschaft-mit Deutsch als Fremd-und Zweitsprache*, 127.

Bovolenta, G., & Husband, E. M. (2022). Structural prediction during language comprehension revealed by electrophysiology: Evidence from Italian auxiliaries. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advance online publication. https://doi.org/10.1037/xlm0001115

Bovolenta, G., & Marsden, E. (2021). Expectation violation enhances the development of new abstract syntactic representations: Evidence from an artificial language learning study. *Language Development Research*, *1*(1), 193-243.

Braem, S., Coenen, E., Bombeke, K., Van Bochove, M. E., & Notebaert, W. (2015). Open your eyes for prediction errors. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 374-380.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Branigan, H. P., & Messenger, K. (2016). Consistent and cumulative effects of syntactic experience in children’s sentence production: Evidence for error-based implicit learning. *Cognition*, *157*, 250–256. https://doi.org/10.1016/j.cognition.2016.09.004

Branigan, H. P., Pickering, M. J., & McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 468–481. https://doi.org/10.1037/0278-7393.31.3.468

Brod, G. (2021). Predicting as a learning strategy. *Psychonomic Bulletin & Review*, *28*, 1839-1847. https://doi.org/10.3758/s13423-021-01904-1

Brod, G., & Breitwieser, J. (2019). Lighting the wick in the candle of learning: Generating a prediction stimulates curiosity. *NPJ Science of Learning*, *4*(1), 1–7. https://doi.org/10.1038/s41539-019-0056-y

Brod, G., Greve, A., Jolles, D., Theobald, M., & Galeano-Keiner, E. M. (2022). Explicitly predicting outcomes enhances learning of expectancy-violating information. *Psychonomic Bulletin & Review*, 1–10. https://doi.org/10.3758/s13423-022-02124-x

Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction*, *55*, 22–31. https://doi.org/10.1016/j.learninstruc.2018.01.013

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*, 203–216. https://doi.org/10.1016/j.jml.2016.10.002

Brown, D. (2016). The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis. *Language Teaching Research*, *20*(4), 436–458.

Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology*, *73*(9), 1431–1443. https://doi.org/10.1177/1747021820916726

Bultena, S., Danielmeier, C., Bekkering, H., & Lemhöfer, K. (2017). Electrophysiological correlates of error monitoring and feedback processing in second language learning. *Frontiers in Human Neuroscience*, *11*, 29.

Bultena, S., Danielmeier, C., Bekkering, H., & Lemhöfer, K. (2020). The role of conflicting representations and uncertainty in internal error detection during L2 learning. *Language Learning*, *70*(S2), 75–103. https://doi.org/10.1111/lang.12401

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491–1494. https://doi.org/10.1037/0278-7393.27.6.1491

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*(1), 69–84. https://doi.org/10.1007/s11409-006-6894-z

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569.

Carroll, S., & Swain, M. (1993). Explicit and Implicit Negative Feedbac|. *Studies in Second Language Acquisition*, *15*(3), 357–386.

Chambers, C. G., & Cooke, H. (2009). Lexical competition during second-language listening: Sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1029–1040. https://doi.org/10.1037/a0015901

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Chen, B., Jia, Y., Wang, Z., Dunlap, S., & Shin, J.-A. (2013). Is word-order similarity necessary for cross-linguistic structural priming? *Second Language Research*, *29*(4), 375–389. https://doi.org/10.1177/0267658313491962

Chun, E. (2020) L2 prediction guided by linguistic experience. *English Teaching*, *75*(1), 79-103.

Chun, E., Chen, S., Liu, S. & Chan, A. (2021) Influence of syntactic complexity on second language prediction. In T. Grüter & E. Kaan (Eds.), *Prediction in second language processing and learning* (pp. 69 – 90). John Benjamins.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cohen Hoffing, R., & Seitz, A. R. (2015). Pupillometry as a glimpse into the neurochemical basis of human memory encoding. *Journal of Cognitive Neuroscience*, *27*(4), 765-774.

Coumel, M., Ushioda, E., & Messenger, K. (2022). Learning multiple L2 syntactic structures via chat-based alignment: What is the role of learners' prior knowledge and conscious decisions?. *System*, *109*, 102869.

Curcic, M., Andringa, S., & Kuiken, F. (2019). The role of awareness and cognitive aptitudes in L2 predictive language processing. *Language Learning*, *69*(S1), 42–71. https://doi.org/10.1111/lang.12321

Davidson, D. J., & Indefrey, P. (2011). Error-related activity and correlates of grammatical plasticity. *Frontiers in Psychology*, *2*, 219.

de Haas, M., Vogt, P., & Krahmer, E. (2020). The Effects of Feedback on Children's Engagement and Learning Outcomes in Robot-Assisted Second Language Learning. *Frontiers in Robotics and AI*, *7*, 101. https://doi.org/10.3389/frobt.2020.00101

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B*, *369*(20120394). https://royalsocietypublishing.org/doi/full/10.1098/rstb.2012.0394

Diedrichsen, J., White, O., Newman, D., & Lally, N. (2010). Use-Dependent and Error-Based Learning of Motor Behaviors. *Journal of Neuroscience*, *30*(15), 5159–5166. https://doi.org/10.1523/JNEUROSCI.5406-09.2010

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, *20*(5), 917–930. https://doi.org/10.1017/S1366728916000547

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2019). Prediction and integration of semantics during L2 and L1 listening. *Language, Cognition and Neurosience*, *34*(7), 881–900.

Doughty, C., & Varela, E. (1998). Communicative focus on form. In C. Doughty & J. Williams (Eds.), *Focus on Form in Classroom Second Language Acquisition* (pp. 114–138). CUP.

Dussias, P. E., Kroff, J. R. V., Tamargo, R. E. G., & Gerfen, C. (2013). When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition*, *35*(2), 353–387. https://doi.org/10.1017/S0272263112000915

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

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Ellis N. C. (2005) At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*(2), 305–352.

Ellis, N. C. (2006). Selective Attention and Transfer Phenomena in L2 Acquisition: Contingency, Cue Competition, Salience, Interference, Overshadowing, Blocking, and Perceptual Learning. *Applied Linguistics*, *27*(2), 164–194. https://doi.org/10.1093/applin/aml015

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition*, *33*(4), 589-624.

Ellis, R. (2007). The differential effects of corrective feedback on two grammatical structures. In *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 339–360). OUP.

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

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.

Ellis, R., & Sheen, Y. (2006). Reexamining the role of recasts in second language acquisition. *Studies in Second Language Acquisition*, *28*(4), 575–600. https://doi.org/10.1017/S027226310606027X

Ellis, R., Tanaka, Y., & Yamazaki, A. (1994). Classroom interaction, comprehension, and the acquisition of L2 word meanings. *Language Learning*, *44*(3), 449–491. https://doi.org/10.1111/j.1467-1770.1994.tb01114.x

Emberson, L. L., Richards, J. E., & Aslin, R. N. (2015). Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months. *PNAS*, *112*(31), 9585–9590.

Ernst, B., & Steinhauser, M. (2012). Feedback-related brain activity predicts learning from feedback in multiple-choice testing. *Cognitive, Affective, & Behavioral Neuroscience*, *12*, 323–336.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455. https://doi.org/10.1016/0013-4694(91)90062-9

Favier, S., Meyer, A. S., & Huettig, F. (2021). Literacy can enhance syntactic prediction in spoken language processing. *Journal of Experimental Psychology: General*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/xge0001042

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science*, *7*(11), 180877. https://doi.org/10.1098/rsos.180877

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, *16*, 88–92.

Ferreira, V. S., & Bock, K. (2006). The functions of structural priming. *Language and Cognitive Processes*, *21*(7–8), 1011–1029. https://doi.org/10.1080/01690960600824609

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: a unified JavaScript framework for online studies. In *International Conference on Computational Social Science* (Cologne).

Frenck-Mestre, C., Kim, S. K., Choo, H., Ghio, A., Herschensohn, J., & Koh, S. (2019). Look and listen! The online processing of Korean case by native and non-native speakers. *Language, Cognition and Neuroscience*, *34*(3), 385–404. https://doi.org/10.1080/23273798.2018.1549332

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200–214. https://doi.org/10.1016/j.jml.2017.04.007

Fritzsche, T. (2021, July 2). *Pupillometry (with a focus on children and language)* [PowerPoint Slides]. https://www.ling.uni-potsdam.de/~fritzsche/assets/files/Pupillometry.pdf

Frömer, R., Nassar, M. R., Bruckner, R., Stürmer, B., Sommer, W., & Yeung, N. (2021). Response-based outcome predictions and confidence regulate feedback processing and learning. *Elife*, *10*.

Fulga, A., & McDonough, K. (2016). The impact of first language background and visual information on the effectiveness of low-variability input. *Applied Psycholinguistics*, *37*(2), 265-283.

Gambi, C., & Messenger, K. (2021). The role of prediction errors in 4-year-olds' learning of English direct object datives. *Book of Abstracts*. EDLL 2021.

Gambi, C., Pickering, M. J., & Rabagliati, H. (2021). Prediction error boosts retention of novel words in adults but not in children. *Cognition*, *211*, 104650. https://doi.org/10.1016/j.cognition.2021.104650

Gass, S. M. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics*, *9*(2), 198–217. https://doi.org/10.1093/applin/9.2.198

Gass, S. M. (1997). *Input, interaction, and the second language learner*. Erlbaum.

Gass, S. M. (2003). Input and interaction. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 224–256). Wiley.

Gass, S. M., & Mackey, A. (2014). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 180–206). Routledge.

Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research*. Routledge.

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385-390.

Gerlanc, D. & Kirby, K. (2021*). bootES: Bootstrap Confidence Intervals on Effect Sizes*. R package version 1.2.1. https://CRAN.R-project.org/package=bootES

Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge.

Godfroid, A. (2019). Investigating instructed second language acquisition using L2 learners' eye-tracking data. In R. P. Leow (Ed.), *The Routledge handbook of second language research in classroom learning* (pp. 44-57). Routledge.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*(3), 483–517. https://doi.org/10.1017/S0272263113000119

Goethe Institut (2020) Einstufungstest [Placement test]. Munich: Goethe Institut. Available at: https://www.goethe.de/en/spr/kup/tsd.html (accessed January 2020).

Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. *Studies in Second Language Acquisition*, *34*(3), 445–474.

Goo, J., & Mackey, A. (2013). The case against the case against recasts. *Studies in Second Language Acquisition*, *35*(1), 127–165. https://doi.org/10.1017/S0272263112000708

Granena, G., & Yilmaz, Y. (2019). Corrective feedback and the role of implicit sequence-learning ability in L2 online performance. *Language Learning*, *69*, 127-156.

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, *94*, 149–165. https://doi.org/10.1016/j.jml.2016.11.001

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem?. *Second Language Research*, *28*(2), 191-215.

Grüter, T., Rohde, H., & Schafer, A. J. (2017). Coreference and discourse coherence in L2: The roles of grammatical aspect and referential form. *Linguistic Approaches to Bilingualism*, *7*(2), 199-229.

Grüter, T., Zhu, Y. A., & Jackson, C. N. (2021). Forcing prediction increases priming and adaptation in second language production. In T. Grüter & E. Kaan (Eds.), *Prediction in second language processing and learning* (pp. 207 – 231). John Benjamins.

Gurzynski-Weiss, L., & Révész, A. (2012). Tasks, teacher feedback, and learner modified output in naturally occurring classroom interaction. *Language Learning*, *62*(3), 851–879.

Han, Z. (2002). A study of the impact of recasts on tense consistency in L2 output. *TESOL Quarterly*, *36*(4), 543–572. https://doi.org/10.2307/3588240

Hartsuiker, R. J., Beerts, S., Loncke, M., Desmet, T., & Bernolet, S. (2016). Cross-linguistic structural priming in multilinguals: Further evidence for shared syntax. *Journal of Memory and Language*, *90*, 14–30. https://doi.org/10.1016/j.jml.2016.03.003

Hartsuiker, R. J., & Bernolet, S. (2017). The development of shared syntax in second language learning. *Bilingualism: Language and Cognition*, *20*(2), 219–234. https://doi.org/10.1017/S1366728915000164

Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition*, *75*(2), B27–B39. https://doi.org/10.1016/S0010-0277(99)00080-3

Havron, N., Carvalho, A. de, Fiévet, A.-C., & Christophe, A. (2019). Three- to four-year-old children rapidly adapt their predictions and use them to learn novel word meanings. *Child Development*, *90*(1), 82–90. https://doi.org/10.1111/cdev.13113

Hebb, D. O. (1949). *The organisation of behaviour: A neuropsychological theory*. Science Editions New York.

Henry, N. (2022). The offline and online effects of processing instruction. *Applied Psycholinguistics*, *43*(4), 945-971.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460–474. https://doi.org/10.1037/0096-1523.34.2.460

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1352–1374. https://doi.org/10.1037/xlm0000388

Hochmann, J. R., & Papeo, L. (2014). The invariance problem in infancy: A pupillometry study. *Psychological Science*, *25*(11), 2038-2046.

Holroyd, C. B., Krigolson, O. E., Baker, R., Lee, S., & Gibson, J. (2009). When is an error not a prediction error? An electrophysiological investigation. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 59–70. https://doi.org/10.3758/CABN.9.1.59

Hopman, E. W., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science*, *29*(6), 961-971.

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29*(1), 33–56.

Hopp, H. (2015a). Semantics and morphosyntax in predictive L2 sentence processing. *International Review of Applied Linguistics in Language Teaching*, *53*(3), 277–306. https://doi.org/10.1515/iral-2015-0014

Hopp, H. (2015b). Individual differences in the second language processing of object–subject ambiguities. *Applied Psycholinguistics*, *36*(2), 129-173.

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*, *32*(2), 277–307.

Hopp, H. (2018). The bilingual mental lexicon in L2 sentence processing. *Second Language*, *17*, 5–27.

Hopp, H. (2020). Morphosyntactic adaptation in adult L2 processing: Exposure and the processing of case and tense violations. *Applied Psycholinguistics*, *41*(3), 627-656.

Hopp, H., & Lemmerth, N. (2018). Lexical and syntactic congruency in L2 predictive gender processing. *Studies in Second Language Acquisition*, *40*(1), 171-199.

Huettig, F. (2015). Four central questions about prediction in language processing—ScienceDirect. *Brain Research*, *1626*(11), 118–135.

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*, *31*(1), 80–93. http://dx.doi.org/10.1080/23273798.2015.1047459

Husband, E. M., & Bovolenta, G. (2020). Prediction failure blocks the use of local semantic context. *Language, Cognition and Neurosience*, *35*(3), 273–291.

Inagaki, K., & Hatano, G. (1977). Amplification of cognitive motivation and its effects on epistemic observation. *American Educational Research Journal*, *14*(4), 485–491.

