

FOREIGN TACT TRAINING AND RETENTION OF EMERGENT FOREIGN VOCABULARY

by John R. Wooderson

Thesis submitted in fulfilment of the requirements for
the degree of

Doctor of Philosophy

under the supervision of

Dr Kirsty Young
Professor Lewis Bizo

University of Technology Sydney
Faculty of Arts and Social Sciences

May 2025

Certificate of original authorship

I, John Wooderson, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the Faculty of Arts and Social Sciences at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 16/4/2025

Publications during candidature

Peer-reviewed papers¹

Wooderson, J. R., Bizo, L. A., & Young, K. (2024). Extracting published graphical data: A guide for researchers wanting to synthesize single-case data. *Experimental Analysis of Human Behavior Bulletin*, 34, <https://doi.org/10.17605/osf.io/rfhjx>

Wooderson, J. R., Bizo, L. A., & Young, K. (2022). A systematic review of emergent learning outcomes produced by foreign language tact training. *The Analysis of Verbal Behavior*, 38, 157–178. <https://doi.org/10.1007/s40616-022-00170-z>

Wooderson, J. R., Bizo, L. A., & Young, K. (2023). Retention of emergent Korean vocabulary following foreign tact training and overlearning. [Manuscript submitted for publication]

Wooderson, J. R., Bizo, L. A., & Young, K. (2024). More is not always better: The testing effect on retention of emergent Korean. [Manuscript submitted for publication]

Conference abstracts

Wooderson, J. R., Bizo, L. A., & Young, K. (2024, May). *Retention of emergent Korean vocabulary following foreign tact training and overlearning* [Paper presentation]. Association for Behavior Analysis International (ABAI) Conference 2024, Philadelphia, United States of America. <https://www.abainternational.org/events/program-details/summary.aspx?intConvId=116&a=r>

¹ These four papers were published or submitted to American journals that use American English spelling, which has been maintained in their associated chapters in the thesis. All other chapters were written using Australian English.

Publications included in this thesis

Wooderson, J. R., Bizo, L. A., & Young, K. (2024). Extracting published graphical data: A guide for researchers wanting to synthesize single-case data. *Experimental Analysis of Human Behavior Bulletin*, 34, <https://doi.org/10.17605/osf.io/rfhjx>

– incorporated as Chapter 3.

Contributor	Statement of contribution
Author John Wooderson (Candidate)	Designed experiments (90%) Wrote and edited paper (90%)
Author Dr Lewis Bizo	Designed experiments (5%) Wrote and edited paper (5%)
Author Dr Kirsty Young	Designed experiments (5%) Wrote and edited paper (5%)

Wooderson, J. R., Bizo, L. A., & Young, K. (2022). A systematic review of emergent learning outcomes produced by foreign language tact training. *The Analysis of Verbal Behavior*, 38, 157–178. <https://doi.org/10.1007/s40616-022-00170-z>

– incorporated as Chapter 4.

Contributor	Statement of contribution
Author John Wooderson (Candidate)	Designed experiments (90%) Wrote and edited paper (90%)
Author Dr Lewis Bizo	Designed experiments (5%) Wrote and edited paper (5%)
Author Dr Kirsty Young	Designed experiments (5%) Wrote and edited paper (5%)

Wooderson, J. R., Bizo, L. A., & Young, K. (2023). Retention of emergent Korean vocabulary following foreign tact training and overlearning. [Manuscript submitted for publication] – incorporated as Chapter 6.

Contributor	Statement of contribution
Author John Wooderson (Candidate)	Designed experiments (90%) Wrote and edited paper (90%)
Author Dr Lewis Bizo	Designed experiments (5%) Wrote and edited paper (5%)
Author Dr Kirsty Young	Designed experiments (5%) Wrote and edited paper (5%)

Wooderson, J. R., Bizo, L. A., & Young, K. (2024). More is not always better: The testing effect on retention of emergent Korean. [Manuscript submitted for publication] – incorporated as Chapter 8.

Contributor	Statement of contribution
Author John Wooderson (Candidate)	Designed experiments (90%) Wrote and edited paper (90%)
Author Dr Lewis Bizo	Designed experiments (5%) Wrote and edited paper (5%)
Author Dr Kirsty Young	Designed experiments (5%) Wrote and edited paper (5%)

Contributions by others to the thesis

No contributions by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

Acknowledgements

I am profoundly grateful to all those who provided their invaluable support and encouragement throughout my PhD journey. First and foremost, I would like to thank my advisors, Dr. Kirsty Young and Dr. Lewis Bizo, for their guidance, encouragement, and expertise.

I am especially grateful for the opportunities you gave me to grow as a researcher and explore areas beyond the familiar field of behaviour analytic research. Thank you for your mentorship.

I am also thankful to Dr. Mary Foster, Dr. Angelika Anderson, and Dr. Tina McAdie for their insightful feedback and constructive criticism during the stage assessment milestones of my candidature. Thank you, Dr. James Foster who also provided valuable advice. Your input has significantly enhanced the quality of my work.

To Oliver Roschke and Dr. Megan Borlase, thank you for your meticulous assistance with reliability checks and interobserver agreement. Your dedication ensured the accuracy of my research findings.

I am especially grateful to my wife, Belle, and children, Sora and Tae Il. Thank you for your love, encouragement, and patience as I spent many evenings locked up with my laptop.

Without your support, this valuable accomplishment would not have been possible.

Thank you.

Certificate of original authorship	I
Publications during candidature	II
Publications included in this thesis	III
Acknowledgements	VI
List of Abbreviations used in the thesis	XIV
Abstract	XV
Chapter 1: Introduction	1
1.1 Introduction and background to the research problem	1
1.1.1 Foreign language learning	1
1.1.2 Vocabulary	1
1.1.3 Vocabulary size	2
1.1.4 Instructional strategies	4
1.1.5 Behaviour analysis	6
1.1.6 Emergent learning and verbal operants	7
1.1.7 Foreign tact training	8
1.1.8 Retention	11
1.2 Statement of the problem	12
1.3 Research questions	13
Chapter 2: Theoretical Framework	16
2.1 Introduction	16
2.2 Historical perspectives of emergent learning	16
2.2.1 Cognitive psychology	16

2.2.2 Gestalt psychology	17
2.2.3 Behaviour analysis	18
2.2.4 Stimulus generalisation	18
2.2.5 Derived relational responding	19
2.3 Theories related to emergent learning	20
2.3.1 Stimulus equivalence	20
2.3.2 Naming theory	23
2.3.3 Relational frame theory	24
2.4 Applications of stimulus equivalence to language learning	25
2.5 Introduction to Chapters 3 and 4	27
Chapter 3: Tutorial: Extracting Published Graphical Data Using Digitizeit: A Step-By-Step Guide for Behavior Analysts Wanting to Reanalyze Single-Case Data	29
3.1 Abstract	29
3.2 Introduction	29
3.3 Method	33
3.31 Importing the graph into the program	35
3.32 Defining the XY axes	35
3.33 Plotting data points and creating datasets	37
3.34 Automatic data plotting	37
Manual data plotting	40
3.35 Creating datasets	41
3.4 Exporting data	41

3.5 Conclusion	43
Chapter 4: A Systematic Review of Emergent Learning Outcomes Produced by Foreign-	
Language Tact Training	44
4.1 Abstract	44
4.2 Introduction	44
4.3 Method	49
4.31 Literature search procedure	49
4.32 Data categorization	52
4.33 Data extraction for meta-analysis	53
4.34 Interobserver agreement	55
4.4 Results	55
4.41 Participant demographics	55
4.42 Target foreign languages, training conditions, mastery criteria, and emergent learning relations	59
4.43 FTT's emergent learning outcomes	59
4.44 Meta-analysis	62
4.5 Discussion	65
4.6 Conclusion and recommendations for future research	70
4.7 Supplementary material	72
Chapter 5: Overlearning, precision teaching and fluency building	82
5.1 Introduction	82
5.2 Summary of systematic review findings	82

Retention of emergent foreign vocabulary

5.2.1 Emergent relations _____	83
5.2.2 Target language _____	83
5.2.3 Mastery criteria _____	84
5.2.4 Emergent relations and retention _____	84
5.3 Overlearning _____	85
5.3.1 Definition _____	85
5.3.2 Overlearning and retention _____	86
5.3.3 Overlearning and emergent learning _____	87
5.4 Precision teaching and fluency-building _____	88
5.4.1 Definition _____	88
5.4.2 Fluency and retention _____	89
5.5 Conclusion _____	90
5.6 Introduction to the first experiment _____	90
Chapter 6: Study 3—Retention of emergent Korean vocabulary following foreign tact training and overlearning _____	92
6.1 Abstract _____	92
6.2 Introduction _____	92
6.3 Method _____	98
6.31 Participants and setting _____	98
6.32 Response Measurement and Dependent Variables _____	101
6.34 Experimental Design and Procedures _____	102
Pre-assessments _____	102

Pre-tests _____	103
Baseline and Instruction _____	103
Accuracy training _____	104
Fluency training _____	104
Post-tests and follow-up tests _____	105
6.35 Interobserver Agreement and Treatment Integrity _____	105
6.4 Results _____	106
6.41 Baseline and Training _____	106
6.42 Pre- and Post-Tests _____	109
6.43 Follow-up Tests _____	110
6.5 Discussion _____	111
Chapter 7: Fluency-building and the testing effect _____	115
7.1 Introduction _____	115
7.2 Summary of Study 3 _____	115
7.2.1 Retention in Study 3 _____	116
7.2.2 Response rate or. extended practice? _____	116
7.3 Free-operant versus restricted-operant training _____	117
7.4.1 Free-operant training _____	117
7.4.2 Restricted-operant training _____	117
7.4.3 Studies with animals _____	118
7.4.4 Studies with human participants _____	119
7.5 Conclusion _____	120

7.6 Introduction to the second and third experiments	121
Chapter 8: Studies 4 and 5 - More is not always better: The testing effect on retention of emergent Korean vocabulary	123
8.1 Abstract	123
8.2 Introduction	124
Experiment 2	126
8.3 Method	126
8.31 Participants and setting	126
8.32 Materials and stimulus sets	127
8.33 Pre-experimental conditions	129
8.4 Experimental design and procedures	129
8.41 Dependent variables	130
8.42 General procedure	130
8.43 Baseline	131
8.44 No-delay training	132
8.45 Delay training	132
8.46 Post-tests	133
8.47 Interobserver agreement and treatment integrity	133
8.5 Results and discussion	134
8.51 Training	134
8.52 Native-to-foreign intraverbal post-tests	136
8.53 Tact post-tests	138

8.54 Foreign-to-native intraverbal and listener post-tests	140
8.55 Post-hoc analysis of target words	141
8.56 Preferences	141
8.6 Summary	141
Experiment 3	141
8.7 Method	142
8.71 Participants and setting	142
8.72 Materials and stimulus sets	142
8.73 Response measurement and dependent variables	143
8.74 Experimental design and procedures	143
8.75 Baseline	143
8.76 Retraining	144
8.77 Interobserver agreement and treatment integrity	145
8.8 Results and discussion	145
8.81 Retraining	146
8.82 Native-to-foreign intraverbal post-tests	146
8.83 Tact post-tests	148
8.9 General discussion	149
8.91 Limitations	152
8.92 Conclusion	153
Chapter 9: General discussion, implications, recommendations, and conclusions	154
9.1 General Discussion	154

9.11 Summary of the studies and their key findings	154
9.2 Implications for emergent learning programming	162
9.21 Recommended FTT protocol	163
9.22 Costs and benefits of fluency-building	165
9.23 The pronunciation guide's effect on spacing	166
9.24 Response rate and retention	167
9.3 General limitations and recommendations for future research	168
9.4 Conclusion	170
References	172
Appendix 1	198
Ethics information and consent forms	198
Study 3	198
Study 4 and 5	204
Appendix 2	210
Research protocol for Study 3	210
Instructions for participants in Study 4	219

List of Abbreviations used in the thesis

EFL	English as a foreign language
FL	foreign language
FNI	foreign-to-native intraverbal (see foreign word/ say English word)
FTT	foreign tact training
NFI	native-to-foreign intraverbal
SCED	single-case experimental design

Abstract

The research on emergent learning principles and their application to foreign language (FL) vocabulary instruction has steadily grown over the last decade. Although this is a relatively new field, existing evidence highlights its potential for efficient FL vocabulary acquisition. This thesis presents a series of experiments that evaluate procedural modifications to foreign tact training (FTT), a key emergent learning procedure, focusing on enhancing learners' retention of emergent vocabulary. The first study established a protocol for extracting single-case graphical data from the research literature. This protocol was then applied in the second study, where a meta-analysis evaluated and compared FTT's outcomes with those of other emergent learning procedures, including foreign-to-native intraverbal (FNI), native-to-foreign intraverbal (NFI), mand, and listener training. In Study 3, five English-speaking adults were taught to tact Korean words using two training conditions: one incorporating a modified FTT protocol with fluency-based overlearning trials, and one without. The retention of untrained emergent NFI relations was evaluated up to six months after FTT. In Study 4, the experimental design was modified to hold total training time constant across two conditions, further investigating the effects of extended practice observed in Study 3. The final experiment, Study 5, built on the earlier studies with additional procedural modifications, including spaced weekly practice tests and corrective feedback. Across all three experiments, FTT produced high rates of untrained emergent relations immediately following training. The meta-analysis findings indicated that FTT resulted in significantly more untrained FL relations than intraverbal (FNI and NFI) or listener training, and it was more efficient than FNI training. In Study 3, adding fluency-based overlearning trials to FTT increased the retention of emergent NFI relations. In Study 4, when total training time was held constant, the modified FTT protocol produced similar levels of retention to the traditional

discrete-trial-based FTT protocol, with neither condition reaching criterion-level performance during delayed retention tests. The procedural modifications in Study 5 led to improved retention levels, comparable to immediate post-training levels, after implementing spaced practice-test trials and corrective feedback. Overall, these studies introduced novel procedural modifications that enhanced FTT's retention outcomes, contributing valuable insights into the literature on emergent learning and FL vocabulary acquisition.

Chapter 1: Introduction

1.1 Introduction and background to the research problem

This thesis investigated the most effective and efficient methods for learners to acquire and retain emergent foreign language (FL) vocabulary. To address FL vocabulary retention, it modified instructional programming to evaluate which procedural variations produced improvements in learning outcomes up to six months after training. The methodology employed was quantitative, including a meta-analysis, and primarily focused on studies utilising single-case experimental design (SCED).

1.1.1 Foreign language learning

Studying an FL offers many advantages for the learner, including greater employment opportunities (New American Economy, 2017), and positive impacts on cognition (Antoniou et al., 2013; Cheng et al., 2015), as well as emotional and social well-being (Klimova et al., 2021). Developing fluency in FL, however, can be difficult and costly, particularly for adults (Hartshorne et al., 2018), with some estimates indicating that languages such as Chinese and Korean require at least 2200 hours of study (88 weeks) for native speakers of English (U.S. Department of State, n.d.). Despite the clear advantages of FL proficiency, the journey to fluency is fraught with challenges. This research aimed to address these challenges by exploring innovative instructional strategies that could improve the efficiency and retention of FL vocabulary.

1.1.2 Vocabulary

Vocabulary is a critical and widely acknowledged component of all aspects of FL learning and performance; fluency and mastery require learners to acquire as much of the lexicon

as possible (Yousefi & Biria, 2018). Successful communication—whether speaking, listening, reading, or writing depends on having a sufficient working vocabulary (Susanto, 2017). Uchihara and Saito (2019) showed that productive vocabulary knowledge was predictive of learners' oral fluency and argued that having a larger vocabulary reduces the likelihood of pauses or repetitions during speech. Similarly, Wallace (2020) demonstrated the importance of vocabulary to listening performance, with learners possessing greater vocabulary knowledge better able to comprehend FL speech. Additionally, Schmitt et al. (2011) found a direct relationship between learners' vocabulary knowledge and their level of comprehension of text. The more words they knew, the better their understanding of the text. Furthermore, as Nation (2022) contends, FL learners' vocabulary is an important measure of the quality of their written work. This was evident when Alderson (2005) compared learners' scores on vocabulary tests and found high correlations with scores from tests of other language components. Alderson, along with the studies described above, showed that learners' vocabulary knowledge contributes greatly to their broader FL performance.

Vocabulary is the cornerstone of FL learning and performance, yet many learners struggle to acquire and retain sufficient vocabulary for effective communication. This thesis's research targeted this critical issue by investigating methods to enhance vocabulary acquisition and retention.

1.1.3 Vocabulary size

Although the importance of vocabulary is clearly acknowledged, acquiring sufficient vocabulary is one of the primary difficulties learners experience (Morgan-Short & van Hell, 2023). In examining the challenges associated with FL vocabulary acquisition, researchers attempted to estimate the number of words learners must know. For example, Schmitt (2010)

estimated that native speakers possess a vocabulary of approximately 16000-20000 English word families. Word families are groups of words that typically comprise six or more related words, including the root form, inflections, and derivatives. To be comparable to native speakers, learners must attain a considerable vocabulary size.

However, the goal of acquiring a native-level size vocabulary may not be practical or necessary for every learner. In terms of written vocabulary, Schmitt et al. (2011) recommend learners to be familiar with 98% of the vocabulary in a text to comprehend it. Using the 98% coverage target, Nation (2006) calculated that English as a foreign language (EFL) learners seeking to engage with a broad range of reading material must be familiar with approximately 8,000–9,000 word families.

In comparison, spoken texts usually comprise more high-frequency words and require a slightly smaller but still substantial vocabulary (6000–7000-word families; Nation, 2006). More recently, Ha (2021) determined that EFL learners had to master the 8000 most common word families to be able to comprehend 98% of academic texts. These targets fall to 5000 word families when a more conservative 95% requirement is applied. Similarly, Van Zeeland and Schmitt (2013) estimated that EFL learners require a vocabulary of 2000–3000 word families to sufficiently comprehend 95% of spoken texts. Regardless of the modality or coverage target applied, these studies illustrate the daunting task that learners face in attaining sufficient vocabulary to communicate successfully in an FL.

The extent of this challenge is evident in studies that identified apparently substantial deficits in the vocabulary sizes of FL learners, including tertiary-level learners from a range of settings (e.g., Barrow et al., 1999; Engku Ibrahim et al., 2013; Nurweni & Read, 1999). Given the noted deficits, it is also concerning that learners often lack knowledge of successful

acquisition strategies (e.g., Romagnoli & Conti, 2019). This current study aimed to bridge the gap by exploring techniques that can help learners increase their vocabulary knowledge in more efficient ways.

1.1.4 Instructional strategies

Vocabulary instruction is often differentiated as incidental or intentional (Jonathans et al., 2021). The former approach involves the implicit acquisition of vocabulary learning, typically through reading texts, watching videos, or playing games. Incidental vocabulary learning approaches, including watching movies and reading books in the FL, although beneficial, tend to be slow and are criticised by some as unsystematic and ineffective for many learners when used as the sole acquisition strategy (Paribakht & Wesche, 1999). Intentional vocabulary instruction, on the other hand, involves explicit programming, such as flashcards. Norris and Ortega (2005) reported in their meta-analysis that intentional teaching methods were more effective than incidental approaches. The current study evaluates an approach that combines intentional and incidental learning in one procedure.

Several studies examined the effectiveness of various approaches to FL vocabulary instruction and found considerable variability in learning gains. Webb et al. (2020), for example, reviewed 100 effect sizes from 22 studies that used flashcards, word lists, writing, or fill-in-the-blanks, and found learning gains of up to 60.1% on average from intentional vocabulary instructional procedures; however, results varied substantially among learning activities, with mean values ranging from 18.4% to 77% on post-tests conducted immediately following training. In their review, flashcards and wordlists appeared to be the most efficacious strategies based on averaged learning gains. Webb et al. also noted that learner gains varied considerably

across studies even when they compared the same strategy, suggesting that not all learners benefited from what appeared to be the most effective strategies.

Webb et al. (2020) highlight the inconsistent outcomes for FL learners and show that the effectiveness of strategies for acquiring vocabulary can vary considerably among individual learners. Zhang and Graham (2020), for example, compared the learning gains associated with three types of attention-enhancing strategies and a control group. This study involved 137 senior high-school EFL learners in China who were divided into four treatment groups receiving different types of post-listening vocabulary explanations (L2-only, codeswitched, contrastive focus-on-form, or none). The researchers compared the effects of these treatments on both vocabulary acquisition (using pre-, post-, and delayed post-tests) and listening comprehension (using pre- and post-tests). All four groups achieved statistically significant gains, and the results indicated large effect sizes between the pre- and immediate post-training vocabulary tests, including the control group. Individual learner's results within and across groups showed sizeable variability in performance, though, and scores worsened for at least one of the learners. These differences at the individual level are important data that may reveal critical learning variables. It is difficult to determine the source of this variability from the aggregated data reported by the study's group-design, though (Johnston & Pennypacker, 2010), and more SCED research could address this gap.

In reviewing the existing FL literature, it is evident that few studies have employed SCED. Although statistical and meta-analytical findings based on aggregated group data provide valuable information about the general effectiveness of instructional strategies at the group level, these results are not generalisable to individual learners (Johnston & Pennypacker, 2010). Therefore, more FL research using SCED is needed, in addition to research using group designs,

to develop a robust approach to vocabulary instruction. Such research must thoroughly examine the critical variables that affect outcomes for all learners. This thesis addresses this gap by systematically examining the impact of emergent learning procedures—combining intentional and incidental learning—at both the group and individual learner levels, an approach employed by behaviour analysts.

1.1.5 Behaviour analysis

Behaviour analysis has much to offer those looking to optimise instructional programs as the history of the field is replete with demonstrations of measurably effective, empirically derived instructional technologies (Binder & Watkins, 1990; Vargas, 2020). One of the cornerstones of behaviour-analytic research is the use of SCED. Single-case experimental design is characterised by ongoing, repeated measurement and replication across conditions or participants, which are critical concerns in studying the behaviour of individuals (Kazdin, 2021).

One of the many benefits of SCED is precise examination of behavioural variability at the individual level. In comparison, group-level designs focus on the effect of an intervention at the population level, which may fail to detect potentially important sources of variability (Johnston & Pennypacker, 2010). Further, well-designed SCEDs are true experiments that allow researchers to control for multiple threats to internal validity and draw causal inferences. Finally, SCEDs provide researchers with viable alternatives to large group studies without the need for costly or time-intensive randomised control trials (Lobo et al., 2017).

Although still an emerging field of study in behaviour analysis, research output in FL instructional programs has grown substantially in the past 10 years. Several recent studies focused on evaluating behaviour analytic procedures to teach FL vocabulary instruction with children (Cao & Greer, 2018; Cortez et al., 2020, 2021; Matter et al., 2020; May et al., 2016,

2019; Petursdottir et al., 2014) and adults (Daly & Dounavi, 2020; Dounavi 2014; Wu et al., 2019). Defining characteristics of this literature are its use of SCED and its emphasis on emergent learning as an efficient approach to FL instruction. Specifically, what are the conditions that reliably occasion emergent learning within applied learning settings? This thesis employed SCEDs alongside statistical analyses to generate further insights into individual learning processes, building on the growing behavior-analytic literature and complementing the broader group-level studies in applied linguistics, while also addressing their limitations.

1.1.6 Emergent learning and verbal operants

Behaviour-analytic approaches to teaching FL vocabulary have typically employed emergent learning principles and verbal operants (Daly & Dounavi, 2020). Emergent learning, occasionally termed generative learning in behaviour-analytic literature, describes the acquisition of knowledge without the necessity for direct experiences (Critchfield & Twyman, 2014). This process results in some learning occurring ‘for free’, where training one set of skills leads to proficiency in related, untrained skills (Critchfield, 2018). Such ‘free’ learning offers significant benefits for both learners and instructors (Critchfield & Twyman, 2014). Importantly, since the amount of time available for instruction is always constrained, educators must aim to achieve the most effective outcomes with the resources available.

Verbal operants are specific verbal behaviours defined not by their form but by their function and the conditions that occasion them (Skinner, 1957). The primary verbal operants include mands, tacts, intraverbals, echoics, textuials, and listener behaviour. Mands, which function as requests, and tacts, which operate as names or labels, are productive verbal behaviour occasioned by non-verbal stimuli (e.g., states of deprivation and the presence of objects and events). In comparison, echoics, textuials, and intraverbals are controlled by other verbal

behaviour (e.g., speech or written texts) and characteristically produce generalised social reinforcement from members of the learner's linguistic community. Finally, listener behaviour is a class of receptive verbal operant in which the learner follows directions given by others.

There is growing interest in emergent learning in behaviour-analytic research and practice, but the field has not yet developed to the point of being considered a robust technology (Blair & Shawler, 2019; Critchfield, 2018; Rehfeldt, 2011). The predominant research in this area is limited to demonstration studies. According to Critchfield, the prevailing position is that when implementing an emergent learning program, emergence may or may not be observed at some point in the future. Basic behavioural research has identified a range of conditions under which emergent learning is more likely; however, it is difficult for instructors to apply what has been learned in these very controlled studies to applied instructional programs. Research to date primarily demonstrates the possibility of learning outcomes that exceed what is achieved by traditional programs targeting only abilities that are directly taught. Critchfield suggests that the lack of a robust technology of emergent learning is due to a predominant focus on basic laboratory research rather than typical learning settings. This study sought to advance the understanding and use of emergent learning principles within an applied FL learning context.

1.1.7 Foreign tact training

Foreign language vocabulary acquisition programs that employ emergent learning procedures have the potential to increase receptive and productive language skills while reducing the time required to learn in comparison with traditional language programs (Matter et al., 2020). Foreign tact training (FTT), an emergent learning procedure, teaches learners to vocalise FL vocabulary when presented with corresponding visual stimuli (e.g., picture cards) and assesses the emergence of untrained verbal operants.

The standard FTT instructional procedure consists of discrete trials with picture flashcards (Matter et al., 2020; May et al., 2019; Wu et al., 2019). The instructor presents individual picture cards to the learner, prompting them to name the object depicted on the card using the appropriate FL word. If the learner fails to respond correctly within three (or five) seconds, the instructor models the correct response. Following a correct response, the instructor acknowledges the learner's response before moving on to present the next picture card. This sequence of trials continues until all cards have been presented at least once or the allocated instructional time expires. At the end of each trial block, the instructor provides the learner with feedback on their accuracy, e.g., 'well done, you got 80% correct.' The instructor might also encourage the learner to continue to improve their performance to achieve a mastery criterion score, e.g., 'Keep it up, you almost scored 100%. Let's see if you can get there during the next session.'

Matter et al. (2020) demonstrated that a single emergent learning procedure comprising FTT alone was just as effective and more efficient than a conventional procedure focused on directly teaching all of the relations (i.e. intraverbals, listener, and tact responses). Matter and colleagues used an adapted alternating treatments design to assess training in Spanish language targets with English-speaking children. For most acquisition targets, FTT achieved mastery criteria with fewer sessions than the multi-component procedure. Foreign tact training alone also resulted in acquisition of almost all emergent untrained receptive and productive relations for all participants, without receiving any direct training in these relations.

May et al. (2019) investigated the outcomes of a group-based emergent learning program with six English-speaking Welsh primary school children. The children were first taught tacts in their native language using pictures of common objects (e.g., 'what do you call this in English?')

until the group's performance met mastery criterion. Using a similar procedure, the children were taught to tact each object in Welsh (e.g., 'what do you call this in Welsh?'). Subsequent post-testing probed for emergence of intraverbal Welsh (e.g., "What is this English word in Welsh?") and intraverbal English responding (e.g., "What is this Welsh word in English?"). The results demonstrated substantial improvements in emergent intraverbal relations for three of the six participants. In other words, half of the participants could translate words between the two languages without explicit training. The other three participants' results demonstrated much lower levels of accuracy. The authors noted that disparities between the pre- and post-testing procedures—specifically the use of both Tact-English and Tact-Welsh trials during post-tests—affected the accuracy of participants' emergent responding. The authors recommended more robust testing procedures in future studies. They also reported that participants who did not maintain high accuracy on the directly taught tact responses (specifically Tact-Welsh) also tended to exhibit lower levels of emergent responding. In other words, failures in emergent learning appeared to be linked to a failure to retain the directly taught tacts. Maintenance was tested during follow-up probes with each participant two weeks after post-testing. The results indicated relatively stable performance, but the authors recommended that future research extend the testing interval to examine more long-term outcomes.

In one of the few behaviour-analytic studies to date to implement emergent learning procedures with adults learning an Asian FL, Wu et al. (2019) compared learners' acquisition of Chinese words using four verbal operants—mand, tact, and two intraverbal training procedures. Overall, mand and tact training produced the most emergent responses. Wu et al. noted one of the limitations of the study was the lack of retention measures, and it is unclear whether responding maintained beyond the immediate post-training test period.

Foreign tact training has shown promise in enhancing FL vocabulary acquisition. This thesis evaluated FTT's effectiveness and explored procedural modifications to maximise its impact, addressing a significant gap in the existing literature concerning learners' retention of emergent FL vocabulary.

1.1.8 Retention

Retention refers to the level of performance demonstrated by the learner following a period (i.e., retention interval) without practice or reinforcement (Binder, 1996). Maintenance, on the other hand, describes learners' performance levels outside of training but within a context that provides the learner with opportunities to practice the acquired material. Robust levels of retention are critical learning outcomes because changes in performance that fail to persist beyond the classroom/instructional context or in the absence of ongoing practice are unlikely to produce meaningful changes in the learner's life. Similarly, FL learners who want to retain vocabulary learning without access to a linguistic community to practice vocabulary naturally are limited to the opportunities within formal instructional programs or self-study (Petursdottir & Oliveira, 2023). Australians learning Korean language vocabulary at home, for example, are unlikely to have ongoing access to a natural Korean language environment. In these circumstances, retention is particularly important for the learner to continue to develop the size and breadth of their vocabulary.

Although interest in emergent learning and FTT is on the rise, the retention of knowledge acquired through FTT is not yet well understood. Only a handful of applied studies have investigated maintenance following FTT (Cortez et al., 2020, 2022; Daly & Dounavi, 2020; Matter et al., 2020; May et al., 2019), and no research has assessed the long-term retention of vocabulary.

According to Johnson and Layng (1996, p. 285) a sufficiently long-term retention interval is “a month or more.” The few studies to date that have demonstrated robust long-term retention come from the broader emergent learning and non-applied research literature (i.e., arbitrary or nonsense words) which raises concerns about the utility of emergent learning procedures for developing socially meaningful outcomes (Eilifsen, & Arntzen, 2017). Regaço et al. (2023) highlighted these concerns when they identified just 24 papers among the broader stimulus equivalence literature examining maintenance of emergent learning outcomes. They recommended further research to identify the critical variables, including training procedures, that impact retention. Retention is a crucial yet often overlooked aspect of emergent FL learning. This thesis assessed not only the immediate effectiveness of FTT but also its long-term retention outcomes, providing a comprehensive evaluation of FTT’s practical utility.

1.2 Statement of the problem

In summary, the contextual factors and issues related to this thesis’ research questions were as follows:

- a) FL learners must acquire a substantial vocabulary size to be able to communicate effectively;
 - b) FL vocabulary acquisition and retention are challenging for learners not immersed in a natural linguistic community;
 - c) a nascent but growing field of research points to the potential benefits of emergent learning procedures for producing effective and efficient FL vocabulary instruction;
- but

- d) emergent learning outcomes are highly variable across learners, retention of learning is rarely evaluated in applied research studies, and it is unclear what training procedures reliably produce long-term learning outcomes.

An initial scoping of the emergent learning literature identified FTT as a promising instructional procedure for FL vocabulary acquisition; however, it was difficult to evaluate FTT's effectiveness in comparison with other instructional procedures as there were no systematic reviews of FL emergent learning programs available at the outset of this thesis. Additionally, little was known about the long-term learning outcomes (i.e. longer than one month) associated with FTT, or the procedural variables that impact retention, or how FTT might be improved to optimise those outcomes. This study sought to address these gaps by systematically investigating the efficacy of FTT and its impact on long-term vocabulary retention.

1.3 Research questions

This thesis sought to investigate the critical issue of vocabulary retention in FL learning through a series of empirical studies within the context of applied learning activities. By examining procedural modifications in FTT, this research aimed to identify strategies that enhance the retention of emergent vocabulary learning, thus providing valuable contributions to the field of behaviour analysis and FL instruction. The main research question was:

What procedural modifications to the design and implementation of FTT contribute to improved retention of emergent Korean vocabulary?

The research methodology included a systematic review, meta-analysis, and a series of applied experimental studies conducted at the individual learner level of analysis. In addition to examining the main research question regarding the impact of a modified FTT protocol on the retention of emergent learning outcomes, this thesis investigated the following research sub-questions:

1. Chapter 3 (Technical paper; Study 1). This chapter considered the technical aspects of computer-assisted data-extraction and developed a protocol for use when conducting systematic reviews with behaviour-analytic literature:
 - a) *How does one extract graphical data for re-analysis from published SCED graphs using a data extraction software program?*
2. Chapter 4 (Systematic review and meta-analysis; Study 2). In this chapter, systematic review examined the effects of FTT on emergent learning outcomes:
 - a) *What are the effects of FTT on emergent learning outcomes in the published literature to date?*
 - b) *How do FTT acquisition, emergence, and overall efficiency compare with other verbal operant training procedures?*
 - c) *Does FTT produce higher levels of emergent responding for adults or children?*
3. Chapter 6 (Empirical paper, Study 3): This chapter evaluated a modified FTT protocol that incorporated fluency-based overlearning trials. The key experimental research question was:

- a) Does repeated practice of foreign tact relations beyond initial mastery (i.e., accuracy criterion) using fluency-building procedures affect the retention accuracy of derived intraverbal relations during testing?*

4. Chapter 8 (Empirical paper, Studies 4 and 5). This chapter incorporates two experiments, the first of which attempted to evaluate the effects of response rate and extended practice while controlling for total training duration:

- a) Does an intertrial interval during training affect the retention accuracy and response rate of emergent and directly trained learning during post-testing?*

The final experiment further modified the FTT protocol to include spaced practice and reassessed retention outcomes:

- b) What are the effects of weekly practice tests with and without feedback on retention of emergent and directly trained learning?*

Chapter 2: Theoretical Framework

2.1 Introduction

Emergent learning² is a term used in behaviour analysis to describe learning that is not the result of direct experience (Critchfield & Twyman, 2014). For instance, when learning a foreign language, ‘direct experience’ might involve explicitly practicing each target word, while emergent learning involves the indirect or incidental acquisition of language. The phenomenon of emergent learning has intrigued countless researchers and educators, and several fields of psychological study have attempted to explain how emergent learning occurs. This chapter presents an overview of the theoretical framework and a historical perspective on the development of emergent learning principles. By exploring these foundational theories, this research aimed to build upon established knowledge and contribute to the growing understanding of how the principles of emergent learning can be used to improve FL vocabulary acquisition and retention.