Ishida, M. (2004). Effects of recasts on the acquisition of the aspectual form -te i-(ru) by learners of Japanese as a foreign language. *Language Learning*, *54*(2), 311–394. https://doi.org/10.1111/j.1467-9922.2004.00257.x

Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition*, *21*(2), 251–264. https://doi.org/10.1017/S1366728917000050

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. https://doi.org/10.1017/S0272263103000019

Jackson, C. N., Dussias, P. E., & Hristova, A. (2012). Using eye-tracking to study the on-line processing of case-marking information among intermediate L2 learners of German. *International Review of Applied Linguistics in Language Teaching*, *50*(2), 101-133.

Jackson, C. N., & Hopp, H. (2020). Prediction error and implicit learning in L1 and L2 syntactic priming. *International Journal of Bilingualism*, *24*(5-6), 895-911.

Jackson, C. N., & Ruf, H. T. (2017). The priming of word order in second language German. *Applied Psycholinguistics*, *38*(2), 315-345.

Jackson, C. N., & Ruf, H. T. (2018). The importance of prime repetition among intermediate-level second language learners. *Studies in Second Language Acquisition*, *40*(3), 677-692.

Jaeger, F. T., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*(2), 257–282. https://doi.org/10.1075/lab.4.2.05kaa

Kaan, E. (2015). Knowing without predicting, predicting without learning. *Linguistic Approaches to Bilingualism*, *5*(4), 482–486. https://doi.org/10.1075/lab.5.4.07kaa

Kaan, E., & Chun, E. (2018). Priming and adaptation in native speakers and second-language learners. *Bilingualism: Language and Cognition*, *21*(2), 228–242. https://doi.org/10.1017/S1366728916001231

Kaan, E., Futch, C., Fuertes, R. F., Mujcinovic, S., & de la Fuente, E. Á. (2019). Adaptation to syntactic structures in native and nonnative sentence comprehension. *Applied Psycholinguistics*, *40*(1), 3-27.

Kaan, E., & Grüter, T. (2021). Prediction in second language processing and learning. Advances and directions. In E. Kaan & T. Grüter (Eds.), *Prediction in second language processing and learning* (pp. 1-24). John Benjamins Publishing Company.

Kafkas, A. (2021). Encoding-linked pupil response is modulated by expected and unexpected novelty: Implications for memory formation and neurotransmission. *Neurobiology of Learning and Memory, 180,* 107412.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research 32*, *32*, 37–55.

Kang, M. J., Hsu, M., Kraijbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The Wick in the Candle of Learning: Epistemic Curiosity Activates Reward Circuitry and Enhances Memory -. *Psychological Science*, *20*(8), 963–973.

Kartchava, E. (2019). *Noticing oral corrective feedback in the second language classroom: Background and evidence*. Rowman & Littlefield.

Kartchava, E., & Ammar, A. (2014). Learners' beliefs as mediators of what is noticed and learned in the language classroom. *Tesol Quarterly*, *48*(1), 86-109.

Kaschak, M. P. (2007). Long-term structural priming affects subsequent patterns of language production. *Memory & Cognition*, *35*, 925–937.

Kaschak, M. P., Kutta, T. J., & Coyle, J. M. (2014). Long and short term cumulative structural priming effects. *Language, Cognition and Neuroscience*, *29*(6), 728–743. https://doi.org/10.1080/01690965.2011.641387

Kaschak, M. P., Kutta, T. J., & Jones, J. L. (2011). Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic Bulletin & Review*, *18*, 1133–1139.

Keppenne, V., Hopman, E. W., & Jackson, C. N. (2021). Production-based training benefits the comprehension and production of grammatical gender in L2 German. *Applied Psycholinguistics*, *42*(4), 907-936.

Kim, Y., & McDonough, K. (2008). Learners' Production of Passives during Syntactic Priming Activities. *Applied Linguistics*, *29*(1), 149–154. https://doi.org/10.1093/applin/amn004

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*(1), 95–127. https://doi.org/10.1016/j.cognition.2004.03.002

Kootstra, G. J., & Doedens, W. J. (2016). How multiple sources of experience influence bilingual syntactic choice: Immediate and cumulative cross-language effects of structural priming, verb bias, and language dominance. *Bilingualism: Language and Cognition*, *19*(4), 710–732. https://doi.org/10.1017/S1366728916000420

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccesfsul retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294. https://doi.org/10.1037/a0037850

Kruschke, J. K. (2011). Models of attentional learning. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches to Categorization* (pp. 120 – 152). Psychology Press.

Laufer, B. (1998). Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics*, *19*(2), 255–271. https://doi.org/10.1093/applin/19.2.255

Lavin, C., San Martín, R., *&* Rosales Jubal, E. *(*2014*)*. Pupil dilation signals uncertainty and surprise in a learning gambling task. *Frontiers in Behavioral Neuroscience*, *7*, 218.

Lee, S. W., O'Doherty, J. P., & Shimojo, S. (2015). Neural Computations Mediating One-Shot Learning in the Human Brain. *PLOS Biology*, *13*(4), e1002137. https://doi.org/10.1371/journal.pbio.1002137

Leeman, J. (2003). Recasts and second language development: Beyond negative evidence. *Studies in Second Language Acquisition*, *25*(1), 37–63. https://doi.org/10.1017/S0272263103000020

Lemhöfer, K., Schriefers, H., & Hanique, I. (2010). Native language effects in learning second-language grammatical gender: A training study. *Acta Psychologica*, *135*(2), 150-158.

Leow, R. P. (1997). Attention, awareness, and foreign language behavior. *Language Learning*, *47*(3), 467-505.

Leow, R. P. (2001). Attention, awareness, and foreign language behavior. *Language Learning*, *51*, 113-155.

Leow, R. P. (2019). *The Routledge Handbook of Second Language Research in Classroom Learning*. Routledge.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, *63*(4), 447–464. https://doi.org/10.1016/j.jml.2010.07.003

Li, S. (2010). The Effectiveness of Corrective Feedback in SLA: A Meta-Analysis. *Language Learning*, *60*(2), 309–365.

Li, S. (2014). The interface between feedback type, L2 proficiency, and the nature of the linguistic target. *Language Teaching Research*, *18*(3), 373–396.

Li, S., & Vuono, A. (2019). Twenty-five years of research on oral and written corrective feedback in System. *System*, *84*, 93–109.

Lightbown, P. M., & Spada, N. (1999). *How languages are learned*. OUP.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., ... & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*(3), 530-566.

Litcofsky, K. A., & Van Hell, J. G. (2019). Bi-directional evidence linking sentence production and comprehension: A cross-modality structural priming study. *Frontiers in Psychology*, *10*: 1095.

Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics*, *41*(5), 791–824. https://doi.org/10.1515/ling.2003.026

Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition*, *27*(3), 361–386. https://doi.org/10.1017/S0272263105050163

Loewen, S. & Erlam, R. (2007). Corrective feedback in the chatroom: an experimental study. *Computer Assisted Language Learning, 19*(1), 1-14.

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). OUP.

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

Long, J. A. (2019*). interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions*. R package version 1.1.0. https://cran.r-project.org/package=interactions

Long, M. H. (1980). Input, interaction, and second language acquisition. *Unpublished doctoral dissertation*. University of California, Los Angeles.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Long, M. H. (2007). *Problems in SLA*. Erlbaum.

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. https://doi.org/10.1111/j.1540-4781.1998.tb01213.x

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

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

Lyster, R., & Izquierdo, J. (2009). Prompts versus recasts in dyadic interaction. *Language Learning*, *59*(2), 453–498.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, *19*(1), 37–66. https://doi.org/10.1017/S0272263197001034

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*(2), 265–302. https://doi.org/10.1017/S0272263109990520

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1–40. https://doi.org/10.1017/S0261444812000365

Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21*(4), 557–587. https://doi.org/10.1017/S0272263199004027

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, *27*(3), 405–430. https://doi.org/10.1093/applin/ami051

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*(4), 471–497. https://doi.org/10.1017/S0272263100004010

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). OUP.

Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *The Modern Language Journal*, *82*(3), 338–356.

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27. https://doi.org/10.1016/j.jml.2016.03.009

Marsh, E. J., Lozito, J. P., Umanath, S., Bjork, E. L., & Bjork, R. A. (2012). Using verification feedback to correct errors made on a multiple-choice test. *Memory*, *20*(6), 645–653. https://doi.org/10.1080/09658211.2012.684882

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports*, *8*(1), 1-9.

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, *50*(1), 94-106.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562. https://doi.org/10.1080/03640210802035357

McClelland, J. L. (2006). How far can you go with Hebbian learning, and when does it lead you astray. *Processes of change in brain and cognitive development: Attention and performance xxi*, *21*, 33-69.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McDonough, K. (2004). Learner-learner interaction during pair and small group activities in a Thai EFL context. *System*, *32*(2), 207–224.

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.

McDonough, K. (2007). Interactional feedback and the emergence of simple past activity verbs in L2 English. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 323–338). OUP.

McDonough, K., & Fulga, A. (2015). The Detection and Primed Production of Novel Constructions. *Language Learning*, *65*(2), 326–357. https://doi.org/10.1111/lang.12103

McDonough, K., & Kim, Y. (2016). Working memory and L2 English speakers' primed and subsequent production of passives. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 205-222). John Benjamins.

McDonough, K., & Mackey, A. (2006). Responses to recasts: Repetitions, primed production, and linguistic development. *Language Learning*, *56*(4), 693–720.

McDonough, K., & Nekrasova-Becker, T. (2014). Comparing the effect of skewed and balanced input on English as a foreign language learners' comprehension of the double-object dative construction. *Applied Psycholinguistics*, *35*(2), 419-442.

McDonough, K., & Trofimovich, P. (2008). *Using Priming Methods in Second Language Research*. Routledge. https://doi.org/10.4324/9780203880944

McDonough, K., & Trofimovich, P. (2013). Learning a novel pattern through balanced and skewed input. *Bilingualism: Language and Cognition*, *16*(3), 654-662.

McDonough, K., & Trofimovich, P. (2015). Structural priming and the acquisition of novel form-meaning mappings. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based persepctives on second language learning* (pp. 105-123). De Gruyter Mouton.

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2016). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology, 54*, 193–203. doi: 10.1111/psyp.12772

Meara, P. M., & Rogers, V.E. (2021). *The LLAMA Tests v3. LLAMA F3: grammar rules*. Retrieved from https://www.lognostics.co.uk/tools/LLAMA_3/

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Mill, R. D., O'Connor, A. R., *&* Dobbins, I. G. *(*2016*)*. Pupil dilation during recognition memory: Isolating unexpected recognition from judgment uncertainty. *Cognition*, *154*, 81–94.

Miller, A. L., & Unsworth, N. (2019). Variation in attention at encoding: Insights from pupillometry and eye gaze fixations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000797

Miller, A. L., & Unsworth, N. (2021). Attending to encode: The role of consistency and intensity of attention in learning ability. *Journal of Memory and Language*, *121*, 104276.

Mitsugi, S., & Macwhinney, B. (2016). The use of case marking for predictive processing in second language Japanese. *Bilingualism: Language and Cognition*, *19*(1), 19–35. https://doi.org/10.1017/S1366728914000881

Muranoi, H. (2000). Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning*, *50*(4), 617–673. https://doi.org/10.1111/0023-8333.00142

Murayama, K. (2022). A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic–extrinsic rewards. *Psychological Review*, *129*(1), 175.

Murphy, P. R., O'connell, R. G., O'sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping*, *35*(8), 4140-4154.

Musan, R. (2002). *The German perfect: its semantic composition and its interactions with temporal adverbials* (Vol. 78). Springer Science & Business Media.

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. (2017). The effectiveness of extensive versus intensive recasts for learning L2 grammar. *The Modern Language Journal*, *101*(2), 353-368.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*(7), 1040-1046.

Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2017). Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. *BioRxiv*, 111807.

Opitz, B., Ferdinand, N. K., & Mecklinger, A. (2011). Timing matters: the impact of immediate and delayed feedback on artificial language learning. *Frontiers in Human Neuroscience*, *5*, 8.

Ortega, L. (2009). *Understanding second language acquisition*. Routledge.

Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, *8*(1), 89. https://doi.org/10.1186/1471-2202-8-89

Otten, M., & Van Berkum, J. J. A. (2008). Discourse-Based Word Anticipation During Language Processing: Prediction or Priming? *Discourse Processes*, *45*(6), 464–496. https://doi.org/10.1080/01638530802356463

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 3–8.

Peter, M., Chang, F., Pine, J. M., Blything, R., & Rowland, C. F. (2015). When and how do children develop knowledge of verb argument structure? Evidence from verb bias effects in a structural priming task. *Journal of Memory and Language*, *81*, 1–15. https://doi.org/10.1016/j.jml.2014.12.002

Phillips, C., & Ehrenhofer, L. (2015). The role of language processing in language acquisition. *Linguistic Approaches to Bilingualism*, *5*(4), 409–453. https://doi.org/10.1075/lab.5.4.01phi

Philp, J. (2003). Constraints on "noticing the gap": Nonnative speakers' noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition*, *25*(1), 99–126. https://doi.org/10.1017/S0272263103000044

Pica, T. (1992). The textual outcomes of native speaker-non-native speaker negotiation: What do they reveal about second language learning? *Text and Context: Cross-Disciplinary Perspective on Language Study*, 198–237.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427–459.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Pickering, M. J., & Garrod, S. (2013). How tightly are production and comprehension interwoven? *Frontiers in Psychology*, *4*: 238. doi: 10.3389/fpsyg.2013.00238

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110. https://doi.org/10.1016/j.tics.2006.12.002

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912.

Porretta, V., Kyröläinen, A., van Rij, J., & Järvikivi, J. (2016). *VWPre: Tools for Preprocessing Visual World Data*. R package version 1.2.4, updated 2020-11-28. https://CRAN.R-project.org/package=VWPre

Pothos, E. M., & Wills, A. J. (2011a). *Formal Approaches in Categorization*. Cambridge University Press.

Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1023–1041. https://doi.org/10.1037/xlm0000637

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology*, *143*(2), 644–667.

Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., & Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, *27*(1), 91–109. https://doi.org/10.1207/s15326985ep2701_7

Preuschoff, K., Marius't Hart, B., & Einhäuser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, *5*, 115.

Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Retrieved from https://support.pstnet.com/.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rabagliati, H., Gambi, C., & Pickering, M. J. (2016). Learning to predict or predicting to learn? *Language, Cognition and Neuroscience*, *31*(1), 94–105. https://doi.org/10.1080/23273798.2015.1077979

Raisig, S., Welke, T., Hagendorf, H., & van der Meer, E. (2010). I spy with my little eye: Detection of temporal violations in event sequences and the pupillary response. *International Journal of Psychophysiology*, *76*(1), 1-8.

Reitter, D., Keller, F., & Moore, J. D. (2011). A Computational Cognitive Model of Syntactic Priming. *Cognitive Science*, *35*(4), 587–637. https://doi.org/10.1111/j.1551-6709.2010.01165.x

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current Research and Theory*, 64–99.

Révész, A. (2012). Working Memory and the Observed Effectiveness of Recasts on Different L2 Outcome Measures. *Language Learning*, *62*(1), 93–132.

Roberts, L., & Liszka, S. A. (2013). Processing tense/aspect-agreement violations on-line in the second language: A self-paced reading study with French and German L2 learners of English. *Second Language Research*, *29*(4), 413-439.

Rothstein, B. (2008). *The perfect time span: On the present perfect in German, Swedish and English* (Vol. 125). Benjamins.

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning*, *45*(2), 283–331. https://doi.org/10.1111/j.1467-1770.1995.tb00441.x

Robison, M. K., & Unsworth, N. (2019). Pupillometry tracks fluctuations in working memory performance. *Attention, Perception, & Psychophysics*, *81*(2), 407-419.

Romberg, A., & Saffran, J. (2013). Expectancy learning from probabilistic input by infants. *Frontiers in Psychology*, *3*, 610. https://doi.org/10.3389/fpsyg.2012.00610

Sabah, K., Dolk, T., Meiran, N., & Dreisbach, G. (2019). When less is more: Costs and benefits of varied vs. fixed content and structure in short-term task switching training. *Psychological Research*, *83*(7), 1531-1542.

Sabah, K., Dolk, T., Meiran, N., & Dreisbach, G. (2021). Enhancing task-demands disrupts learning but enhances transfer gains in short-term task-switching training. *Psychological Research*, *85*(4), 1473-1487.

Sagarra, N., & Abbuhl, R. (2013). Optimizing the noticing of recasts via computer-delivered feedback: Evidence that oral input enhancement and working memory help second language learning. *The Modern Language Journal*, *97*(1), 196-216.

Sagarra, N., & Ellis, N. C. (2013). From seeing adverbs to seeing verbal morphology: Language experience and adult acquisition of L2 tense. *Studies in Second Language Acquisition*, *35*(2), 261-290.

Saito, K., & Lyster, R. (2012a). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning*, *62*(2), 595–633. https://doi.org/10.1111/j.1467-9922.2011.00639.x

Saito, K., & Lyster, R. (2012b). Investigating the pedagogical potential of recasts for L2 vowel acquisition. *TESOL Quarterly*, *46*(2), 387–398.

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nature Reviews Neuroscience*, *10*(3), 211-223.

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: the locus coeruleus mediates cognition through arousal. *Neuron*, *76*(1), 130-141.

Sauro, S. (2021). Computer-mediated corrective feedback and the development of L2 grammar. *UMBC Education Department Collection*.

Scheepers, C., & Crocker, M. W. (2004). Constituent order priming from reading to listening: A visual-world study. In M. Carreiras & C. Clifton Jr. (Eds.), *The on-line study of sentence comprehension: eyetracking, ERP, and beyond* (pp. 167–185). Psychology Press.

Scheepers, C., Mohr, S., Fischer, M. H., & Roberts, A. M. (2013). Listening to limericks: a pupillometry investigation of perceivers' expectancy. *Plos one*, *8*(9), e74986.

Schmidt, R. W. (1990). Role of Consciousness in Second Language Learning. *Applied Linguistics*, *11*(2), 129–158.

Schmidt, R. W. (2001). Attention. In P. R. Robinson, M. H. Long, & J. C. Richards (Eds.), *Cognition and second language instruction* (pp. 3–32). CUP.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207–217.

Schmidtke, J. (2018). Pupillometry in linguistic research: An introduction and review for second language researchers. *Studies in Second Language Acquisition*, *40*(3), 529–549. https://doi.org/10.1017/S0272263117000195

Schmuck, M. (2013). *Relevanzgesteuerter verbalmorphologischer Umbau: eine kontrastive Untersuchung zum Deutschen, Niederländischen und Schwedischen*. G. Olms.

Schoonbaert, S., Hartsuiker, R. J., & Pickering, M. J. (2007). The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language*, *56*(2), 153-171.

Schremm, A., Hed, A., Horne, M., & Roll, M. (2017). Training predictive L2 processing with a digital game: Prototype promotes acquisition of anticipatory use of tone-suffix associations. *Computers & Education*, *114*, 206-221.

Seabrooke, T., Mitchell, C. J., Wills, A. J., Waters, J. L., & Hollins, T. J. (2019). Selective effects of errorful generation on recognition memory: The role of motivation and surprise. *Memory*, *27*(9), 1250-1262.

Sebastian-Gallés, N., Rodríguez-Fornells, A., de Diego-Balaguer, R., & Díaz, B. (2006). First-and second-language phonological representations in the mental lexicon. *Journal of Cognitive Neuroscience*, *18*(8), 1277-1291.

Sekicki, M., & Staudte, M. (2018). Eye'll help you out! How the gaze cue reduces the cognitive load required for reference processing. *Cognitive science*, *42*(8), 2418-2458.

Segaert, K., Kempen, G., Petersson, K. M., & Hagoort, P. (2013). Syntactic priming and the lexical boost effect during sentence production and sentence comprehension: An fMRI study. *Brain and Language*, *124*(2), 174–183. https://doi.org/10.1016/j.bandl.2012.12.003

Sheen, Y. (2004). Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Language Teaching Research*, *8*(3), 263–300.

Sheen, Y. (2006). Exploring the relationship between characteristics of recasts and learner uptake. *Language Teaching Research*, *10*(4), 361–392. https://doi.org/10.1191/1362168806lr203oa

Sheen, Y. (2007). The effects of corrective feedback, language aptitude, and learner attitudes on the acquisition of English articles. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 301–322). OUP.

Sheen, Y. (2008). Recasts, language anxiety, modified output, and L2 learning. *Language Learning*, *58*(4), 835–874.

Sheen, Y. (2010). Differential effects of oral and written corrective feedback in the ESL classroom. *Studies in Second language Acquisition*, *32*(2), 203-234.

Sheen, Y. (2011). *Corrective feedback, individual differences and second language learning*. Springer.

Sheen, Y., & Ellis, R. (2011). Corrective feedback in language teaching. *Handbook of Research in Second Language Teaching and Learning*, *2*, 593–610.

Shin, J.-A., & Christianson, K. (2012). Structural Priming and Second Language Learning. *Language Learning*, *62*(3), 931–964.

Sibley, C., Foroughi, C. K., Moclaire, C. M., King, K. M., Brown, N. L., & Coyne, J. T. (2019, November). Additional Evidence for Pupil Size as a Measure of Within-Task Learning. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 54-58). SAGE Publications.

SR Research Ltd. (2018). *EyeLink Data Viewer software*, *version 3.2.1*.

Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *The Canadian Modern Language Review*, *50*(1), 158–164. https://doi.org/10.3138/cmlr.50.1.158

Swain, M. (2005). *The Output Hypothesis: Theory and research*. Routledge.

Theobald, M., Galeano-Keiner, E. M., & Brod, G. (2022). Predicting vs. Guessing: The role of confidence for pupillometric markers of curiosity and surprise. *Cognition and Emotion*, *36*(4), 731–740.

Thomas, K. E. (2018). Comparing explicit exemplar-based and rule-based corrective feedback: Introducing analogy-based corrective feedback. *The Modern Language Journal*, *102*(2), 371–391.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1-42.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, *16*(2), 183–203.

Trofimovich, P., & McDonough, K. (2011). *Applying priming methods to L2 learning, teaching and research.* John Benjamins Publishing.

University of Wisconsin Testing and Evaluation. (2006). German Language Placement Test. Madison,WI: University of Wisconsin.

Van Bergen, G., & Flecken, M. (2017). Putting things in new places: Linguistic experience modulates the predictive power of placement verb semantics. *Journal of Memory and Language*, *92*, 26-42.

van der Helden, J., Boksem, M. A. S., & Blom, J. H. G. (2010). The importance of failure: Feedback-related negativity predicts motor learning efficiency. *Cerebral Cortex*, *20*(7), 1596–1603. https://doi.org/10.1093/cercor/bhp224

van Vugt, F. T., & Tillmann, B. (2015). Auditory feedback in error-based learning of motor regularity. *Brain Research*, *1606*, 54–67. https://doi.org/10.1016/j.brainres.2015.02.026

VanPatten, B., Williams, J., Rott, S., & Overstreet, M. (2004). *Form-meaning connections in second language acquisition*. Routledge.

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

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, *19*, 899–905.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1): 79-95 https://doi.org/10.1017/S0272263108080042

Weber, K., Christiansen, M. H., Indefrey, P., & Hagoort, P. (2019). Primed from the start: Syntactic priming during the first days of language learning. *Language Learning*, *69*(1), 198-221.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. https://doi.org/10.1162/0898929041920487

Williams, J. N., & Cheung, A. (2011). Using priming to explore early word learning. In P. Trofimovich & K. McDonough (Eds.), *Applying priming methods to L2 learning, teaching and research: Insights from psycholinguistics* (pp. 73 – 104). John Benjamins.

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*(2), 235–263.

Yilmaz, Y., & Granena, G. (2016). The role of cognitive aptitudes for explicit language learning in the relative effects of explicit and implicit feedback. *Bilingualism: Language and Cognition*, *19*(1), 147-161.

Yilmaz, Y., & Granena, G. (2019). Cognitive individual differences as predictors of improvement and awareness under implicit and explicit feedback conditions. *The Modern Language Journal*, *103*(3), 686-702.

Yilmaz, Y., & Granena, G. (2021). Implicitness And Explicitness In Cognitive Abilities And Corrective Feedback: A Double Dissociation?. *Studies in Second Language Acquisition*, *43*(3), 523-550.

Zhang, F., & Emberson, L. L. (2020). Using pupillometry to investigate predictive processes in infancy. *Infancy*, *25*(6), 758–780. https://doi.org/10.1111/infa.12358

Zhang, F., Jaffe-Dax, S., Wilson, R. C., & Emberson, L. L. (2018). Prediction in infants and adults: A pupillometry study. *Developmental Science*, *22*(4), e12780.

Zhao, Y., & Ellis, R. (2020). The relative effects of implicit and explicit corrective feedback on the acquisition of 3rd person-s by Chinese university students: A classroom-based study. *Language Teaching Research*, 1362168820903343.

Zeileis, A., and Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software, 14*(6), 1-27. doi:10.18637/jss.v014.i06

Zirnstein, M., van Hell, J. G., & Kroll, J. F. (2018). Cognitive control ability mediates prediction costs in monolinguals and bilinguals. *Cognition*, *176*, 87–106. https://doi.org/10.1016/j.cognition.2018.03.001

Appendix A

Language Background Questionnaire Experiment 1 and 2

Q1
Part I.
The first section of the questionnaire deals with some of your demographics and language exposure.

Q3 Gender:

____

Q4 Age (in years):

____

Q9 List your academic major(s). Separate with semi-colons, if necessary.

____

____

____

____

____

Q10 Native country / countries (please check all that apply):

- ☐ United States (1)
- ☐ Other (please specify): (2) ________________

Q11 Native language(s) (please check all that apply):

- ☐ English (1)
- ☐ Other (please specify): (2) ________________

Q35 Were you exposed to any languages other than English before the age of 5?

- ○ No (1)
- ○ Yes (2)

Display This Question:

If Were you exposed to any languages other than English before the age of 5? = Yes

Q48
Please specify which languages you were exposed to before the age of 5, including extent and environment of exposure. Separate with semi-colons, if necessary.

Example: Spanish, daily, kindergarten; German, monthly, grandparents

____

____

____

____

Page Break

Q13 Please list all of the languages you know in order of ***dominance*** (i.e. how often you use each language), including your native language. Then please list what percentage of the time you ***currently*** and on ***average*** use each of these languages (your percentages should add up to 100%).

	Language (4)	Average use (% current use) (5)
1: (1)		
2: (2)		
3: (3)		
4: (4)		
5: (5)		

Page Break

Q14 ***Part II. The next section of the questionnaire deals with your foreign language learning experience.***

Q15 If you have learned any foreign language(s), please write the name of the language next to the context in which you learned it and indicate the starting age and length of study for that language. For example, next to "High School" you could write something like "Spanish, 14, 3 years."

	Language(s), starting age, and length of study (1)
Home/Outside of School: (1)	
Elementary School: (2)	
Middle School: (3)	
High School: (4)	
University: (6)	

Q18 Please list the most advanced course(s) in a language besides your native language you have completed at the university level. Example: SPAN 3, GER 1 (if none, indicate "None")

____

Q21 Have you ever studied and/or lived abroad for more than four consecutive weeks?

○ No (1)

○ Yes (2)

Display This Question:

If Have you ever studied and/or lived abroad for more than four consecutive weeks? = Yes

Q48 Where and when did you study and/or live abroad, for how long, and what language(s) did you speak?

	Country (1)	Approximate dates (2)	Length of stay (3)	Language(s) (4)
Stay 1: (1)				
Stay 2: (2)				
Stay 3: (3)				
Stay 4: (4)				

Page Break

Q23 ***Part III.***
The next section of the questionnaire deals with your <u>English language skills</u>. Please rate yourself on each measure by selecting the appropriate number.

Q24 Your reading proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Q25 Your spelling proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Q26 Your writing proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Q53 Your speaking ability. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Q49 Your listening comprehension ability. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Page Break

Q35 ***Part IV.***
The next section of the questionnaire deals with your <u>*German language skills*</u>*. Please rate yourself on each measure by choosing the appropriate number.*

Q50 Your reading proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Q51 Your spelling proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Q52 Your writing proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Q27 Your speaking ability. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Q54 Your listening comprehension ability. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Page Break

Q41
Part V.
This next section of the questionnaire deals with your <u>skills in any additional languages</u> you may know.

Q36 Do you know another language besides English and German?

○ No (1)

○ Yes (2)

Display This Question:

If Do you know another language besides English and German? = Yes

Q37 Please specify one other language you know.

____

Display This Question:

If Do you know another language besides English and German? = Yes

Q38 ***Please rate yourself on each measure for the language you just indicated by choosing the appropriate number.***

Display This Question:

If Do you know another language besides English and German? = Yes

Q55 Your reading proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Display This Question:

If Do you know another language besides English and German? = Yes

Q56 Your spelling proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Display This Question:

If Do you know another language besides English and German? = Yes

Q57 Your writing proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Display This Question:

If Do you know another language besides English and German? = Yes

Q58 Your speaking ability. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Display This Question:

If Do you know another language besides English and German? = Yes

Q59 Your listening comprehension ability. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Page Break

Q39
Part VI.
This next section of the questionnaire deals with your <u>skills in any additional languages</u> you may know.