2.2 Historical perspectives of emergent learning

2.2.1 Cognitive psychology

Cognitive psychologists argue that emergent learning occurs due to the learner engaging in what is termed ‘meaning making’ through a process of critical reflective thinking (Critchfield, 2018; Zittoun & Brinkmann, 2012). Additionally, cognitive psychologists assert that learners produce new knowledge beyond what they already know by mentally manipulating the material using logical reasoning (Seel, 2012). For example, the EFL learner who ‘knows’ that a ‘bird’ is a winged animal, and a ‘house’ is a dwelling, uses this knowledge to conclude, without direct

² Other related terms include derived relations, stimulus equivalence, and generativity (Lafrance and Tarbox, 2020).

instruction, that a 'birdhouse' is a dwelling designed to accommodate winged animals. While this perspective highlights the role of cognitive processes in learning, it underestimates the importance of social and environmental factors. Learning is not solely an internal, cognitive activity; it is also influenced by interactions with others, cultural contexts, and the availability of external resources and experiences (Vygotsky, 1978). Therefore, focusing exclusively on mental manipulation and logical reasoning provides an incomplete picture of how emergent learning occurs.

2.2.2 Gestalt psychology

Gestalt psychologists hold a similar position to cognitive psychologists and argue that the learner constructs meaning through insight (Ash et al., 2012). Köhler (1925) was one of the first to explore the notion of insight through a series of well-known experiments in which chimpanzees were placed in an environment with food in their line of sight but not within reach. The apes attempted a range of previously employed problem-solving strategies, including climbing, and using sticks. When these attempts failed, Köhler observed the apes stop engaging in all overt trial and error problem-solving actions for a period. Then, suddenly, some of the apes solved the problem by stacking crates and climbing on top. Köhler concluded that the successful apes had abandoned searching for solutions in the external physical environment and shifted their efforts to internal cognitive trial-and-error strategies (Johnson & Street, 2004). Gestalt psychologists refer to this internal cognitive learning approach as 'insight learning' (Ash et al., 2012, p. 6). Behaviour analysts argue that viewing these phenomena as internal processes only is unhelpful as it fails to provide a means by which the learner who lacks insight or the ability to produce meaning making can develop these apparent spontaneous untaught capabilities (Critchfield, 2018).

2.2.3 Behaviour analysis

Epstein et al. (1984) aimed to extend Köhler's (1927) work and provide a behavioural interpretation of 'insight' (Johnson & Street, 2004). Epstein et al. taught pigeons a series of three problem-solving strategies: pushing a small box around, stepping on a box, and pecking at an object. The experimenters taught each strategy separately, and systematically varied instruction for the pigeons. Consequently, only the birds that were taught all three problem-solving strategies were able to successfully solve the problem. Epstein et al. contended that the performance of this group of pigeons demonstrated "...suddenness, directness and continuousness" (p. 62) and satisfied "...Köhler's criteria for 'genuine' or 'insightful' solutions" (p. 62). Consequently, Epstein argued that the pigeons' ability to perform these novel solutions was a function of previous experience, not 'insight'. Epstein (1991) later referred to this phenomenon, in which combinations of existing repertoires spontaneously emerge in the form of novel behaviour, as 'generativity'. Emergent learning and other related behaviour-analytic terms, including stimulus equivalence and derived stimulus relations, are forms of generativity; characterised by the emergence of novel behaviour, they encompass all behaviour that is not reinforced or directly trained (Lafrance and Tarbox, 2020). By grounding emergent learning in observable behavioural processes, this perspective offers practical strategies for promoting such learning in FL instruction.

2.2.4 Stimulus generalisation

Despite the ubiquity of emergent learning and other forms of generative behaviour, it has long been considered one of the most difficult of psychological phenomena to account for and is the subject of countless theories and investigations over the last 100 years, including those

described above (Critchfield et al., 2018). For example, in a paper published in 1939, Hull asked the question, “How can we account for the fact that a stimulus will sometimes evoke a reaction to which it has never been conditioned?” (p. 9). At the time, Hull contended that this phenomenon is due to stimulus generalisation.

Stimulus generalisation describes learners’ tendency to produce similar responses to stimuli with comparable physical dimensions (e.g., Reynolds, 1961). In other words, a response which has been reinforced in the presence of a particular controlling antecedent stimulus, comes to be evoked by other physically similar stimuli (Cooper et al., 2007). This is evident when a learner is taught to say “개” (Korean word for dog) when referring to his teacher’s German Shephard, and then uses the same phrase without prompting to tact his neighbours’ Siberian Husky. In this example, tacting the Korean word for dog was generalized to novel, but physically similar antecedent stimuli – German Shephard to Siberian Husky. Yet, stimulus generalisation does not fully explain emergent learning.

2.2.5 Derived relational responding

Stimulus generalisation does not explain humans’ ability to produce similar responses to stimuli with dissimilar physical properties (Dougher et al., 2014). For instance, a Korean language learner who is taught to say '의자' when shown a picture of a chair, and then independently points to the picture of the chair when he hears '의자', is relating two stimuli that have no physical similarities (Sidman, 2009; Stewart, 2018). The visual image of a chair lacks any formal correspondence with the spoken word—one is a sound, and the other is a visual image. Their relation is arbitrary (Stewart, 2018).

In contrast to stimulus generalisation, derived relational responding is a form of equivalence between two or more stimuli that emerges without the need for explicit instruction or similar physical properties (Stewart, 2017). Countless examples of this phenomenon are observable in a variety of everyday circumstances. For example, when told that Seonhwa is older than Dongil and Chloe is younger than Dongil, most typically developing school-age learners can correctly identify that Chloe is the youngest and Seonhwa is the oldest of the three, despite not seeing what they look like or knowing their specific ages; the learner has derived the arbitrary equivalence relations between the stimuli. This ability to derive relationships between stimuli that do not share physical attributes highlights the complexity of emergent learning, setting the context for a deeper exploration of the theories that aim to explain this phenomenon.

2.3 Theories related to emergent learning

Three main theories have been proposed to explain the conditions that give rise to derived stimulus relations and emergent learning (Lafrance & Tarbox, 2020; Rehfeldt, 2011). These three theories: stimulus equivalence (Sidman, 1971), naming theory (Horne & Lowe, 1996), and relational frame theory (RFT; Hayes et al., 2001), all focus on the learner's experiences and history of reinforcement of verbal behaviour. Language as verbal behaviour is like all other behaviour in that its development is impacted by the individual's learning history with behaviour-environment relations (Lafrance & Tarbox, 2020).

2.3.1 Stimulus equivalence

In his seminal study on stimulus equivalence, Sidman (1971) discovered that when certain stimulus-response relations are taught under the right conditions, other untrained (i.e. emergent) stimulus-response relations may emerge. The learner in Sidman's experiment could not read and









was subsequently taught to match 20 spoken and printed words. Following training, they matched printed words to related pictures and said their names aloud, without prior training in these specific stimulus-response relations. Sidman referred to this as an equivalence relation. Equivalence relations are said to have emerged when each of the stimuli that comprise the putative equivalence class are substitutable with the other stimuli within the class (Bortoloti & de Rose, 2011).

It was Sidman and colleagues who first operationally defined criteria and behavioural tests for verifying equivalence relations (Sidman & Tailby, 1982). Specifically, they proposed that an equivalence relation established between dissimilar stimuli must demonstrate reflexivity, symmetry, and transitivity (Sidman, 2018; Sidman & Tailby, 1982).

Table 1 shows examples of Sidman's stimulus equivalence relations (Sidman & Tailby, 1982). Reflexivity refers to the condition in which the learner can match a stimulus with itself ($A=A$); the learner who can match a picture of a dog (A) with a copy of the same picture (A) demonstrates the property of reflexivity. Symmetry, on the other hand, is the condition in which two different stimuli are explicitly taught to be related ($A=B$) and the learner then independently matches the second stimulus with the first ($B=A$). For example, when shown a picture containing a dog (A), the learner is taught to select the text “개” (B) and when later shown the text “개”, independently matches it to the picture of a dog. Transitivity requires a third stimulus. First, two different stimuli are related, and at the same time a third stimulus is explicitly related to the second ($A=B$ and $B=C$). The learner then independently matches the first and third stimuli ($A=C$). As shown in the example, after learning to match the picture with the referent Korean text ($A\rightarrow B$), the learner is then taught to say “geh” when shown the written text ($B\rightarrow C$). Transitivity is observed when the learner is then able to independently, without further training,

say “geh” when shown the picture of a dog (A=C) and point to the picture of a dog when he hears “geh” (C=A)

Table 1*Stimulus equivalence relations*

	Stimuli		Emergent stimulus relations	
Reflexivity	<div style="border: 1px solid black; padding: 10px; text-align: center;"> A (object/picture)  </div>		<div style="display: flex; align-items: center; justify-content: center;"> <div style="border: 1px solid black; padding: 10px; text-align: center;"> A (object/picture)  </div> = <div style="border: 1px solid black; padding: 10px; text-align: center;"> A (object/picture)  </div> </div>	
Symmetry	<div style="border: 1px solid black; padding: 10px; text-align: center;"> A (object/picture)  </div>	→ <div style="border: 1px solid black; padding: 10px; text-align: center;"> B (foreign text) 가 </div>	<div style="border: 1px solid black; padding: 10px; text-align: center;"> B (foreign text) 가 </div>	= <div style="border: 1px solid black; padding: 10px; text-align: center;"> A (object/picture)  </div>
Transitivity	<div style="border: 1px solid black; padding: 10px; text-align: center;"> A (object/picture)  </div>	→ <div style="border: 1px solid black; padding: 10px; text-align: center;"> B (foreign text) 가 </div>	<div style="border: 1px solid black; padding: 10px; text-align: center;"> A (object/picture)  </div>	= <div style="border: 1px solid black; padding: 10px; text-align: center;"> C (spoken word) “geh” </div>
	<div style="border: 1px solid black; padding: 10px; text-align: center;"> B (foreign text) 가 </div>	→ <div style="border: 1px solid black; padding: 10px; text-align: center;"> C (spoken word) “geh” </div>	<div style="border: 1px solid black; padding: 10px; text-align: center;"> C (spoken word) “geh” </div>	= <div style="border: 1px solid black; padding: 10px; text-align: center;"> A (object/picture)  </div>

Note. The foreign text “가”, pronounced ‘geh’, meaning dog, is written using the Korean writing system, Hangul.

Sidman (1986) stated that the key contribution arising from the study of stimulus equivalence relations is the mechanism it provides for identifying and producing the conditions that occasion emergent learning:

"By reacting to a word as an equivalent stimulus - the meaning of a word - a person can behave adaptively in an environment without having previously been exposed to it. The emergence of equivalence from conditionality permits Behavior Analysis to account for the establishment of at least simple semantic correspondences without having to postulate a direct reinforcement history for every instance. Instead of appealing to cognitions, representations, and stored correspondences to explain the initial occurrence of appropriate new behaviour, one can find a complete explanation in the...units that are the prerequisites for the emergent behaviour" (p. 236).

Sidman's work laid the foundation for understanding how these untrained relations emerge, leading to the development of other key theories in this area.

2.3.2 Naming theory

The second key theory, naming, involves the combination of listener and speaker behaviours following multiple examples and social reinforcement typically experienced during the formative years of child development (Horne & Lowe, 1996). Naming is developed as a generalised operant through a history of reinforcement for engaging in multiple exemplars of appropriate combinations of speaker-listener responses and is considered a key skill in the development of numerous academic skills (Miguel, 2016). Naming is also a pivotal skill for vocabulary acquisition, enabling learners to acquire language without direct instruction (Horne & Lowe, 1996).

This process typically begins during the early stages of infancy when children start to attend to their parents' vocal utterances as well as the ways in which they interact with the environment. Consequently, objects in their environment begin to take on discriminative functions (e.g., balls evoke rolling, bouncing, and kicking behaviours). As parents start labelling and requesting items from their child and differentially reinforcing non-verbal responses, the child develops a listener repertoire. In turn, the child develops a repertoire of speaker responses as her own vocal utterances contact reinforcement from her parents. As the child engages in these vocal utterances, she may respond to her own speaker behaviour with appropriate listener responses (and vice versa), which is said to be when she starts to develop the ability "...to listen with understanding" (Miguel, 2016, p. 128). When the learner develops the ability to engage in bidirectional naming of an object following simply listening to others tact the object, a robust bidirectional naming repertoire is developed (Miguel, 2016). This theory highlights the importance of integrating both speaker and listener behaviours in FL instruction to promote emergent vocabulary acquisition.

2.3.3 Relational frame theory

In contrast to naming theory, RFT expands on Sidman's stimulus equivalence theory and explains emergent learning as a repertoire of generalised verbal operants, which refers to verbal behaviour that has become controlled by a variety of stimuli and reinforcers. Multiple exemplar training in contextually cued patterns of stimulus relations leads to relational responding (Stewart, 2018). When an individual learns to relate stimuli in specific ways, and a pattern of generalised relational responding is developed, she can then apply those same relational patterns to novel stimuli and relevant contextual cues. For example, the relational terms hotter (☞

뜨겁다) and colder (더 차갑다) may be developed through discriminative learning and multi-exemplar training, e.g., repeated discrimination trials with prompting - which is hotter/ colder?

When the learner has had sufficient opportunities to engage in and be reinforced for discriminating hotter or colder stimuli when provided with these verbal contextual cues, the relational response may be applied to a range of stimuli that she has no direct experience or training with because the verbal cues (e.g., hotter/colder) occasion this pattern of relational responses. For example, when the learner is told that the sun is hotter than the surface of the earth, and the surface of the earth is hotter than the surface of the moon, they can then derive several untrained relations—sun is hotter than moon, moon is colder than sun, moon is colder than earth, earth is hotter than moon, and earth is colder than the sun. Relational frame theory extends stimulus relations beyond equivalence, meaning that a potentially unlimited number of relational frames are possible (Barnes-Holmes et al., 2018).

2.4 Applications of stimulus equivalence to language learning

As noted earlier in section 1.16, emergent learning procedures have the potential to make vocabulary acquisition more efficient. All three related theories—stimulus equivalence, naming theory, and RFT—share the common premise that language learners can develop vocabulary without direct exposure to each response form of a word (Matter et al., 2020). That is, rather than teaching all relations directly, establishing a single relation (e.g., learning to tact ‘geh’ in the presence of a dog) may result in several other verbal operant relations without instruction (e.g., intraverbal and listener relations; Petursdottir & Oliveira, 2023). This potential for emergent learning to improve vocabulary acquisition efficiency underscores the importance of this research in identifying and refining effective instructional strategies.

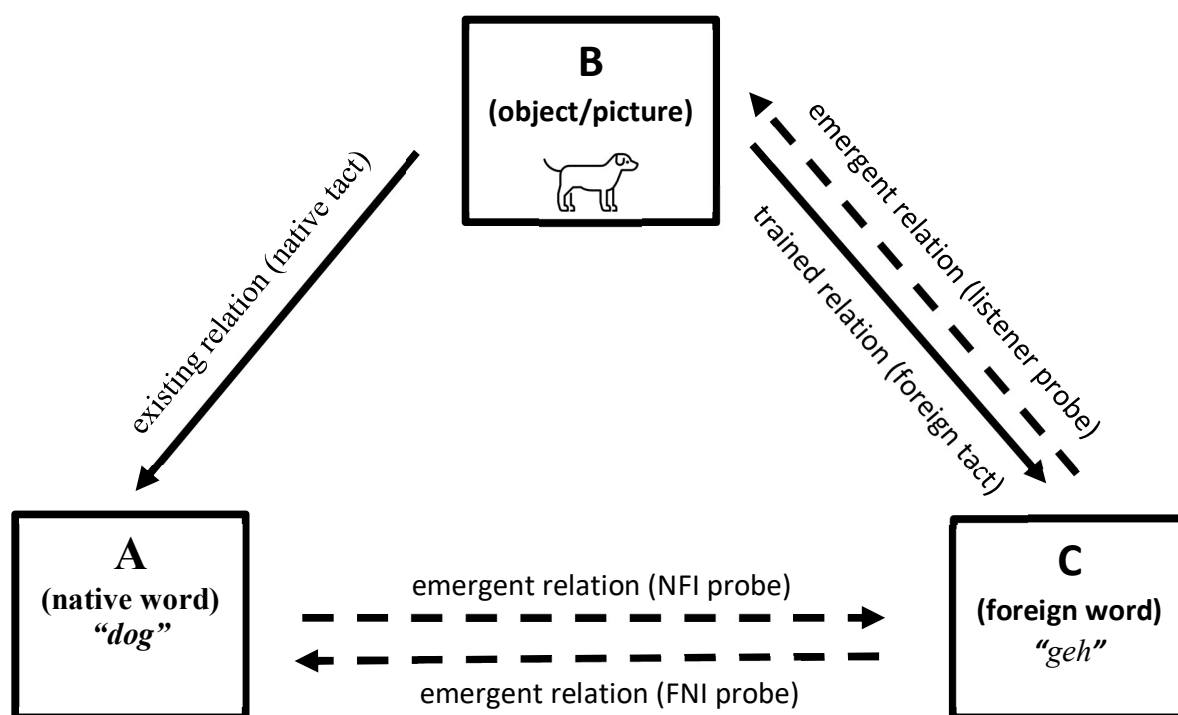
Figure 1 illustrates the FTT process by showing how several emergent relations are derived from existing and directly trained tact relations. Foreign language learners can potentially derive equivalence between pictures, the related native words, and foreign words (Daly & Dounavi, 2020; May et al., 2013). According to Sidman’s stimulus equivalence, relating stimulus A (‘dog’) with B (picture of dog), and B (picture of dog) with C (“개”), is likely to result in several untrained relations (Sidman, 2018).

Similarly, naming theory emphasises the integration of listener and speaker behaviours, suggesting that learners who develop bidirectional naming can acquire new vocabulary without direct instruction in each form. Relational frame theory extends these ideas by proposing that relational frames, once established, can be applied to novel stimuli and relevant contextual cues.

Commonly, FL learners can name (tact) each of the target pictures in their own language (e.g., Daly & Dounavi, 2020; Dounavi, 2011, 2014; Matter et al., 2020; Petursdottir & Haflíðadóttir, 2009; Petursdottir et al., 2008; Wu et al., 2019). Therefore, following successful FTT, learners can relate stimulus A (native-language word) to B (picture), and B (picture) to C (foreign-language word). Instructors then test learners’ emergent responding, and probe for untaught relations: FNI trials probe for the untrained C–A relations, NFI trials probe test for A–C relations; and listener probes test for C–B relations. Contextual cues are used to evoke learners’ emergent responses, e.g.— “What is ‘geh’ in English?”, “How do you say ‘dog’ in Korean?”, “Point to ‘geh’”, “What is the Korean name for this?”

Figure 1

Existing, trained, and emergent relations following foreign tact training



Note. NFI = native-to-foreign intraverbal, FNI = foreign-to-native intraverbal. Adapted from Wooderson et al. (2022).

2.5 Introduction to Chapters 3 and 4

This chapter presented an overview of the theoretical foundations and historical perspectives of emergent learning, highlighting its potential in FL instruction. Understanding these theories is crucial as they provide a framework for developing effective instructional strategies using emergent learning principles to enhance vocabulary acquisition. The insights gained from these theories are essential for designing interventions that maximise learning efficiency and facilitate the acquisition of new language skills without requiring extensive direct instruction.

The next two chapters focused on the methodological aspects of this thesis. Chapter 3 investigated techniques for extracting graphical data from SCED research. These data extraction techniques were then utilised in the systematic review presented in Chapter 4. This review focused on FTT, which has been described as potentially the most productive of the various verbal operant emergent learning procedures (Matter et al., 2020). However, determining the generality of FTT outcomes is difficult because current research is limited to SCED studies. To address this gap in the literature, aggregate data from the research on FTT was extracted and analysed to determine its impact on emergent learning outcomes at the group level of analysis. Finally, the meta-analysis evaluated the extent to which FTT acquisition, emergence, retention, and overall efficiency compared to other emergent learning procedures.

Chapter 3: Tutorial: Extracting Published Graphical Data Using Digitizeit: A Step-By-Step Guide for Behavior Analysts Wanting to Reanalyze Single-Case Data

Peer reviewed paper: Wooderson, J. R., Bizo, L. A., & Young, K. (2024). Extracting published graphical data: A guide for researchers wanting to synthesize single-case data. *Experimental Analysis of Human Behavior Bulletin*, 34, <https://doi.org/10.17605/osf.io/rfhjx> ³

3.1 Abstract

Graphical data displays and visual analysis are cornerstones of behavior-analytic research. However, graphical data present challenges when conducting analyses across published studies. Specifically, single-case experimental design graphs frequently employ different axis scales across studies and sometimes within studies; and researchers often do not publish the raw data. Consequently, researchers wanting to compare or reanalyze data across published literature often need to extract data from plotted line graphs. In cases where raw data is inaccessible, data extraction software programs provide a valid and reliable method for digitizing graphed data sets from published single-case experimental design studies. Our purpose in writing this article is to demonstrate how to extract graphical data from published articles using the software program Digitizeit™. We present a series of task analyses and an example for readers to follow to import single-case graphs into the program, plot data points, and export the numerical data to a spreadsheet program.

Keywords: single-case experimental designs, data extraction, task analysis, visual analysis

3.2 Introduction

Single-case experimental designs (SCED) are characterized by ongoing, repeated measurement and replication across conditions or participants—which are critical concerns in

³ The paper was published in an American journal and uses American English spelling throughout.

studying the behavior of individual organisms (Kazdin, 2021). One of the many benefits of SCEDs is that they allow for precise examination of behavioral variability at the level of the individual. In comparison, group-level designs focus on the effect of an intervention at the group level, which may fail to detect potentially important sources of variability among individuals (Johnston & Pennypacker, 2010). Further, well-designed SCEDs allow researchers to control for multiple threats to internal validity and draw causal inferences. Consequently, SCEDs provide researchers with viable alternatives to large group studies (Lobo et al., 2017). Due to these methodological advantages, researchers and practitioners in applied behavior analysis and related fields primarily use SCEDs and visual analysis methods to evaluate intervention outcomes (Horner & Swoboda, 2014; Wolfe et al., 2019). Specifically, they use visual inspection of data plotted on line graphs to make decisions about whether to adjust treatment during intervention and determine the strength of causal relations between behavior and environmental variables (Kratowill et al., 2013; Ledford et al., 2018).

This reliance on visual analysis presents some challenges when comparing data across studies. Specifically, SCED graphs frequently employ different axis scales across studies and, sometimes, within studies. Also, SCED graphs may contain more than one independent variable (e.g., alternating treatments designs), which may confound visual comparisons with other studies. As a result, several supplemental statistical measures exist for estimating effect sizes for SCED data (Maggin et al., 2017). Still, visual analysis remains the most appropriate methodology for examining variability in SCED data stability, level, and trend (Ledford et al., 2018). Researchers and practitioners wanting to use visual analysis to evaluate the impact of interventions by comparing data patterns across studies must first extract the data from the published literature.

For example, Wooderson et al. (2022) recently conducted a meta-analysis comparing FL vocabulary training procedures across seven studies and 23 learners. One of the study's aims was to compare the effectiveness of several verbal operant training procedures. In addition to statistical measures, the researchers employed descriptive visual analysis by extracting data from the training phases of each study and graphing the data on standardized panels. Figure 1 shows graphs of two training phases included in the systematic review and a synthesized graph showing both phases in one panel. Even though the top two graphs were from the same study (Dounavi, 2014; p. 168), their x-axes used different scales and were difficult to compare using visual inspection. We found it easier to compare the data after plotting them on one graph (bottom panel, figure 1).

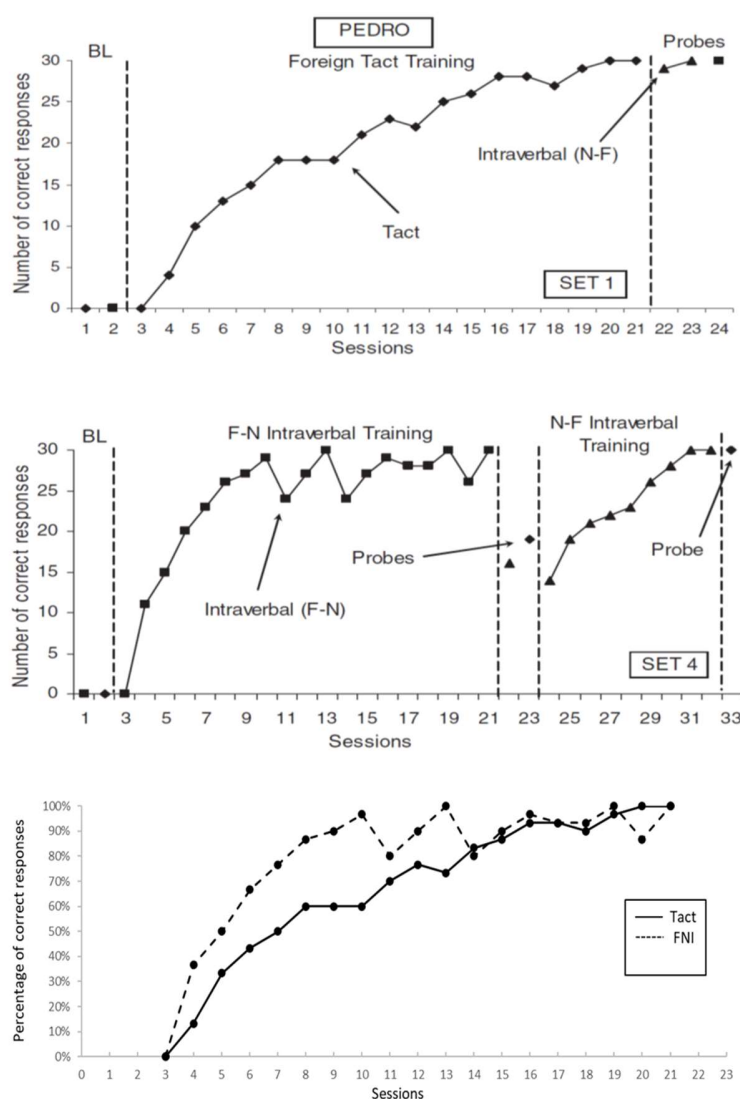
Although it would have been preferable to regraph the studies' raw data, Wooderson et al. (2022) could access the data directly from the authors of four of the seven papers, and could not obtain any data for three studies. Unfortunately, requesting data from researchers often does not work, as it may not be possible to contact them, they do not respond to requests, or advise that they no longer have access to the datasets (Van Der Zee & Reich, 2018). Although the advent of the Open Science movement and the capabilities afforded by digital technologies create the potential for greater access to empirical data, uptake by researchers is limited (Robson et al., 2021). Moreover, there is abundant basic and applied behavioral data not stored on digital repositories but published in the extant SCED literature.

In addition to the example above, there are other reasons why practitioners may want to extract and reanalyze published research data. First, engaging with and conducting research allows practitioners to approach problems in ways in which they have yet to be directly trained. Sidman (2011) considered the development of practitioners into scientist-practitioners a critical

imperative within the field of behavior analysis. He encouraged practitioners to engage in basic and translational research because it improves their practice and offers a “whole new slant on behavior analysis” (Sidman, 2011, p. 976). Meta-analyses can be useful tools for practitioners to

Figure 1

Example graph panels (top 2 panels) with differing x-axis scales (reprinted, with permission, from Dounavi, 2014 © John Wiley and Sons) and synthesised graph (bottom panel) containing both the Foreign Tact and Foreign-Native Intraverbal training data extracted from Dounavi (2014; p. 168;)



pose their own research questions and assess the strength of evidence for a given practice, particularly in the case of emerging or promising practices that have yet to be evaluated in the published literature. Second, essential criteria for behavioral interventions include the requirement that they are effective (Baer et al., 1968). Practitioners evaluating whether a given treatment intervention produced enough of a behavior change to be considered effective might compare their results with those in the published literature. Again, visual analysis is useful because it allows the practitioner to examine the data across interventions according to level, trend, and variability.

Difficulty accessing raw data is an issue for researchers and practitioners who want to reanalyze SCED data from the extant literature. Behavior-analytic literature employs graphical data presentations almost exclusively, and raw data or effect sizes are rarely presented within publications. In cases where raw data is inaccessible, data extraction software programs provide a valid and reliable method for digitizing graphed data sets from published SCED studies (Aydin & Yassikaya, 2022; Drevon et al., 2017; Flower et al., 2016; Rakap et al., 2016). These data are available for reanalysis or meta-analysis if one knows how to use tools like DigitizeIt™.

3.3 Method

This technical article demonstrates how to extract data from published SCED graphs using the data extraction software program. We recognize that readers may experience difficulties when using these procedures, so we included a simple example for readers to follow. However, the supplemental material also includes a demonstration video (<https://youtu.be/3XVkyUEWxkY>) and troubleshooting guide for more complex situations (e.g., plots with multiple y-axes). The basic procedure, however, is similar for most extraction programs and includes five key steps (Moeyaert et al., 2016; Rakap et al., 2016):

1. Import the graph into the program
2. Define XY axes
3. Plot data points
4. Create datasets
5. Export data

The task analyses below were performed using a registered copy of *DigitizeIt*TM (Version 2.5.3; Bormann, 2020). We selected *DigitizeIt*TM for this demonstration because it is the only program we could identify that includes the capability to automatically plot data points based on their shape (i.e., symbols). Thus, *DigitizeIt*TM potentially significantly reduces the time required to extract data from SCED graphs. We tested other software programs (Biosoft, 2004; Geomatrix, 2021; Rohatgi, 2020; Tummers, 2015) that include automated extraction tools, but they are limited to tracing lines or curves rather than matching symbols.

*DigitizeIt*TM was downloaded from <http://www.digitizeit.xyz/> and installed on a desktop computer running Windows 10. At the time of writing, the unregistered version was available to download and use for evaluation purposes for up to 21 days. The reader should install *Adobe Acrobat*TM on their computer if following the steps that describe importing graphs. We used the free version, *Adobe Acrobat*TM Reader DC, downloaded from <https://get.adobe.com/reader/>. We also used the online version of *Google Sheets*TM (<https://docs.google.com/spreadsheets/>) for the final set of steps regarding exporting data. *Google Sheets*TM is free to use but requires a Google account, which is free to create. The reader may also download or access all three programs following Google searches keywords ‘DigitizeIt’, ‘Adobe Acrobat Reader’, and ‘Google Sheets’.

3.31 Importing the graph into the program

First, the reader should copy the graph image and import it into *DigitizeIt™*. Adobe Acrobat's *snapshot tool* provides a convenient method for copying graph images directly from .pdf files.

1. Open the research article in *Adobe Acrobat™* and navigate to the page that displays the graph panel from which you intend to extract the data. Our example uses the top panel in Figure 1 – ‘Foreign Tact Training’ (Dounavi, 2014; p. 168).
2. Next, click on the EDIT menu, select MORE, then TAKE A SNAPSHOT. Position the cursor at the top-left corner of the graph's image, then press and hold the left mouse button while dragging the cursor to highlight a bounding box around the graph (Figure 2). Ensure that the selection window includes all necessary information, including the x- and y-axes and the graph's key if it has one. After releasing the left mouse button, a dialogue box should open and state that *the selected area has been copied*. At this point, you should click OK, then open and maximize the *DigitizeIt™* program.
3. From within *DigitizeIt™*, select EDIT, then PASTE GRAPH to import the graph into the workspace.

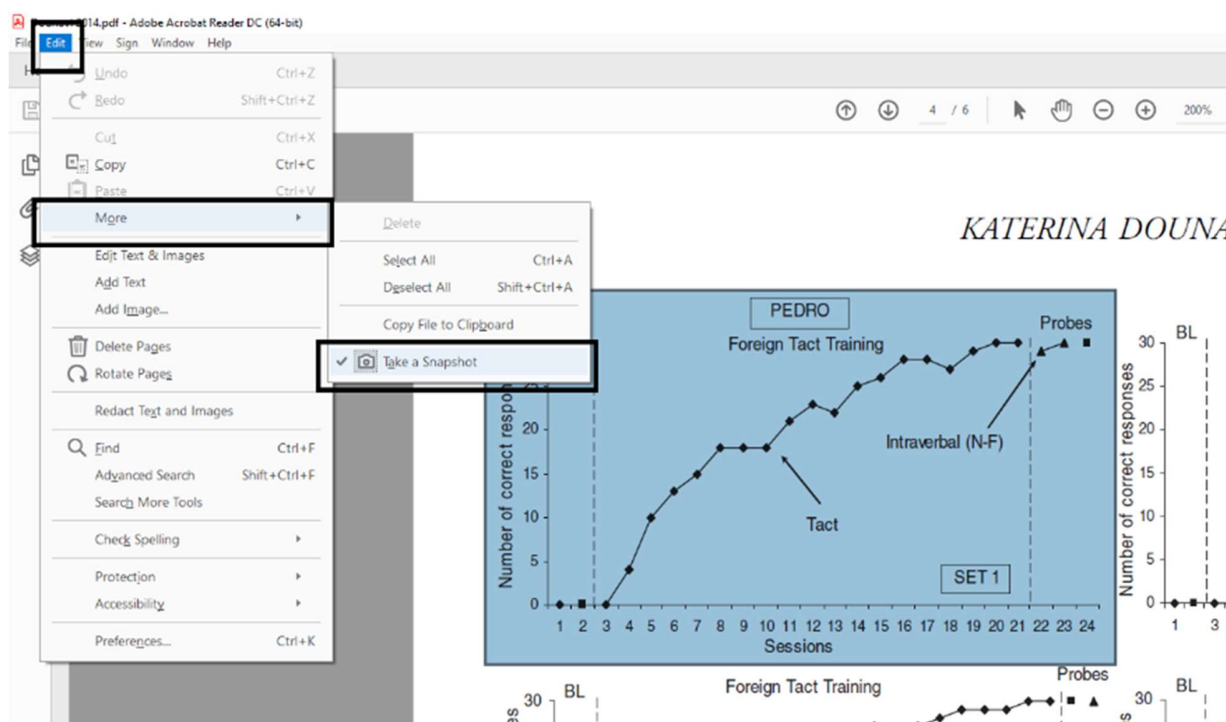
3.32 Defining the XY axes

The following task analysis describes the steps required to calibrate and align the coordinate system with the imported graph's axes. The program requires four coordinates to define the XY axes – x min, x max, y min, and y max. The reader must complete these steps accurately before moving on to plotting data points.