Q40 Do you know another language besides English, German, and the language you just rated?

○ No (1)

○ Yes (2)

Display This Question:

If Do you know another language besides English, German, and the language you just rated? = Yes

Q41 Please specify one other language you know.

____

Display This Question:

If Do you know another language besides English, German, and the language you just rated? = Yes

Q42 ***Please rate yourself on each measure for the language you just indicated by choosing the appropriate number.***

Display This Question:

If Do you know another language besides English, German, and the language you just rated? = Yes

Q43 Your reading proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Display This Question:

If Do you know another language besides English, German, and the language you just rated? = Yes

Q44 Your spelling proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Display This Question:

If Do you know another language besides English, German, and the language you just rated? = Yes

Q45 Your writing proficiency. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Display This Question:

If Do you know another language besides English, German, and the language you just rated? = Yes

Q46 Your speaking ability. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

Display This Question:

If Do you know another language besides English, German, and the language you just rated? = Yes

Q47 Your listening comprehension ability. (1=very poor and 10=very good)

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)
Rating: (1)	○	○	○	○	○	○	○	○	○	○

End of Block: Part 1

Appendix B

Proficiency Test Experiment 1 and 2

1. “Gestern war ich krank.”
“__ du auch Fieber?”

a. Hättest
b. Hattest
c. Habest
d. Hast

3. „Arbeitet er morgen?“
„Ja, er __.“

a. müsste
b. muss
c. musst
d. musste

5. Du gibst bald eine Party! __ willst du denn einladen?

a. Wer
b. Wem
c. Wen
d. Wo

7. „Hast du einen Plattenspieler?“
„Nein, ich habe __.“

a. nicht
b. keinen
c. kein
d. nicht einen

9. „Was habt ihr denn gestern gemacht?“

a. „Wir gehen ins Kino.“
b. „Wir möchten ins Kino gehen.“
c. „Wir sind ins Kino gegangen.“
d. „Wir waren ins Kino gegangen.“

11. „Was für ein Mann war denn das?“
„Es war __ dicker.“

a. ein
b. einen
c. einer
d. einem

2. Er ist __, denn er hat seine Kamera vergessen.

a. zurückfahren
b. zurückzufahren
c. zurückgefahren
d. zurückfahrend

4. Ich habe mit __ Freund getanzt.

a. mein
b. meinem
c. meinen
d. meiner

6. „Was trinkst du, Cola oder Tee?“
„Ich trinke __ Cola als Tee.“

a. gern
b. am liebsten
c. lieb
d. lieber

8. Für __ist das zu früh.

a. ihm
b. ihr
c. Ihnen
d. ihn

10. Heute ist Montag, ___.

a. am ersten Juni.
b. ersten Juni.

c. den ersten Juni.
d. der erste Juni.

12. „Was hast du am Montag gemacht?“
„Am Montag ___ zum Arzt gegangen.“

a. ich bin
b. bin ich
c. ich war
d. war ich

13. „Weißt du, was er gesagt hat?“
„Nein, daran kann ich ___.“

a. nicht mehr erinnern
b. dich nicht erinnern
c. mich nicht erinnern
d. mir nicht erinnern

15. Hans ist schon wieder zu Hause. Er __ gestern Abend aus dem Urlaub zurück.

a. kommt
b. komme
c. kam
d. kamt

17. Die Dame kam ins Geschäft, sah den blauen Pullover und __ ihn sofort.

a. kaufen
b. kaufte
c. kauft
d. kauftet

19. „Wer ist denn das?“
„Das ist die Frau __.“

a. meines Arztes
b. meinen Arzt
c. meinem Arzt
d. mein Arzt

21. „Was wollt ihr zu Weihnachten bekommen?“
„Wir ___ viele schöne Geschenke bekommen!“

a. mögen
b. mochten
c. wollten
d. möchten

23. „Was machen wir heute Abend?“
„Gehen wir zu Meyers. __ denen wird Geburtstag gefeiert.“

a. An
b. Bei
c. Zu
d. Auf

14. Heute kann ich das verstehen, aber gestern __ ich es nicht.

a. könne
b. konnte
c. kann
d. könnte

16. Jens war so dumm. Er __ die Antwort nicht.

a. weißt
b. wißt
c. wusste
d. weiß

18. Er braucht Hilfe mit seinen Hausaufgaben. Kannst du __?

a. ihm mit ihnen helfen
b. ihnen damit helfen
c. ihn mit ihnen helfen
d. ihm damit helfen

20. „Wie lange seid ihr schon hier?“
„Wir warten schon seit fünf Uhr __euch.“

a. für
b. bis
c. auf
d. vor

22. „Diese Brücke ist neu, nicht wahr?“
„Nein, sie __.“

a. wurde 1937 gebaut
b. wird 1937 gebaut
c. waren 1937 gebaut
d. sind 1937 gebaut

24. „Warum ist Fritz nicht hier?“
„Er ist noch in Bonn und __.“

a. er erst später kommt
b. er erst später ankommt
c. erst später kommt an
d. kommt erst später an

25. „Wie kommt ihr nach Berlin?“
„Wir fahren ___.“

a. mit dem Zug dahin
b. da bei dem Zug
c. da mit dem Zug
d. bei dem Zug dahin

27. Der Komponist, __ heute spielt, kommt aus Berlin.

a. den
b. der
c. dem
d. des

29. „Warum bist du gestern nicht zur Party gekommen?“
„Gestern ___.“

a. ich habe nicht kommen können
b. ich kann nicht kommen
c. könnte ich nicht kommen
d. konnte ich nicht kommen

26. Komm zu Tisch! Vater hat __ schon hingesetzt.

a. ihm
b. ihn
c. sich
d. dich

28. Wenn ich Geld __, würde ich nach Österreich fliegen.

a. habe
b. gehabt hätte
c. hätte
d. hatte

30. Ein Student studiert, ___.

a. eine Stelle zu bekommen
b. zu bekommen eine Stelle
c. eine Stelle bekommen
d. um eine Stelle zu bekommen

<u>Wie lange kann man Informationen behalten?</u>

Was hat (31) ____ Ihre Zahnärztin über Ihre Zähne erzählt? Sie wissen (32) ____ nicht mehr ganz genau? Das ist ganz normal, aber nur dann, wenn das Gespräch (33) ____ Ihrer Ärztin schon länger als 24 Stunden zurückliegt. Denn die Hälfte aller Informationen, (34) ____ wir hören, werden nach einem Tag vergessen. Und nach 48 Stunden haben wir nur (35) ____ ein Viertel des Gesprächs im Kopf.

31.	32.	33.	34.	35.
a. dir	a. er	a. mit	a. das	a. noch
b. Ihnen	b. es	b. über	b. der	b. über
c. Sie	c. sie	c. von	c. die	c. weniger

36. Du willst in Berlin arbeiten? Hast Du dich denn schon ---- eine Stelle bemüht?
    a. an
    b. für
    c. so
    d. um

37. Leider waren die Ferien schon zu Ende, sonst ---- ich mit den Kindern länger geblieben.
    a. hätte
    b. wäre
    c. werde
    d. würde

38. ---- nach Qualität kosten die Jacken Euro 100.- bis Euro 210.-.
    a. Entlang
    b. Je
    c. Jede
    d. Mal

39. Trotz ---- Bemühungen des technischen Personals mussten die Fluggäste stundenlang warten.
    a. viel
    b. viele
    c. vieles
    d. vieler

40. Die Universitätsgebäude sind zwar hässlich, ---- ist aber das Studienangebot sehr vielfältig.
    a. dafür
    b. damit
    c. dazu
    d. hierfür

41. Bei dem neuen Autotyp ist technisch viel verbessert ----- .
    a. geworden
    b. werden
    c. worden
    d. wurde

42. Der kleine Junge hatte lange Haare, ---- ihn viele für ein Mädchen hielten.
    a. darum
    b. deshalb
    c. denn
    d. weshalb

Appendix C

Materials for Experiment 1 and 2

Experiment 1

Target verbs

Target verb	Translation	Auxiliary	Participle	Participle type	Type
essen	eat	haben	gegessen	strong	trained
lesen	read	haben	gelesen	strong	trained
schlafen	sleep	haben	geschlafen	strong	trained
spielen	play	haben	gespielt	weak	trained
arbeiten	work	haben	gearbeitet	weak	trained
rauchen	smoke	haben	geraucht	weak	trained
kochen	cook	haben	gekocht	weak	trained
lernen	learn/study	haben	gelernt	weak	trained
singen	sing	haben	gesungen	strong	untrained
waschen	wash	haben	gewaschen	strong	untrained
angeln	fish	haben	geangelt	weak	untrained
fahren	ride	sein	gefahren	strong	trained
laufen	run	sein	gelaufen	strong	trained
klettern	climb	sein	geklettert	weak	trained
wandern	hike	sein	gewandert	weak	trained
schwimmen	swim	sein	geschwommen	strong	untrained
surfen	surf	sein	gesurft	weak	untrained
gehen	walk	sein	gegangen	strong	untrained

Sentence fragments

#	Proper noun	Auxiliary	Adverbial 1	Adverbial 2	Participle
1	Tina	hat/ist	gestern Abend	spät	…
	Tina	has	yesterday night	late	
2	Tina	hat/ist	gestern Morgen	früh	…
	Tina	has	yesterday morning	early	
3	Tina	hat/ist	letzten Monat	oft	…
	Tina	has	last month	often	
4	Tina	hat/ist	letzte Woche	immer	…
	Tina	has	last week	always	
5	Tina	hat/ist	letztes Wochenende	häufig	…
	Tina	has	on the weekend	often	
6	Tina	hat/ist	gestern Nachmittag	viel	…
	Tina	has	yesterday afternoon	a lot	

Experiment 2

Target verbs

Target verb	Translation	Auxiliary	Participle	Participle type	Type
essen	eat	haben	gegessen	strong	trained
lesen	read	haben	gelesen	strong	trained
schlafen	sleep	haben	geschlafen	strong	trained
spielen	play	haben	gespielt	weak	trained
arbeiten	work	haben	gearbeitet	weak	trained
rauchen	smoke	haben	geraucht	weak	trained
kochen	cook	haben	gekocht	weak	trained
lernen	learn/study	haben	gelernt	weak	trained
singen	sing	haben	gesungen	strong	trained
waschen	wash	haben	gewaschen	strong	trained
angeln	fish	haben	geangelt	weak	trained
duschen	shower	haben	geduscht	weak	trained
bügeln	iron	haben	gebügelt	weak	trained
packen	pack	haben	gepackt	weak	trained
putzen	clean	haben	geputzt	weak	trained
weinen	cry	haben	geweint	weak	trained
fahren	ride (a bike)	sein	gefahren	strong	trained
laufen	run	sein	gelaufen	strong	trained
klettern	climb	sein	geklettert	weak	trained
wandern	hike	sein	gewandert	weak	trained
schwimmen	swim	sein	geschwommen	strong	trained
surfen	surf	sein	gesurft	weak	trained
gehen	walk	sein	gegangen	strong	trained
reiten	ride (a horse)	sein	geritten	strong	trained

Sentence fragments

#	Proper noun	Auxiliary	Adverbial 1	Adverbial 2	Participle
1	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	gestern Abend	spät	…
	The girl/woman/boy/man	has	yesterday night	late	
2	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	gestern Morgen	früh	…
	The girl/woman/boy/man	has	yesterday morning	early	
3	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	letzten Monat	oft	…
	The girl/woman/boy/man	has	last month	often	

4	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	letztes Wochenende	häufig	…
	The girl/woman/boy/man	has	on the weekend	often	
5	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	gestern Nachmittag	wahrscheinlich	…
	The girl/woman/boy/man	has	yesterday afternoon	probably	
6	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	im Jahr 2021	unter der Woche	…
	The girl/woman/boy/man	has	in 2021	during the week	
7	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	letzten Frühling	öfter	…
	The girl/woman/boy/man	has	last spring	often	
8	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	letzten Samstag	morgens	…
	The girl/woman/boy/man	has	last Saturday	morning	
9	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	letzten Montag	wohl	…
	The girl/woman/boy/man	has	last Monday	most likely	
10	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	letztes Jahr	jede Woche	…
	The girl/woman/boy/man	has	last year	every week	
11	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	letzten Donnerstag	nachmittags	…
	The girl/woman/boy/man	has	last Thursday	afternoon	
12	Das Mädchen/Die Frau/Der Junge/Der Mann	hat/ist	letzten Sommer	am Wochenende	…
	The girl/woman/boy/man	has	last summer	on the weekends	

Appendix D

Experiment 1 Session 2 Task Instructions

<u>Welcome</u>

Welcome to Session 2 of the German Language Experiment!

This session consists of two parts.

- Part 1: You will imagine a scenario and complete simple tasks.
- Part 2: You will complete a questionnaire.

The whole session takes approximately 70 minutes.

While you complete all tasks, please do not use any outside sources or tools (like dictionaries or translation websites) or write anything down. We are interested in what you know without any help!

Click "Next" to receive instructions for the first part.

Next

Training Block 1 Semantics Feedback

Instructions

Imagine you are in Germany and receive a phone call from a detective to schedule an interview. They would like to ask you about your roommate Tina, who is suspected of having caused a car accident.

You should remember Tina from the first session.

The detective would like to know if you can confirm Tina's whereabouts and activities leading up to the night of the accident. You know that Tina is innocent, but you can't quite recall what she did these past couple of days.

Next

On the day of the interview, the detective shows you incomplete sentences of what Tina said she was doing, along with two pictures. Only one of these pictures fits with what she said she did.

Your task is to guess everything Tina said she did by choosing a picture and completing the incomplete sentences. The closer your guesses are to what she said she did, the more convinced the detective will be that Tina is innocent!

Back Next

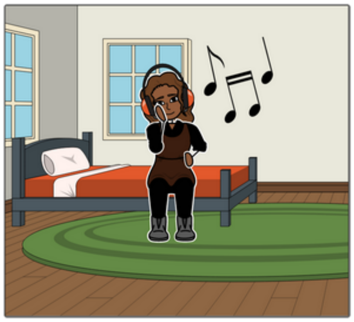
Each time you choose a picture and complete an incomplete sentence, the detective will let you know whether it fits with what Tina said she did.

Let’s try this with a warm-up scenario.

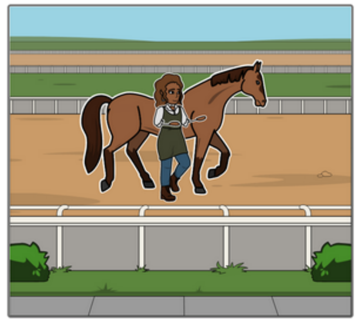
Back Next

+

Tina hat letzten Monat ...