1. Select AXIS, then X MIN, and you should see the cursor change to a crosshair.

Figure 2

Copying graph images directly from pdf files using the SNAPSHOT tool in Adobe Acrobat™



2. Position the crosshair at the lowest labelled point on the graph's x-axis, click the left mouse button, and enter the corresponding x-axis value into the *Axis value* dialogue box that appears. If following our example, click on the center of the first data point and enter 1 as the *x min* value. After you click on OK, a red crosshair should then appear, marking the *x min* position.
3. Now select **AXIS**, then **X MAX**, click on the highest labelled point on the x-axis and enter its value into the *Axis value* dialogue box. In our example, you should click on the midpoint between the final two tick marks along the x-axis and enter 24 as the *x max* value. A horizontal red line will appear along the x-axis connecting the first and second crosshairs after clicking on OK.

4. Repeat the previous two steps by selecting AXIS, followed by Y MIN and then Y MAX to enter the *y min* and *y max* values accordingly. Following along with our example, position and set the *y min* value at 0 (i.e., the origin) and the *y max* value at 30 (i.e., the top of the y-axis; Figure 3) along the y-axis. Once this is done, a vertical red line connecting the *y min* and *y max* points should be visible.

3.33 Plotting data points and creating datasets

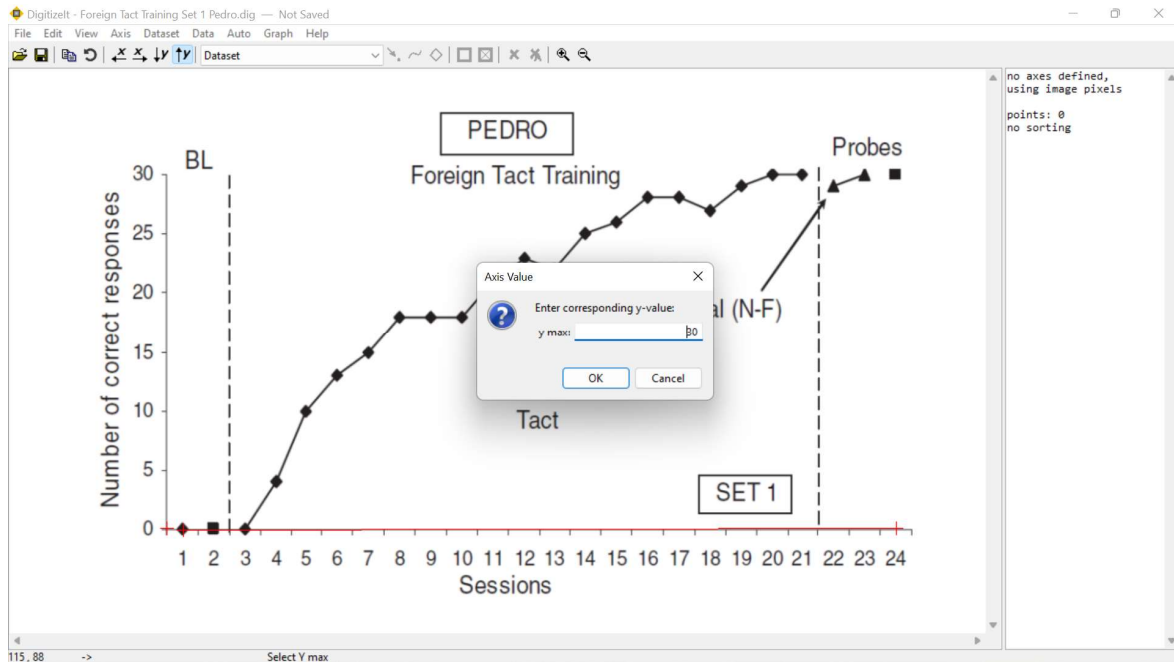
A powerful feature of *DigitizeIt*TM is the ability to automatically match and digitize symbols within graphs. This feature sets *DigitizeIt*TM apart from other data extraction programs that lack this capability. Automatic digitization (i.e., data plotting) can reduce the effort and time needed to extract data from SCED graphs; however, *DigitizeIt*TM sometimes fails to identify and match all symbols. In such cases, *DigitizeIt*TM also includes the capability to plot data points manually. The task analysis below describes both procedures, including steps for adjusting settings and defining search regions to improve automatic digitization.

3.34 Automatic data plotting

1. If your graph includes more than one phase or condition, you may choose to restrict *automatic digitization* to a specified region of the graph by defining a *search region*. To do this, click on AUTO and select SEARCH REGION. Then click and drag a rectangle with the mouse around the data points you intend to include in the search region. Try to exclude any unwanted data points or text. Following our example below, create a search

Figure 3

Defining the XY axes in DigitizeIt™

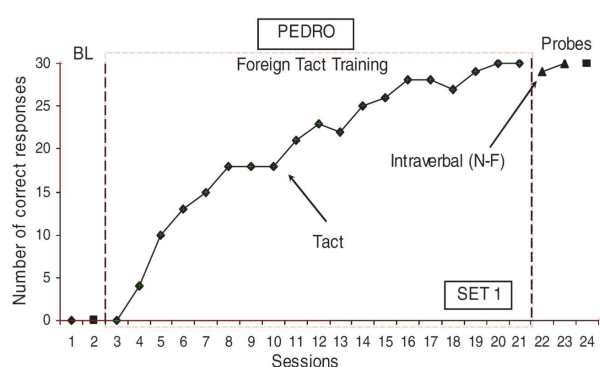


2. region around the middle phase containing sessions 3 – 21 (i.e., ‘Foreign Tact Training phase’). If you wish to include the whole graph, skip this step and proceed to step 2.
3. To automatically digitize the data points, click on AUTO, then select FIND SYMBOLS, and click on one of the symbols that you want to digitize within the graph panel. In the case of our example, click on one of the Tact symbols (i.e., filled black diamond) from within the search region. Now *DigitizeIt™* will try to find all similar symbols and put them into a new *dataset*. If successful, you will see a green crosshair positioned at each symbol’s center (Figure 4). You may also see a popup “New User Tip” asking if you would like to “Change symbol finder parameters”; dismiss these tips as they pop up.

If DigitizeIt™ fails to match all the symbols correctly, try adjusting the similarity settings in the automatic digitizing dialogue box. Before doing this, clear the existing data points by selecting DATASET, then DELETE to prevent the program from digitizing the same data points twice.

Figure 4

Screenshot from DigitizeIt™ showing the search region for the foreign tact training dataset and automatically digitized data points (Reprinted, with permission, from Dounavi, 2014 © John Wiley and Sons)



Then, open the AUTO menu and click on OPTIONS to access the automatic digitizing settings. You can adjust the similarity settings by moving the SYMBOL MATCHING IN% slider control up or down. This setting's value determines how well the selected symbol must match the one you want to digitize to be considered the same. If DigitizeIt™ does not find all the data points you want to digitize, try a lower value. If it finds too many (i.e., other symbols or text), try a higher value. After adjusting the slider, try clicking on one of the symbols again. You may need to make several adjustments; remember to clear the existing data points before each attempt.

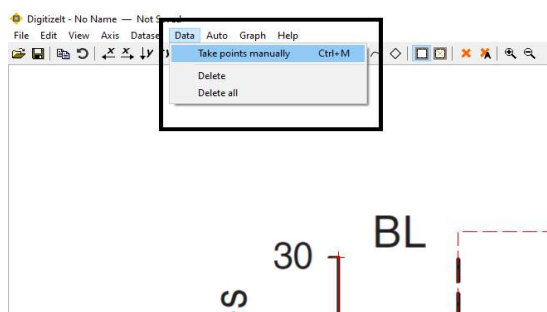
In some cases, DigitizeIt™ appears to have difficulty detecting symbols if the graph's image is too uniform. The software may perform better with 'noisy' images containing slight variations between the symbols. If you have attempted all the above adjustments and DigitizeIt™ does not find any matched symbols, try importing a screenshot of the graph with a lower image resolution. One way to achieve this is by zooming out in Adobe Acrobat™ before taking a snapshot and importing it into DigitizeIt™.

Manual data plotting

1. You can add data points manually if *automatic digitization* fails to find all of them. To do this, click on DATA, and then TAKE POINTS MANUALLY (Figure 5).
2. Then, click on the center of each data point that you want to add to the dataset.
3. To remove unwanted data points, click on DATA and select DELETE. Then, click on each data point you want to delete from the dataset.

Figure 5

Selecting the TAKE POINTS MANUALLY and DELETE points tools in DigitizeIt™



3.35 Creating datasets

1. After digitizing, *rename* the *dataset*, so it is easy to identify later when exporting the data.

To *rename* a *dataset*, click on DATASET, then RENAME, type the name into the *dataset* window that pops up, and click on OK. In our example, we named the first *dataset* “Foreign Tact Training

2. Set 1 Pedro” (Figure 6).

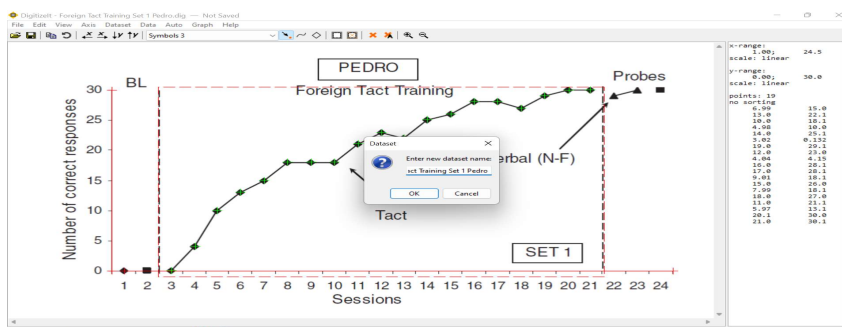
3. Repeat the above steps, digitizing and creating datasets as necessary for each dataset you intend to digitize.
4. To switch between *datasets* in *DigitizeIt™*, click on the dropdown menu on the command ribbon (located to the right of the SET Y MAX button (i.e., the upward arrow and a “y”). At this point, any unwanted datasets should be deleted. For our example, switch to the empty *dataset* named Dataset, then click on DATASET, then DELETE.

3.4 Exporting data

The final set of steps below describes how to export digitized data out of the program. The registered version of *DigitizeIt™* supports export to text via the clipboard or .csv (comma-

Figure 6

Naming a dataset using the RENAME tool in DigitizeIt™



separated values) file, which can be used with most spreadsheet programs, including *Microsoft Excel*TM or *Google Sheets*TM. The following steps demonstrate to the reader how to export the data to a .csv file. Note that the unregistered version of *DigitizeIt*TM does not support data exportation.

Exporting to .csv

1. By default, the digitized data are unsorted. Before exporting, arrange the data, smallest to largest, based on the data points' x-axis values. To do this, select DATASET, then SORT, and ASCENDING.
2. Then select FILE, then EXPORT ALL AS CSV. In the SAVE AS window that opens, enter a file name, choose a location to save the file, and click on SAVE.
3. Next, open *Google Sheets*TM and click on + to create a new blank spreadsheet. Then, open the .csv file you created in *DigitizeIt*TM by clicking on FILE, selecting OPEN, and UPLOAD. In the UPLOAD window, click on SELECT A FILE FROM YOUR DEVICE, and locate the .csv file.
4. After opening the file, you will see that the data appear in scientific notation format. For ease of use, change the format in *Google Sheets*TM from scientific notation to rounded whole numbers (Figure 7). First, click and drag the mouse button to select all the numerical data. Then, click on FORMAT, then NUMBER, and select NUMBER once again. With the numerical data still highlighted, click twice on the DECREASE DECIMAL PLACES button located on the command ribbon to round the data to whole numbers.

Figure 7

Sample showing extracted data points (x and y values) formatted as rounded whole numbers in Google Sheets™

	A	B
1	Foreign Tact Training Set 1 Pedro x	Foreign Tact Training Set 1 Pedro y
2	3	0
3	4	4
4	5	10
5	6	13
6	7	15
7	8	18
8	9	18
9	10	18
10	11	21
11	12	23
12	13	22
13	14	25
14	15	26
15	16	28
16	17	28
17	18	27
18	19	29
19	20	30
20	21	30

3.5 Conclusion

This paper demonstrated how to extract graphical data from published SCED articles using the *DigitizeIt*™ software program (Version 2.5.3; Bormann, 2020). We acknowledge that several other data extraction programs are available to the reader to perform these tasks; however, we believe *DigitizeIt*™ to be the most efficient due to its automatic digitizing tools. After extracting the data, the reader may conduct statistical analyses or graph and reanalyze it using their preferred graphing software. Readers are encouraged to apply the above procedures to examine their own practice or research questions through reanalysis or metanalysis of empirical data from the research literature.

Chapter 4: A Systematic Review of Emergent Learning Outcomes Produced by Foreign-Language Tact Training

Peer reviewed paper: Wooderson, J. R., Bizo, L. A., & Young, K. (2022). A systematic review of emergent learning outcomes produced by foreign language tact training. *The Analysis of Verbal Behavior*, 38, 157–178. <https://doi.org/10.1007/s40616-022-00170-z>⁴

4.1 Abstract

This systematic review evaluated the effects of foreign tact training on emergent learning outcomes in 10 published studies. We also conducted a meta-analysis of aggregate data from seven studies comparing outcomes of foreign tact training with other verbal operant procedures. The preliminary findings indicated foreign tact training produced criterion-level responses in 84 of 106 (79.2%) post-test probes across 37 learners and 55 evaluations of foreign tact training. The meta-analysis results revealed significantly higher within-subjects mean levels of emergent responding following foreign tact training than foreign-to-native intraverbal, native-to-foreign intraverbal, and foreign listener training. Emergent outcomes for adults were not significantly greater than for children. Finally, foreign tact training was slightly more efficient than the other verbal operant procedures, although most of the differences were not statistically significant.

Keywords: emergent learning, foreign language learning, second language learning, tact training

4.2 Introduction

Learning a foreign language is thought to provide a range of cognitive (Antoniou et al., 2013; Cheng et al., 2019), emotional (Klimova et al., 2021) and financial (New American Economy, 2017) benefits. Learning a foreign language may be costly and time-consuming, with

⁴ The paper was published in an American journal and uses American English spelling throughout.

some languages requiring at least 2200 hours (88 weeks) of study to develop fluent performance (U.S. Department of State, n.d.). Furthermore, some programs use considerable education resources. South Korea, for example, spent 40% (12 billion dollars) of its public education budget on English language programs in 2009, and private education costs were estimated to be even higher (Piller, 2016). In the European Union, up to 95% of students in upper secondary education study a foreign language (European Commission, 2020). Given the potential cost of foreign language study, educators must optimize learning by making instruction efficient. In this regard, behavior analysis has much to offer as the field's history is replete with empirical demonstrations of evidence-based instructional procedures (Binder & Watkins, 1990; Vargas, 2020). This review examines foreign tact training (FTT)—a promising behavior-analytic procedure for efficient foreign language learning.

Traditional language theories view verbal operants, such as speaking and listening behaviors, as innately interdependent (e.g., Chomsky, 1957; Kuhl, 2004). However, various behavior-analytic accounts contend that these operants are initially independent but may become 'joined' through repeated incidental experiences, modeling, and direct reinforcement (Greer & Speckman, 2009). Furthermore, the learner's integration of these capabilities represents a generalized verbal operant that allows for potentially unlimited patterns of emergent responding and generalized language development. Three main theories—stimulus equivalence (e.g., Sidman, 1971), naming theory (e.g., Horne & Lowe, 1996), and relational frame theory (RFT, e.g., Barnes-Holmes et al., 2018)—have been developed to understand the conditions that occasion derived stimulus relations and emergent learning (Critchfield et al., 2018; Lafrance & Tarbox, 2020; Rehfeldt, 2011). Sidman's (1971) influential study on stimulus equivalence discovered that untrained relations could emerge following the teaching of certain stimulus-

response relations. Relational frame theory further builds upon stimulus equivalence by conceptualizing equivalence and other stimulus relations as classes of generalized relational operants, which are referred to within RFT as relational frames. Engaging in relational responding is occasioned by contextual cues that function as discriminative stimuli for previously established patterns of relational responding (Barnes-Holmes et al., 2018). According to RFT proponents, learners develop relational frames due to a reinforcement history of relational exemplars. In naming theory (Horne & Lowe, 1996), naming refers to the learner's combination of speaker and listener behaviors. These three theories have generated extensive research and a broad range of empirically validated language development and learning procedures.

A growing field of study has emerged in the behavior-analytic literature examining the efficacy of behavior-analytic based procedures for foreign language learning. This literature applies emergent learning practices and verbal operants to foreign vocabulary training (Daly & Dounavi, 2020). The languages taught include Native American (Haegele et al., 2011), Japanese (Petursdottir et al., 2014), French (Daly & Dounavi, 2020; Polson et al., 1997; Polson & Parsons, 2000), German (Rocha e Silva & Ferster, 1966), Spanish (Joyce & Joyce, 1993; Matter et al., 2020; Petursdottir et al., 2008; Ramirez et al., 2009), Italian (Petursdottir & Haflíðadóttir, 2009), Chinese (Wu et al., 2019), Welsh (May et al., 2019; May et al., 2016), and English (Cortez et al., 2020, 2021; Dounavi, 2011, 2014; Rosales et al., 2011, 2012). This literature's defining feature is its focus on emergent learning as a critical outcome of effective foreign language instruction.

Behavior analysts value emergent learning because it represents what might be characterized as 'free' knowledge or skills that do not require direct experience (e.g., Critchfield et al., 2018; Critchfield & Twyman, 2014). However, if learning goals are limited to what may

only be explicitly taught, then the scope and breadth of outcomes are also limited (Critchfield, 2018). Instead, the instructor expects untrained operants to emerge following a carefully selected subset of learning content (Critchfield, 2018; Dixon & Stanley, 2020). Studies in this field have implemented training procedures involving a range of verbal operants, including listener behavior (e.g., Rocha e Silva & Ferster, 1966), echoics (e.g., Petursdottir et al., 2014), mands (e.g., Wu et al., 2019), native-to-foreign intraverbals (NFI; e.g., Petursdottir & Haflíðadóttir, 2009), foreign-to-native intraverbals (FNI; e.g., Polson & Parsons, 2000), and tacts (e.g., Petursdottir et al., 2008) and tested for the emergence of untrained verbal operants.

Skinner (1957, p. 83) considered the tact the most important verbal operant because mands, intraverbals, and listener relations often depend on the learner's ability to reference a wide range of environmental stimuli (Sundberg, 2015). Consequently, a strong tact repertoire is vital to social and academic success (Bak et al., 2021; Lalonde et al., 2020). Foreign tact training involves teaching learners to tact environmental stimuli using appropriate foreign language referents. Following FTT, learners may acquire several untrained relations, including listener responses, intraverbals, and mands in addition to the trained tacts. Among the various teaching procedures, FTT may be the most productive; several studies have noted its superior efficiency (e.g., Cortez et al., 2020; Cortez et al., 2022; Daly & Dounavi, 2020; Dounavi, 2011; Matter et al., 2020). In emergent learning, efficiency means the amount of and ease with which learners acquire the trained and untrained material (Dounavi, 2011).

Recently, Matter et al. (2020) showed that FTT alone was more efficient than a traditional multi-component procedure comprising four verbal operants (tact, FNI, NFI, and listener training). Using an adapted alternating treatment design, the authors provided Spanish-language training to four English-speaking children. The results showed FTT required fewer sessions to

mastery than the multi-component procedure and resulted in almost all learners acquiring emergent receptive and productive relations despite not receiving any training in the FNI, NFI, and listener relations. In addition, FTT produced more efficient emergent FNI and NFI responses than listener training with Portuguese-speaking Brazilian children learning English in studies by Cortez et al. (2020, 2021). However, the authors noted FNI and NFI relations did not always emerge at comparable levels. Dounavi (2011, 2014) conducted two methodologically similar studies with adult native-Spanish speakers. Both studies compared FTT with FNI training and NFI training. In the earlier study (Dounavi, 2011), FTT achieved higher levels of emergent responding and required fewer training trials than FNI or NFI training for both participants. In Dounavi (2014), on the other hand, NFI relations took fewer trials to achieve mastery criterion than the foreign tact relations; so NFI training was the most efficient condition. When Daly and Dounavi (2020) systematically replicated and extended Dounavi (2014), they used a modified concurrent multiple probe design to improve internal validity. Their results were comparable with Dounavi (2014); FTT produced more emergent responses than FNI or NFI training. However, FTT needed fewer trials to criterion. Furthermore, probes at four weeks post-training showed better maintenance of emergent responses following FTT than the two intraverbal conditions.

Foreign tact training is not successful for all learners, though. For example, Wu et al. (2019) compared the effects of FTT, FNI training, NFI training, and mand training in Mandarin Chinese vocabulary. They found FTT was the most efficient procedure for only one of the four participants. Also, May et al. (2019) reported equivocal results—robust increases in derived intraverbal relations after FTT for only half of the children in their study. Some researchers (e.g., Daly & Dounavi, 2020; Dounavi, 2014; Petursdottir & Haflíðadóttir, 2009) suggest young

children are less likely to produce emergent responses because they are less verbally competent than adults. However, we could find no studies directly comparing emergent foreign language learning outcomes between adults and children.

In summary, emergent learning and FTT offer considerable potential for optimizing foreign language programs. However, it is difficult to determine the generality of FTT outcomes as the available research is limited to single-case experimental studies. Thus, it is unclear whether FTT is more efficient than other verbal operant training procedures at the group level analysis. This paper aimed to extract and analyze aggregate data from the literature on FTT use. In doing this, we considered the following three questions. First, what are the effects of FTT on emergent learning outcomes in the published literature to date? Second, how do FTT acquisition, emergence, and overall efficiency compare with other verbal operant training procedures? Finally, does FTT produce higher levels of emergent responding for adults or children?

4.3 Method

4.31 Literature search procedure

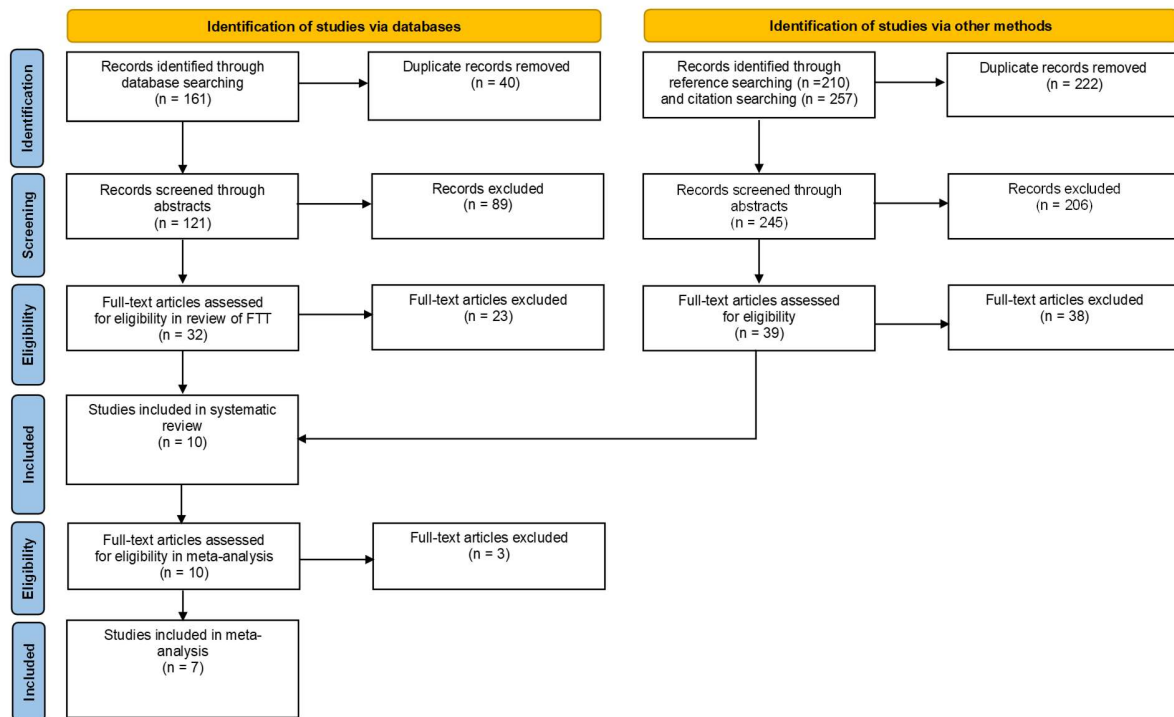
The search included APA PsycINFO (EBSCOhost), Medline (EBSCOhost), ERIC (EBSCOhost), CINAHL (EBSCOhost), APA PsycArticles (EBSCOhost), Psychology and Behavioral Sciences Collection (EBSCOhost), SocINDEX (EBSCOhost), and Web of Science electronic databases for English language studies published in peer-reviewed journals, with no limit specified regarding the year of publication. In addition, we combined various keyword terms related to emergent learning (emerg*, derive*, equivalenc*, generative*), foreign language learning (foreign language, second language), and verbal operant training (mand, tact, intraverbal, echoic, textual, dictation, autoclitic, verbal behavi*, verbal operant, match-to-

sample, conditional discrimination, multiple exemplar). Finally, the wildcard * expanded the search to include all variants of the keywords.

The search sequence (Figure 1) initially identified 161 articles—121 after removing duplicates. The first author then reviewed the abstracts of all 121 unique articles and removed all non-English articles, non-empirical papers (review, policy, position, commentary, or conceptual articles), and studies focused on language use (linguistic, diagnostic, textual, historical, cultural, psychometric, phonological, content, orthographic, or discourse analyses). We screened the remaining 31 full-text articles for three inclusion criteria: The experimenters focused on observable and measurable foreign language targets; the experiment included at least one standalone FTT procedure; the procedures involved at least one pre- and post-test for untrained emergent relations. We excluded studies with native-, contrived-, artificial-, nonsense-, or non-language learning targets and studies that combined FTT with other verbal operant procedures. We allowed, however, studies with native-tact pre-training trials—checks to see whether learners could tact the stimuli in their native language.

Figure 1

PRISMA chart showing systematic literature search sequence (Page et al., 2021)



After initial full-text eligibility screening, we identified nine articles that met the criteria (Cortez et al., 2020; Cortez et al., 2022; Daly & Dounavi, 2020; Dounavi, 2014; Matter et al., 2020; May et al., 2019; Petursdottir & Hafliðadóttir, 2009; Petursdottir et al., 2008; Wu et al., 2019). We then conducted reference and citation searches using Google Scholar and Web of Science. These searches returned a further 245 potential papers, which we also assessed for eligibility—yielding one additional article (Dounavi, 2011). In total, 10 articles were included that contained 55 distinct evaluations of FTT. We excluded Matter et al.’s (2020) ‘mixed’ training evaluation from our sample because these trials combined FTT, NFI, FNI, and listener training. Furthermore, ‘mixed’ training post-tests in Matter et al. (2020) evaluated directly trained relations only.

We also evaluated a subset (seven) of the 10 FTT studies through meta-analysis (Cortez et al., 2020; Cortez et al., 2022; Daly & Dounavi, 2020; Dounavi, 2011, 2014; Petursdottir & Haflíðadóttir, 2009; Wu et al., 2019). We only included studies in the meta-analysis if they contained at least one within-subject evaluation comparing the emergent learning outcomes produced by FTT with at least one other verbal operant training condition. Consequently, we excluded Petursdottir et al. (2008) from the meta-analysis because it did not contain any within-subject evaluations of training conditions. In addition, we excluded Matter et al. (2020) and May et al. (2019) because neither study compared the emergent learning outcomes following FTT with those produced by other verbal operant training procedures. As noted above, Matter et al. (2020) taught all target relations in the ‘mixed’ training condition directly, meaning they could only test for emergent relations following FTT; May et al. (2019) implemented FTT only.

4.32 Data categorization

The 10 articles were categorized according to participant demographic data (age, gender, native language, setting), target foreign language, types of training conditions employed, and mastery criteria for instructional and emergent learning outcomes. Each of the 55 FTT evaluations was coded according to whether it produced criterion-level responses in post-training probes. All experiments probed two or more distinct types of emergent relations; we evaluated each relation separately, where appropriate. The post-test results for each emergent relation were categorized as either achieving or not achieving criterion levels. If studies stated no specific mastery criteria, we set a criterion of 100%. Some evaluations included more than one post-test probe per emergent relation—we only included the highest post-test score recorded for each relation.

We evaluated the quality of each study using criteria as recommended by Schlosser and Sigafoos (2007): experimental design; follow-up data collected after three months, at minimum, for at least 90% of the participants; appropriate and independently assessed reliability measures; and counterbalancing or random allocation of stimuli to training conditions. We also evaluated the studies against the Council for Exceptional Children (CEC, 2014) quality standards for evidence-based practices. The CEC standards include 22 indicators for assessing the quality of single-case experimental studies, which can be used to determine whether an instructional procedure qualifies as an evidence-based practice.

4.33 Data extraction for meta-analysis

The meta-analysis evaluated training acquisition rates, emergent post-test scores, and the overall efficiency of each verbal operant training procedure. The first author extracted data from the seven papers' graphs and tables using DigitizeIt (Bormann, 2020). Concurrently, we emailed the corresponding author of each study once and requested the training and post-test data to conduct our analyses. We received written responses from six authors—one of whom stated they had not retained the data, and another noted the data were not immediately available. We did not receive a response from one author. Our requests resulted in raw data for four papers (Cortez et al., 2020; Daly & Dounavi, 2020; Dounavi, 2011, 2014). We did not send any follow-up requests; rather, we utilized the software-extracted data only for the remaining three papers.

Following data extraction, we regraphed the acquisition curves from each study on standardized panels and compared acquisition rates using descriptive visual analysis methods. Then, we calculated standardized acquisition rates (SAR), which represent the average number of training trials needed per word learned. To calculate SARs, we multiplied the number of trial blocks by the number of trials per block and divided by the number of items trained and the

terminal percent correct; $SAR = (\text{number of trial blocks} * \text{number of trials per block}) / (\text{number of items per training set}) / (\text{terminal\% correct}) * 100$). The smaller the resulting value, the better the SAR. By including ‘terminal% correct’ in the calculation, we could weight scores and compare training evaluations with different mastery criteria. Furthermore, we could compare training evaluations that researchers discontinued before the learner reached the mastery criterion.

We then compared FTT emergent post-test results with FNI, NFI, listener, and mand training post-tests. We did this by converting all post-test scores to percentages and calculating mean scores for each training evaluation within and across each study. Then, we conducted within-subjects statistical analyses using mean post-test scores for each learner and each training condition in which they participated. The analyses comprised Wilcoxon signed-rank non-parametric dependent-samples tests conducted in Jamovi (The jamovi project, 2020). We included all post-tests for emergent tact, FNI, NFI, and mand relations but excluded all tests for emergent listener relations (six scores) from the analysis due to the potential confounds of comparing unbounded scores with scores bounded by chance (Petursdottir & Haflíðadóttir, 2009). Three studies implemented reverse intraverbal training with participants following initial post-test probes (Daly & Dounavi, 2020; Dounavi, 2011, 2014). Consequently, we only included post-test data from the initial training sequence to control confounds associated with potential sequencing effects. We also used a Mann-Whitney U non-parametric independent-samples test to compare children’s mean FTT post-test scores (under 18 years) and adults (18 years and older). Lastly, we evaluated the overall efficiency of each verbal operant procedure by calculating an efficiency index score (EIS) using the SAR and mean post-test scores described above; $EIS = \text{mean post-test} / SAR$. The larger the resulting value, the better the EIS. We then analyzed the EIS data using Wilcoxon signed-rank non-parametric dependent-samples tests.

4.34 Interobserver agreement

The first author and an independent rater (BCBA-D®) read the full text of 31 articles and evaluated their eligibility based on the inclusion and exclusion criteria. The mean agreement was 100%. The first author also compared the data from four articles (Cortez et al., 2020; Daly & Dounavi, 2020; Dounavi, 2011, 2014), extracted using DigitizeIt, to the raw data provided by the authors. In total, we evaluated 88 (57.1%) post-test scores and 688 (80.0%) training trial scores. The mean agreement was 100%.

4.4 Results

4.41 Participant demographics

Table 1 summarizes the demographic data, types of emergent relations tested, and mastery criteria (if any) stated by the authors. Across the 10 studies, 26 participants were children, and 11 were adults. Eight studies reported data on individual participant age; the mean participant age across 27 children and adults was 15.8 years (range: 4 – 40 years). The mean age of children ($n = 16$) was 5.0 years (range: 4 – 6 years), and adults ($n = 11$) was 31.5 years (range: 23 – 40 years). The remaining two studies reported the range of participants' ages only ($n = 10$; range: 7-9 years). Just five studies directly reported participant gender, including six females and nine males. Participants' native language was reported as English ($n = 17$), Portuguese ($n = 10$), Spanish ($n = 4$), or Icelandic ($n = 6$). The studies occurred in various settings, with the highest number ($n = 4$) conducted in learners' homes.