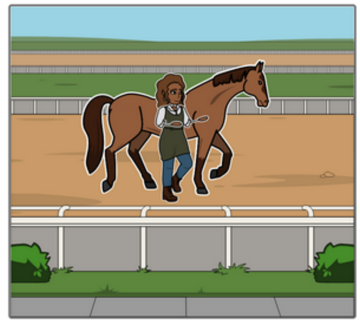
Musik hören



reiten

Press either the F- or J-key to choose the picture that best fits the incomplete sentence and pay attention to the verb that you chose. Once you press a key, you will automatically get to the next screen.

Type the appropriate form of the verb underneath the picture you just chose.
When finished, press the "Enter" key to move to the next screen.



A green frame around one of the pictures indicates what Tina said she did.

That's all there is to it.

Once you click "Next", the interview with the detective will begin.

When you see the crosshair (+) on your screen, place your fingers on the F- and J-keys.

Remember to be as accurate and quick as possible in guessing what Tina said she did to convice the detective that she is innocent!

Next

Place your fingers on the F- and J-keys!

Training Block 1 Form Feedback

(identical to Training Block 1 Semantics Feedback except for the feedback slide)

Musik gehört

This tells you what Tina said she did.

<u>Training Block 1 End</u>

15%

The detective decides to take a short break. While they leave to go get you a glass of water, you decide to go over in your head what Tina might have been doing.

Next

<u>Interim Test Picture Description</u>

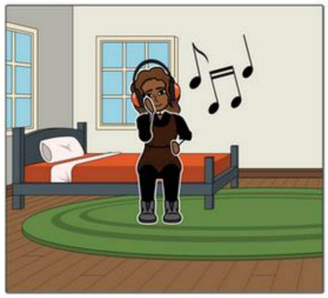
You start of by thinking about what you might tell the detective when they come back.

Your task now is to look at pictures similar to what you were just shown in the interview and to come up with your own descriptions of each picture.

On your screen, you will see pictures of Tina, and underneath each picture, you will also see a verb in German that corresponds to what Tina might have been doing. Your task is to describe each picture and what Tina might have been doing over the last month with one sentence in German.

Next

Your sentences should look something like this.



Musik hören

Tina hat Musik gehört. !

Back Next

Once you click "Next", you will start going over what Tina might have done by creating your own picture descriptions.

Remember to press "Enter" after you have described a picture to move on to the next one.

Back Next

Interim Test Picture Description End

30%

You decide that you now have a good idea of what Tina might have done. Before the detective returns, however, you also decide to practice choosing between two pictures.

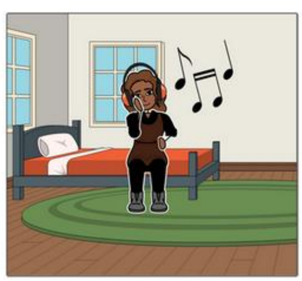
Next

Interim Test Picture Matching

Your screen will look something like this.

Tina hat letzten Monat ...

reiten



Musik hören

Next

To choose the left picture, press "F". To choose the right picture, press "J". As soon as you press one of those keys, the next screen will automatically open.

Remember to be as accurate and quick as possible!

Once you click "Next", you will start going over choosing pictures that the detective might present to you.

Back Next

Place your fingers on the F- and J-keys!

Interim Test Picture Matching End

45%

You are now ready for the detective to return and continue the interview.

Next

Training Block 2 Semantics Feedback

When the detective returns with the glass of water, they decide to ask you more questions. Once again, you will be given excerpts of what Tina said and have to guess what she did by choosing a picture and completing her sentences.

However, the detective has decided to change their strategy. This time, they will give you a different kind of feedback to see if that changes your story.

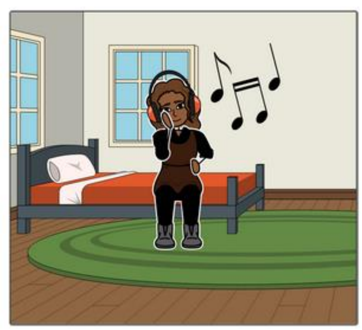
Let's try another warm-up scenario.

Next

+

Tina hat letzten Monat ...

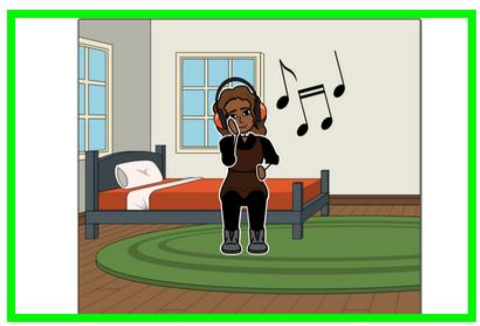
reiten



Musik hören

Press either the F- or J-key to choose the picture that best fits with the incomplete sentence and pay attention to the verb that you chose. Once you press a key, you will automatically get to the next screen.

Type the appropriate form of the verb underneath the picture you just chose.
When finished, press the "Enter" key to move to the next screen.



A green frame around one of the pictures indicates what Tina said she did.

Once you click "Next", the interview with the detective will continue.

When you see the crosshair (+) on the screen, place your fingers on the F- and J-keys.

Remember to be as accurate and quick as possible in guessing what Tina said she did to convice the detective that she is innocent!

Next

Training Block 2 Form Feedback

(identical to Training Block 2 Semantics Feedback except for the feedback slide)

Musik gehört

This tells you what Tina said she did.

Training Block 2 End

60%

The interview with the detective is now over. Still, you decide to review what you learned about Tina's activities leading up to the night of the alleged accident.

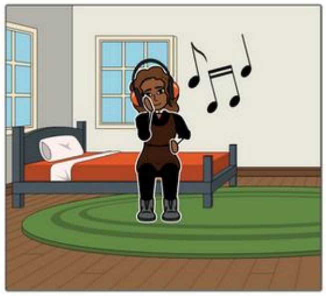
Next

Immediate Posttest Picture Description

Once again, you will see pictures of Tina on your screen, and underneath each picture, you will also see a verb in German that corresponds to what Tina was doing. Your task is to describe each picture and what Tina was doing over the last month with one sentence in German.

Next

Your sentences should look something like this.



Musik hören

Tina hat Musik gehört. !

Back Next

Once you click "Next", you will start going over what you learned about Tina's activities over the last month.

Remember to press "Enter" after you have described a picture to move on to the next one.

Back Next

Immediate Posttest Picture Description End

75%

You are now done describing pictures of what Tina did over the last month, but you decide to also review the detective's strategy of making you match your account of the past to Tina's.

Next

Immediate Posttest Picture Matching

Your screen will look something like this.

Tina hat letzten Monat ...

reiten

Musik hören

Next

To choose the left picture, press "F". To choose the right picture, press "J". As soon as you press one of those keys, the next screen will automatically open.

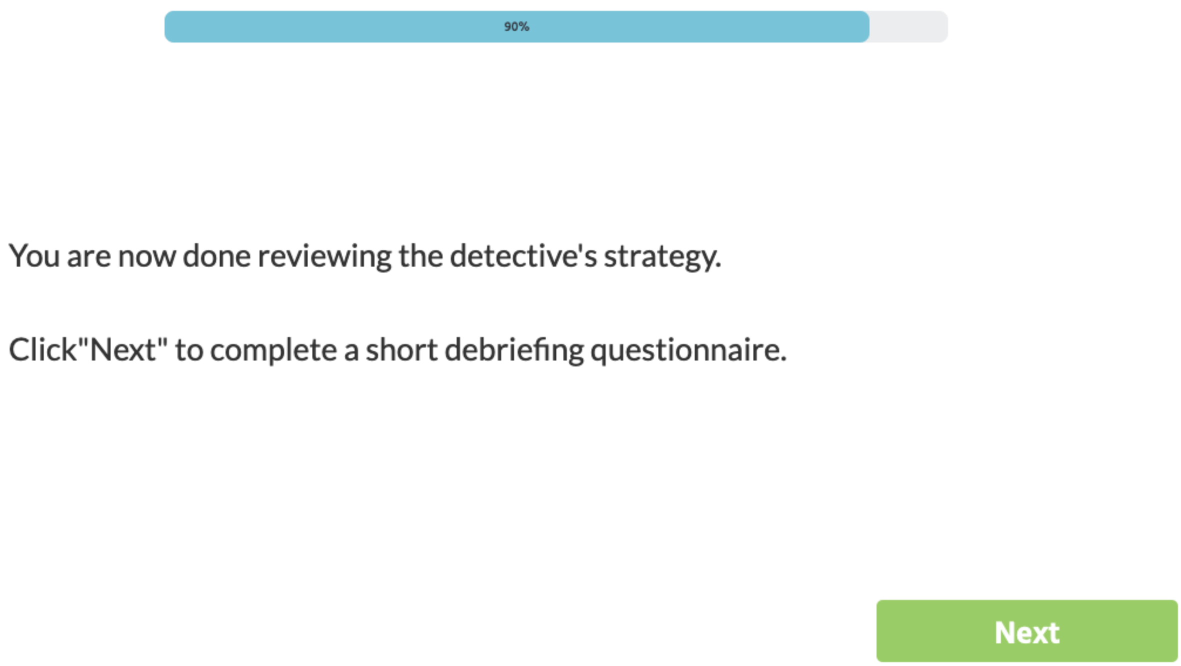
Remember to be as accurate and quick as possible!

Click "Next" to review the detective's strategy and to choose the pictures that fit with the incomplete sentence best.

Back Next

Place your fingers on the F- and J-keys!

Immediate Posttest Picture Matching End



Appendix E

Debriefing Questionnaires Experiment 1 and 2

Experiment 1

Debriefing Questionnaire 1

This questionnaire was administered after the immediate posttest in Session 2.

Question 1
On a scale from 1 (no effort) to 7 (a lot of effort), how much effort did you put into trying to successfully match your accounts of the past with what Tina said she did?

Question 2
On a scale from 1 (not difficult at all) to 7 (very difficult), how difficult was it to match your accounts of the past with what Tina said she did?

Question 3
Did you use any particular strategies to try to successfully match your accounts of the past with what Tina said she did?

Question 4
On a scale from 1 (not at all) to 7 (very much), how much did you enjoy the task of trying to successfully match your accounts of the past with what Tina said she did?

Question 5
On a scale from 1 (not curious at all) to 7 (very curious), how curious were you to find out whether you correctly matched your accounts of the past with what Tina said she did?

Question 6
When you received feedback from the detective, did you use that feedback to become more accurate in matching your accounts of the past with what Tina said she did?

No Yes

Question 6a (conditional; appears if Question 6 was answered with yes)
Describe how you used the feedback to become more accurate in matching your accounts of the past with what Tina said she did.

Question 7
Which type of feedback from the detective did you find more useful when trying to match your accounts of the past with what Tina said she did?

ge-words pictures with a green frame equally useful

Debriefing Questionnaire 2

This questionnaire was administered after the delayed posttest in Session 3.

Question 1
In your opinion, what was the goal of making you try to match your accounts of the past with what Tina said she did?

Question 2
Did you notice anything about the sentences you were given when trying to match your accounts of the past with what Tina said she did?

Question 3
Did you notice that some sentences had a form of *haben* and others had a form of *sein*?

No Yes

Question 3a (conditional; appears if Question 3 was answered with yes)
Did this impact your picture choice at all?

No Yes

Question 3b (conditional; appears if Question 3a was answered with yes)
How did this impact your picture choice?

Question 4
Did you notice anything about the *ge*-words you were supposed to type when trying to match your accounts of the past with what Tina said she did?

Question 5
On a scale from 1 (not important at all) to 7 (very important), how important is it to you personally to be grammatically correct in German?

Question 6
On a scale from 1 (not important at all) to 7 (very important), how important do you personally think it is to be grammatically correct to be understood in German?

<u>Experiment 2</u>

Debriefing Questionnaire 1

This questionnaire was administered after the immediate posttest in Session 2.

<u>Question 1</u>
On a scale from 1 (not difficult at all) to 7 (very difficult), how difficult was the first task you completed today?

<u>Question 2</u>
Did you know all the words that were used in the task?

No Yes

<u>Question 2a</u> (conditional; appears if Question 2 was answered with no)
Can you remember the words that you didn't know?

Debriefing Questionnaire 2

This questionnaire was administered after the delayed posttest in Session 3.

<u>Question 1</u>
Did you notice anything about the sentence fragments you were presented with during the task?

No Yes

<u>Question 1a</u> (conditional; appears if Question 1 was answered with yes)
What specifically did you notice about the sentence fragments?

<u>Question 2</u>
Was there anything about the sentence fragments you heard that influenced your picture choice during the task?

<u>Question 3</u>
Did you notice that some sentences had a form of *haben* and others had a form of *sein*?

No Yes

<u>Question 3a</u> (conditional; appears if Question 3 was answered with yes)
Did this impact your picture choice at all?

No Yes

Question 3b (conditional; appears if Question 3a was answered with yes)
How did this impact your picture choice?

Question 4
Do you feel like you waited until the end of the sentence fragment to choose a picture?

No Yes

Question 4a (conditional; appears if Question 4 was answered with no)
What helped you make your choice before the end of the sentence fragment?

Question 5
When you received feedback during the task, did you use that feedback to become more accurate in choosing pictures later in the task?

No Yes

Question 5a (conditional; appears if Question 5 was answered with yes)
Describe how you used feedback to become more accurate later in the task.

Appendix F

Untrained Generalization Items

In order to investigate whether learning of the auxiliary alternation also extended to untrained items — testing for generalizability of the learned knowledge — the same set of analyses as for trained items were performed on six untrained generalization items (3 *haben*, 3 *sein*) and are presented below.

F.1 Learning and the Role of Feedback Ordering

F.1.2 Analyses

To analyze whether learning also occurred for untrained items and differed by group, data from the picture matching and picture description task were analyzed using mixed-effects logistic regression analyses, mirroring the model building procedures described in Section 3.3.2.1, with an interaction term between testing time, coded using a backward-difference coding matrix, and group, sum-coded -0.5 and 0.5, as well as proficiency as a covariate.

F.1.3 Results

F.1.3.1 Picture Matching Task

Descriptive results for the picture matching task can be found in Table F.1. The results are also visually represented in Figures F.1 and F.2.

Table F.1

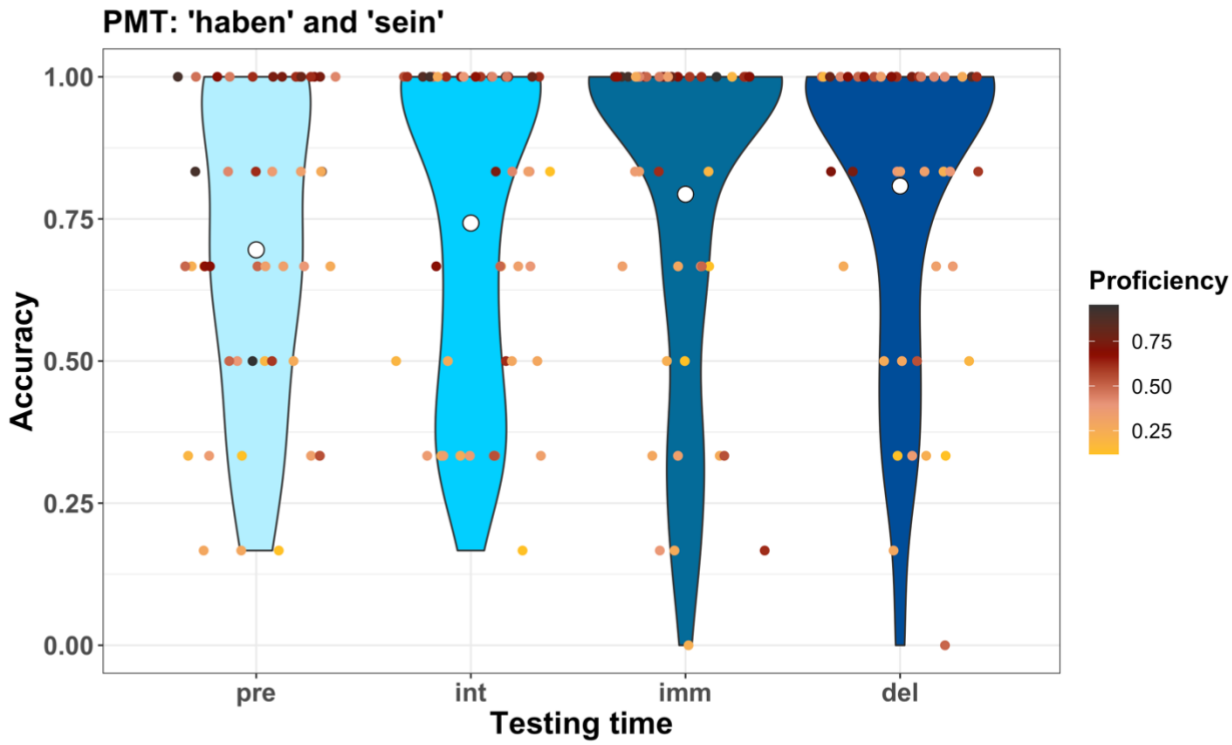
Means for auxiliary accuracy for untrained generalization items on the Picture Matching Task with bootstrapped 95% CIs

	Form-focused feedback first		Semantic-focused feedback first		Across groups	
	M (*SD*)	95% CI	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Pretest	0.67 (0.47)	[0.55,0.78]	0.72 (0.43)	[0.60,0.80]	0.70 (0.46)	[0.62,0.77]
Int. test	0.72 (0.45)	[0.57,0.84]	0.76 (0.43)	[0.65,0.85]	0.74 (0.44)	[0.66,0.82]
Imm. posttest	0.80 (0.40)	[0.63,0.89]	0.79 (0.41)	[0.65,0.88]	0.79 (0.41)	[0.70,0.87]
Del. posttest	0.81 (0.39)	[0.68,0.89]	0.81 (0.40)	[0.65,0.90]	0.81 (0.39)	[0.71,0.87]

The final generalized linear mixed effects model included random intercepts only for subjects. The model (see Table F.2) revealed a significant main effect of testing time, such that accuracy across groups significantly increased from pretest to interim test. While numerically the accuracy also increased from interim test to immediate posttest and from immediate to delayed posttest, these changes remained non-significant. These effects can also be traced in Figure F.1, where the white dots represent mean accuracy at each testing time across groups.

Figure F.1

Accuracy for untrained generalization items on the Picture Matching Task by testing time

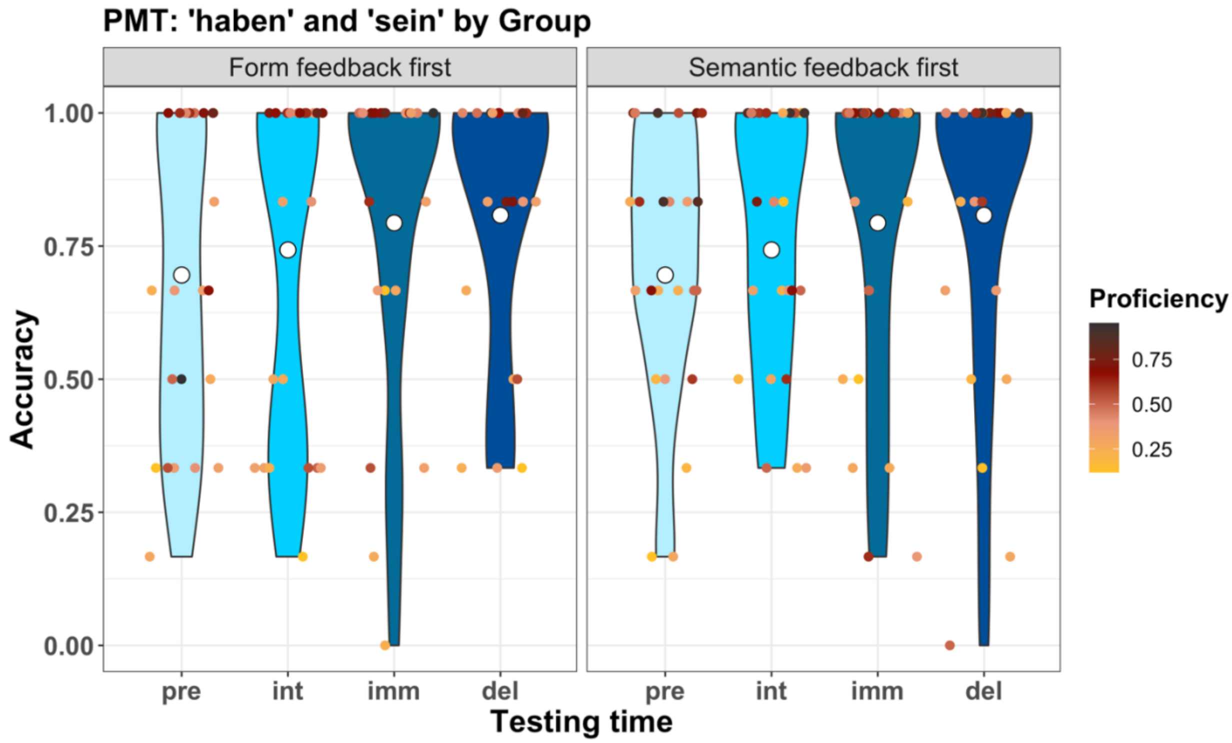


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

There were no main effects of group or any significant interactions between group and testing time. This can also be seen in Figure F.2, where the facets for each group look remarkably similar except for slightly different degrees of within-group variation at each testing time. Finally, the analysis also revealed a main effect of proficiency, showing that more proficient participants were more accurate than less proficient participants across all testing times.

Figure F.2

Accuracy for untrained generalization items on the Picture Matching Task by group and testing time



Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Table F.2

Model output for auxiliary accuracy for untrained generalization items on the Picture Matching Task as a function of group

						95% Confidence Interval	
Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		Lower Bound	Upper Bound
(Intercept)	0.166	0.456	0.365	0.715		-0.727	1.059
Pre-Int	1.706	0.401	4.255	2.09E-05	***	0.920	2.492
Int–Imm	-0.482	0.394	-1.223	0.221		-1.253	0.290
Imm–Del	0.633	0.405	1.564	0.118		-0.160	1.425
SemFirst	-0.326	0.814	-0.400	0.689		-1.922	1.270
Proficiency	2.663	0.500	5.329	9.87E-08	***	1.684	3.642
Pre–Int:Sem	0.682	0.791	0.863	0.388		-0.867	2.232
Int–Imm:Sem	-0.683	0.787	-0.868	0.385		-2.225	0.859
Imm–Del:Sem	-0.622	0.807	-0.770	0.441		-2.203	0.960

Random Effects				
Groups	Names	Variance	Std. Deviation	Correlation
Subject	Intercept	5.6779	2.3828	
Item	Intercept	0.1192	0.3453	

Model fit		
	Marginal	Conditional
R^2	0.459	0.804

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

Model equation: Accuracy ~ TestingTime*Group + Proficiency + (1 |Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

F.1.3.2 Picture Description Task (Auxiliaries)

Descriptive results for auxiliary accuracy on the picture description task can be found in Table F.3. The results are also visually represented in Figures F.3 and F.4.

Table F.3

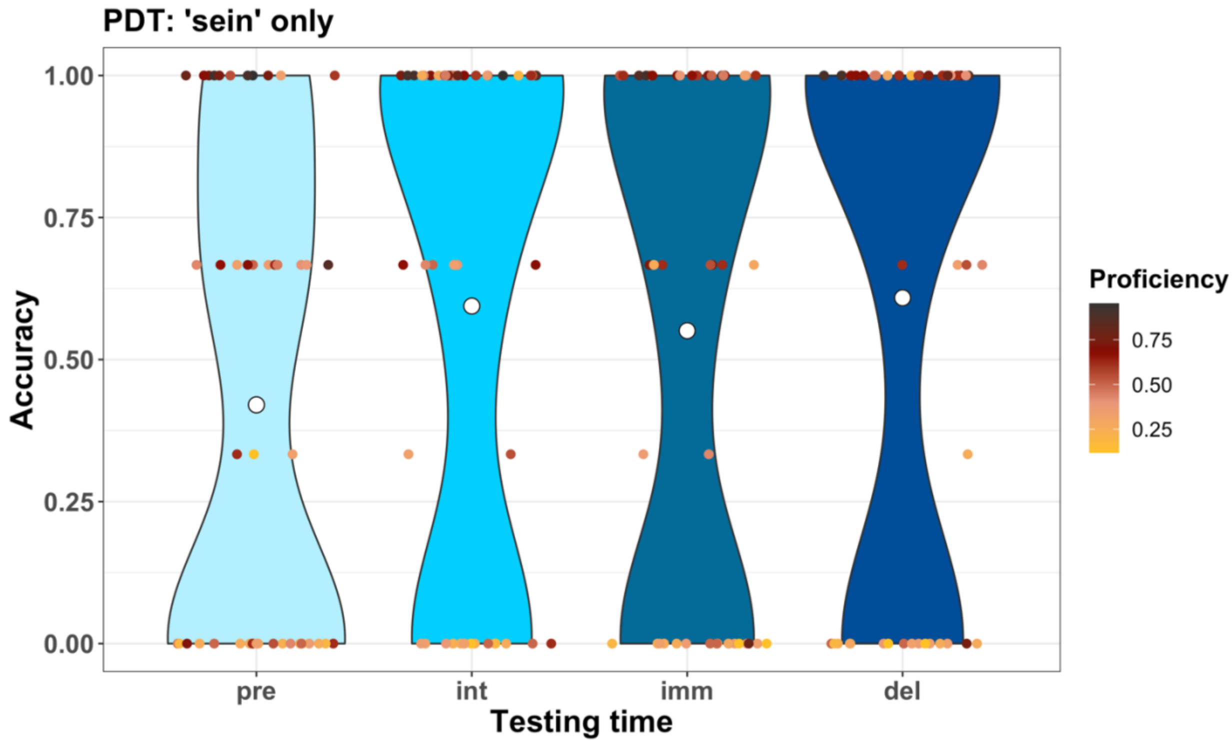
Means for auxiliary accuracy for untrained generalization items on the Picture Description Task with bootstrapped 95% CIs

	Form-focused feedback first		Semantic-focused feedback first		Across groups	
	M (*SD*)	95% CI	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Pretest	0.44 (0.50)	[0.24,0.59]	0.40 (0.49)	[0.24,0.57]	0.42 (0.50)	[0.30,0.54]
Int. test	0.61 (0.49)	[0.39,0.74]	0.58 (0.50)	[0.36,0.75]	0.59 (0.49)	[0.44,0.71]
Imm. posttest	0.59 (0.50)	[0.38,0.76]	0.51 (0.50)	[0.32,0.68]	0.55 (0.50)	[0.40,0.68]
Del. posttest	0.68 (0.47)	[0.47,0.83]	0.54 (0.50)	[0.33,0.71]	0.61 (0.49)	[0.47,0.73]

The final generalized linear mixed effects model included random intercepts for subjects and items and a random slope for testing time by subject. The model (see Table F.4) revealed a significant main effect of testing time, such that accuracy across groups significantly increased from pretest to interim test. While numerically the accuracy decreased from interim test to immediate posttest and increased again from immediate to delayed posttest, these changes remained non-significant. These effects can also be traced in Figure F.3, where the white dots represent mean accuracy at each testing time across groups.

Figure F.3

Auxiliary accuracy for untrained sein *generalization items on the Picture Description Task by testing time*

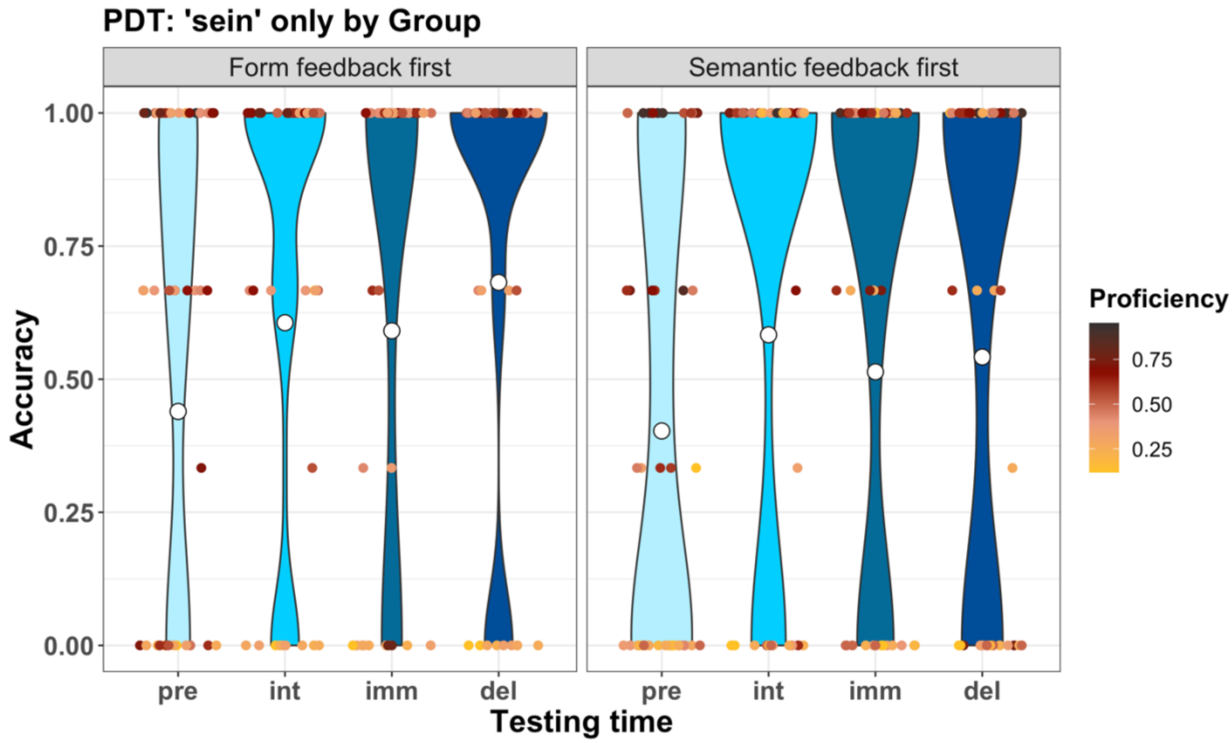


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

There were no main effects of group or any significant interactions between group and testing time. While there seem to be group differences in Figure F.4 going from interim test to immediate posttest as well as from immediate to delayed posttest, these visual differences are mostly driven by some more participants in the semantic-focused feedback first group scoring between 0.33 and 0.66 than in the form-focused feedback first group, while the majority of participants in both groups is either at ceiling or at floor. Finally, the analysis also revealed a main effect of proficiency, showing that more proficient participants were more accurate than less proficient participants across all testing times.