Table 1 Data extraction of studies included in the systematic review

Paper	Participants	Setting and target foreign language	Instructional conditions	Types of emergent relations probed post-FTT	Instructional mastery criteria	Mastery criteria for emergent post-test probes	Number of criterion level emergent relations for FTT	Number of non-criterion level emergent relations for FTT
Cortez et al. (2020)	Six Portuguese-speaking children. Age range 7–9 years	Lab room English	FTT, listener behavior training	FNI, NFI	100% correct responses in three consecutive trial blocks	None stated	11* (5 NFI, 6 FNI)	1* (NFI)
Cortez et al. (2021)	Four Portuguese-speaking children. Age range 7–9 years	Lab room English	FTT, listener behavior training	FNI, NFI	100% correct responses in three consecutive trial blocks	None stated	7* (4 NFI, 3 FNI)	1* (FNI)
Daly and Dounavi (2020)	Three English speaking adults – 31M, 33F and 40F	Home French	FTT, intraverbal training (FNI, NFI)	FNI, NFI	100% correct responses in two consecutive trial blocks	10 out of 10 correct responses in one probe. A second probe session was conducted if a participant scored less than 10 but at least 7	6 (3 NFI, 3 FNI)	0

Table 1 (continued)

Paper	Participants	Setting and target foreign language	Instructional conditions	Types of emergent relations probed post-FTT	Instructional mastery criteria	Mastery criteria for emergent post-test probes	Number of criterion level emergent relations for FTT	Number of non-criterion level emergent relations for FTT
Dounavi (2011)	Two adult native Spanish speakers – 36M and 39M	Home English	FTT, intraverbal training (FNI, NFI)	FNI, NFI	100% correct responses in two consecutive trial blocks	30 out of 30 correct responses in one probe A second probe session was conducted if a participant scored less than 30 but at least 27	8 (4 NFI, 4 FNI)	0
Dounavi (2014)	Two adult native Spanish speakers – 37M and 29F	Home English	FTT, intraverbal training (FNI, NFI)	FNI, NFI	100% correct responses in two consecutive trial blocks	30 out of 30 correct responses in one probe A second probe session was conducted if a participant scored less than 30 but at least 27	8 (4 NFI, 4 FNI)	0
Matter et al. (2020)	Four 4-year-old English speaking children	School Spanish	FTT, mixed training (FTT, FNI, NFI, listener behavior)	Listener, FNI, NFI (only following FTT)	83.3% correct responses for two consecutive trial blocks	10 out of 12 correct responses in one probe	16 (4 NFI, 5 FNI, 7 Listener)	5 (3 NFI, 2 FNI)

Table 1 (continued)

Paper	Participants	Setting and target foreign language	Instructional conditions	Types of emergent relations probed post-FTT	Instructional mastery criteria	Mastery criteria for emergent post-test probes	Number of criterion level emergent relations for FTT	Number of non-criterion level emergent relations for FTT
May et al. (2019)	Six English speaking children (two 5-year-olds, four 6-year-olds)	School Welsh	FTT (group choral) only	FNI+NFI (mixed intra-verbal trials)	89% correct responses for one trial block	None stated	10* (mixed intraverbals - NFI+FNI)	7* (7 mixed intraverbals - NFI+FNI)
Petursdottir and Hafþíðdóttir (2009)	Two 5-year-old native-Icelandic speakers	Pre-school Italian	FTT, intraverbal training (FNI, NFI), listener behavior training	Listener, FNI, NFI	83.3% correct responses for two consecutive trial blocks	10 out of 12 correct responses in one probe	3 (1 NFI, 2 Listener)	3 (1 NFI, 2 FNI)
Petursdottir et al. (2008)	Four 5-year-old native-Icelandic speakers	Pre-school Spanish	FTT, listener	FNI, NFI	100% correct responses in three consecutive trial blocks	None stated	7* (4 NFI, 3 FNI)	1* (FNI)
Wu et al. (2019)	Four native English-speaking adults 26M, 26M, 26M, 23F	University-based clinic and home Mandarin Chinese	FTT, intraverbal training (FNI, NFI), mand training	Mand, NFI, FNI	83.3% correct responses for two consecutive trial blocks	None stated	8* (2 NFI, 4 FNI, 2 Mand)	4* (2 NFI, 2 Mand)

FTT, foreign tact training, NFI, native-to-foreign intraverbal, FNI, foreign-to-native intraverbal, M, male, F, female.

*Denotes studies that did not state mastery criteria for emergent relations—for which, we set a criterion of 100%

4.42 Target foreign languages, training conditions, mastery criteria, and emergent learning relations

The 10 studies targeted six foreign languages— four trained English vocabulary to non-English speaking learners (Cortez et al., 2020; Cortez et al., 2022; Dounavi, 2011, 2014). Other than the one study that focused on Mandarin Chinese words (Wu et al., 2019), all target foreign languages were European: English (n = 4), Spanish (n = 2), French (n = 1), Italian (n = 1), and Welsh (n = 1).

In addition to FTT, studies included a range of verbal operant training procedures: FNI, NFI, listener behavior, and mands training. Instructional mastery criteria ranged from 83.3% correct responses across two consecutive sessions (Matter et al., 2020; Petursdottir & Haflíðadóttir, 2009; Wu et al., 2019) to 100% correct responses across three consecutive sessions (Cortez et al., 2020; Cortez et al., 2022; Petursdottir et al., 2008). Additionally, FTT studies tested a range of untrained relations: FNI, NFI, listener, and mands. Notably, all 10 studies tested for emergent intraverbal (FNI and NFI) relations post FTT. Although only five studies stated specific mastery criteria for emergent relations, the reported standards varied from 83.3% (Matter et al., 2020; Petursdottir & Haflíðadóttir, 2009) to 100% correct (Daly & Dounavi, 2020; Dounavi, 2011, 2014). Five studies did not specify any mastery criteria for emergent relations; in which case, we set a conservative criterion of 100% correct (Cortez et al., 2020; Cortez et al., 2022; May et al., 2019; Petursdottir et al., 2008; Wu et al., 2019)

4.43 FTT's emergent learning outcomes

Table 1 also shows FTT's emergent learning outcomes in each of the 10 studies (55 FTT evaluations). In total, 84 (79.2%) post-test probes scored at or above criterion level responding,

and 22 (20.8%) scored below. Overall, FNI relations (84.2%) emerged at mastery criterion levels slightly more often than NFI (81.6%). Furthermore, FTT produced criterion-level emergent listener relations for all six learners in the two studies with listener probes (Matter et al., 2020; Petursdottir & Haflíðadóttir, 2009). However, chance-level responding for listener probes was 33% (Petursdottir & Haflíðadóttir, 2009). In contrast, FTT produced criterion-level mand relations in only 50% of probes, although only one study included tests for emergent foreign mands (Wu et al., 2019).

Table 2 shows that the studies achieved most of the quality standards recommended by Schlosser and Sigafos (2007), except for follow-up data and experimental design. For example, although Matter et al. (2020) included long-term follow-up data beyond three months post-training, they only conducted sessions with two of the four participants. Additionally, most studies employed robust experimental designs to evaluate the effects of training procedures on the trained relations, but only five studies used control conditions or multiple-baseline designs when evaluating emergent relations (Matter et al., 2020; May et al., 2019; Petursdottir and Haflíðadóttir, 2009; Petursdottir et al., 2008; Wu et al., 2019). Three of the 10 studies met all 22 CEC quality indicators (Matter et al., 2020; May et al., 2019; Wu et al., 2019). Most studies that did not meet all 22 quality indicators failed to include an evaluation of treatment integrity (Cortez et al., 2020; Daly and Dounavi, 2020; Dounavi, 2011; Dounavi, 2014; Petursdottir and Haflíðadóttir, 2009; Petursdottir et al., 2008). Other reasons studies fell short of the CEC standards included not having at least three data points in post-test phases or robust controls for threats to internal validity (e.g., control conditions or multiple-baseline designs). Based on these results and the CEC (2014) standards, the review's findings indicate FTT is a potentially evidence-based practice.

Table 2 Analysis summary of key research paper methodological design and analytic elements

Paper	Experimental design (trained relations)	Experimental design (emergent relations)	Follow-up data	Reliability measures	Counterbalancing or random allocation of stimuli to conditions	CEC standards outcomes (total 22)
Cortez et al. (2020)	AATD	Pre- post-test probes	No	Yes	No	17
Cortez et al. (2021)	AATD	Pre- post-test probes	Yes, at two weeks or one month for two of the four participants	Yes	Yes	20
Daly and Dounavi (2020)	Modified MBL probe	Pre- post-test probes	Yes, at four weeks post- training	Yes	Yes	18
Dounavi (2011)	MBL across participants and stimulus sets	Pre- post-test probes	No	Yes	No	16
Dounavi (2014)	MBL across participants and stimulus sets	Pre- post-test probes	No	Yes	Yes	16
Matter et al. (2020)	AATD embedded within MBL and a control condition	Pre- post-test probes with control conditions	Yes, at 2- and 4-months post-training for two of the four participants	Yes	Yes	22
May et al. (2019)	Concurrent multiple- probe design across stimulus sets	Concurrent multiple- probe design across stimulus sets	Yes, at two weeks follow- ing post-testing for each participant stimulus set	Yes	No	22
Petursdottir and Hafþadóttir (2009)	MBL across participants with an embedded AATD	Pre- post-test probes with control conditions	No	Yes	No	19
Petursdottir et al. (2008)	MBL across stimulus sets	MBL design across stimulus sets	No	Yes	Yes	19
Wu et al. (2019)	MBL across participants with an embedded AATD	Pre- post-test probes with control conditions	No	Yes	Yes	22

AATD, adapted alternating treatments design, MBL, multiple baseline.

Based on criteria as recommended by Schlosser and Sigafos (2007) and Council for Exceptional Children (2014)

4.44 Meta-analysis

Visual analysis (available in the supplemental materials at the end of this chapter) did not reveal consistent differences in acquisition curves. In other words, some participants acquired foreign tacts faster than other relations, but not all. Table 3 shows the mean acquisition rates (SARs) for the studies in the meta-analysis. Foreign tact training produced the lowest SAR within just one of the seven studies (Dounavi, 2011)—most participants in this study acquired trained foreign tacts faster than FNI responses. On the other hand, FTT produced the highest SAR in two studies (Cortez et al., 2020; Dounavi, 2014), which meant that participants generally acquired foreign tacts slower than the listener, FNI, or NFI relations. The SAR for FTT was neither the lowest nor the highest in four studies (Cortez et al., 2022; Daly & Dounavi, 2020; Petursdottir & Haflíðadóttir, 2009; Wu et al., 2019). For example, all three Daly and Dounavi (2020) participants acquired the trained foreign tact relations in fewer trials than FNI relations. Still, only one participant acquired foreign tact relations in fewer trials than NFI relations. Wu et al.'s (2019) mand and FNI training conditions produced lower SARs than FTT; however, the SAR for FTT was superior to that of NFI training. Overall, mand training (18.8) produced the lowest average SAR, followed by FNI training (20.2), FTT (22.1), NFI training (22.6), and listener training (23.8).

Table 3 Average standard acquisition rate (SAR) (lower value is better) and post-test scores (higher value is better) for each training condition

Study		FTT (n = 23)	NFI (n = 13)	Listener (n = 12)	FNI (n = 13)	Mand (n = 4)
Cortez et al. (2020)	SAR	26.3		24.3		
	Post-test	98.6%		55.6%		
Cortez et al. (2021)	SAR	28		28		
	Post-test	97.9%		72.9%		
Daly and Dounavi (2020)	SAR	13	10		15.7	
	Post-test	98.6%	94.4%		70.0%	
Dounavi (2011)	SAR	12.3	17.5		17	
	Post-test	99%	81.7%		63.3%	
Dounavi (2014)	SAR	21.8	18.5		17.5	
	Post-test	99.3%	96.7%		64.2%	
Petursdottir and Hafidadóttir (2009) *	SAR	27.5	43.1	14	30	
	Post-test	58.3% (70.8%*)	58.3% (69.4%*)	58.3%	14.6% (37.5%*)	
Wu et al. (2019)	SAR	24	26.3		21.8	18.8
	Post-test	85.3%	62.1%		54.7%	82.4%
Overall average SAR	Mean (SD)	22.1 (9.8)	22.6 (14.7)	23.8 (12.0)	20.2 (8.2)	18.8 (6.2)
Overall average post-test score	Mean (SD)	93.6% (16.3), 93.7%*	79.0% (31.1), 79.6%*	61.5 (21.7)	55.8% (27.5), 57.4%*	82.4% (26.2)

FTT, foreign tact training, NFI, native-to-foreign intraverbal, FNI, foreign-to-native intraverbal.

*Denotes scores with emergent listener post-tests included in the dataset

Foreign tact training achieved the highest mean post-test scores in all seven studies (Table 3). The within-subjects tests (Table 4) revealed participants' mean FTT post-test scores were significantly higher than NFI, FNI, and listener training. FTT produced slightly higher mean scores than mand training, but the difference was not statistically significant. However, Wu et al. (2019) conducted mand training with the item to be requested in view of the participant, meaning it was a combination of foreign mand and tact relations under convergent multiple control (Michael et al., 2011). The Mann-Whitney U independent samples t-test indicated no significant differences in mean FTT post-test scores for children ($Mdn = 100$) and adult participants ($Mdn = 100$), $U = 417$, $p = .203$.

Table 4

Within-Subjects Comparisons Between Post-Test Scores for Foreign Tact Training (FTT) and All Other Conditions

Training condition comparisons	n	Mean difference	Wilcoxon W statistic	p-value	Effect size (Rank biserial correlation)
FTT > NFI	13	8.9%	81	*0.007	0.780
FTT > Listener	12	29.5%	75	*0.003	0.923
FTT > FNI	13	35.2%	78	*0.001	1.00
FTT > Mand	4	3.71%	4	0.395	0.333

Note. NFI = native-to-foreign intraverbal, FNI = foreign-to-native intraverbal.

* Denotes a statistically significant result

Foreign tact training produced the highest average EIS (5.1), followed by NFI training (4.9), mand training (4.7), listener training (3.3), and FNI training (3.1). Statistical analysis revealed no significant differences between FTT and the other training conditions—except FNI training ($W = 77, p = .002$).

4.5 Discussion

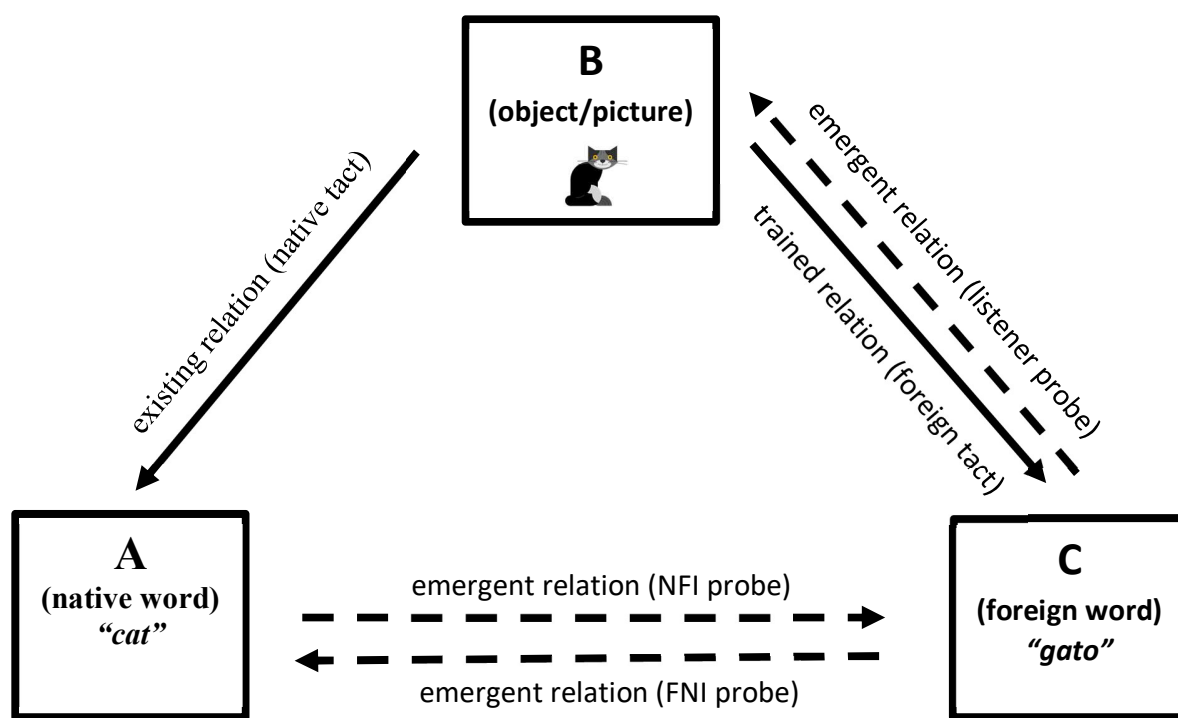
This review adds to the growing literature on emergent foreign language learning. We found FTT produced high levels of emergent verbal relations for most participants. An explanation for the emergence of untaught NFI relations is that they share common stimulus and response topographies (covert native word and overt foreign word) with trained foreign tact relations (Petursdottir et al., 2008). According to naming theory, FTT stimuli are likely to evoke covert native responses and overt foreign vocal responses in verbally competent learners. It is difficult to determine if this occurred, as covert vocalizations are private events. Also, no authors reported learners' overt native tacts during FTT.

An alternative explanation (Figure 2) is FTT learners derived equivalence relations between the native word, the object/picture, and the foreign word (Daly & Dounavi, 2020; May et al., 2013). Stimulus equivalence theory states that when stimulus A (native word) is related to B (object/picture), and B (object/picture) to C (foreign word), several relations may emerge without further training (Sidman, 2018). In all but one study (Cortez et al., 2020), experimenters ensured that participants could tact each target in their language either before (Cortez et al., 2022, Daly & Dounavi, 2020; Dounavi, 2011, 2014; Matter et al., 2020; Petursdottir & Haflíðadóttir, 2009; Petursdottir et al., 2008; Wu et al., 2019) or during training (May et al., 2016). Therefore, participants could relate stimulus B (object/picture) to A (native word) and B (object/picture) to C (foreign word) following FTT. Experimenters then tested participants'

emergent responses, demonstrating a range of equivalence relations: NFI probes tested for the emergence of untrained A–C equivalence relations; FNI probes tested C–A, and listener probes C–B. Then, the contextual cues that likely occasioned participants’ derived equivalence responses were the experimenters’ vocal stimuli— “What is the Spanish word for cat?”, “How do you say Gato in English?”, “Point to Gato”, “What do you call this in Spanish?” Pure mands, on the other hand, are evoked by motivating operations, not discriminative stimuli (Skinner, 1957)—the mands in Wu et al. (2019) were multiply controlled and probably tested B (object/picture) to C (foreign word) relations. However, the authors did not provide contextual cues consistently between FTT trials and mand post-tests, which may have caused the low levels of foreign manding following FTT. It is also possible that the tacts Wu et al. (2019) taught during FTT failed to emerge as mands because the tact training stimuli did not function as reinforcers (Wallace et al., 2006).

Figure 2

Existing, trained, and emergent relations following foreign tact training



Note. NFI = native-to-foreign intraverbal, FNI = foreign-to-native intraverbal

The meta-analysis compared emergent learning outcomes from FTT with outcomes from other verbal operant training procedures; FTT occasioned a significantly higher mean number of untrained verbal responses than intraverbal (FNI or NFI) or listener training and was more efficient than FNI training. Foreign tact training also produced a higher efficiency score (EIS) than NFI, mand, and listener training, but the differences were not statistically significant. Although results are preliminary due to the small number of studies, the aggregated data support the findings of several single-subject studies (e.g., Cortez et al., 2020, 2021; Daly & Dounavi, 2020; Dounavi, 2011). Furthermore, the findings suggest that teaching foreign language speaker skills is more efficient than teaching receptive skills, consistent with research on emergence in

language programming. For example, Contreras et al. (2020) found tact or intraverbal training produced more emergent responses than listener training. In the present review, tact training was the most efficient condition; listener training and FNI were the least efficient conditions. Cortez et al. (2020) suggested that FTT is often effective at producing emergent foreign language responding because it provides opportunities to practice and reinforce the spoken foreign word.

We found no statistically significant difference in emergent responses between adults and children. Several researchers have previously posited a difference (e.g., Cortez et al., 2020; Daly & Dounavi, 2020; Dounavi, 2014; Petursdottir & Haflíðadóttir, 2009); however, our examination of aggregate data did not confirm this position. If differences exist, individual learning histories likely impact learners' ability to derive emergent relations, as derived relations are learned behavior resulting from a history of multiple-exemplar instruction (Barnes-Holmes et al., 2018; Rehfeldt, 2011). Multiple-exemplar instruction may improve emergent learning outcomes by developing and reinforcing a repertoire of derived relations in less verbally competent learners for whom derived relations do not consistently or readily emerge (Lafrance & Tarbox, 2020). It is also likely that training arrangements, including mastery criteria, affected emergent outcomes.

Instructional mastery criteria varied across studies, and several experimenters (Cortez et al., 2020; Matter et al., 2020; May et al., 2019; Petursdottir & Haflíðadóttir, 2009) discontinued training phases before participants attained criterion-level responding, which may have affected emergent outcomes. Although not directly examined by the studies in this review, researchers have found that variability in training criteria can impact the emergence and maintenance of derived relations. For example, Fienup and Brodsky (2017) compared the levels of emergent learning resulting from two different training mastery criteria. Their results showed that more stringent training criteria produced higher levels of emergent responding. Similarly, the two

studies in our meta-analysis with the lowest instructional mastery criteria (Petursdottir & Haflíðadóttir, 2009; Wu et al., 2019) produced the lowest average FTT post-test scores. These findings suggest that the production and retention of emergent relations depend on the strength of directly trained relations. In other words, it is the strength of participants' trained skills that determines the strength and longevity of untrained skills (Critchfield & Twyman, 2014).

Lastly, the reviewed studies failed to examine response maintenance consistently. Maintenance of trained and untrained emergent responses are vital components of any emergent learning program (Wu et al., 2019), yet less than half of the studies reported any follow-up data (Cortez et al., 2022; Daly & Dounavi, 2020; Matter et al., 2020; May et al., 2019). Evaluating emergent outcomes requires a rigorous empirical assessment of learning maintenance over the long term.

This review has some limitations that the reader should consider. First, we excluded several studies that did not include standalone FTT conditions but did evaluate emergent foreign language learning outcomes (e.g., Cao & Greer, 2018; Haegele et al., 2011; May et al., 2016; Petursdottir et al., 2014; Polson & Parsons, 2000; Rosales et al., 2011, 2012) because we aimed to evaluate FTT's outcomes, which required studies with at least one standalone FTT procedure to avoid the risk that combined procedures might produce confounding effects. Also, we chose not to search grey literature, which limited the number of included studies to peer-reviewed ones only. We consider our results preliminary data due to the small number of eligible studies. Second, we evaluated overall training efficiency based on the number of training trials conducted, not the duration of training, because no studies reported the total time required for each condition, and only two studies reported approximate session length (Cortez et al., 2022; Wu et al., 2019). A final limitation, common to any literature review, concerns the

acknowledged bias within publications towards studies that produce positive findings (May et al., 2016; Torgerson, 2006). There is less potential for publication bias to negatively impact the results of the current meta-analysis, though, as we only included studies that directly compared at least two verbal operant training conditions. As such, studies showing negative FTT results would be just as likely to be published as studies showing positive results.

4.6 Conclusion and recommendations for future research

This review examined the effects of tact training on emergent foreign language learning outcomes. The key observation from these preliminary data was that FTT produced higher levels of emergent foreign language learning than other verbal operant procedures. This review raises several questions that warrant further research. First, why is FTT readily acquired for some learners but not all? Future research should consider what procedural variations might improve acquisition (e.g., number of stimuli; Kodak et al., 2019).

Second, why does FTT fail to produce emergence for some learners? It is possible that learners fail to emit emergent responses based on an insufficient reinforcement history of relational exemplars. Future FTT studies could include pre-assessment of learners' relational responses and, if necessary, provide multiple-exemplar instruction before the commencement of the study. Pre-assessment of learners' relational skills and selection of participants with similar pre-assessment results better controls for confounds associated with learner histories.

Third, how do FTT instructional mastery criteria affect emergence? Researchers should examine the preliminary finding that less stringent instructional mastery criteria negatively impacted emergent outcomes by using a within-subjects experimental design (e.g., an adapted alternating treatments design with different criteria assigned to each condition and counterbalanced across participants).

Fourth, what are the long-term learning outcomes associated with FTT? Further, what variables impact maintenance, and how might FTT be combined with other instructional procedures to improve outcomes (e.g., precision teaching; Critchfield & Twyman, 2014)? Researchers should look beyond the accuracy-based mastery criteria commonly employed in these studies to other mastery measures such as those employed within precision teaching's fluency-based free-operant response and measurement systems (Johnson & Layng, 1996; Bucklin et al., 2000).

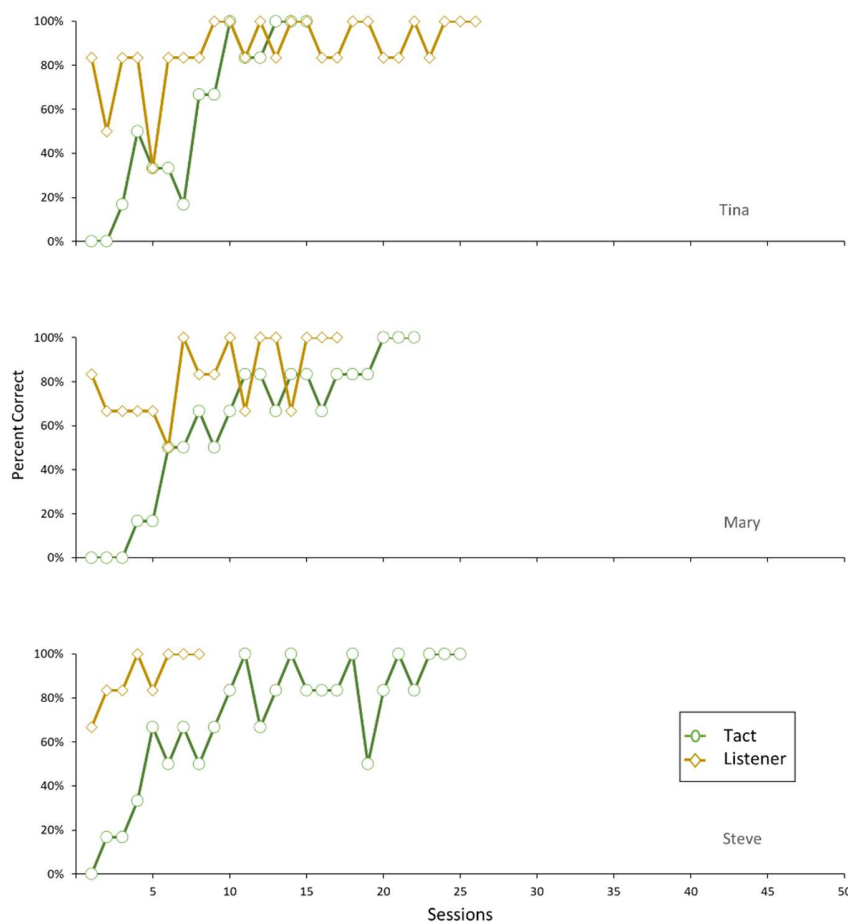
Finally, we recommend future FTT research target a broader range of languages and instructional settings. Although the literature within the field is small, it highlights the considerable potential benefits that behavior analysis offers for optimizing foreign language learning programs.

4.7 Supplementary material

The following figures are supplementary material for the paper by Wooderson, J.R., Bizo, L. A., & Young, K. A systematic review of emergent learning outcomes produced by foreign language tact training. *The Analysis of Verbal Behavior*.

Figure 1

Acquisition Curves (Part A) Adapted from Cortez et al. (2020)



Note. This figure illustrates the differences in acquisition rates across tact and listener training conditions for Tina, Mary, and Steve in Cortez et al. (2020). The authors conducted training sessions in 6-trial blocks—one trial per target word. The mastery criterion was 100% correct responses in three consecutive trial blocks.

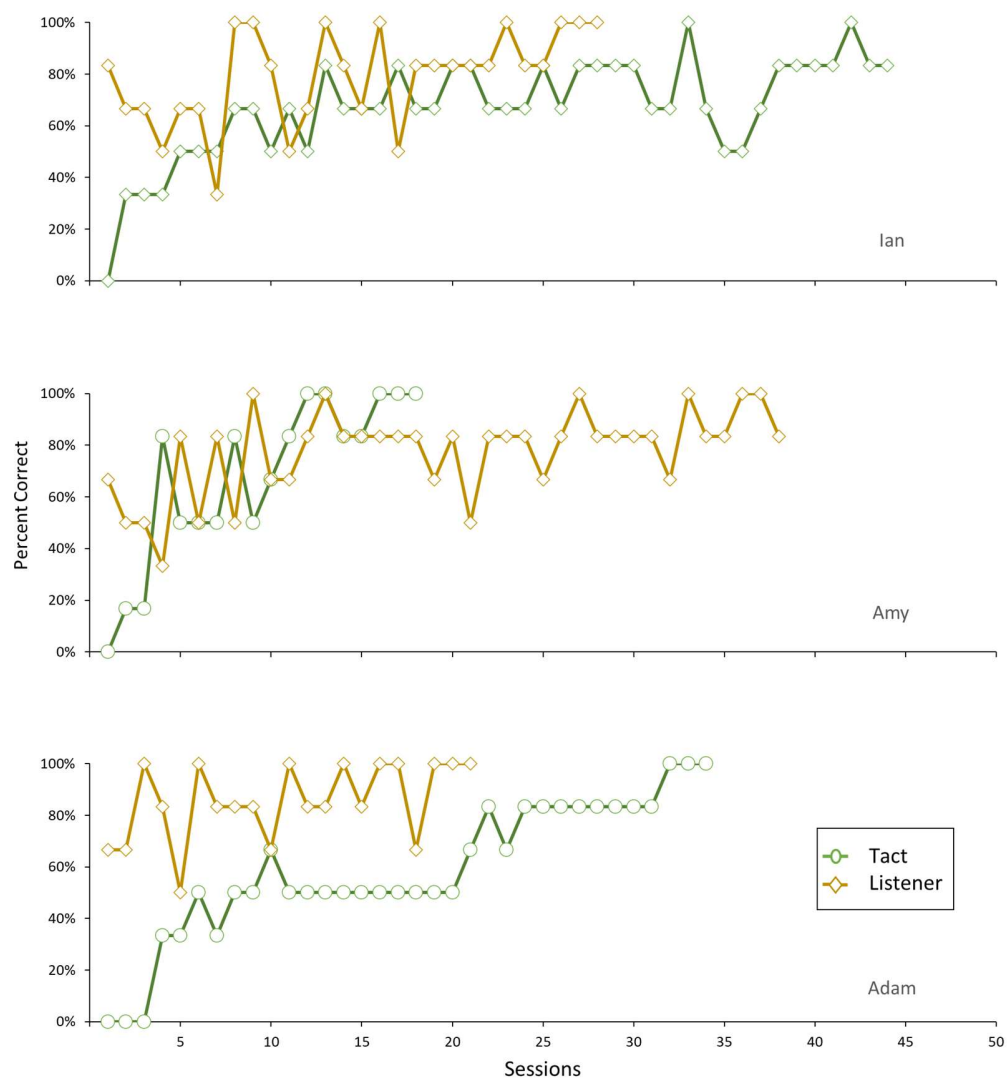
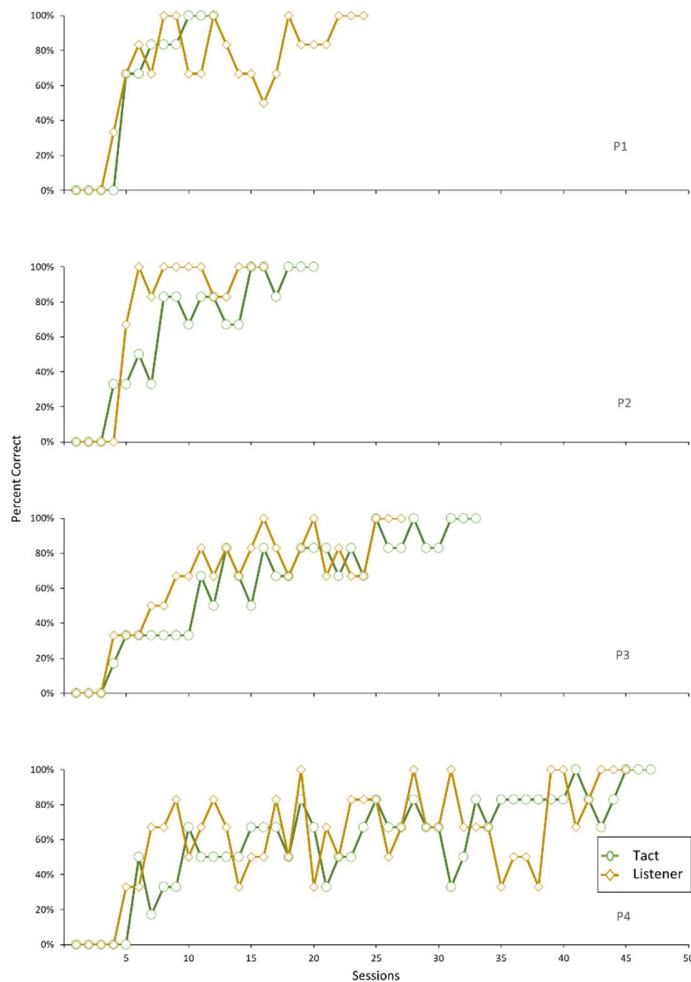
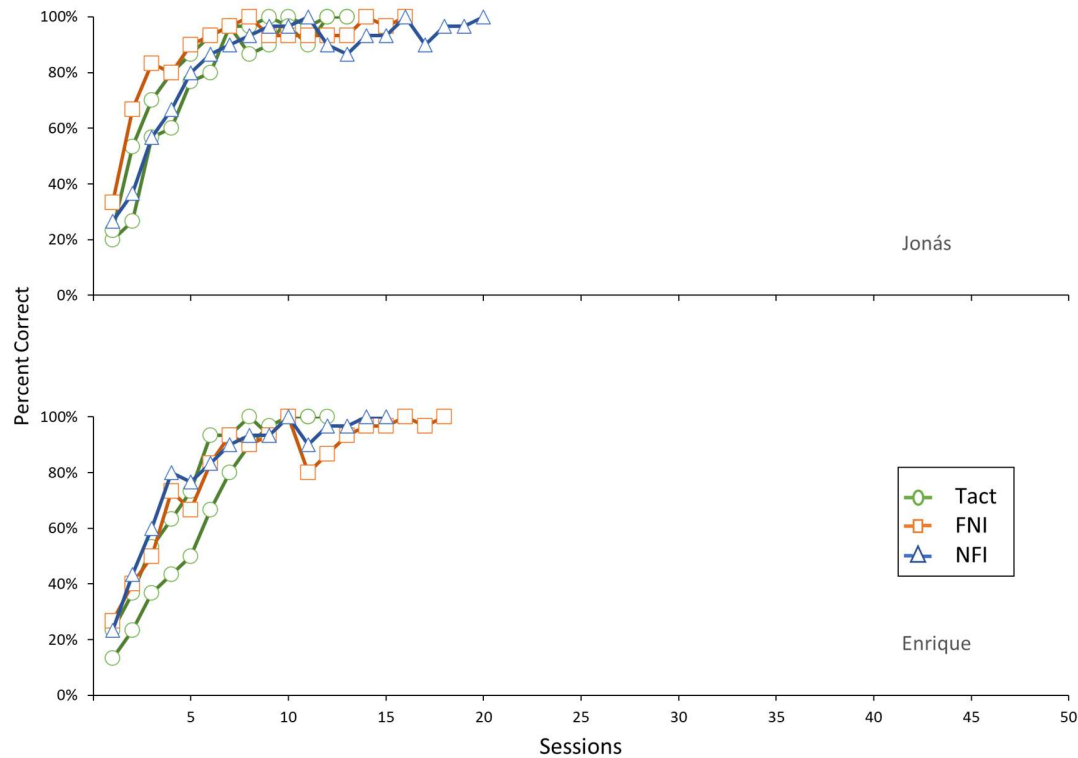
Figure 2*Acquisition Curves (Part B) Adapted from Cortez et al. (2020)**Note.* This

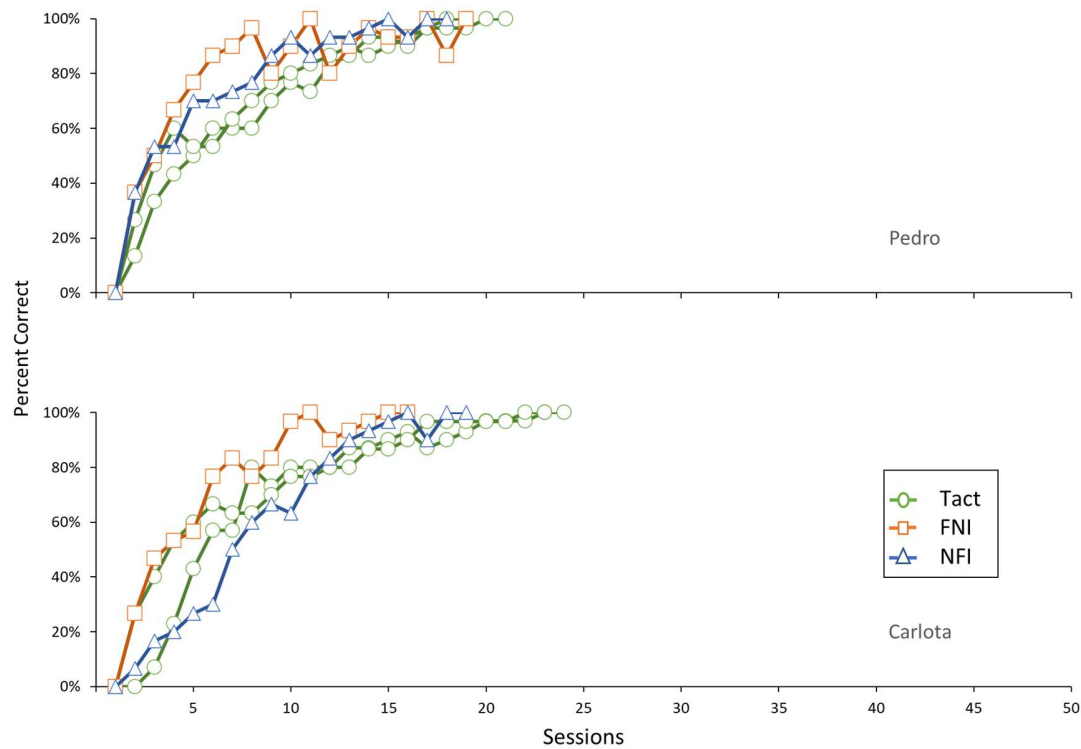
figure illustrates the differences in acquisition rates across tact and listener training conditions for Ian, Amy, and Adam in Cortez et al. (2020). The authors conducted training sessions in 6-trial blocks—one trial per target word. The mastery criterion was 100% correct responses in three consecutive trial blocks.