Figure F.4

Auxiliary accuracy for untrained sein *generalization items on the picture description task by group and testing time*



Note. White dots represent group means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

Table F.4

Model output for auxiliary accuracy for untrained generalization items on the Picture Description Task as a function of group

Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		95% Confidence Interval Lower Bound	Upper Bound
(Intercept)	0.647	0.883	0.732	0.464		-1.084	2.377
Pre-Int	1.700	0.805	2.107	0.035	*	0.118	3.273
Int–Imm	-0.464	0.848	-0.547	0.584		-2.126	3.200
Imm–Del	1.727	1.549	1.115	0.265		-1.308	4.763
Proficiency	5.128	1.405	3.651	2.61E-04	***	2.375	7.881
Pre–Int:Sem	0.939	1.405	0.668	0.504		-1.815	2.693
Int–Imm:Sem	1.566	2.097	0.747	0.455		-2.545	5.677
Imm–Del:Sem	4.033	4.166	0.968	0.333		-4.133	12.199

Random Effects

Groups	Names	Variance	Std. Deviation	Correlation		
Subject	Intercept	20.1061	4.4840			
	Pre-Int	12.2983	3.5069	0.66		
	Int-Imm	11.6637	3.4152	0.62	0.24	
	Imm-Del	20.5576	4.5341	0.70	0.87	0.04
Item	Intercept	0.2868	0.5355			

Model fit

	Marginal	Conditional
R^2	0.488	0.952

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

Model equation: Accuracy ~ Proficiency*TestingTime + (1+TestingTime|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

F.1.4 Summary: Learning and the Role of Feedback Ordering

The results from the picture matching task and the picture description task suggest that learning occurred across both groups, with accuracy at delayed posttest being about 11% higher than at pretest in the picture matching task (see Table F.1) and about 19% higher in the picture description task (see Table F.3). In both tasks, significant main effects of testing time going from pretest to interim test show that the largest amount of learning was achieved after the first round of training. Unlike the results for trained items on the picture matching task, there was no descriptive or statistically significant decrease from interim test to immediate posttest on the picture matching task for untrained items. However, such a decrease now descriptively appears in the picture description task, unlike for trained items where accuracy consistently increased. More proficient learners were more accurate at all testing times in both tasks, which is unsurprising, seeing that they have more experience with the target structure itself. There were no other significant effects, indicating that the type of feedback provided as well as the feedback ordering may not play a role in generalizing knowledge from trained items to untrained items.

F.2 Learning and the Role of Proficiency

F.2.1 Analyses

To address the potential role of proficiency in learning the untrained generalization items, data from the picture matching and picture description task were also analyzed using mixed-effects logistic regression analyses, mirroring the model building procedures described in Section 3.3.3.1 with an interaction term between testing time, coded using a backward-difference coding matrix, and proficiency, which was transformed to *z*-scores and centered around 0.

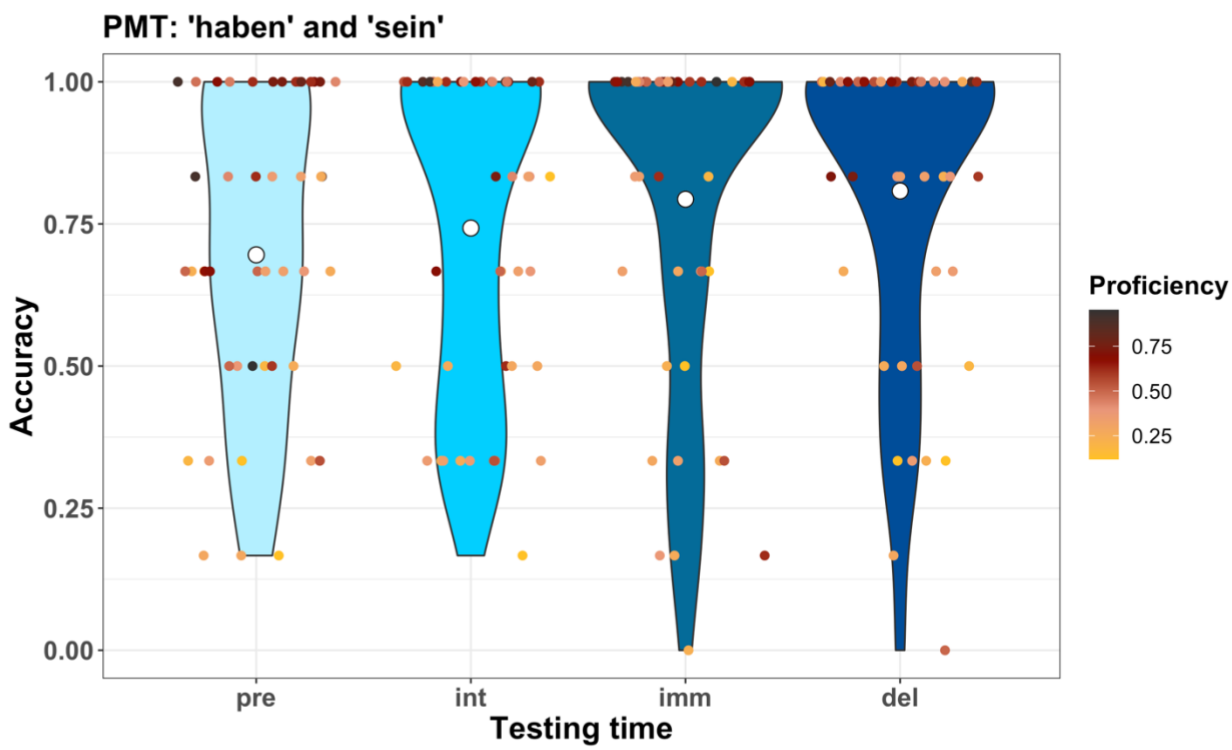
F.2.2 Results

F.2.2.1 Picture Matching Task

As proficiency was coded as a continuous variable, descriptive results are only graphically represented in Figure F.5, which corresponds to Figure F.1. Figure F.5 shows increases in accuracy from each testing time to the next, with participants clustering at ceiling from the interim test onwards, as indicated by high density of data points at ceiling.

Figure F.5

Accuracy for untrained generalization items on the Picture Matching Task by testing time

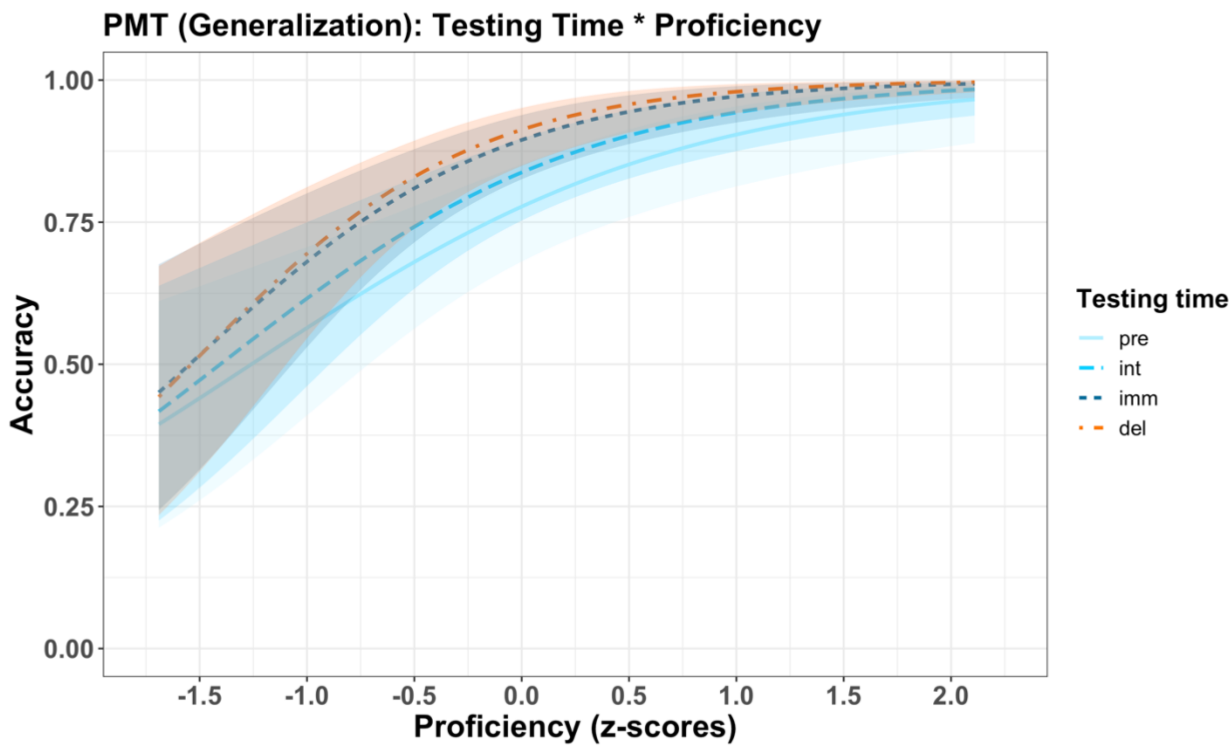


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

The final generalized linear mixed effects model included random intercepts only for subjects. The model (see Table F.5) revealed a marginally significant main effect of testing time, such that accuracy significantly increased from interim test to immediate posttest. However, the confidence intervals for this effect cross zero, indicating that the effect might be spurious. While numerically the accuracy increased from pretest to interim test as well as from immediate to delayed posttest, these changes remained non-significant. This result is similar to the first set of analyses for the picture matching task, except that the (marginally) significant effect appears from interim test and immediate posttest instead of from pretest and interim test. The analysis also revealed a main effect of proficiency, showing that more proficient participants were more accurate than less proficient participants across all testing times. However, there were no significant interactions between proficiency and testing time. Figure F.6 visualizes the absence of significant interactions — despite the fact that the distance in accuracy between testing times seems larger around the center of the proficiency scale, the error bands consistently overlap.

Figure F.6

Visualization of interaction terms between proficiency and testing time for untrained generalization items on the Picture Matching Task



Note. Error bands represent 95% confidence intervals.

Table F.5

Model output for auxiliary accuracy for untrained generalization items on the Picture Matching Task as a function of proficiency

						95% Confidence Interval	
Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		Lower Bound	Upper Bound
(Intercept)	1.845	0.223	8.267	2E-16	***	1.407	2.282
Pre-Int	0.390	0.246	1.581	0.114		-0.093	0.872
Int–Imm	0.499	0.285	1.750	0.080	†	-0.060	1.059
Imm–Del	0.208	0.322	0.645	0.519		-0.423	0.840
Proficiency	1.268	0.224	5.657	1.54E-08	***	0.829	1.707
Pre–Int:Prof	0.175	0.262	0.668	0.504		-0.338	0.688
Int–Imm:Prof	0.215	0.305	0.710	0.480		-0.382	0.813
Imm–Del:Prof	0.142	0.343	0.414	0.679		-0.530	0.813

Random Effects

Groups	Names	Variance	Std. Deviation	Correlation
Subject	Intercept	1.391	1.179	

Model fit

	Marginal	Conditional
R^2	0.281	0.495

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

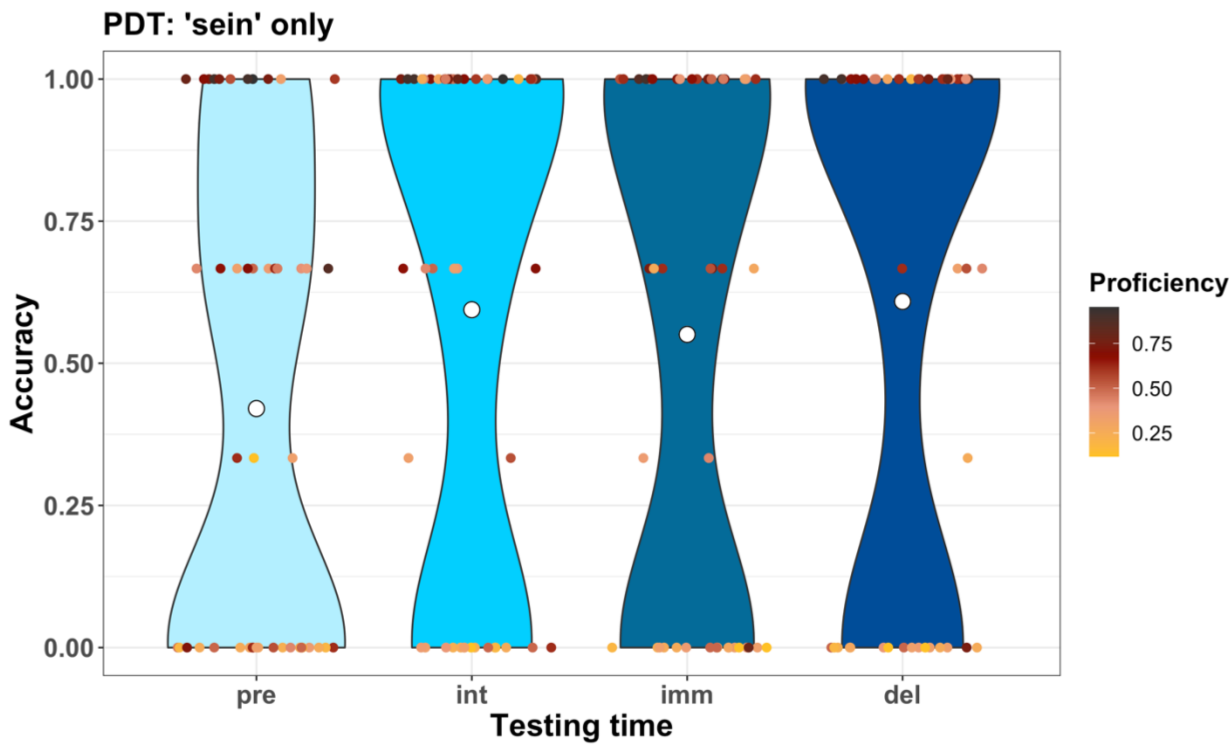
Model equation: Accuracy ~ Proficiency*TestingTime + (1 |Subject), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa)

F.2.2.2 Picture Description Task (Auxiliaries)

Again, descriptive results are graphically represented in Figure F.7, which corresponds to Figure F.3. Figure F.7 shows an increase from pretest to interim test as well as from immediate to delayed posttest. However, accuracy also decreases from interim test to immediate posttest. While the density of data points at pretest is highest at floor, the largest density consistently moves upward towards ceiling, such that the distribution of data points becomes almost bimodal.