Figure 3*Acquisition Curves Adapted from Cortez et al. (2021)*

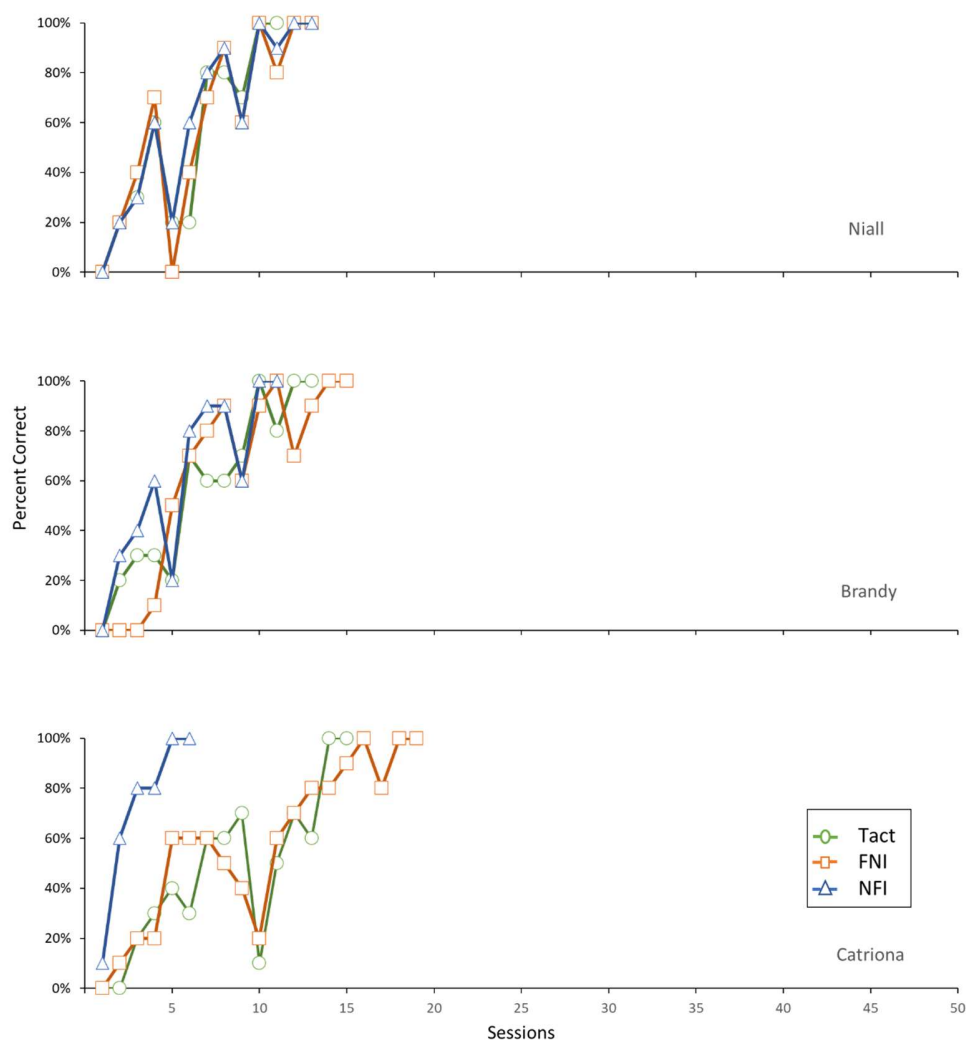
Note. This figure illustrates the differences in acquisition rates across tact and listener training conditions for Participants 1-4 in Cortez et al. (2021). The authors conducted training sessions in 6-trial blocks—one trial per target word. The mastery criterion was 100% correct responses in three consecutive trial blocks.

Figure 4*Acquisition Curves Adapted from Dounavi (2011)*

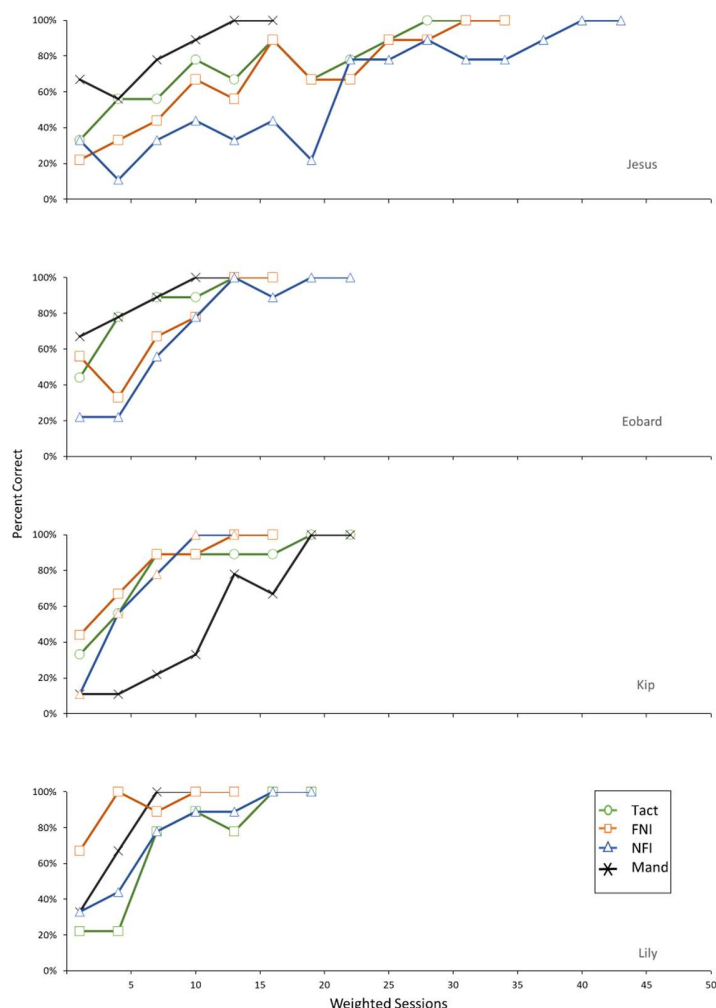
Note. This figure illustrates the differences in acquisition rates across tact, foreign-native intraverbal (FNI), and native-foreign intraverbal (NFI) training conditions for Jonás and Enrique in Dounavi (2011). The author conducted training sessions in 30-trial blocks—one trial per target word. The mastery criterion was 100% correct responses in two consecutive trial blocks.

Figure 5*Acquisition Curves Adapted from Dounavi (2014)*

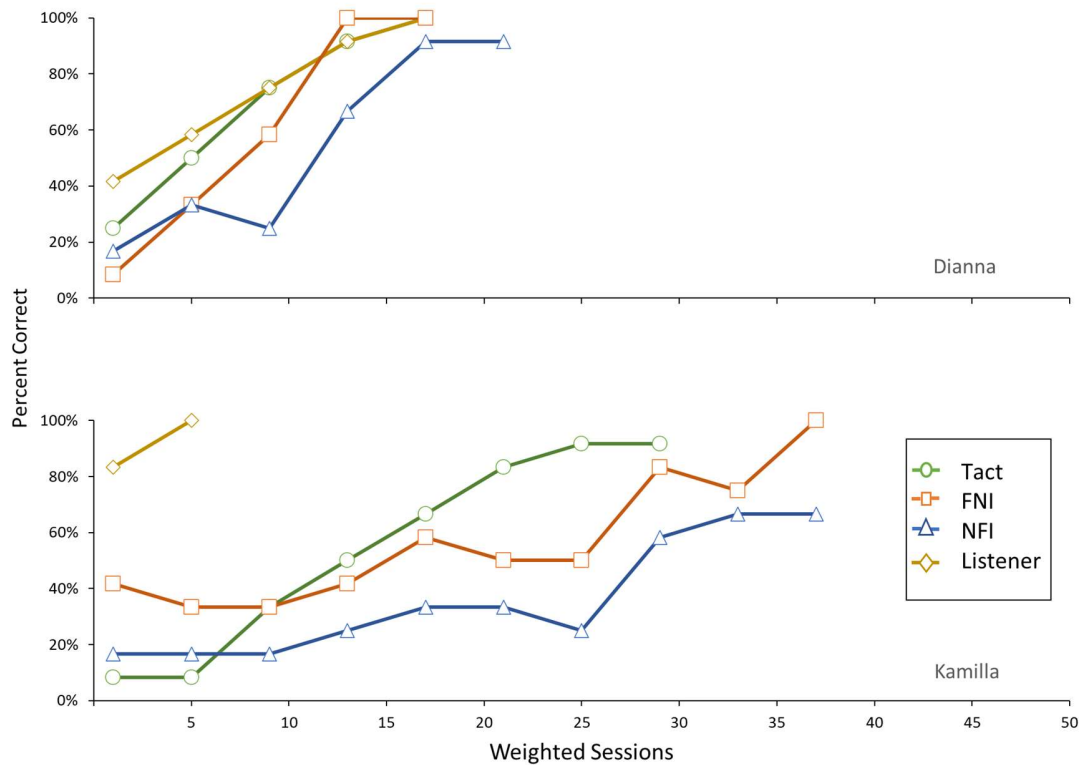
Note. This figure illustrates the differences in acquisition rates across tact, foreign-native intraverbal (FNI), and native-foreign intraverbal (NFI) training conditions for Pedro and Carlota in Dounavi (2014). The author conducted training sessions in 30-trial blocks—one trial per target word. The mastery criterion was 100% correct responses in two consecutive trial blocks.

Figure 6*Acquisition Curves Adapted from Daly & Dounavi (2020)*

Note. This figure illustrates the differences in acquisition rates across tact, foreign-native intraverbal (FNI), and native-foreign intraverbal (NFI) training conditions for Niall, Brandy, and Catriona in Daly & Dounavi (2020). The authors conducted training sessions in 10-trial blocks—one trial per target word. The mastery criterion was 100% correct responses in two consecutive trial blocks.

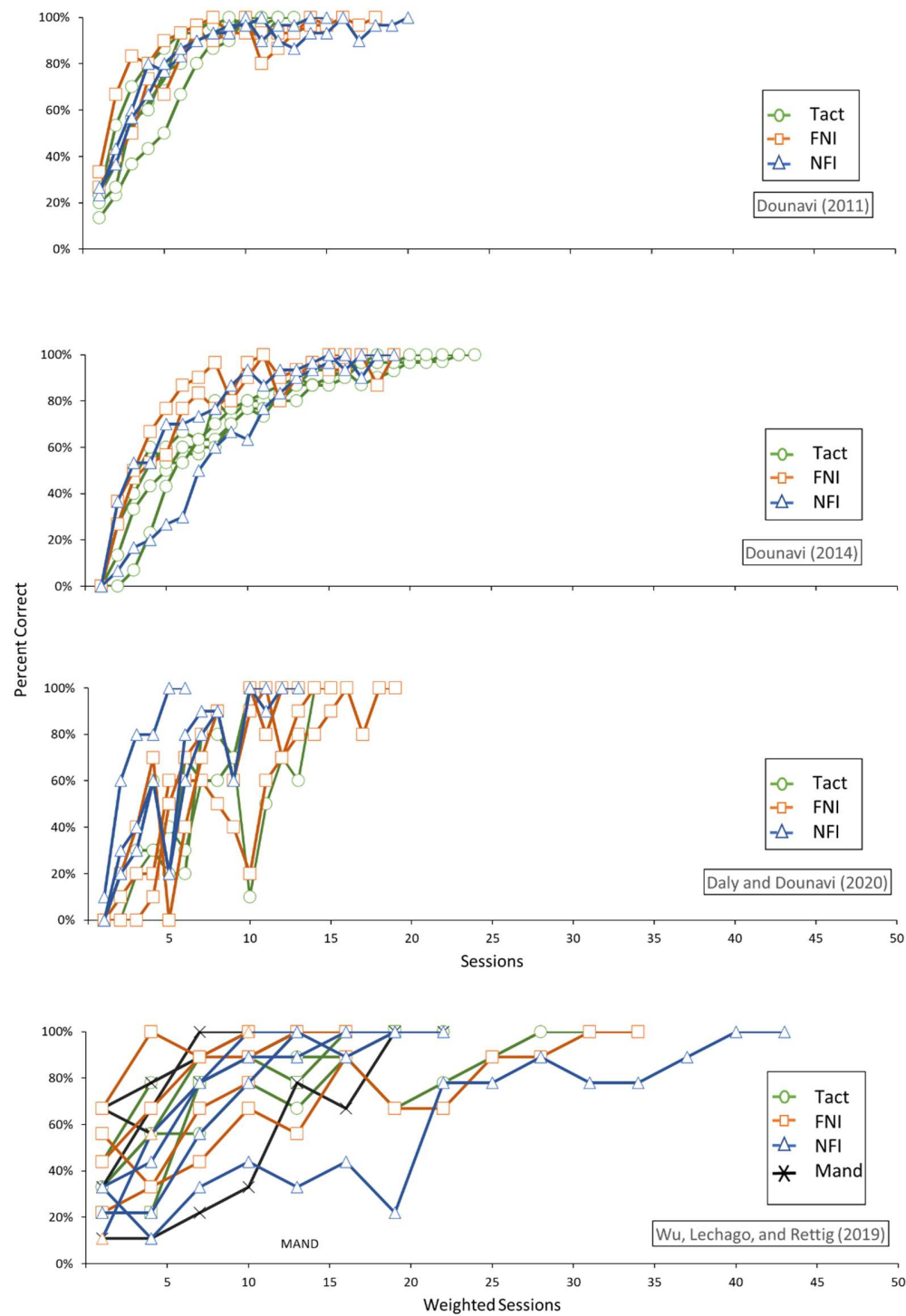
Figure 7*Acquisition Curves Adapted from Wu et al. (2019)*

Note. This figure illustrates the differences in acquisition rates across tact, foreign-native intraverbal (FNI), native-foreign intraverbal (NFI), and mand training conditions for Jesus, Eobard, Kip, and Lily in Wu et al. (2019). The authors conducted training sessions in 9-trial blocks—three trials per target word. We weighted session counts by multiplying the number of sessions by three so that the data could be compared with the other studies in which each session comprised one trial per target word. The mastery criterion was 83.3% correct responses for two consecutive trial blocks.

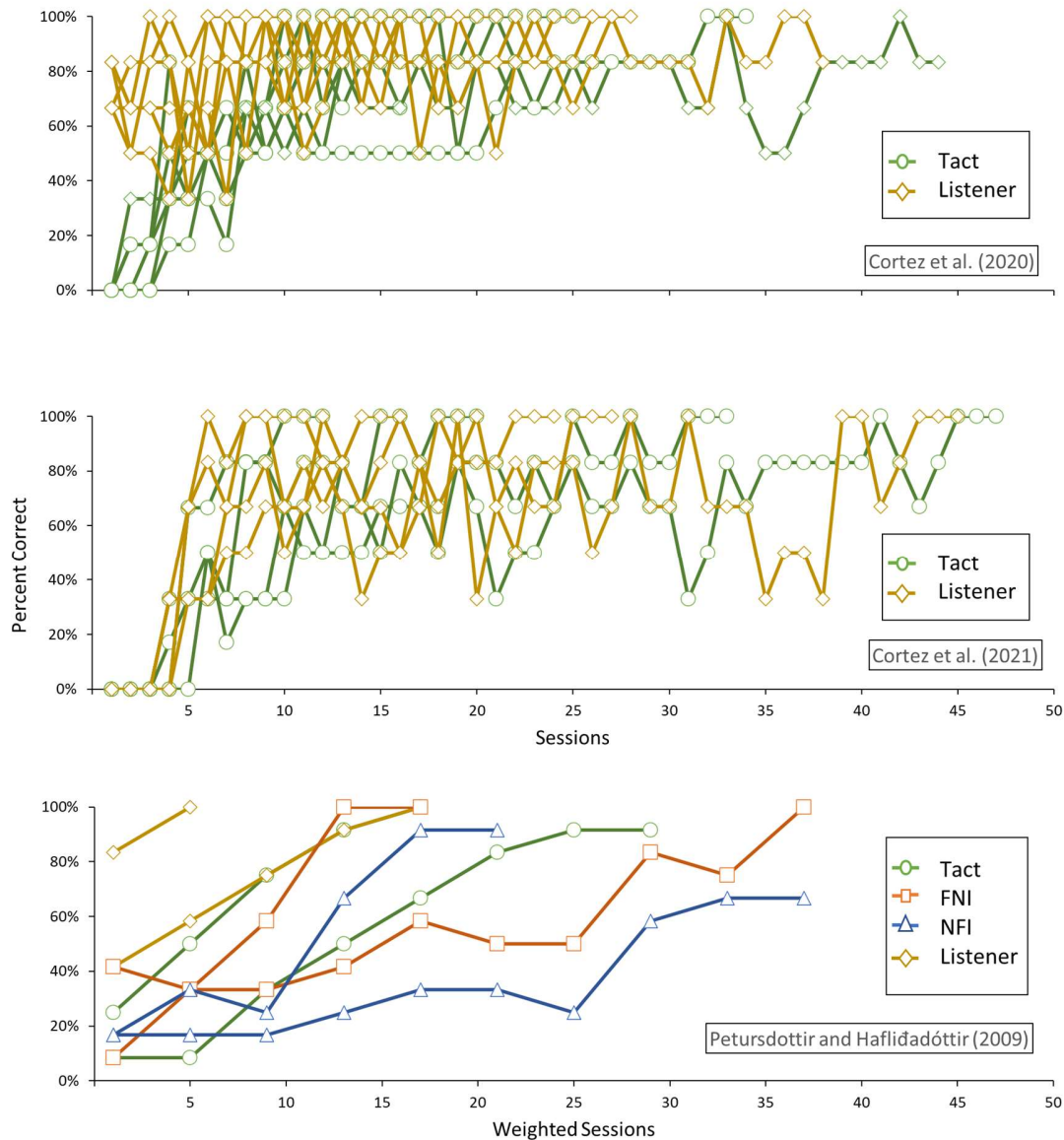
Figure 8*Acquisition Curves Adapted from Petursdottir & Haflíðadóttir, (2009)*

Note. This figure illustrates the differences in acquisition rates across tact, foreign-native intraverbal (FNI), native-foreign intraverbal (NFI), and listener training conditions for Dianna and Kamilla in Petursdottir & Haflíðadóttir (2009). The authors conducted training sessions in 48-trial blocks—four trials per target word. We weighted session counts by multiplying the number of sessions by four so that the data could be compared with the other studies in which each session comprised one trial per target word. The mastery criterion was 83.3% correct responses for two consecutive trial blocks.

Figure 9
Combined Acquisition Curves for Each Study



Retention of emergent foreign vocabulary



Note. This figure illustrates the differences in acquisition rates across tact, foreign-native intraverbal (FNI), native-foreign intraverbal (NFI), mand, and listener training conditions for all participants in each of the seven studies in the meta-analysis (Cortez et al., 2020; Cortez et al., 2022; Daly & Dounavi, 2020; Dounavi, 2011, 2014; Petursdottir & Haflíðadóttir, 2009; Wu et al., 2019).

Chapter 5: Overlearning, precision teaching and fluency building

5.1 Introduction

This chapter builds on the findings from the systematic review and meta-analysis reported in the previous chapter by exploring additional instructional strategies that may further enhance the retention of emergent FL vocabulary. Specifically, it presents an overview of two key learning approaches—overlearning and precision teaching—that were implemented in the first experiment (Study 3, Chapter 6) to improve retention of emergent FL vocabulary.

The systematic review and meta-analysis (Study 2, Chapter 4) highlighted the effectiveness of FTT as a procedure for acquiring emergent FL vocabulary. However, the review also identified a gap in the literature concerning the long-term retention of these learning outcomes. To address this gap, this chapter examines how overlearning and precision teaching, particularly fluency-building procedures, might be integrated into FTT to enhance retention. These approaches are explored in the context of their potential to produce more durable learning outcomes, which is critical for FL learners, especially those not immersed in a natural linguistic environment.

The insights gained from this exploration set the stage for the experimental work detailed in the following chapter, where these concepts are applied and tested in the context of FL vocabulary acquisition.

5.2 Summary of systematic review findings

In Chapter 4, the systematic review and meta-analysis (Wooderson et al., 2022) examined the effectiveness of FTT, a prominent emergent learning procedure. As stated in the review and in Chapter 1.1.7, FTT has been shown to be an effective procedure for acquiring emergent FL

vocabulary. In contrast to more traditional acquisition procedures, FTT teaches one set of responses from which the learner can derive other responses not yet taught (Matter et al., 2020).

This section first summarizes the results of the systematic review and then examines related research on overlearning, precision teaching, and fluency building to explore how these concepts might enhance the effectiveness and retention of emergent learning procedures.

5.2.1 Emergent relations

Foreign tact training produced high levels of emergent learning in all 10 studies included in the review. Following training, 84 (79.2%) emergent learning probes met or exceeded the criterion level. This finding supports Matter et al.'s (2020) contention that FTT may result in more efficient learning than teaching all foreign-language relations at the same time. FTT is used to train tact relations only, while other verbal operants emerge in the learner's repertoire without any direct training. Criterion-level emergent responding emerged for learners during 50 - 100% of post-test probes: listener behaviours (100%, $n = 9$); foreign-to-native intraverbals (FNI, 84.2%, $n = 37$); native-to-foreign intraverbals (NFI, 81.6%, $n = 36$); and mands (50%, $n = 2$).

5.2.2 Target language

Only Wu et al. (2019) targeted a non-European FL, Mandarin Chinese, which the U.S. Department of State's (n.d.) Foreign Service Institute lists as a Category IV language. They describe Category IV languages as exceptionally difficult for native English learners. The four other languages they include in their list of Category IV languages are: Arabic, Chinese – Cantonese, Japanese, and Korean.

5.2.3 Mastery criteria

The studies with the lowest training mastery criteria reported the lowest mean post-test scores (Petursdottir & Haflíðadóttir, 2009; Wu et al., 2019). This suggests that the use of more stringent instructional criteria may have strengthened the retention of emergent responses (Critchfield & Twyman, 2014). In other words, the extent of participants' mastery of the trained skill influenced the strength and durability of untrained emergent skills.

5.2.4 Emergent relations and retention

While the research to date is promising, the data is limited, particularly with respect to retention. Post-training performance with emergent relations should be investigated further because retention is critical to FL learning (Wu et al., 2019). Only four of the ten studies reported follow-up data (Cortez et al., 2022; Daly & Dounavi, 2020; Matter et al., 2020; May et al., 2019) and none systematically examined retention.

In May et al. (2019), follow-up probes produced relatively stable results, albeit over a short retention interval—two weeks after training. Daly and Dounavi (2020) discovered that overall response maintenance was low, and only one of the three participants demonstrated criterion-level emergent responding. The emergent relations were better maintained than the trained relations, but this could have been due to sequencing effects because FTT probes were administered before emergent intraverbal probes. Cortez et al. reported mixed results after examining maintenance of emergent English vocabulary with two of the four children in their study: one 14 days, and the other 30 days after training. Matter et al. (2020) also found mixed results with more delayed follow-up testing two and four months after training, during which

time at least one participant had opportunities to practice with Spanish language users in their own home.

All four studies included only one or two follow-up probes to assess maintenance. It is possible that participants may have performed better (or worse) on subsequent probes. Without at least three probe data points, it is not possible to assess the trend and stability of participants' performance.

In summary, the research on FL vocabulary acquisition and FTT is nascent but promising. The results of these studies demonstrate the effectiveness of emergent learning procedures in producing short-term mastery and maintenance in directly taught and emergent relations. A thorough examination of long-term retention with emergent relations is missing, though. This means that the actual benefits for FL vocabulary learners not immersed in the natural linguistic environment are unclear. As noted above, the degree to which participants learn the directly trained foreign tact relations likely impacts the emergent relations. Stated more directly, the strength of participants' tacting behaviour (i.e., response strength) influences retention with the untaught emergent relations. Response strength and retention are topics that have been explored in the literature on overlearning. The following section of this chapter investigates this literature and the potential impact of overlearning on retention in FTT.

5.3 Overlearning

5.3.1 Definition

Overlearning is defined as the continued practice of a skill beyond the point at which the learner first meets the mastery criterion (Binder, 1996). Overtraining is another term that is often used synonymously with overlearning, and it too implies additional practice beyond predetermined performance criteria; however, overlearning is also used to refer to the results

produced by additional practice (Dougherty & Johnston, 1996). The amount of overlearning used in studies is typically determined as a percentage of the number of trials used to achieve the initial criterion score (Driskell et al., 1992). For example, 50% overlearning might comprise 10 trials to achieve the criterion score of 100% accuracy and five additional overlearning trials. The exact mechanisms that make overlearning effective are not certain; however, recent studies examining overlearning's effects on neurochemical activity and retention indicate that training beyond initial acquisition criteria stabilises new learning, preventing its disruption by other learning (e.g., Shibata et al., 2017).

5.3.2 Overlearning and retention

Within the extensive literature on overlearning, there is widespread agreement that it enhances retention (Colman, 2015; Driskell et al., 1992). Driskell et al. stated that it was unknown what degree of overlearning (e.g., 50%, 100% or 150%) was necessary to attain the purported retention benefits and set about determining this through meta-analysis. Their results found overall positive effects of overlearning on retention, with the degree of overlearning strongly correlated with observed effect sizes—the higher the percentage of overlearning, the greater the effect size.

The benefits of overlearning for directly trained skills are not always long-lasting, though (Driskell et al., 1992; Rohrer et al., 2005). Rohrer et al. found overlearning effects decreased substantially three weeks after training. The differences in the mean levels of retention were significant one-week post-training but decreased dramatically at longer retention intervals. The authors claimed that the discrepancies between their results and the findings of other overlearning research was due to the retention interval and that most studies showing improved retention following overlearning employed retention intervals of one week or less. Driskell et al.

also noted that cognitive skills (e.g., verbal learning tasks) were more susceptible to deterioration than physical skills, and the longer the retention period, the less likely learners were to recall the material.

5.3.3 Overlearning and emergent learning

More applied studies are needed to evaluate whether the expected benefits of overlearning trained material extend to emergent material; related basic research studies suggest they might (e.g., Bortoloti et al., 2013; Bucklin et al., 2000). Bortoloti et al. investigated the effects of overlearning (the authors used the term overtraining) on the strength of emergent relations between abstract symbols and photos of human expressions (happy, neutral, and angry). Their study included two groups of undergraduate students, with one of the groups completing twice as many acquisition trials as the other. The results appeared to indicate that more trials improved the overtraining group's ability to discriminate stimuli accurately, which strengthened the emergent classes developed during training. Similarly, Bortoloti and de Rose (2009) determined that as nodal distance increased (i.e., the extent to which relations between stimuli are derived from intermediate emergent relations), alignment between comparisons of directly trained and derived stimuli decreased, which supported the authors' claim that strengthening stimulus control through trial exposure strengthened the emergent relations.

Bucklin et al. (2000) are one of the few studies to investigate overlearning and the retention of emergent learning over extended intervals. In their research, Bucklin and colleagues utilised emergent learning principles combined with discrete trial and fluency-training procedures to teach participants to associate Hebrew symbols and Arabic numerals with nonsense syllables. The initial instructional phase focused on accuracy using discrete trials, then the overlearning group was provided additional fluency-based trials (timed drills) to achieve the

fluency criterion. Following the training, the experimenters evaluated both the overlearning and accuracy-only training groups on a composite emergent skill. This task required participants to derive the untrained relationships between the Arabic numerals and Hebrew symbols to correctly calculate a sum. Retention was assessed at intervals of two or four weeks, ranging from four to twenty weeks post-training. The probes revealed significantly higher retention rates and accuracy for both emergent and directly trained relations in the overlearning group. Notably, the accuracy-only training group's performance declined substantially just four weeks after training, whereas the overlearning group maintained their performance levels. The authors concluded that overlearning enhances retention of both directly trained and untrained emergent skills.

Contrary to Rohrer et al. (2005) and Driskell et al. (1992) results, overlearning in Bucklin et al. showed benefits that were retained up to 20 weeks post-training. Bucklin et al. also differed from Rohrer et al. (2005) and Driskell et al. (1992) in that the overlearning trials incorporated fluency-building procedures derived from precision teaching practices.

5.4 Precision teaching and fluency-building

5.4.1 Definition

Gist and Bulla (2020) define precision teaching as an instructional measurement technology that comprises four key components. The first component involves establishing precise measurable learning goals, which typically describe fluency criteria (i.e., corrects per minute) and learning channels (the form in which the learner responds). The learning channel used in FTT is a spoken response to a visual stimulus—SEE picture/SAY foreign word.

The second component is use of the standard celeration chart to graph learners' performance. Daily graphing on the standard celeration chart allows teachers to make data-driven decisions about learning and instruction.

The third component refers to the subsequent actions teachers take based on the data collected, which generally are either: a) continue providing practice opportunities where fluency aims are not yet achieved; b) complete the current instructional program because the learner has met the learning target; c) or make some changes to the instructional procedures because the learner is not making adequate progress.

The final component in this recursive process is noted as 'keep going', in which case, the teacher continues to steer the learner towards the learning target and adjust the learning program as needed.

Fluency-building, also sometimes described as rate-building, refers to procedures aimed at encouraging learners to achieve fluent performance characterised by speed and accuracy of responding (Binder, 1996). While conventional educational programmes typically focus on the accuracy of learners' performances, precision teachers target fluency, a combination of accuracy and speed. A prevalent misconception is that fluency-building focuses on fast performance. Proponents of behavioural fluency do not target the speed of responding alone, though. Instead, their focus is "...speed that characterizes competent performance" (Binder, 1996, p. 164), "...doing the right thing without hesitation" (Binder, 1996, p. 164), and "...behavior that is flowing, effortless, well-practiced, and accurate" (Johnson & Layng, 1996, p. 281).

5.4.2 Fluency and retention

Fluent performance is expected to produce a range of benefits to the learner, including improvements in retention, endurance, and application to new settings and stimuli (Johnson & Street, 2004; Pennypacker et al., 2003). Support for the claim that fluency improves retention is found in both basic and applied research (e.g., Porritt et al., 2009; Lee & Singer-Dudek, 2012; Singer-Dudek & Greer, 2005). However, some researchers argue that the existing evidence is

inconclusive and advocate for further investigation (e.g., Doughty et al., 2004; Peladeau et al., 2003).

5.5 Conclusion

In conclusion, the systematic review's results (Wooderson et al., 2022) identified a paucity of experimental research evaluating long-term retention of emergent FL vocabulary learning. Consequently, little is known about the instructional arrangements that promote retention. Overlearning has been shown to positively impact the retention of directly trained learning material (Driskell et al., 1992), but its advantages over more traditional training procedures may dissipate over the longer term (Rohrer et al., 2005). Few studies to date have evaluated the effects of overlearning on long-term retention (i.e., one month or more) of emergent learning material. The few that have (e.g., Bucklin et al., 2000) provide promising, albeit basic research demonstrations of fluency-building procedures' potential to produce long-term retention of emergent learning material. More applied research that includes repeated post-test measures and longer retention intervals is needed. Furthermore, existing FTT research to date focused solely on accuracy criteria (e.g., 100 percent correct for two consecutive trials). Researchers studying overlearning procedures have proposed that training beyond traditional mastery criteria improves retention of both directly trained and emergent learning (e.g., Bucklin et al., 2000), but further applied research is needed.

The insights gained from this review set the stage for the experimental work detailed in the following chapter.

5.6 Introduction to the first experiment

Chapter 6 presents this thesis' first experiment, which evaluated a modified FTT protocol with fluency-based overlearning trials aimed at improving learners' long-term retention of

emergent FL intraverbals. The study employed repeated post-training probes of emergent intraverbal relations. Korean was selected as the target FL due to its level of difficulty (U.S. Department of State, n.d.) and the limited number of studies to date involving non-European languages. This applied experiment extended Bucklin et al.'s (2000) demonstration study by comparing the emergent learning outcomes of the modified FTT protocol with a more traditional discrete trial-based approach. The main research question was: *Does repeated practice of foreign tact relations beyond initial mastery (i.e., accuracy criterion) using fluency-building procedures affect the retention accuracy of derived intraverbal relations during testing?*

Chapter 6: Study 3—Retention of emergent Korean vocabulary following foreign tact training and overlearning

Wooderson, J. R., Bizo, L. A., & Young, K. (2023). Retention of emergent Korean vocabulary following foreign tact training and overlearning. [Manuscript submitted for publication]⁵

6.1 Abstract

Overlearning refers to the repeated practice of material beyond initial mastery (Binder, 1996). In this preliminary study, we evaluated the effects of overlearning on the retention of emergent intraverbal Korean vocabulary with five adults. The participants completed two training conditions (regular training and overlearning), involving tacting visual stimuli in Korean. In the overlearning condition, participants were set additional fluency criteria and continued to practice beyond the initial accuracy mastery criterion. Both conditions generated high levels of emergent intraverbal responding post-training; however, data indicated that retention of emergent relations was greater following overlearning.

6.2 Introduction

The question of 'when should training end?' is critical for any learning program. The amount of time learners can spend practicing material is always limited, and educators must carefully consider at what point further training is unlikely to provide additional benefit to the learner. Behavior-analytic researchers and educators often use a predetermined mastery criterion to decide when training is complete or when to shift focus to a different learning target (Fuller & Fienup, 2018), however, there is limited empirical research to guide the selection of the mastery criterion (McDougale et al., 2020; Fienup & Carr, 2021).

⁵ The paper was submitted to an American journal and uses American English spelling throughout.

In a series of recent studies examining the mastery criterion, researchers found that the higher the criterion, the more likely learners were to retain what was taught (Fuller & Fienup, 2018; Longino et al., 2021; Pitts & Hoerger, 2021; Richling et al., 2019). Fuller and Fienup (2018) taught children spelling skills to three levels of performance criterion: 50%-, 80%-, or 90%-correct for a single session. Their results showed that the higher (i.e., 90%) criterion level produced the highest levels of maintenance 3 to 4 weeks post-training. Richling et al. (2019) conducted three parametric studies similar in design to Fuller and Fienup (2018). The first two studies (Experiments 2 and 3) compared three criteria: 60%-, 80%- and 100%-correct across three consecutive sessions during four weekly maintenance probes. The final study (Experiment 4) evaluated 80%-, 90%- and 100%-correct across three consecutive sessions during a single follow-up test one week after training. In all three experiments, the 100% criterion produced the best (or equal best) results across all but 2 of the 108 maintenance probes, whereas the authors asserted that the 60%, 80% and 90% accuracy criteria did not produce sufficient skill maintenance for most of the participants. When Longino et al. (2021) systematically replicated this study, the results were similar, except that the 90% criterion was sufficient for producing skill maintenance at comparable and sometimes higher levels than the 100% criterion. In contrast, Pitts & Hoerger (2021) found that the 100%-accuracy criterion reliably produced better skill maintenance than the 80% or 90% criterion levels. These studies' findings add to a body of evidence suggesting that higher mastery criteria result in more effectively maintained skills. However, one inherent limitation of performance criteria based on percent correct measures is that 100% is the highest level that can be achieved. As a result, it is unclear whether there are any additional benefits for learners above this threshold. According to Binder (2004), percentage correct measures are insufficiently sensitive to answer this question. That is, the number of

accuracy trials required to reach a 100% criterion for any learner does not predict the number of additional trials required for that learner to achieve a given retention level.

Fluency criteria derived from precision teaching practices, are an alternative to accuracy-based performance criteria and offer a potential solution to the 100% correct measurement ceiling. While traditional educational programs are often concerned with the accuracy of students' performances alone, precision teachers are concerned with accuracy and speed, which they refer to as fluency. This enables the instructor to establish a mastery criterion that is not restricted by percentage correct measures. One common misconception about fluency-building is that it emphasizes speed (Binder, 2004). However, advocates of behavioral fluency do not target 'fast' performance. Instead, emphasis is placed on developing accurate performance at "...speed that characterizes competent performance" (Binder, 1996, p. 164), "...doing the right thing without hesitation" (Binder, 1996, p. 164), and "...behavior that is flowing, effortless, well-practiced, and accurate" (Johnson & Layng, 1996, p. 281). Fluent performance is associated with improved learning retention, endurance, and adaptability to novel environments and stimuli (Johnson & Street, 2004; Pennypacker et al., 2003). Although there is empirical support for precision teachers' claims that fluency-building improves retention in the basic (e.g., Porritt et al., 2009) and applied literature (e.g., Singer-Dudek & Greer, 2005; Lee & Singer-Dudek, 2012, Quigley et al., 2018), some researchers (e.g., Doughty et al., 2004; Peladeau et al., 2003) argue that the evidence is inconclusive and lacks a convincing argument for its effectiveness. Others have argued that fluency-building improves retention because it is a form of overlearning in which students practice learning material beyond the 100 percent correct accuracy criterion until they achieve the desired response rate (Bucklin et al., 2000). Overlearning is thought to improve learning retention, and there is a large body of research evaluating its outcomes (Driskell et al.,

1992). Recent research into the effects of overlearning on changes in neurochemical processing and skill performance suggests that overlearning may work by stabilizing new learning and protecting it from disruption by subsequent learning (Shibata et al., 2017). Despite the reported advantages of overlearning on directly trained material, more research is needed to determine whether these advantages extend to emergent learning material.

Emergent learning, sometimes referred to as generative learning in behavior analytic literature, is learning that does not require direct experience to acquire (Critchfield & Twyman, 2014). The resulting outcome is that some learning occurs 'for free'. Typically, this means that direct training in one set of skills produces competency in a related but untaught set of skills. This 'free' learning presents apparent advantages for learners and those responsible for designing and delivering educational programs. Notably, instructional time is always limited, and educators must consider achieving the most impactful outcomes with available resources. The results of the few studies that have examined the effects of fluency-based overlearning on emergent learning suggest it might positively impact the retention of emergent skills and those trained directly. For example, Bucklin et al. (2000) examined the outcomes for directly trained and emergent learning skills using a between-groups experimental design to compare differences in retention rates up to 20 weeks post-training with 30 college student participants. Using emergent learning procedures and accuracy-based mastery criteria, participants in both groups learned to associate a set of nonsense syllables with a) Hebrew Symbols and b) Arabic numerals. Following initial accuracy training, participants in the overlearning group received additional fluency-based training (five one-minute timed drills per session) until they achieved the fluency criterion (i.e., number of responses per minute). After training, all participants were tested on an emergent skill that required the addition of two Hebrew symbols. Participants had to be able to derive the untrained

emergent relations between the Hebrew symbols and Arabic numerals to complete the task correctly. Their retention was then tested at two or four-week intervals between 4 and 20 weeks after training. The retention probe results showed that the overlearning group maintained significantly higher relative levels of accuracy and fluency for the emergent relations and the trained relations. Interestingly, the group that did not complete the overlearning trials showed a significant decrease in performance after only four weeks, but not the overlearning group. These findings indicate that fluency-based overlearning can provide improved retention over the longer term. The authors recommended that further studies be conducted with different participants, tasks, and settings to determine whether the results generalize to other areas of learning. If other studies demonstrate similarly positive effects on retention accuracy from fluency-building, this might assist with the development of more efficient and effective training procedures. We set out in the current paper to evaluate the effects of fluency-building procedures on foreign language vocabulary learning.

Several recent studies demonstrate the potential benefits of emergent-learning approaches to deliberate vocabulary learning (Cao & Greer, 2018; Cortez et al., 2020, 2022; Daly & Dounavi, 2020; Matter et al., 2020; May et al., 2019; Wu et al., 2019). Indeed, instruction based on emergent learning principles and verbal operant procedures may be more efficient than training all relations directly (Matter et al., 2020). Foreign tact training (FTT) is one such verbal operant procedure that teaches the learner to vocalize the appropriate foreign word in the presence of a corresponding visual stimulus and then tests for the emergence of one or more untrained verbal operants. Recent systematic reviews demonstrate FTT's effectiveness in evoking untrained foreign language repertoires (Melvin-Brown et al., 2022; Wooderson et al., 2022). Wooderson et al. reviewed 10 FTT studies and concluded that FTT produced a range of derived

stimulus relations and significantly higher levels of emergent responding than foreign-to-native intraverbal (emitting the appropriate native-language word in the presence of the foreign word), native-to-foreign intraverbal (NFI; emitting the appropriate foreign word in the presence of the native-language word), and foreign listener training (selecting an appropriate referent in the presence of the foreign word). Melvin-Brown et al. (2022) also found that FTT produced a greater number of emergent relations in comparison with other verbal operant training procedures. Although the growing research interest in this area is promising, little is known about the retention of the emergent learning outcomes produced by FTT. Wooderson et al. reported that only four (Cortez et al., 2022; Daly & Dounavi, 2020; Matter et al., 2020; May et al., 2019) of the 10 studies they reviewed conducted follow-up tests, and no studies demonstrated consistently robust long-term retention of vocabulary learning. Melvin-Brown et al.'s (2022) findings were similar: only 38% of the reviewed studies reported follow-up probes. Without retention, language learners are unlikely to develop the depth of vocabulary necessary to engage in successful communication (Schmitt, 2010).

The present study compared the emergent learning outcomes produced by overlearning (accuracy and fluency criteria) with regular training (accuracy criterion only) using FTT to examine learners' retention of emergent Korean intraverbals up to 33 weeks after training. The primary research question in our study was: Does repeated practice of foreign tact relations beyond initial mastery (i.e., accuracy criterion) using fluency-building procedures affect the retention accuracy of derived intraverbal relations during testing?

6.3 Method

6.31 Participants and setting

The study included five adults known to the first author (P1 to P5; two males, three females; Table 1) whose ages ranged from 26 to 52 years (mean=34.8 years). Two participants held graduate degrees, while three held undergraduate degrees. All participants spoke English, and three had prior experience with a second language but not Korean.

Table 1

Participants' gender, age, ethnicity, education level, and experience with a second language.

Participant ID	Gender	Age	Ethnicity	Education level	Second language experience
P1	Male	41	European Australian	Graduate	Spanish, Japanese
P2	Female	26	Japanese Australian	Undergraduate	Japanese
P3	Male	26	European Australian	Undergraduate	None
P4	Female	29	European Australian	Graduate	None
P5	Female	52	Malay	Undergraduate	Indonesian

The training was conducted online via Microsoft Teams®, and participants connected to sessions from their homes using personal computer equipment. Before each session, the experimenter instructed participants to close all software programs other than Microsoft Teams on their computers and clear any materials from their work area. Instructional sessions lasted

approximately 1-3 minutes each. Each participant undertook 10 sessions per day for up to 10 weekdays, depending on whether they achieved the mastery criterion before the end of the training phases. If the participant met the criterion before then, the next phase began the following weekday. The training was limited to a maximum of 10 days per participant due to participant time constraints. All sessions were recorded using Microsoft Teams' video recording function. All participants signed a written informed consent form before the study.

During training, the experimenter shared their screen showing the learning materials, Microsoft PowerPoint® slides via Microsoft Teams. Participants could not access learning materials outside of sessions. Furthermore, participants were instructed not to practice and refrain from using the words taught during sessions outside of the training and testing sessions.

We assigned two stimulus sets (Table 2) for each participant (one set per training condition). The tact training stimuli presented during accuracy training included 20 color pictures (10 per set) centered on a white background in Microsoft PowerPoint™ and resized to approximately 50% of the available screen space. During fluency training, the 10 pictures in each set were resized and randomly arranged in a 5 x 2 grid so that all 10 stimuli could be seen on screen simultaneously. All pictures were obtained using Google™ image searches. The NFI stimuli presented during pre-, post-training, and follow-up phases were the corresponding 20 written English nouns on a white background using a black-colored Calibri font at 96 point font size in Microsoft PowerPoint™.

Table 2

Target words in Korean (foreign language) and English (native language) for both stimulus sets

Set 1		Set 2	
Korean	English	Korean	English
Sajin	Photo	Janggap	Gloves
Gawi	Scissors	Naembi	Pot
Oi	Cucumber	Gamja	Potato
Chima	Skirt	Chimdae	Bed
Usan	Umbrella	Begae	Pillow
Baji	Pants	Jido	Map
Ageo	Crocodile	Subak	Watermelon
Chamgmun	Window	Namu	Tree
Sagwa	Apple	Sangeo	Shark
Yeonpil	Pencil	Chiyak	Toothpaste

To equate the difficulty of each set, we employed the following procedures (Shepley et al., 2020; Cariveau et al., 2021):

1. Each set comprised 10 common noun targets.
2. Each target word contained two syllables, and the syllables were audibly distinct from each other.
3. Each set contained an equal number of words beginning with the same initial sound, and the configuration of each word was different.
4. All visual stimuli were composed of contrasting shapes and colors.

5. Each participant was randomly assigned sets using an online sorting software (Picker Wheel, n.d.). Consequently, P1 and P2 were assigned set 1, and P3, P4, and P5 were assigned set 2 in the overlearning condition.
6. All participants completed pre-assessment tests to ensure they could a) pronounce each of the Korean words, and b) tact the words in English before pre-testing commenced.
7. The order of presentation of sets to participants was counterbalanced, and the order of presentation of stimuli during trials was randomized using a custom Microsoft Excel TM macro written by the first author.
8. A baseline condition with 2-6 trial blocks was conducted with each participant before commencing instruction.
9. Procedural fidelity checks to examine the level of adherence to the described experimental procedures were conducted across both study conditions.
10. Reliability checks to assess the level of agreement between independent observers scoring the dependent variables was also conducted.

6.32 Response Measurement and Dependent Variables

The primary dependent variable was correct responses. Incorrect responses consisted of the participant saying: the wrong word, the target word in the wrong language, they did not know, or no response within 5 seconds of trial initiation (Wu et al., 2019). Observers scored a response correct if the participant vocalized the Korean referent after being shown a) the equivalent English written word during pre-and post-test sessions or b) a picture representing the word during tact training sessions. Participants were required to articulate the Korean targets

accurately; a phonetic dictionary with Romanized terms verified by a native Korean speaker was used to assess pronunciation.

6.34 Experimental Design and Procedures

We employed an adapted alternating treatments design (Sindelar et al., 1985) to compare the effects of two tact-training conditions—1) regular training (accuracy criterion only) and 2) overlearning (accuracy and fluency criteria)—on participants' retention of emergent Korean vocabulary. The study's phases were implemented in the following order: pre-assessments, native-to-foreign intraverbal (NFI) pre-tests, baseline, instruction, NFI post-tests, and follow-up tests. The pre-assessment, accuracy, pre-, post-, and follow-up testing procedures were adapted from Matter et al. (2020). All procedures performed in the study were approved by the University of Technology Sydney Human Research Ethics Committee (ETH22-6997).

Pre-assessments

The two pre-assessment procedures, Korean pronunciation and native tact pre-assessments, were conducted on consecutive days at the commencement of the study. During both pre-assessment sessions, all forty stimuli were tested once each. Pronunciation pre-assessments involved the experimenter vocalizing and asking participants to repeat each target word in Korean; feedback followed each response. If participants closely approximated the target word, they were asked to repeat the word until they could pronounce it correctly, and if they were unable to pronounce the word in Korean correctly, it was discarded and replaced.

During native tact pre-assessments, the experimenter asked participants to label pictures of each target word in English. This procedure aimed to remove any ambiguity regarding the stimuli used in the study and adjust the materials to include the English referents preferred by the participants. If a participant's response was synonymous with the expected noun (e.g., 'quilt'

instead of 'blanket'): the experimenter asked the participant to select between the two; the participant's preference was noted, and learning materials were adjusted accordingly.

Pre-tests

Before training, the experimenter tested the participants on both stimulus sets using the NFI relations. Each pre-test phase comprised four sessions, in which we tested both sets of words and all stimuli twice—eighty trials. At the beginning of each pre-test, the experimenter told the participant, "I will ask you what an English word is in Korean. I'm not going to tell you if you're right or wrong, but I want you to give it a try". Finally, the experimenter provided neutral comments (e.g., "okay") as feedback for correct and incorrect responses.

Baseline and Instruction

Using tact relations, we conducted baseline (the first 2-6 sessions) and instructional phases for both stimulus sets. All baseline sessions followed accuracy procedures except that the experimenter only provided neutral feedback—like the pre-tests. We implemented the same accuracy procedures for both stimulus sets during the first phase of instruction until participants achieved the accuracy mastery criterion—100% correct responding across two consecutive ten-trial blocks— or completed twenty-five ten-trial blocks, whichever occurred first. If the participant failed to attain the accuracy criterion within twenty-five trial blocks, we proceeded to the next phase of the study. During the next phase, the experimenter commenced weekly post-tests with the regular training stimulus set and continued training with the overlearning stimulus set using the fluency training procedures. Post-testing with the overlearning stimulus set started when the participant either achieved the fluency criterion or completed 50 one-minute timed trials.

Accuracy training

During the accuracy training phase, we conducted trials by presenting pictures of each target word (10 per set) one at a time in a randomized order and providing immediate feedback by acknowledging a correct response or error correction. Sessions started with the experimenter telling the participant, "I will show you slides with pictures and ask you to label each picture you see in Korean. After you label a picture, I will tell you if you are right or wrong. If you are unsure about any of the pictures, you can say, 'I don't know' or guess. Then, I will tell you the correct answer before moving to the next slide." The experimenter waited five seconds for the participant to respond before providing feedback and presenting the next trial until all 10 stimuli in the set had been presented. Immediately following the last accuracy training trial, the experimenter conducted the first post-test with the regular training set and commenced fluency training with the overlearning set.

Fluency training

The overlearning condition added fluency-building procedures and focused on increasing the participants' response rate to attain the fluency criterion—100-80 words per minute with no errors (K. Johnson, personal communication, February 23, 2021). This criterion rate was comparable to that of a native Korean speaker known to the first author. Whereas the accuracy training procedures were conducted with both stimulus sets, only one stimulus set (overlearning) was trained using the fluency procedure.

Before each session, the experimenter provided the following instructions, "I will show you slides with pictures of the words you have learned, and you will have one minute to label as many pictures in Korean as you can. When the timer starts, and I say, 'Please begin,' label each picture from left to right, starting with the top row and moving to the bottom. Then, we will

move to the next slide. If you are unsure about any of the pictures, you can say, 'don't know' or guess. Do not skip any pictures. When the timer ends, I will tell you how well you did."

During fluency training, the experimenter only provided feedback at the end of the one-minute session (i.e., when the timer ended) and not after every tact. Following each one-minute timing, the experimenter scored the correct responses and errors and calculated the participant's response rate. The experimenter then gave feedback to the participants on their performance and corrected any errors the participant made. The participant was also encouraged to increase their response rate if they had not yet attained the target rate. For example, "You scored 60 corrects per minute during the last trial. Your target is at least 80 corrects per minute. Try to increase your rate of correct responses during the next trial".

Post-tests and follow-up tests

Post-tests followed the same procedure as pre-tests, except post-tests were repeated weekly over four weeks. Using the same procedures, the experimenter implemented a one-off follow-up test with both stimulus sets at 27-33 weeks post-training, depending on the availability of each participant.

6.35 Interobserver Agreement and Treatment Integrity

A second independent observer (BCBA®) watched 29% (range, 25% to 37%) of the session recordings across all baseline, instruction, pre-, post-, and follow-up test phases while independently collecting data on participant responses during trial presentations. We then compared the data collected by both observers on a trial-by-trial basis. If both observers recorded an incorrect or correct response for the same trial, we scored this as an agreement; otherwise, we scored the trial as a disagreement. We calculated interobserver agreement by dividing the number of agreements by the total number of agreements and disagreements and multiplying by

100. The mean interobserver agreement for sessions during pre-and post-testing was 99.5% (range, 98.8% to 100%), baseline was 100%, and training was 97.4% (range, 95.9% to 100%).

The same independent observer collected data on the experimenter's implementation of the study procedures. During baseline, pre-, and post-test phases, the second observer scored whether the experimenter: 1) presented the discriminative stimulus, 2) waited up to 5 seconds for the participant to respond, and 3) provided neutral feedback for all learner responses. During the accuracy phase, the steps were the same, except the observer checked whether the experimenter provided acknowledgment of a correct response or error correction rather than neutral feedback. Finally, during the fluency phase, the observer scored whether the experimenter provided feedback to the participant on the number of correct responses only after each one-minute trial. We calculated treatment integrity by dividing the total number of correctly implemented steps by the total number of correctly and incorrectly implemented steps and multiplying by 100. Mean treatment integrity was 100% across all observed sessions.

6.4 Results

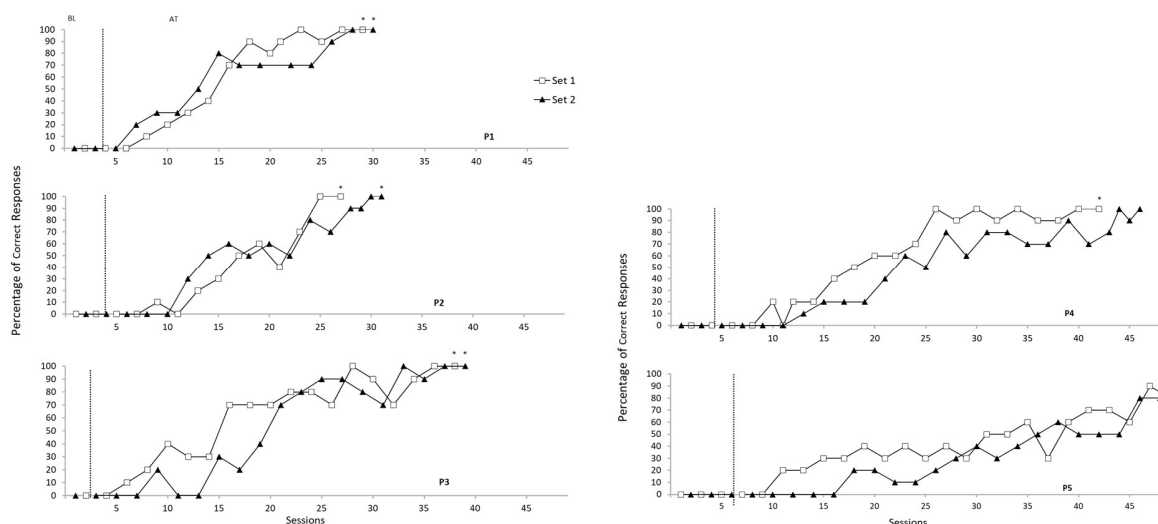
6.41 Baseline and Training

Figure 1 shows the five participants' baseline and training data during the accuracy training phase. No participants responded correctly during baseline trials. P1, P2, and P3 steadily improved their performance during the accuracy phase until they achieved the accuracy mastery criterion with both stimulus sets. P4 met the mastery criterion for set 1 (regular training set) but not set 2 (overlearning set), and P5 failed to achieve the mastery criterion for either set before the end of the accuracy training phase. This potentially confounded and negatively impacted P5's post-test results; however, her performance was similar across both sets during the regular training condition. Further, it is unlikely that this significantly impacted P4's post-test results

because she achieved mastery with the regular training stimulus set and then participated in additional fluency-based trials with the overlearning set.

Figure 1.

Accuracy Training Outcomes for P1, P2, P3, P4, and P5



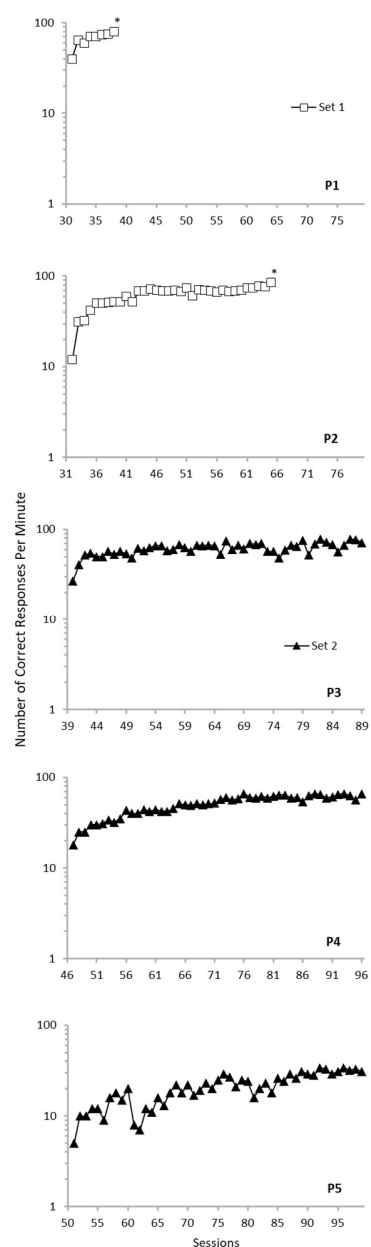
Note. Asterisks note the session in which the participant met the accuracy mastery criterion. BL = baseline; AT = accuracy training

P1 and P2 participated in the fluency phase with the stimuli from set 1 (Figure 2). P1 attained the fluency criterion of 80 or more correct tact responses per minute with no errors within eight sessions—P2 achieved it within 34 sessions. P3, P4, and P5 completed the fluency phase with set 2 but failed to attain the target rate before the end of the phase. Of these three participants, P3 came closest to achieving the fluency criterion with 78 correct responses per minute; P4's best rate was 66, and P5's was 34. Both P1 and P2, spoke Japanese as a second language, which shares similarities with Korean as both languages are Ural-Altaic language

systems (Martin, 1966), which might explain why they achieved the fluency criterion, but the other participants did not.

Figure 2.

Fluency Training Outcomes for P1, P2, P3, P4, and P5



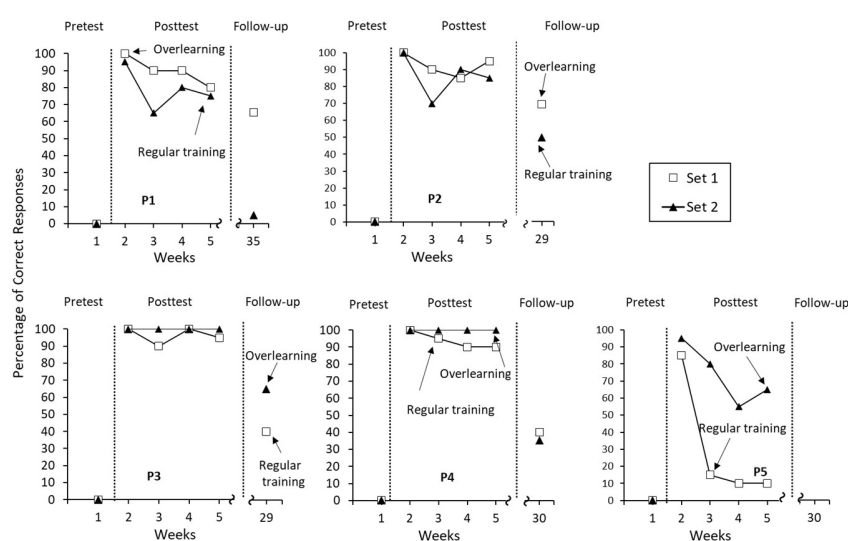
Note. Asterisks note the session in which the participant met the fluency mastery criterion. The y-axis uses a logarithmic scale.

6.42 Pre- and Post-Tests

All five participants completed four weeks of post-testing. Figure 3 shows the participants' performances during pre- and post-training tests for the untrained emergent NFI relations. Similar to FTT baseline sessions, none of the participants responded correctly during NFI pre-test trials. However, both training conditions produced high levels of emergent responding immediately following training. All participants scored at or above 85% correct responding during the initial NFI post-tests, and P2, P3, and P4 scored 100%. Overlearning (mean = 99%) produced slightly higher results during the first post-test than the regular training condition (mean = 96%). Furthermore, visual analysis of the level, trend, and stability of participants' performances over the four weeks of post-testing indicates that all five learners showed better overall retention of emergent responses in the overlearning condition.

Figure 3

Pre, Post, and Follow-up Test Outcomes for P1, P2, P3, P4, and P5



Note. P5's graph does not include follow-up results because she was not available for testing.

The differences between the two conditions increased as post-testing continued. By the fourth week of post-testing, three weeks after training, the mean difference between overlearning (99%) and regular training (96%) increased from 3% to 14% (85% and 71%, respectively). P5's performance with the regular training set appeared to degrade most among the five participants—85% correct responses during the first post-test and 10% three weeks later. P5 scored much higher during the first post-test with the overlearning stimulus set (95%) and the fourth post-test (65%) three weeks later. P3 and P4 maintained 100% correct responding during all four weekly post-tests, but only with the stimuli in the overlearning set. By the fourth post-test, P4's score with the regular training set had dropped to 90%, and P3's score was 95%. On the other hand, P2 and P1's post-test scores were reduced in both conditions. Still, they performed better in the overlearning condition than in the regular training condition.

6.43 Follow-up Tests

P1, P2, P3, and P4 completed follow-up testing between 27-33 weeks after training; P5 was unavailable to participate. P1's follow-up tests were conducted 33 weeks after training, scoring 65% for the overlearning set and 5% for the regular training set. Twenty-seven weeks post-training, P2 scored 70% (overlearning) and 50% (regular training), while P3 scored 65% (overlearning) and 40% (regular training). P4 was the only participant to score higher at follow-up in the regular training condition (40%) than in the overlearning condition (35%). Her follow-up results suggest that her failure to attain the fluency criterion negatively affected her retention at follow-up. Although, it should be noted that P4 was pregnant and gave birth between the post-testing and follow-up sessions, which could have influenced her follow-up results. Still, her overall mean score in the overlearning condition (87%) was slightly higher than in the regular

training condition (83%, Table 3). On the other hand, P3 showed improved retention at follow-up in the overlearning condition despite not achieving the fluency criterion. However, P3's highest response rate (78 corrects per minute) was much closer to the criterion than P4's (66 corrects per minute).

Table 3

Mean post-test and follow-up scores for each participant

Participant	Number of test sessions	Mean score (Overlearning)	Mean score (Regular training)	Range
P1	10	85%	64%	5-100
P2	10	88%	79%	50-100
P3	10	93%	83%	40-100
P4	10	87%	83%	35-100
P5	8	74%	30%	30-95

6.5 Discussion

We compared repeated measures of emergent intraverbal responses following two FTT conditions—regular training and overlearning. Despite no direct training in NFI relations, all participants demonstrated high levels (> 85% correct) of emergent NFI responding following both conditions. The results reflect those of previous studies in which FTT was effective in producing emergent foreign language intraverbals (Cortez et al., 2020, 2022; Daly & Dounavi, 2020; Dounavi, 2011, 2014; Matter et al., 2020; May et al., 2019; Petursdottir & Haflidadóttir, 2009; Petursdottir et al., 2008; Wu et al., 2019).

This study extends previous FTT research by examining the effects of overlearning and is the first to evaluate the impact of adding fluency criteria to regular FTT procedures. All five participant's mean post-test scores were higher in the overlearning condition and showed increased levels of emergent responding compared with training to accuracy criterion alone, which is consistent with previous research on overlearning and emergent relations (e.g., Binder, 1979; Bortoloti et al., 2013; Bucklin et al., 2000).

The results also indicated that the addition of fluency-building procedures in the overlearning condition increased participants' retention of emergent responses. However, it is uncertain whether better retention was a function of increased practice time rather than the fluency criterion (Doughty et al., 2004). Still, once learners achieve 100% accuracy, it remains uncertain how much additional practice might be needed to improve retention (Binder, 2004).

Instead of relying on accuracy criterion alone, researchers and instructors might consider what fluency criterion predicts long-term learning retention (Binder, 2004). Including fluency criteria may assist instructors in pinpointing a measurable target to aim for during overlearning. In our study, the fluency criterion was between 100-80 corrects per minute, with no errors. Because we limited the number of training sessions, only two participants, P1 and P2, achieved this criterion. Still, three of four participants showed improved retention at 6-7 months in the overlearning condition. P4 was the only participant whose follow-up performance was not improved with overlearning; interestingly, her response rate during training was also the lowest among the participants who completed follow-up testing.

To our knowledge, this study is also the first to examine retention following FTT using multiple (i.e., more than two) post-tests. We conducted 8-10 repeated measures post-training with each participant, which allowed us to examine the level and trend of participants' emergent

responses through visual analysis. Notably, the level and trend appeared higher and more stable for the overlearning stimulus sets than the regular training stimuli. Another finding of interest was the observed effect of repeated weekly testing on retention. Cowley et al. (1992) argued that repeated testing of untrained relations can establish more consistent emergence of equivalence relations. In the current study, weekly testing during the first three weeks following training appeared to produce more stable responding. However, we also observed a ceiling effect where performance was measured using percentage correct. More differentiated results may have been noted had we also measured participants' response rates during the post- and follow-up testing phases.