Figure F.7

Auxiliary accuracy for untrained sein *generalization items on the Picture Description Task by testing time*

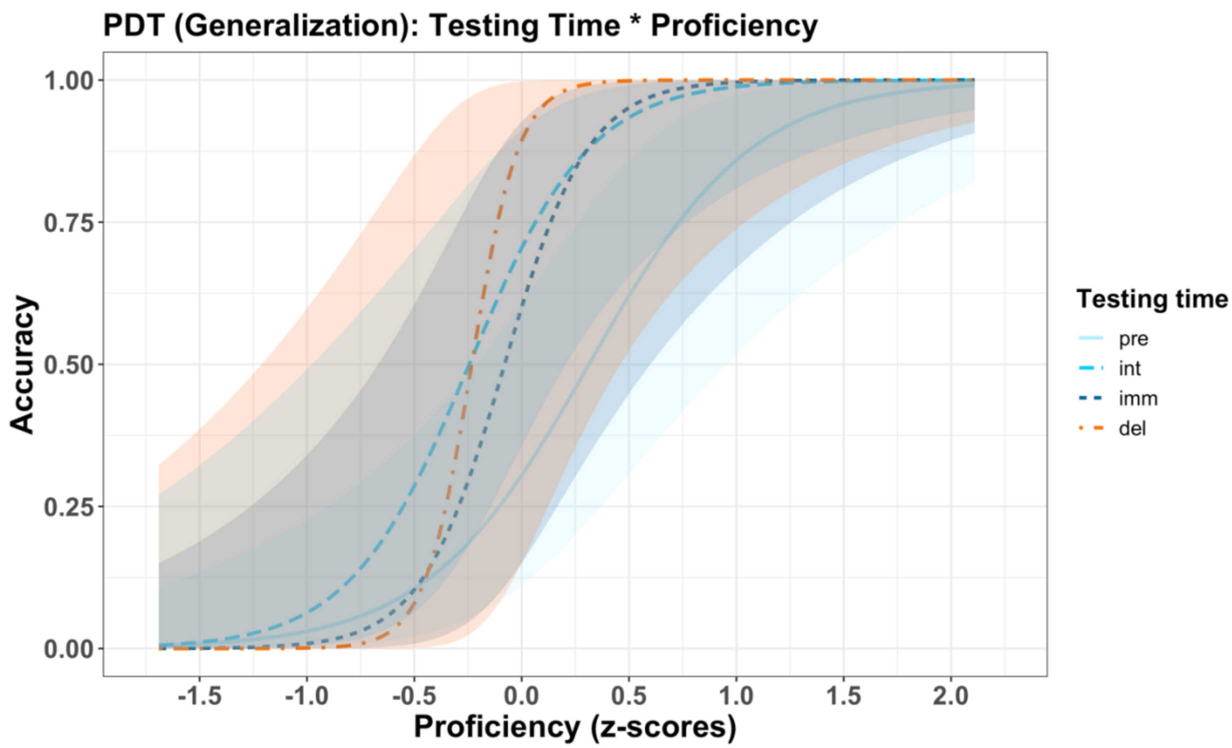


Note. White dots represent overall means, whereas colored dots represent individual participants, with the color of the dot corresponding to participants' proficiency.

The final generalized linear mixed effects model included random intercepts for subjects and items and a random slope for testing time by subject. The model (see Table F.6) revealed a significant main effect of testing time, such that accuracy significantly increased from pretest to to interim test. Other effects of testing rime remained non-significant, mirroring the results for the picture description task from the first set of analyses. The analysis also revealed a main effect of proficiency, showing that more proficient participants were more accurate than less proficient participants across all testing times. However, there were no significant interactions between proficiency and testing time. Figure F.8 again visualizes the absence of significant interactions — despite different patterns emerging for participants below and above proficiency *z*-scores of -0.5, the error bands are extremely wide consistently overlap, indicating that there is a large degree of variation in accuracy even among participants of similar proficiency.

Figure F.8

Visualization of interaction terms between proficiency and testing time for untrained generalization items on the Picture Description Task



Note. Error bands represent 95% confidence intervals.

Table F.6

Model output for auxiliary accuracy for untrained generalization items on the Picture Description Task as a function of proficiency

Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)		95% Confidence Interval Lower Bound	Upper Bound
(Intercept)	0.647	0.883	0.732	0.464		-1.084	2.377
Pre-Int	1.700	0.805	2.107	0.035	*	0.118	3.273
Int–Imm	-0.464	0.848	-0.547	0.584		-2.126	3.200
Imm–Del	1.727	1.549	1.115	0.265		-1.308	4.763
Proficiency	5.128	1.405	3.651	2.61E-04	***	2.375	7.881
Pre–Int:Prof	0.939	1.405	0.668	0.504		-1.815	2.693
Int–Imm:Prof	1.566	2.097	0.747	0.455		-2.545	5.677
Imm–Del:Prof	4.033	4.166	0.968	0.333		-4.133	12.199

Random Effects

Groups	Names	Variance	Std. Deviation	Correlation		
Subject	Intercept	20.1061	4.4840			
	Pre-Int	12.2983	3.5069	0.66		
	Int-Imm	11.6637	3.4152	0.62	0.24	
	Imm-Del	20.5576	4.5341	0.70	0.87	0.04
Item	Intercept	0.2868	0.5355			

Model fit

	Marginal	Conditional
R^2	0.488	0.952

p-values and confidence intervals for fixed effects calculated using the Wald method.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘†’ 0.1 ‘ ’ 1

Model equation: Accuracy ~ Proficiency*Group + (1+TestingTime|Subject) + (1|Item), data = MyData, family = "binomial", glmerControl(optimizer = "bobyqa")

F.2.3 Summary: Learning and the Role of Proficiency

The results of the picture matching task indicate that proficiency plays a role for the ability to generalize learned knowledge to untrained test items. In particular, more proficient learners are again overall more accurate than less proficient learners. This effect mirrors results described in the analysis of group effects in Section F.1.3.1 and is unsurprising, as more proficient learners not only have prior knowledge of but also experience with the target structure and would therefore be able to more easily generalize or reactive prior knowledge of the targeted untrained items. While there is also a marginally significant main effect of testing time going from interim test to immediate posttest, error bands in Figure F.6 as well as zero-crossing confidence intervals suggest that this effect is spurious.

The results for auxiliary production in the picture description task mostly mirror those of the picture matching task. While more proficient participants were, again, more accurate overall, the only other significant effect was a main effect of testing time, with a significant increase in accuracy from pretest to interim test. There were no interactions between testing time and proficiency.

Appendix G

Descriptive Statistics for Pupil Size in Experiment 2

The three tables below include the descriptive statistics for pupil size in response to feedback presentation in Experiment 2. These measures include mean baseline corrected pupil size in pixels (Table G.1), mean baseline correct peak pupil amplitude in pixels (Table G.2), and mean peak latency in ms (Table G.3), all by training block, auxiliary, and error, type. If more data is collected in the future, additional analyses could be run on these measures, for instance comparing overall peak amplitude and latency across conditions.

Table G.1

Mean (baseline corrected) pupil size (pixels) by training block, auxiliary, and error type

	haben		*sein*	
	M	*SD*	*M*	*SD*
Block 1				
All correct	-13.40	111.16	-19.57	90.87
Choice correct, participle incorrect	-24.92	88.06	-20.35	97.36
All incorrect	-11.79	81.32	-24.36	81.08
Block 2				
All correct	0.00	98.43	-44.94	101.36
Choice correct, participle incorrect	-11.40	84.13	-15.56	97.98
All incorrect	-19.53	75.24	-12.35	71.89
Block 3				
All correct	-41.71	85.24	-50.61	88.51
Choice correct, participle incorrect	-3.27	83.13	-28.50	113.24
All incorrect	-25.24	86.63	-17.05	61.71
Block 4				
All correct	-34.41	79.25	-53.59	84.12
Choice correct, participle incorrect	23.79	109.88	-6.49	101.09
All incorrect	-8.42	107.91	1.38	72.28

Table G.2
Mean peak (baseline corrected) pupil amplitude (pixels) by training block, auxiliary, and error type

	haben		*sein*	
	M	*SD*	*M*	*SD*
Block 1				
All correct	45.18	70.94	40.77	47.07
Choice correct, participle incorrect	72.16	69.24	65.63	97.11
All incorrect	63.55	70.21	57.64	51.11
Block 2				
All correct	50.71	75.83	46.45	60.31
Choice correct, participle incorrect	74.76	74.60	77.91	75.31
All incorrect	66.84	57.69	69.09	62.98
Block 3				
All correct	51.30	76.46	49.78	72.09
Choice correct, participle incorrect	94.68	109.99	78.95	134.47
All incorrect	76.73	62.39	75.05	71.02
Block 4				
All correct	83.78	73.25	85.14	74.89
Choice correct, participle incorrect	128.84	157.64	79.06	98.82
All incorrect	82.53	114.01	76.33	84.43

Table G.3
Mean peak latency (in ms) by training block, auxiliary, and error type

	haben		*sein*	
	M	*SD*	*M*	*SD*
Block 1				
All correct	1202.15	846.74	1308.00	915.11
Choice correct, participle incorrect	1061.59	955.42	1195.35	796.63
All incorrect	1266.47	792.89	928.28	726.71
Block 2				
All correct	848.89	900.93	1011.28	971.68
Choice correct, participle incorrect	1175.76	985.00	1003.75	702.63
All incorrect	1239.38	863.62	1196.00	960.73
Block 3				
All correct	964.21	932.43	907.23	882.58
Choice correct, participle incorrect	1320.69	1018.21	1078.10	966.50
All incorrect	891.11	729.19	1118.67	937.45
Block 4				
All correct	882.84	966.34	665.71	832.05
Choice correct, participle incorrect	763.64	625.64	1029.10	1150.53
All incorrect	1048.57	921.61	1187.33	1084.06

Appendix H

LLAMA F and SRT

Participants completed the SRT task (Linck et al., 2013) followed by the LLAMA F task (Meara & Rogers, 2021) as measures of implicit and explicit learning aptitude, respectively. As a reminder, implicit learning aptitude on the SRT task was calculated as the difference in median reaction times between the last sequential block (block five) and the last pseudo-random block (block six). The longer the times on the random block (block six) and the shorter the times on the last sequential block (block 5), the greater the implicit learning effect. For the LLAMA F, explicit learning aptitude scores were calculated as the number of correct items during testing, with 20 being the highest achievable score. The results of both tasks are presented in Table H.1.

Table H.1

Results of the SRT and LLAMA F

	Range (in ms)	*M* (in ms)	*SD*
SRT	9.5 – 64	29.88	20.08
	Range (0 – 20)	*M*	*SD*
LLAMA F	2 – 16	9.77	3.88

VITA - Valerie Keppenne

EDUCATION

2023	**Dual-title Ph.D. in German Applied Linguistics & Language Science,** The Pennsylvania State University, PA
2017	**M.A. in English Linguistics & English Literature,** Universität Trier, Germany
2015	**B.A. in English Language, Literature, and Linguistics & Media Sciences,** Universität Trier, Germany

SELECTED PUBLICATIONS

2022	Piotti, A., Matangos, H. A., Chin, A., Goetze, J., **Keppenne, V**., and Kerschen, K. (Co)Constructing Antiracist Pedagogy: Expanding on Standing Departmental Approaches to Foreign Language Teaching. *Second Language Research and Practice*, *3*(1), 151-163.
2021	**Keppenne, V**., Hopman, E. W., & Jackson, C. N. Production-based training benefits the comprehension and production of grammatical gender in L2 German. *Applied Psycholinguistics, 42*(4), 907-936. doi:10.1017/S014271642100014X

SELECTED CONFERENCE PRESENTATIONS

2022	**Keppenne, V.,** Grüter, T., & Jackson, C. N. *Predicting, but not guessing, coupled with implicit corrective feedback benefits L2 learners*. BilForum 2022, Chicago, IL.
	Keppenne, V., Grüter, T., & Jackson, C. N. *Guessing and/or predicting benefits learners of varying proficiency*. American Association of Applied Linguistics (AAAL) 2022, Pittsburgh, PA.
2020	**Keppenne, V.** & Jackson, C. N. *Training Grammatical Gender in L2 German: Lab-based versus Classroom-based Studies.* Second Language Research Forum (SLRF) 2020, Nashville, TN (virtual).
2019	**Keppenne, V.**, Hopman, E. W. M., & Jackson, C. N. *Production training benefits comprehension of grammatical gender in L2 German.* Paper presented at 12th International Symposium of Bilingualism (ISB 12), Edmonton, Canada.
	Keppenne, V., Kerschen, K. R., Rocker, M. H., & Putnam, M. T. *Investigating the Superiority Effect: A Comparison of English and German.* Paper presented at 25th Germanic Linguistics Annual Conference (GLAC-25), Iowa City, IA.
	Keppenne, V., Hopman, E. W. M., & Jackson, C. N. *Production practice benefits comprehension and production of grammatical gender in L2 German.* Poster session presented at 32nd Annual CUNY Conference on Human Sentence Processing (CUNY 2019), Boulder, CO.

SELECTED GRANTS AND AWARDS

2021 – 2022	Max Kade Graduate Fellowship
2021 – 2022	Dissertation Improvement Grant, *Language Learning*
2021	Harold F. Martin Graduate Assistant Outstanding Teaching Award, Penn State