The present study compared the emergent foreign language learning outcomes from overlearning (accuracy and fluency criteria) with the outcomes produced by regular training (accuracy criterion only). The findings are similar to Bucklin et al. (2000) in that overlearning increased learners' retention of emergent learning. A limitation of this study is that two of the five participants failed to achieve the accuracy criterion before commencing the overlearning phase. As a result, we cannot be certain that this did not affect their post-training results. However, as previously stated, P5's performance was comparable across both sets during the regular training condition. Furthermore, because P4 achieved mastery with the regular training stimulus set rather than the overlearning set, and then participated in additional fluency-based trials with the overlearning set, it is unlikely that her post-test results were affected significantly. Future research should continue training to criterion to directly compare the criterion-accuracy level and criterion-fluency level mastery criteria to verify whether the criterion-fluency level adds any benefits with respect to response maintenance.

Additionally, we did not test tact relations during post-tests to avoid potentially influencing the participant's retention of the derived relations through repeated testing. However, it would be interesting to see whether the effects we observed for the emergent NFI relations are similar to the trained tact relations. Bucklin et al.'s (2000) results suggested that retention increased for both the trained and untrained relations, and future studies should also evaluate both.

Finally, we recommend that future FTT research systematically isolate the effects of fluency criteria on emergent learning outcomes while controlling for practice effects. For example, a potential experiment could compare learning outcomes following a 'fast' training condition—high rate-aim and free-operant responding during timed trials—with a 'slow' training condition—low-rate aim and restricted operant responding using an intertrial delay to depress responding. Although overlearning and fluency-based training procedures may potentially improve retention, it is unclear whether this is a function of the fluency criterion, the increased practice time, or both.

Chapter 7: Fluency-building and the testing effect

7.1 Introduction

This chapter reviews the first experiment (Study 3, Chapter 6), and analyses the factors that influenced the retention outcomes observed across the two training conditions. It discusses the role that response rate played in these outcomes and raises questions about whether the fluency criterion is a required component for improved retention or if the improved retention observed in the fluency-building condition was a function of extended practice testing rather than response rate.

To further understand the mechanisms underlying the improved retention observed in Study 3, this chapter examines whether the findings are characteristic of the testing effect. Additionally, understanding the testing effect can help refine instructional strategies to incorporate optimal retrieval practices, enhancing the retention of emergent language skills.

The discussion then explores how the dosage of practice tests—meaning the frequency and extent of testing—may influence retention outcomes. To investigate this, the chapter proposes a comparison between free-operant training conditions, where learners have continuous opportunities to respond, and restricted-operant training conditions, where the number of response opportunities is limited. By comparing these training conditions, the study aims to determine the optimal practice test dosage for maximising retention and understand how different training methodologies impact emergent learning and memory consolidation.

7.2 Summary of Study 3

7.2.1 Retention in Study 3

Study 3 (Chapter 6) compared the outcomes produced by a modified FTT protocol that included fluency-based overlearning trials with a more traditional discrete-trial based FTT procedure. Repeated post-training tests of emergent intraverbal responding across the two FTT conditions revealed better retention in the modified procedure that included fluency-building, and the differences became more pronounced as post-testing continued. After four weeks, participants stopped practice and testing for at least five months. Tests conducted 6-7 months after training, revealed that three out of four participants showed improved retention following the overlearning trials included in the modified FTT procedure.

7.2.2 Response rate or. extended practice?

It was hypothesised that fluency-building would improve emergent learning outcomes by increasing the response strength (i.e., response rate) of directly trained tacts. During post-testing, fluency-building consistently produced better retention scores than regular training, with the difference between the two conditions increasing over time. Follow-up tests confirmed that overlearning led to more durable retention in three of four participants, as reflected in higher mean scores across participants compared to regular training.

The observed retention improvements align with precision teachers' claims that skills taught to fluency tend to remain in the learner's repertoire for an extended duration (Johnson & Street, 2004; Pennypacker et al., 2003). However, only two participants met the fluency criterion, while the other three participants fell short, suggesting that the fluency criterion was not a required component for improved retention for all participants. It is possible that these improvements resulted from repeated practice rather than from an increased response rate

(Doughty et al., 2004), as training duration was longer, and participants engaged in substantially more trials in the fluency-based overlearning condition. The key question arising from the first experiment was: *What role does extended practice play in retention?*

Fluency-building typically provides more practice opportunities than traditional discrete-trial training approaches because it minimises response constraints and encourages learners to pace themselves and respond to as many stimuli as they can in the time allotted to training (Lindsley, 1996). According to the testing effect, increased practice through repeated testing (i.e., more learning trials) is one of the most effective strategies for improving retention (Carpenter et al., 2022, Polack & Miller, 2022). Learners who practice retrieving previously learned information or skills retain more than learners who do not. During fluency-building procedures like those employed in Study 3's overlearning condition, learners engaged in free-operant trials and high rates of retrieval practice.

7.3 Free-operant versus restricted-operant training

7.4.1 Free-operant training

In precision teaching, free-operant practice is seen as important for developing fluent performance and improving retention (Johnson & Layng, 1996). Free-operant procedures enable participants to complete as many trials as possible within the allotted practice time. Here, the learner controls the pace of responding, eliminating the need to wait for the instructor to present stimuli (e.g., self-paced flashcards).

7.4.2 Restricted-operant training

Restricted-operant training, on the other hand, sees the instructor controlling the presentation of stimuli as well as the time between reinforcement and subsequent presentation of

stimuli (Cooper et al., 2007). In contrast to free-operant training, restricted-operant trials introduce interruptions or limitations on participant performance, reducing the number of presentation trials in each instructional period, and potentially limiting the learner's response rate (e.g., discrete trial training; Lindsley, 1996).

There are few studies to date comparing free- and restricted-operant learning arrangements in the retention literature, and among them the results are mixed (e.g., Kong, 2009; Mathews, 2010; McGregor, 2006; Porritt et al., 2009; Wheatley, 2005). Understanding the effects of these different training conditions on retention could provide valuable insights into optimising instructional strategies for long-term learning.

One relevant study that highlights the differences between these training conditions is Porritt et al.'s pigeon study, which provides valuable data on the impact of free- and restricted-operant practices.

7.4.3 Studies with animals

Porritt et al. (2009) demonstrated positive effects from free-operant practice on pigeons' response rates and retention levels. The study alternated three training conditions: no delay (free-operant), delays within- and between-response chains (restricted-operants). During the within-chains condition, a five-second delay was inserted between every key press. In the between-chains condition, the experimenters added a 15-second delay after the third key press in the chain and before presenting the next discriminative stimulus in the subsequent chain. The experimenter yoked the number of trials in the delay conditions to the amount completed in the no-delay condition. The no-delay condition produced the shortest mean response latency, which meant that the pigeons' reaction time was quickest and response rate highest in the free-operant condition. Porritt et al. attributed this to the free-operant condition's lack of an intertrial delay.

They contended that the intra- and intertrial delays implemented during the restricted-operant conditions limited the pigeons' response rates. Overall, free-operant trials produced significantly higher response rates and levels of retention than the restricted-operant conditions. Porritt et al. cautioned, though, that their findings might not be generalisable to human participants.

7.4.4 Studies with human participants

Several studies with human participants (e.g., Darvell, 2006; Kong, 2009; Mathews, 2010; McGregor, 2006; Wheatley, 2005) evaluated the impact of delays within and between trials but failed to replicate the positive retention results observed by Porritt et al. (2009). Similar to Porritt et al.'s study with pigeons, these studies involving human participants ensured that the total number of trials remained the same across both conditions. Wheatley inserted a one-second intra-trial delay between presentations of match-to-sample stimuli. Retention tests showed mixed results: two participants performed better in the free-operant condition, while the other two performed better in the restricted-operant condition. Wheatley identified the study's short retention interval as a limitation and suggested that future research increase the time between training and testing. Additionally, the results showed no difference in response rates across conditions for three participants, which means that the one-second delay failed to suppress fluency in the restricted-operant condition.

Both McGregor (2006) and Kong's (2009) experiments imposed a three-second intertrial delay. McGregor observed minimal differences in retention accuracy. Ceiling effects may have confounded the results, though, as each stimulus set only contained five number facts and the lowest retention score across all fourteen tests was four out of five correct responses. In other words, the difference between full retention and the lowest score was just one error. If McGregor had employed a longer retention interval, used more number facts in each condition, or measured

response rates during retention tests, the results may have been more differentiated. Kong's findings indicated an increased rate of responding for tertiary students participating in free-operant practice, but relatively stable and high levels of accuracy scores in all conditions from the first to the last post-test. Free-operant training yielded the highest post-test scores but only marginally better than those in the restricted-operant condition. Kong suggested that additional post-tests and longer retention intervals may have produced more substantial differences in retention rates within and across conditions.

Mathews (2010) employed a 10-second delay between flashcard presentations. During retention tests, participants achieved high accuracy levels in both conditions. Mathews did not assess the terminal response rates, though, and fluency levels may or may not have been the same across both conditions.

7.5 Conclusion

Taken together, these studies' findings indicate that free-operant trials produced response rates and retention levels comparable with the same number of restricted-operant trials. Only Porritt et al.'s (2009) findings support the notion that free-operant practice and fluency-building positively impact response rate and retention when the total number of training trials are held constant across conditions. The few similar studies to date conducted with human participants did not yield consistently positive results, casting doubt on the purported advantages of fluency-building and free-operant practice (Darvell, 2006; Kong, 2009; Mathews, 2010; McGregor, 2006; Wheatley, 2005).

It is important to recognise that limitations within these studies make it difficult to draw definitive conclusions about their findings. Specifically, researchers reported undifferentiated increases in some learner's response rates across free- and restricted-operant conditions

(McGregor, 2006; Wheatley, 2005), potential ceiling effects (Mathews, 2010; McGregor, 2006), and relatively short retention intervals (Kong, 2009; Mathews, 2010; Wheatley, 2005).

Considering these limitations, the researchers recommended future studies employ more post-tests and longer retention intervals with the aim of revealing whether differences exist in retention rates across conditions.

A final consideration relates to the decision to equate the number of trials across free- and restricted-operant conditions, which researchers argue controls for confounds associated with practice. Doing this, however, removes the inherent advantage that free-operant training provides over other approaches—more training trials per training time. Additionally, it produces comparatively longer session durations and training for restricted-operant practice; and long-periods of practice may be punishing and aversive for learners (Binder, 1996). The issue of instructional time is an important factor in any learning program, particularly in emergent learning programs, where the primary goal is to achieve the most efficient and impactful results within the constraints of limited instructional time (Critchfield, 2018).

7.6 Introduction to the second and third experiments

The thesis's second experiment (Study 4) was designed to address the two main confounds present in Study 3 (Experiment 1). First, when the procedures limited the overlearning condition to 50 or fewer trial blocks, three out of five participants did not meet the fluency criterion. This may have influenced the results, as the participant who scored the lowest terminal training response rate was the only one whose retention accuracy at follow-up was higher in the regular training condition than in the overlearning condition. This also raises the question of whether fluency criteria are a required component for improved retention. To address these concerns, Study 4 (Experiment 2) continued training in the free-operant condition until all mastery criteria

were met while attempting to suppress the rate of responding in the restricted-operant condition using an intertrial delay.

Second, the extended practice employed during overlearning called to question whether the results were due to practice effects. Previous studies controlled for practice effects by matching the number of trials per condition. However, this usually results in longer session duration during restricted-operant practice conditions. In response to this, Study 4 (Experiment 2) aimed to systematically isolate the effects of fluency criteria while equalising practice duration, which involved directly comparing free-operant and restricted-operant procedures while holding total training time constant.

The specific research question evaluated in Study 4 was: *Does an intertrial interval during training affect the retention accuracy and response rate of emergent and directly trained learning during post-testing?*

It was expected that constraining participants' freedom to respond would limit the number of trials they were exposed to and negatively affect their rate of responding and post-training retention levels.

Study 5 followed Study 4 by examining whether modifying the FTT protocol by changing the testing arrangements during the first month post-training would provide additional improvements in retention. The question posed was: *What are the effects of weekly practice tests with and without feedback on retention of emergent and directly trained learning?*

Chapter 8: Studies 4 and 5 - More is not always better: The testing effect on retention of emergent Korean vocabulary

Wooderson, J. R., Bizo, L. A., & Young, K. (2024). More is not always better: The testing effect on retention of emergent Korean. [Manuscript submitted for publication]⁶

8.1 Abstract

Traditionally, testing is seen as a tool for assessing learning. However, research suggests it also enhances retention. Despite abundant research on the testing effect, its specific role in foreign language vocabulary learning remains unclear, particularly its effect on the retention of emergent, untrained verbal relations. To address this gap, the current study compared the retention outcomes of free-operant (fluency practice) and restricted-operant (accuracy practice) test trials with two English-speaking adults learning Korean vocabulary. Dependent variables included retention accuracy and response rates for emergent and directly trained vocabulary over a six-month period. Results from the first experiment in this study (Experiment 2) revealed no functional relation between testing type and retention. However, the second experiment (Experiment 3) demonstrated that retention improved to criterion levels when practice tests with corrective feedback were spaced weekly. These results suggest that increasing the dosage of practice tests had a minimal impact, while spacing practice sessions enhanced the retention of both directly trained and emergent vocabulary. The findings provide important considerations for foreign-language vocabulary learning.

Keywords

Testing effect, spacing effect, emergent learning, foreign tact training, foreign vocabulary acquisition

⁶ The paper was submitted to an American journal and uses American English spelling throughout.

8.2 Introduction

The ‘testing effect’ is a well-documented phenomenon in which testing not only assesses learning but also enhances it. Learners who engage in tests requiring retrieval of previously learned material (i.e., practice testing; Polack & Miller, 2022) tend to retain information more effectively compared to those who merely reread the material (Roediger & Nestojko, 2015). Despite robust evidence supporting the testing effect across various environments and learners (Carpenter et al., 2022), the role of practice testing in foreign-language vocabulary learning, particularly the types of practice arrangements that optimize learners’ retention of emergent, untrained vocabulary, remains unclear (Wooderson et al., 2022).

Foreign-language learning often involves acquiring both directly trained and emergent vocabulary—verbal relations that are not explicitly taught but are acquired incidentally (Critchfield & Twyman, 2014; Nation, 2022). Recently, emergent learning has become the focus of a growing body of literature on foreign vocabulary acquisition. This literature demonstrates the benefits of procedures targeting incidental learning, including apparent ‘free’ knowledge and skills that do not require direct programming (for reviews see Melvin-Brown et al., 2022; Wooderson et al., 2022). For example, foreign tact training (FTT; see picture/say foreign word) has proven effective in generating emergent intraverbals, mands, and listener behavior in addition to directly trained tacts. However, little is known about the long-term retention of emergent learning following FTT, as existing FTT research primarily focuses on short-term maintenance (i.e., one month or less) and discrete-trial training procedures (Wooderson et al., 2022).

Further research is needed as long-term retention is a critical outcome of effective vocabulary instruction. In an earlier study (Study 3), improved retention of emergent Korean

intraverbals was observed up to six months after training with fluency-based overlearning trials compared to training to an accuracy criterion. It remains unclear, though, whether the improvements resulted from increased practice due to longer training duration and more trials in the overlearning condition.

The current study aims to fill this gap by systematically examining how different practice testing methods affect the retention of emergent Korean vocabulary among English-speaking adults. Specifically, we compare the outcomes of free-operant and restricted-operant testing on the retention of both emergent and directly trained vocabulary over a six-month period. Free-operant practice allows learners to set their own pace and complete as many test trials as possible within the time available for practice, promoting high rates of retrieval practice, fluent performance, and optimizing retention (Evans et al., 2021; Martinho et al., 2021). In contrast, restricted-operant practice involves instructor-controlled presentation of test trials, often resulting in interruptions and fewer practice trials per instructional period (Lindsley, 1996, Johnson & Layng, 1996).

Despite these theoretical advantages, previous research has shown mixed results regarding the retention superiority of free-operant over restricted-operant practice (Darvell, 2006; Kong, 2009; Mathews, 2010; McGregor, 2006; Porritt et al., 2009; Wheatley, 2005). These studies attempted to control the amount of practice by yoking the number of restricted-operant test trials with the number of free-operant trials, resulting in the same number of test trials in both conditions but producing comparatively longer sessions and total training time for restricted-operant practice. This approach prevents a direct evaluation of testing dosage and its impact on retention. Several laboratory-based studies have demonstrated the benefits of increased practice test dosage (e.g., Pavlik & Anderson, 2005; Roediger & Karpicke, 2006; Wheeler & Roediger,

1992). However, few applied studies to date have manipulated test trial quantities, and those that did showed small effect sizes (e.g., Foss & Pirozzolo, 2017; Leeming, 2002), indicating a need for further research examining the effects of practice test dosage on retention (Agarwal et al., 2021).

We hypothesize that free-operant practice testing will lead to better retention due to the increased number of practice test trials in comparison with restricted-operant practice. To our knowledge, no study has directly assessed the effects of free-operant high-density testing on the retention of emergent foreign vocabulary learning. To investigate this hypothesis, we conducted two experiments with native English-speaking adults learning Korean vocabulary. The first experiment compared retention outcomes of free-operant and restricted-operant testing methods with monthly retention probes over six months. The second experiment followed the unexpected results observed in Experiment 2 by evaluating the impact of weekly spaced tests following retraining of vocabulary not retained in the first experiment.

Experiment 2

The research question examined in Experiment 2 was: Does an intertrial interval during training affect the retention accuracy and response rate of emergent and directly trained foreign language vocabulary learning during post-testing? We anticipated that imposing a programmed response constraint during practice sessions would reduce the total number of trials and negatively impact both the response rate and retention levels.

8.3 Method

8.31 Participants and setting

This study involved two native English-speaking adults with graduate-level qualifications (Table 1). Both participants expressed an interest in learning a foreign language and volunteered

for the study. Participant 1 (P1), a 43-year-old male, had previously acquired basic proficiency in Spanish and Japanese while living abroad. Participant 2 (P2), a 27-year-old female, was raised in a bilingual Japanese-speaking household. Notably, neither had formal training in the Korean language, except for their participation in a previous experiment (Study 3). The previous and current experiments were conducted online and recorded using Microsoft Teams. Participants attended the sessions from their home personal computers.

Table 1

Participants' gender, age, ethnicity, education level, and experience with a second language.

Participant ID	Gender	Age	Ethnicity	Education level	Second language experience
P1	Male	43	European Australian	Graduate	Spanish, Japanese
P2	Female	27	Japanese Australian	Graduate	Japanese

8.32 Materials and stimulus sets

The study used 60 Korean words, organized into three categories and six stimulus sets (Table 2). We balanced the difficulty of these sets using procedures similar to those outlined by in Study 3. Each set contained 10 nouns, with an equal distribution of one-, two-, and three-syllable words, and distinct word configurations and contrasting visual shapes and colors. Sets were randomly assigned and counterbalanced across participants to control for any order effects. Additionally, the presentation order of stimuli was randomized using custom JavaScript code (Cariveau et al., 2021; Shepley et al., 2020). Each stimulus set's materials included a

pronunciation guide featuring pictures of the 10 target words alongside their corresponding Romanized Korean text. The guides were verified by a native Korean speaker.

Table 2

Target Korean words and their English equivalents

Body Parts				Animals				Common nouns			
Set 1		Set 2		Set 3		Set 4		Set 5		Set 6	
Korean	English	Korean	English	Korean	English	Korean	English	Korean	English	Korean	English
ohm-jee	thumb	dar-ree	legs	nuk-deh	wolf	sar-sum	deer	gorng	ball	chehk	book
ee-barl	tooth	ohk-geh	shoulder	geh	dog	sor	cow	gee-char	train	johp-shee	plate
parl-goom-chee	elbow	sorn-gar-rark	fingers	dark	chicken	marl	horse	mor-jar	hat	shin-barl	shoes
byarm	cheek	gwee	ear	dweh-jee	pig	yohm-sor	goat	jar-dorng-char	car	jar-john-goh	bike
moo-rup	knee	ip-sool	lips	mool-gor-gee	fish	gor-yarng-ee	cat	wee-jar	chair	yarng-marl	sock
mork	neck	noon	eye	goh-wee	goose	tar-jor	ostrich	york-jor	bath	darm-yor	blanket
ee-mar	forehead	moh-ree	head	nark-tar	camel	kool-bohl	bee	jip	house	gort	flower
sorn	hand	barl	foot	kor-ki-ree	elephant	geh-goo-ree	frog	ar-gee	baby	yohl-seh	key
kor	nose	tohk	chin	yarng	sheep	behm	snake	kar-bamg	bag	goh-ool	mirror
beh	belly	hyoh	tongue	geh-mee	ant	sar-jar	lion	mamg-chee	hammer	pen-chee	pliers

Training and testing materials consisted of three types of digital flashcards: picture cards, English cards (English printed words), and Korean cards (Korean Romanized printed words). In total, there were 180 flashcards, each measuring 400 x 380 pixels. Picture cards featured color images obtained from Google Image Search on a white background. English and Korean cards displayed text in black Arial font at 90-point size, also on a white background. Participants accessed these flashcards through a custom-developed HTML webpage coded in JavaScript and hosted on a private Amazon Web Services server.

The experimental program's on-screen controls included two buttons: 'start timer' and 'next card.' Upon pressing the 'start timer' button, an on-screen 60-second timer commenced. Flashcards were randomly selected from the target stimulus set by the software program.

Participants revealed flashcards by pressing the ‘next card’ button until the timer reached zero, at which point the on-screen controls were disabled. During pre-assessment, baseline, delay training, and the initial phase of no-delay training, only one flashcard at a time was displayed on the screen. During post-tests and the second training phase of the no-delay condition, all 10 flashcards from the target stimulus set were presented simultaneously in a 5x2 grid, arranged in randomized order.

8.33 Pre-experimental conditions

Before the first experiment, we conducted pre-assessment tests to ensure participants could correctly pronounce all 60 Korean words. The experimenter read the words aloud from the pronunciation guides and asked the participants to repeat what was said, without showing the pronunciation guide to them. During a second session the following day, participants were asked to identify the English words they preferred for each picture card. We ensured that the English card materials matched participants’ preferences, thereby mitigating potential confusion or errors during post-testing.

8.4 Experimental design and procedures

The experiment employed a concurrent multiple-probe across stimulus sets with an embedded adapted alternating treatments design. This approach compared the effects of free-operant (no-delay) tact trials and restricted-operant (delay) tact trials on the acquisition and retention of emergent and directly trained foreign language vocabulary. Monthly post-test probes tested retention, with participants instructed not to practice outside of session times. Ethical approval for the study was provided by the University of Technology Sydney Human Research Ethics Committee (ETH22-7405).

8.41 *Dependent variables*

The primary dependent variables were directly trained tacts (see picture card/ say Korean word) and emergent native-to-foreign intraverbal responses (NFI; see English card/ say Korean word). The final post-test session included emergent listener responses (see Korean card/ select English card) and foreign-to-native intraverbal probes (FNI; see Korean card/ say English word).

8.42 *General procedure*

All baseline, training, and post-test sessions were conducted as a series of one-minute trial blocks, referred to throughout the study as timings. Each timing lasted exactly sixty seconds across all conditions, and no feedback was delivered within timings. In the current study, we did not equate the number of trials because our aim was to compare training with more versus fewer practice trials, while keeping instructional time constant.

We conducted twelve to sixteen timings per day on weekdays. Immediately before the first timing of each session, the experimenter read the instructions for that condition (Study 3). Throughout all training sessions, prior to the first daily timing for each stimulus set, the experimenter showed the pronunciation guide onscreen for 30 seconds, then demonstrated and prompted the participant to repeat each Korean word once. At the conclusion of each 60-second timing, the same pronunciation guide was presented for an additional 30 seconds while any errors were reviewed by the experimenter and participant. While reviewing errors, the experimenter pointed to the associated on-screen picture and Romanized Korean text, then vocalized the correct pronunciation. To avoid any direct NFI or FNI training, the English referent was not used. The total duration of time participants spent viewing the pronunciation guides was consistent across both conditions.

We implemented the same number of timings for both training types to ensure equal duration of instruction across conditions. Additionally, we randomly alternated the order of presentation so that on some days, training commenced with the first stimulus set, while on other days, the second stimulus set was trained initially. Set assignment and sequencing were also counterbalanced across participants.

The experimenter continuously recorded correct and incorrect responses on electronic datasheets during baseline, training, and post-test timings. Correct responses were noted when the participant: a) vocalized the target Korean word in the presence of a picture card (tact) or English card (NFI), b) vocalized an equivalent English word in the presence of a Korean card (FNI; synonyms, plural and singular variations were acceptable), or c) selected the equivalent English flashcard when presented with the referent Korean card (listener response). Incorrect responses were recorded if participants responded with the wrong word, 'don't know', the wrong verbal operant (e.g., saying an English word in response to a picture card), or selected the wrong flashcard. The experimenter calculated response rates by counting the number of correct responses recorded per minute; accuracy scores were determined for each one-minute trial block by dividing the number of correct responses by the total number of responses (correct and incorrect), then multiplying the result by 100.

8.43 Baseline

Before training commenced, we conducted probe timings without feedback using NFI relations. Tact relations were not tested during baseline unless the participant emitted the correct NFI response, which none of the participants did. This was particularly important as we were simultaneously conducting tact training with other stimuli and wanted to avoid any potential interference.

8.44 No-delay training

No-delay tests consisted of free-operant procedures without programmed delays between trials. After each timing, the experimenter provided feedback on the number of correct responses (response rate). Participants were reminded of the fluency aim—100-80 tacts per minute with no errors for two consecutive timings (Study 3)—and were encouraged to improve their performance if they had not yet achieved this aim.

Observation indicated that presenting picture cards one at a time limited participants' response rate (Johnson & Layng, 1996) to 70-60 correct tacts per minute during the no-delay tests. Consequently, when performance failed to improve (i.e., less than 1.25 times the participant's previous day's best timing; Binder, 1996), we slightly modified the practice conditions to encourage increased rates of responding. In the adjusted acquisition phase, all 10 flashcards were shown on screen simultaneously within a randomly arranged 5x2 grid, and participants were instructed to tact each picture card from the top left to the bottom right before advancing to the next arrangement of cards.

8.45 Delay training

Delay training timings were restricted-operant procedures and implemented a three-second intertrial delay between each flashcard presentation (Kong, 2009; McGregor, 2006). Throughout delay timings, only one flashcard was presented per trial. After pressing the 'next card' button, the screen went blank for three seconds before the next picture card appeared on screen. Following each timing, the experimenter delivered feedback to the participant on the percentage of correct responses and encouraged them to improve their accuracy if they had not achieved the 100% correct criterion.

While feedback on response rate was not provided to participants during delay training, we recorded rate data for comparison with the no-delay condition. To adjust for the three-second intertrial delay, we used the following formula to calculate the response rate:

$$\text{adjusted rate} = \frac{\text{number of correct responses} \times 60}{60 - 3 \times (\text{number of correct responses} + \text{number of incorrect responses})}$$

This adjustment allowed us to fairly compare the response rates between the delay and no-delay conditions.

8.46 Post-tests

If the number of no-delay and delay timings were equal, the first NFI post-test probes occurred immediately after participants achieved the fluency criteria. If not, the experimenter conducted additional timings using the delay stimuli to ensure an equal number and duration of timings across both types of training. To prevent the pronunciation guide from influencing post-test results, it was omitted after the final training timing.

Immediately after the first NFI timings, one tact timing using the free-operant procedures was conducted to check participants' unconstrained response rates with the delay stimuli. Subsequently, all monthly post-test sessions included three NFI timings and three tact timings, in that order, with no performance feedback provided during or following post-test timings. Post-tests were repeated monthly for six months, with the final post-test session comprising twelve timings per set—3 x NFI, 3 x tact, 3 x FNI, and 3 x listener relations.

8.47 Interobserver agreement and treatment integrity

A second observer independently collected data on 31.6% of the timings across all baseline, training, and post-test phases via video recordings. Agreement between the experimenter and the second observer was scored when both recorded the same response for a

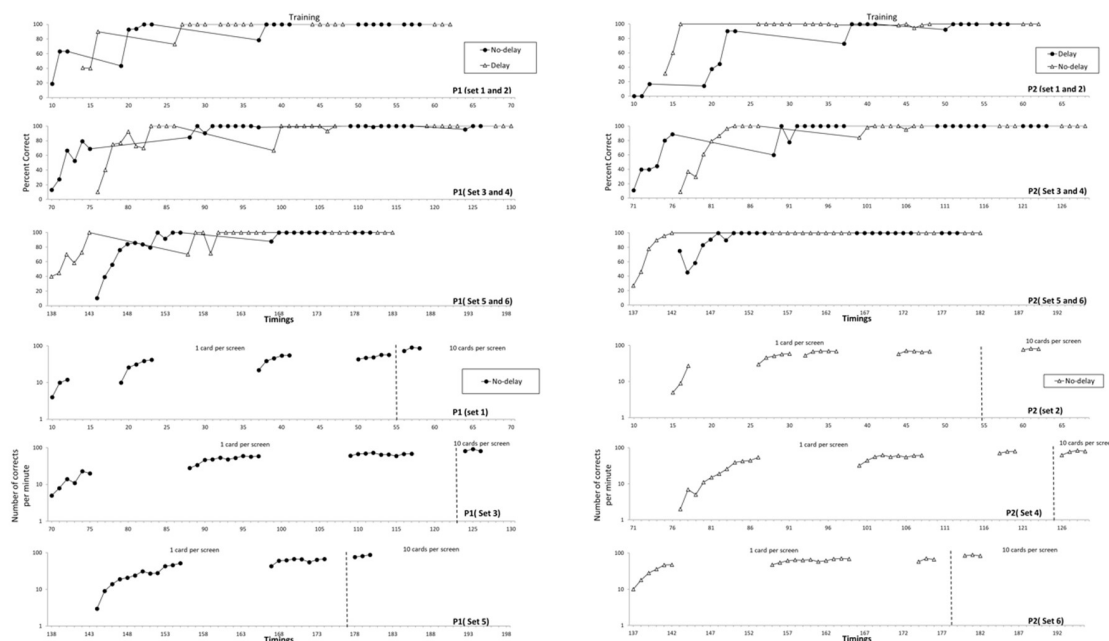
trial. The mean interobserver agreement was 99.7% for baseline and post-test phases, and 99.4% for training.

Treatment integrity was assessed by both observers. The first author reviewed all recordings and scored the experimenter's implementation of the procedures across all phases of the study: timings lasted sixty seconds, no instructional prompts and feedback were given during timings, and feedback was withheld for baseline and post-test timings. Mean treatment integrity across all observed trials was 99.9%, with 99.8% for baseline and post-test timings and 100% for training timings. In one post-test timing, the experimenter incorrectly stated the number of correct intraverbal responses when feedback should have been withheld. The second observer completed the same checks while viewing the interobserver assessment recordings. Agreement between the two observers was 100%.

8.5 Results and discussion

8.51 Training

Figure 1 displays the results of the tact training phases. The top three panels show the percentage of correct responses across both conditions and all six stimulus sets. Participant 1 required 43 timings to meet the initial accuracy criterion across all six sets, while P2 needed 41 timings. The bottom three panels show participants' response rates with the no-delay stimuli. Participant 1 achieved the fluency criteria across all three no-delay stimulus sets in 71 timings, whereas P2 took 72 timings. On average, P1 completed 16 overlearning timings (i.e., timings with 100% accuracy) per set, and P2 completed 16.5 overlearning timings per set.

Figure 1*Tact training phases*

Note. The y-axis for the bottom three panels is logarithmic. There was a phase change in which the number of picture cards increased from one card per screen to ten. Following this change, participants successfully achieved the fluency criteria within the next session.

Table 3 shows the total number of training trials conducted with each set: 1796 tact trials in the delay condition and 7552 for no-delay timings, which is more than four times as many trials overall. Despite this, total training time was the same for both conditions: 143 minutes ($M = 23.8$ minutes per set) for tact timings and 79 minutes ($M = 13.2$ minutes per set) spent viewing the pronunciation guides.

Table 3*Number of tact trials per stimulus set compared with post-test accuracy*

Participant	Training type	Stimulus set	Number of tact training trials	Mean NFI post-test accuracy
P1	No-delay	1	945	34.3%
		3	1525	41.8%
		5	1134	25%
	Delay	2	287	26.4%
		4	365	9.2%
		6	297	26.1%
P2	No-delay	2	1204	62.1%
		4	1350	16.1%
		6	1394	25.3%
	Delay	1	248	36.8%
		3	317	53.5%
		5	282	47%

Note. NFI = Native-to-foreign intraverbal

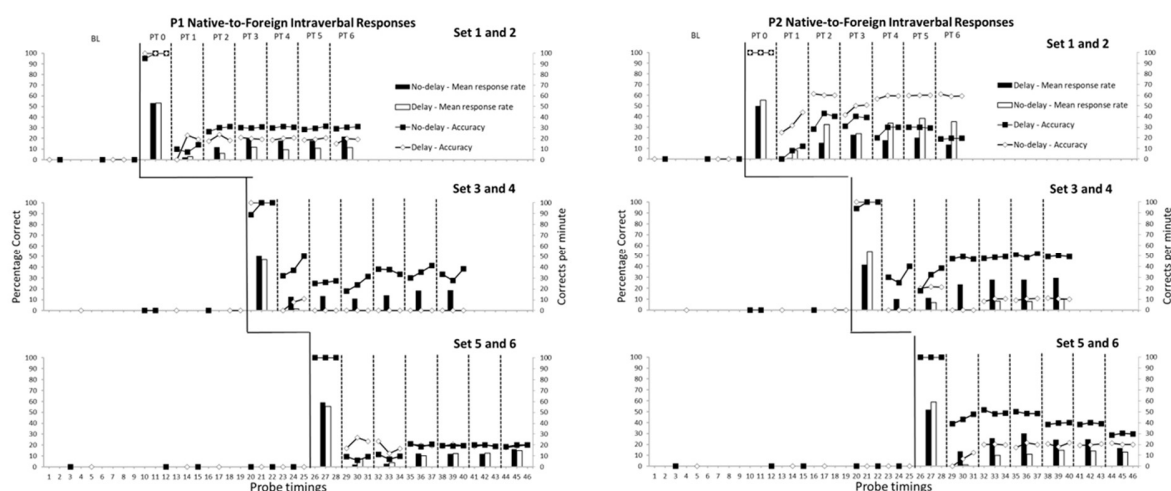
8.52 Native-to-foreign intraverbal post-tests

Neither participant responded correctly during NFI baseline probes (Figure 2). However, accuracy immediately following training was 100% in all but 3 of 36 initial NFI post-test timings. During these initial post-tests, mean response rates were slightly higher with the no-

delay stimuli ($M = 55.2$ corrects per minute) compared to the delay stimuli ($M = 49.9$ corrects per minute).

Figure 2

Baseline and Monthly Post-Test Intraverbal Probes



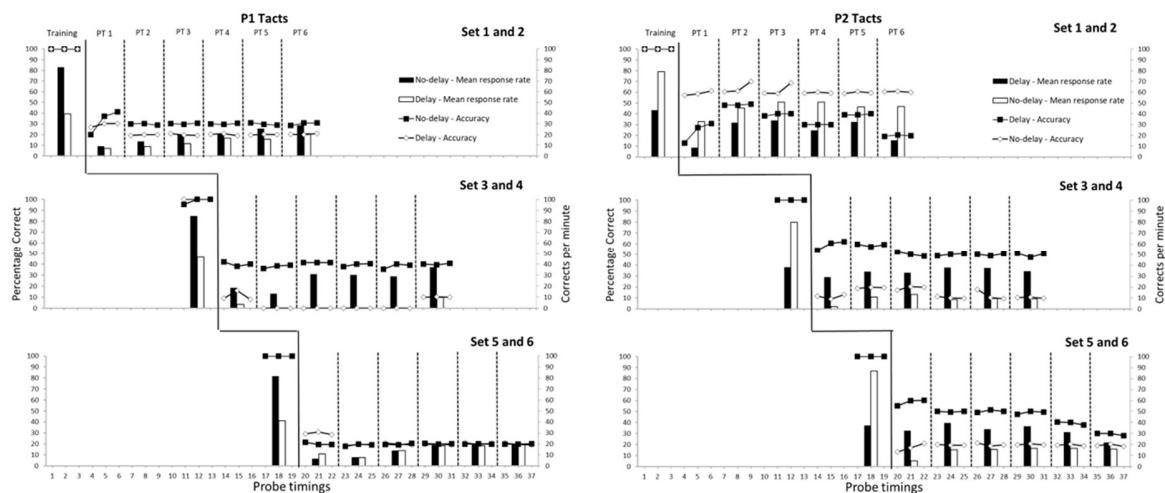
Note. PT = Post-test. BL = Baseline. The data points shown in the line graphs represent native-to-foreign intraverbal accuracy probes and are measured as percentage of correct responses on the left-hand y-axis. Columns represent mean response rate and are measured as corrects per minute on the right-hand y-axis. Post-tests were conducted monthly up to six months post-training.

Both participants' performance levels decreased substantially one month after training, then stabilized at low levels across the remaining retention testing period. Participant 1's mean accuracy for the six months following initial post-tests was 19% (no-delay = 25%, delay = 13%). Visual analysis of P1's retention accuracy and response rate data suggests better performance with Set 1 (no-delay), Set 3 (no-delay), and Set 6 (delay). Participant 2's mean accuracy for the same period was 32% (no-delay = 26%, delay = 37%). Her retention accuracy and response rate were highest for Set 2 (no-delay), Set 3 (delay), and Set 5 (delay).

A Mann-Whitney U independent samples t-test indicated no significant differences between NFI post-test scores for no-delay stimuli ($Mdn = 7.4$) and delay stimuli ($Mdn = 7.8$), $U = 1692, p = .571$. Taken as a whole, the data suggest the absence of a functional relation between training type and retention. Furthermore, Spearman's rank-order test revealed no significant correlation between the quantity of training trials conducted with each stimulus set and the mean percent correct score observed during post-test timings, $r(11) = -.252, p = .792$.

8.53 Tact post-tests

We conducted monthly tact probes following NFI probes. Visual analysis showed that tact post-test performance closely matched NFI post-test scores (Figures 2 and 3). Spearman's rank-order test revealed a significant strong positive correlation between tact and NFI post-test scores, $r(216) = .860, p < .001$. Participants' post-test accuracy was slightly higher with tact relations ($M = 30.2\%$) than with NFI relations ($M = 25.3\%$). A Wilcoxon Signed-Ranks test indicated that this difference was unlikely to be due to chance ($Z = 4.6, p < .001$). Participant 1's mean accuracy score was 23% for tact post-test timings and 19% for NFI post-test timings. Similarly, P2's mean accuracy score was 38% for tacts and 32% for NFI tests.

Figure 3*Training and Monthly Post-Test Tact Probes*

Note. PT = Post-test. The data points shown in the line graphs represent tact accuracy probes and are measured as percentage of correct responses on the left-hand y-axis. Columns represent mean response rate and are measured as corrects per minute on the right-hand y-axis. Training probes were the final three timings conducted with each stimulus set. Post-tests were conducted monthly up to six months post-training.

The only notable differences between no-delay and delay training were observed in the terminal training response rates. The mean response rate recorded with no-delay stimuli during the final three training timings was 82.5 corrects per minute, approximately double the rate observed with delay stimuli (40.9 corrects per minute). However, when we re-tested the delay stimuli using the tact relations with one post-test timing immediately following training—displaying all 10 flashcards on screen simultaneously—much higher response rates were observed (data not shown). In fact, P1’s rate of responding with the delay stimuli met fluency criteria, with a mean of 84 corrects per minute (range: 81–87), comparable to the terminal response rate he achieved with the no-delay stimuli. Participant 2’s tact response rate with the

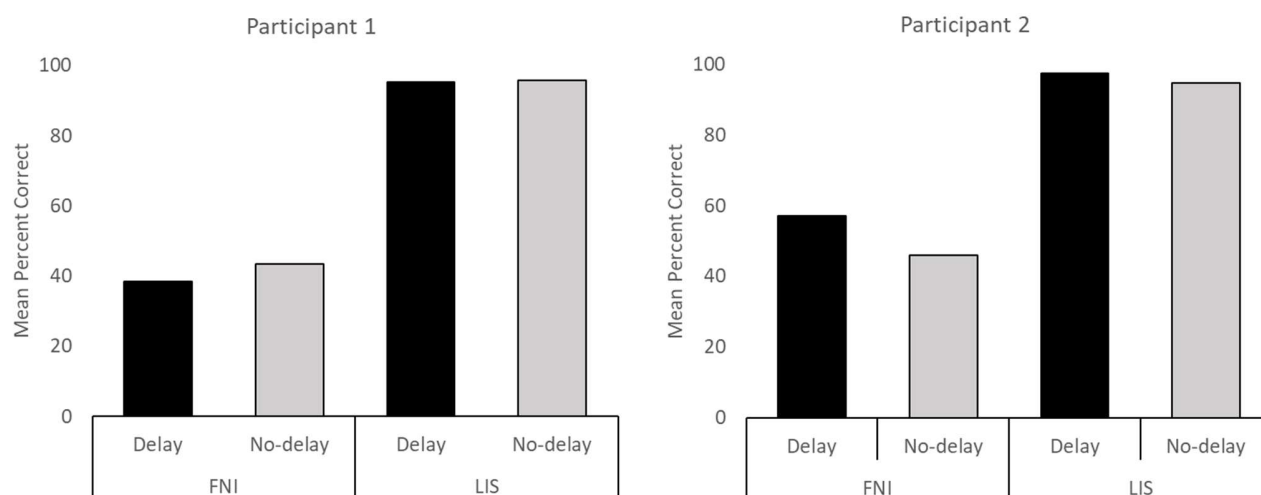
delay stimuli also improved under free-operant conditions but fell short of the fluency criteria, with a mean of 68 corrects per minute (range: 66 – 71).

8.54 Foreign-to-native intraverbal and listener post-tests

During the final post-test session, FNI and listener response probes were added (Figure 4). Six months post-training participants demonstrated high levels of emergent listener responses ($M = 95.8\%$, range: 80% – 100%) without previous direct training or testing with these relations. Foreign-to-native intraverbal responses also emerged without training or prior testing ($M = 46.2\%$, range: 12.9% – 80%). Notably, both participants scored higher in the reverse-intraverbal FNI tests than the NFI and tact timings during the final post-test probes.

Figure 4

Foreign-To-Native Intraverbal Probes and Listener Probes



Note. FNI = Foreign-to-native intraverbal. LIS = Listener response.

8.55 Post-hoc analysis of target words

Finally, we conducted statistical analyses to identify any differences between the experimental stimuli that may have impacted retention. Three non-parametric one-way ANOVA tests were performed to compare the effects of word category (body parts, animals, or common nouns), number of syllables (1,2 or 3), and the initial vowel or consonant sound on participants' accuracy during NFI testing. Kruskal-Wallis tests showed no statistically significant differences between words based on category ($H(2, 120) = 3.36, p = .187$), number of syllables ($H(2, 120) = 1.58, p = .454$), or initial sounds ($H(17, 120) = 21.2, p = .218$).

8.56 Preferences

After the final post-test session, the experimenter asked the participants which condition they preferred. Both indicated a preference for the free-operant (no-delay) condition.

8.6 Summary

In Experiment 2, the inclusion of an intertrial interval during training did not appear to affect the retention accuracy and response rate of emergent and directly trained foreign language vocabulary learning during post-testing. High levels of accurate and fluent performance were observed with directly trained and emergent Korean vocabulary immediately following both FTT conditions. However, subsequent delayed post-tests revealed a substantial drop in performance just one month later. Overall retention levels were low in both conditions throughout the six-month retention testing period.

Experiment 3

Experiment 3 followed up on Experiment 2 by examining whether changes to the testing arrangements during the first month of post-training would improve retention. This approach was

inductive, guided by the earlier finding that weekly post-testing in Experiment 1 produced superior retention outcomes during the first month post-training. Specifically, Experiment 3 evaluated the effects of spaced weekly tests after retraining tact relations that participants failed to retain in the first experiment, addressing the research question: Do weekly practice tests and feedback affect the retention of emergent and directly trained foreign language vocabulary learning?

8.7 Method

8.71 Participants and setting

Experiment 3 commenced approximately one month after Experiment 2's final post-tests. The participants and settings remained unchanged from Experiment 2.

8.72 Materials and stimulus sets

Experiment 3 utilized a subset of the materials from Experiment 2, including 30 words that required retraining (Table 4). Specifically, any Korean word (tact and NFI relations) that a participant could not remember during the final five months of post-testing was included in the pool of potential targets; words emitted correctly on one or more occasions were excluded. Subsequently, we randomly assigned thirty words across three stimulus sets (10 words per set) for each participant.

Table 4*Target Korean (English) words*

Participant 1			Participant 2		
Set 1a (testing with feedback)	Set 2a (testing without feedback)	Set 3a (delayed testing)	Set 1b (testing without feedback)	Set 2b (testing with feedback)	Set 3b (delayed testing)
noon (eye)	behm (snake)	gorng (ball)	ee-barl (tooth)	moo-rup (knee)	ohm-jee (thumb)
gort (flower)	geh (dog)	sorn (hand)	sar-sum (deer)	ee-mar (forehead)	geh-mee (ant)
mork (neck)	jip (house)	marl (horse)	dar-ree (legs)	sar-jar (lion)	yohm-sor (goat)
gwee (ear)	chek (book)	tohk (chin)	goh-wee (goose)	kool-bohl (bee)	mor-jar (hat)
york-jor (bath)	wee-jar (chair)	darm-yor (blanket)	tar-jor (ostrich)	gee-char (train)	johp-shee (plate)
gee-char (train)	marn-g-chee (hammer)	ee-mar (forehead)	marn-g-chee (hammer)	yarn-g-marl (sock)	yohl-seh (key)
goh-wee (goose)	ee-barl (tooth)	johp-shee (plate)	ar-gee (baby)	darm-yor (blanket)	pehn-chee (pliers)
sar-sum (deer)	ar-gee (baby)	kool-bohl (bee)	gort (flower)	sorn (hand)	gwee (ear)
dar-ree (legs)	yohl-seh (key)	dweh-jee (pig)	chek (book)	geh (dog)	sor (cow)
gor-yang-ee (cat)	sorn-gar-rark (fingers)	geh-goo-ree (frog)	sorn-gar-rark (fingers)	jar-john-goh (bike)	geh-goo-ree (frog)

8.73 Response measurement and dependent variables

Experiment 3 employed the NFI and tact response measurement procedures and dependent variables from Experiment 2.

8.74 Experimental design and procedures

A concurrent delayed multiple-baseline design across conditions and stimulus sets was used to assess the effects of weekly tests and corrective foreign tact feedback on retention accuracy and response rates with emergent and directly trained relations.

8.75 Baseline

Prior to retraining, three NFI probes under extinction were conducted with each stimulus set to confirm that participants could not verbally produce the foreign word when presented with

its native equivalent. As with Experiment 2, tact relations were not tested during baseline unless the participant emitted the correct NFI response.

8.76 Retraining

After the baseline phase, participants completed tact retraining using the procedures established in the second acquisition phase of Experiment 2's no-delay training condition, where all 10 picture cards were presented simultaneously on the screen. Unlike Experiment 2, we did not attempt to maintain the same number of timings and training duration across stimulus sets. The retraining mastery criteria were the same across all three comparison conditions. Participants were required to achieve 100-80 tacts per minute with no errors for two consecutive timings with each stimulus set before commencing post-tests.

Post-test procedures

Post-tests were conducted using three procedures: weekly-testing-with-feedback, weekly-testing-without-feedback, and delayed-testing. We randomly assigned stimulus sets and counterbalanced the sequencing for each of these conditions. Each post-test session included three NFI timings followed by three tact timings. The initial post-tests took place immediately after participants achieved the fluency criteria. The final post-tests occurred four weeks later, during which no modelling or feedback was provided to participants.

Weekly-testing-with-feedback

Throughout all timings during the weekly post-test sessions with the weekly-testing-with-feedback condition, the experimenter told the participant their response rate, presented the pronunciation guide for thirty seconds, and modelled any words the participant failed to emit correctly. Consistent with Experiment 2's general procedure, the experimenter provided corrective foreign tact feedback, excluding any English referents to avoid direct training with the

NFI relations. The pronunciation guide was withheld before the first timing to assess retention levels after one week without practice. Post-tests were conducted weekly for four weeks. During the final post-test, the experimenter gave only neutral feedback after each timing, and the pronunciation guide was not shown.

Weekly-testing-without-feedback

Testing-without-feedback sessions were also conducted weekly until the final session, four weeks post-training. Only neutral feedback was given after each timing, and the pronunciation guide was not shown.

Delayed-testing

In the delayed-testing condition, there were just two post-test sessions: the initial post-training tests and the final session four weeks later. Throughout delayed-testing, only neutral feedback was given, and pronunciation guides were not shown to participants.

8.77 Interobserver agreement and treatment integrity

Agreement and integrity checks were conducted in the same manner as Experiment 2 for 30.9% of the timings across all baseline, retraining, and post-test phases. Mean interobserver agreement was 99.1%. Treatment integrity was maintained at 100% across all observed trials, with 100% agreement between the two observers.

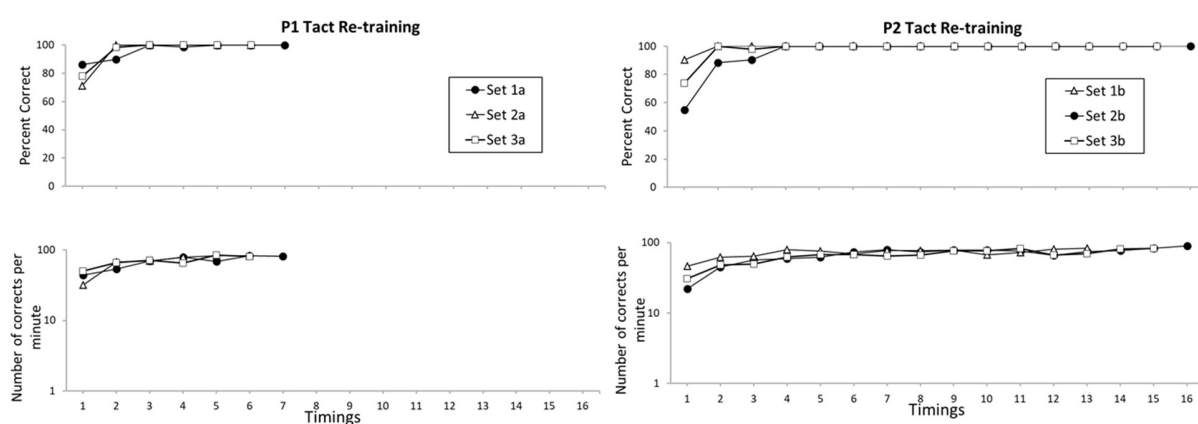
8.8 Results and discussion

8.81 Retraining

Figure 5 depicts the results of the retraining phases. Participant 1 achieved the acquisition criteria with Set 1a in 7 timings (10.5 minutes total training time); Sets 2a and 3a, required 6 timings each (9 minutes). Participant 2 took 16 (24 minutes), 13 (19.5 minutes), and 15 (22.5 minutes) timings respectively for Sets 1b, 2b, and 3b.

Figure 5

Tact retraining accuracy and response rates



Note. The three stimulus sets were taught on consecutive days, not concurrently, but are displayed above to compare the differences in acquisition rates. The y-axis for the bottom panel is logarithmic.

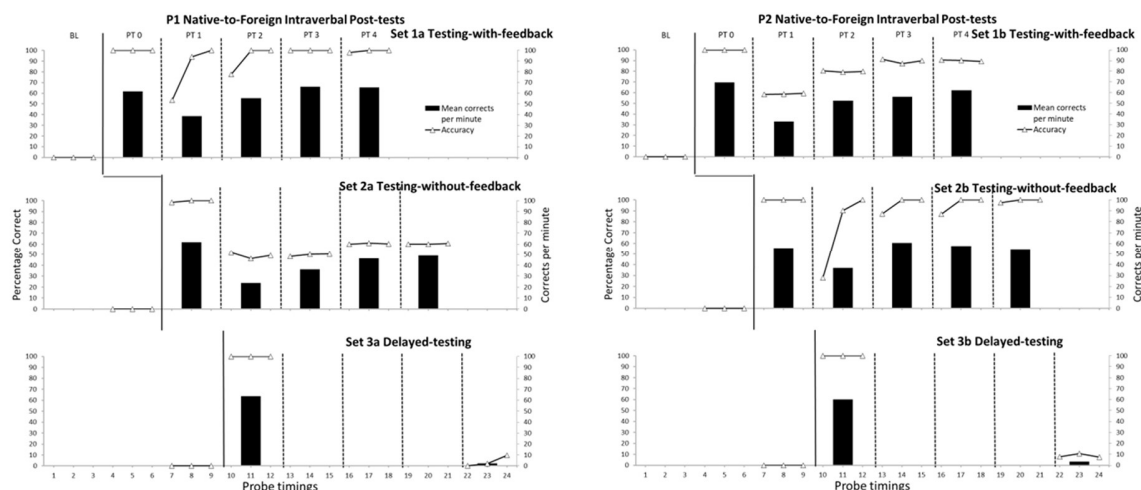
8.82 Native-to-foreign intraverbal post-tests

All baseline NFI probes scored 0% (Figure 6). Immediately after training, performance with the untrained NFI relations improved across all stimulus sets for both participants—17 of the 18 initial NFI probes scored 100%, and the mean response rate was 62.1 corrects per minute (range: 55.7 – 69.7). Similar to Experiment 2, performance decreased substantially at the second post-testing session. Participant 1's accuracy changed from 100% to 53.3% in Set 1a, and 100%

to 51.9% in Set 2a. Participant 2's performance went from 100% down to 58.3% with Set 1b, and 100% down to 28% with Set 2b.

Figure 6

Baseline and Post-Test Intraverbal Probes



Note. PT = Post-test. BL = Baseline. The data points shown in the line graphs represent native-to-foreign intraverbal accuracy probes and are measured as percentage of correct responses on the left-hand y-axis. Columns represent mean response rate and are measured as corrects per minute on the right-hand y-axis. Post-tests were conducted immediately following training then weekly up to four weeks post-training, or once after four weeks.

Over the final three weeks, the two participants' performance levels gradually improved through weekly testing— with and without feedback. Improvements were most evident from the results of the first timing conducted each week, which assessed the number of words each participant retained from the previous week's tests. Participant 1's accuracy with the weekly-testing-with-feedback set increased from 78% to 98%, and from 48% to 60% with the weekly-

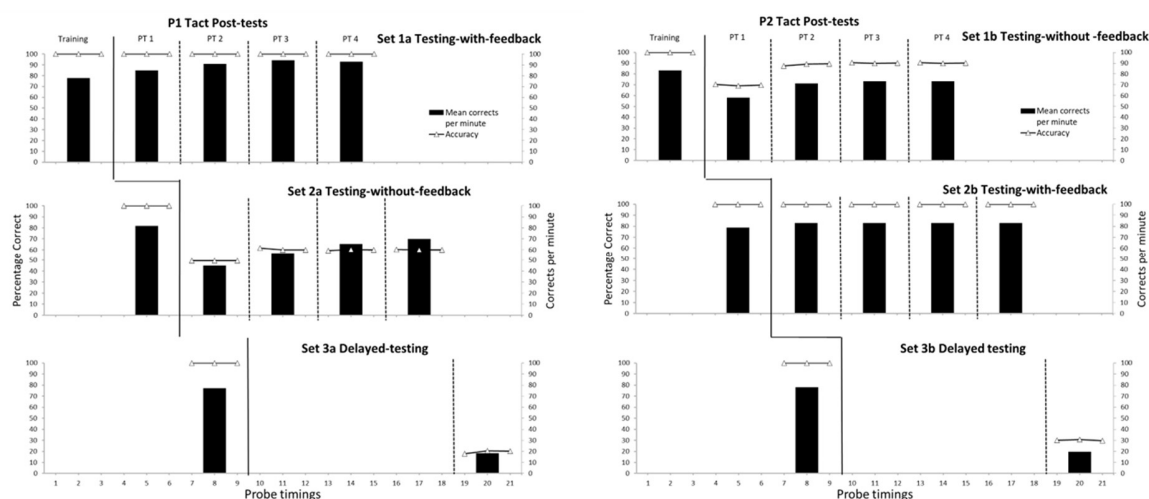
testing-without-feedback set. Participant 2 improved her accuracy with the weekly-testing-with-feedback set from 80% to 90%, and 87% to 98% with the weekly-testing-without-feedback set.

The greatest improvements were observed in the tests that included corrective foreign tact feedback. Consequently, both participants' performance with the weekly-testing-with-feedback stimuli exceeded 97% accuracy in the final week of testing (range: 97.5 – 100) and their response rates were similar to or exceeded the initial post-test results. The results for the delayed-testing stimuli, on the other hand, dropped to levels that were comparable with baseline. By the fourth week after training, in the absence of weekly testing, mean scores for the delayed-testing stimuli had fallen to 6.3% and 2.8 corrects per minute.

8.83 Tact post-tests

Consistent with Experiment 2, post-test tact results (Figure 7) generally mirrored the NFI results, except for participants' performance during the weekly-testing-with-feedback timings. Both participants demonstrated very high levels of accuracy and fluency during these timings because they were conducted immediately after the NFI timings, which included corrective foreign tact feedback. Consequently, the retention interval for tact post-tests during the first three weeks of post-training was a few minutes rather than days. The exception was the final post-test sessions, where all modeling and corrective feedback were withheld. In this session, following a one-week retention interval, both participants demonstrated foreign tact performance that exceeded fluency-criterion levels.

Figure 7

Training and Post-Test Tact Probes

Note. PT = Post-test. The data points shown in the line graphs represent tact accuracy probes and are measured as percentage of correct responses on the left-hand y-axis. Training probes were the final three timings conducted with each stimulus set. Columns represent mean response rate and are measured as corrects per minute on the right-hand y-axis. Post-tests were conducted monthly one to six months post-training.

8.9 General discussion

The goal of Experiment 2 was to compare the emergent-learning retention outcomes of free- and restricted-operant FTT while controlling for total training time. We expected that free-operant FTT would provide more testing opportunities and produce better retention than the same duration of restricted-operant FTT. Overall, there were more than four times as many trials in the free-operant condition, but we found no functional relation between training type and level of retention. Although free-operant FTT increased test dosage, it did not consistently improve retention. Extended testing failed to produce long-term criterion-level retention for any of the stimulus sets. However, retention of some of the words remained stable over the final five months of post-testing, and both participants retained some directly trained tact and emergent

NFI and FNI relations for six months without feedback or correction. Additionally, both participants demonstrated high levels of accurate untrained listener behavior with all stimulus sets six months after FTT. Participants indicated a preference for the free-operant condition despite none of the outcomes being dependent on training type.

We anticipated that the programmed constraints (i.e., the three-second intertrial delay) would suppress participants' response rates with the restricted-operant stimuli, and this was evident for P2. However, removing the intertrial delay and re-testing P1 with the restricted-operant stimuli after training ended, resulted in fluent criterion-level tacting. This observation suggests that as some learners become more proficient, they tend to respond more quickly with both free- and restricted-operant stimuli (Darvell, 2006; Kong, 2009; Wheatley, 2005).

The precise reasons for the lack of a functional relation between training type and retention in Experiment 2 are unclear. We considered the results may have been due to differences between stimulus set pairs (Cariveau et al., 2021; Shepley et al., 2020). However, a range of strategies were used to equate stimulus sets, and participants managed to achieve the accuracy criteria with a similar number of training timings. Furthermore, we conducted statistical analysis to identify potential confounds associated with the stimuli but found no significant differences based on word category, number of syllables, or initial sounds.

Another possible reason for the unexpected results observed in Experiment 2 is that the retention improvements produced by high-density testing in free-operant FTT may have been counterbalanced by spacing effects (Rohrer, 2009). Rohrer argued that while training with extended testing improves immediate recall and short-term retention (i.e., less than one week), these benefits come with a cost. As the rate of testing increases during free-operant timings, trial-spacing decreases. High-density, free-operant practice may enhance immediate performance, but

it can be detrimental to long-term retention if it lacks adequate spacing. According to the distributed practice effect, when learning trials are spread out over a longer period rather than massed together, retention improves (Carpenter et al., 2022). In Experiment 2, free-operant FTT provided high rates of testing and minimal intertrial delay; consequently, testing density gradually increased and spacing between trials became increasingly shorter, potentially negatively affecting retention (Karpicke & Bauernschmidt, 2011; Zigterman et al., 2015). Subsequent changes to testing during Experiment 3 introduced weekly spaced sessions and positively impacted retention.

In Experiment 3, we evaluated the effects of conducting weekly free-operant tests for four weeks after re-training 30 foreign tact relations that participants failed to retain in Experiment 2. The results of Experiment 3, when compared with Experiment 2, demonstrated improved retention of emergent and directly trained vocabulary when free-operant practice was spaced weekly for one month post-training.

When we withheld test opportunities for four weeks post-training, performance levels were comparable to those at baseline. Weekly tests in the absence of feedback, on the other hand, resulted in improved performance above baseline levels but below the criterion. Weekly tests and corrective foreign tact feedback produced the highest levels of retention four weeks post-training; retention accuracy and response rates gradually improved until they were similar with or exceeded the mastery criterion. These results indicate the importance of corrective feedback in enhancing retention.

These findings indicate that spacing testing weekly not only maintained but improved retention of emergent NFI and trained tact relations. Prior to the implementation of spaced tests, retention accuracy decreased by at least 40% one week after participants achieved the mastery

criteria. Both participants' performances incrementally improved with the introduction of tests and a one-week retention interval, suggesting that retention increased as a function of spacing, both with and without feedback. However, we are unsure if these improvements would have continued in the absence of feedback had we extended the study beyond four weeks.

Participants' response rates in the weekly-testing-without-feedback condition maintained an increasing trend at the conclusion of the study, but accuracy appeared to level off.

8.91 Limitations

The studies were conducted with only two participants, which limits the generalizability of the findings. However, the use of single case experimental design and within-subject replication in Experiment 2 allowed for precise examination of behavioral variability at the individual level, an approach that has not previously been employed in the existing FTT literature (Wooderson et al., 2022). Future research should include replication with larger sample sizes to confirm these results and better understand individual variability in response to different testing methods.

While our findings suggest retention of emergent and directly trained foreign vocabulary improves when tests are spaced weekly, we hesitate to draw definitive conclusions due to some limitations in the research design of Experiment 3. Specifically, we intentionally withheld practice-test opportunities for four weeks during the delayed-testing condition to compare the retention levels observed in the absence of weekly testing with those following a weekly schedule. The results were consistent with Experiment 2: words that participants previously failed to retain one-month post-training once again returned to baseline levels after one month of retraining. However, the delayed-testing condition was much shorter than the weekly testing conditions, consisting of just two sessions—only two-fifths of the total post-training test duration provided by each weekly-testing condition. While we do not believe this was the cause of the

differentiated results in Experiment 3, primarily due to the equivocal effects produced with high-density testing in Experiment 2, future research could systematically replicate the current study to control for potential practice effects. One approach would be to increase the duration of testing with the delayed-testing stimuli during the first post-test so that it provides the same total duration of practice-testing as the other two conditions.

8.92 Conclusion

Experiment 3 was informed by and built upon the attempts in Experiment 2 to understand the role of testing in the retention of emergent foreign language vocabulary learning. Experiment 2 revealed that free-operant high-density testing had an equivocal impact on retention of emergent and directly trained Korean vocabulary. Experiment 3 demonstrated improved retention following the introduction of spaced testing sessions. These results suggest that merely increasing the number of tests without considering other factors like trial spacing may not enhance long-term retention of vocabulary. This challenges the longstanding convention that extended practice is a key variable for retention. More practice may not always be better, and extended practice may not be a sufficient cause that is primary over other variables such as spacing of trials and feedback frequency. The practical implications of these findings for educators include the recommendation to implement weekly spaced testing sessions and corrective feedback to enhance the long-term retention of vocabulary in foreign language curricula. This approach can be more effective than increasing the number of tests within a short period, which could lead to improved learning outcomes and more efficient use of instructional time. We recommend that future studies further evaluate the potential benefits of free-operant spaced practice and assess the durability of retention improvements over longer periods and in other educational contexts.

Chapter 9: General discussion, implications, recommendations, and conclusions

9.1 General Discussion

This thesis explored the most effective and efficient methods for learners to acquire and retain emergent FL vocabulary. The research focused on improving emergent vocabulary retention by modifying instructional programming to assess which procedural variations to foreign tact training led to better learning outcomes, measured up to six months after training.

9.11 Summary of the studies and their key findings

To address this overarching objective, a series of technical, descriptive, and applied experimental investigations were conducted to explore the main research question: *What procedural modifications to the design and implementation of FTT contribute to improved retention of emergent Korean vocabulary?*

In conducting these investigations, a pragmatic inductive approach was taken, where each investigation's findings informed the next. Chapters 3 and 4 addressed the first four research sub-questions:

- a) *How does one extract graphical data for re-analysis from published SCED graphs using a data extraction software program?*
- b) *What are the effects of FTT on emergent learning outcomes in the published literature to date?*
- c) *How do FTT acquisition, emergence, and overall efficiency compare with other verbal operant training procedures?*
- d) *Does FTT produce higher levels of emergent responding for adults or children?*

The systematic review (Study 2, Chapter 4) used the protocols developed in Study 1 (Chapter 3) to extract data for the purpose of evaluating FTT outcomes in the published

literature. The data included several positive short-term outcomes of FTT: Learners acquired a range of emergent relations without the need for direct training, and FTT produced higher levels of emergent learning more efficiently than other verbal operant procedures.

As noted throughout the thesis, emergent relations are a form of incidental learning acquired along with intentional learning during direct training. The key benefit of FTT is its efficiency in producing a range of verbal vocabulary responses for ‘free’ (Critchfield 2018). These efficiencies can mitigate the considerable challenges encountered by FL learners, as highlighted in the introduction to this thesis, that FL learners must acquire a substantial vocabulary to communicate effectively in the FL.

The review’s (Study 2) findings showed that FTT produced criterion-level performance in 84 of 106 (79.2%) post-tests following training and was the most effective and efficient of the emergent learning procedures evaluated in the study. The meta-analysis revealed that FTT led to significantly higher within-subject levels of emergent responding than other verbal operant procedures, as well as slightly greater efficiency for both children and adults.

The primary data absent from the literature concerned long-term learning outcomes; specifically, all 10 FTT studies in the review failed to evaluate retention. Four studies examined maintenance of emergent learning outcomes with mixed results, but maintenance differs from retention in that learners have opportunities to practice learning in the natural environment under maintenance conditions. For all, particularly those individuals not immersed in the natural FL environment, retention is a key learning outcome.

A noteworthy preliminary finding from the meta-analysis suggested a potential link between acquisition criteria and retention. Specifically, the two studies with the least stringent acquisition criteria (Petursdottir & Haflídadóttir, 2009; Wu et al., 2019) resulted in the lowest

levels of emergence. This finding generated the focus for the subsequent investigations: Does production and retention of emergent relations depend on the strength of directly trained relations? Despite the limited research in this area, existing studies indicate that implementing stricter mastery criteria results in higher emergent responding during immediate post-testing (e.g., Fienup & Brodsky, 2017) and at follow-up (e.g., Bucklin et al., 2000). It follows that adjusting acquisition criteria might enhance the response strength of directly trained foreign tacts and improve the amount of emergent learning acquired and retained over time. Teaching aims to change the likelihood of a learner engaging in a particular response under certain circumstances - e.g. when a student sees an object, she labels it using the correct FL word. Vargas and McLaughlin (1977) argued that likelihood (or inclination) cannot be measured, so we examine how frequently learners respond—for example, by measuring how fast they can tact. When the student is able to label objects correctly and more quickly, she is said to have improved her ability to perform tacting. This area of inquiry has been seldom explored, but it was a significant focus of the current thesis.

Consequently, three experiments were conducted to investigate the hypothesis that acquisition criteria and response strength of directly trained vocabulary impact retention of emergent foreign vocabulary. Although the results across these experiments were at times conflicting, they collectively suggest a promising direction for future research. This research could fundamentally transform how emergent learning procedures are developed and implemented by identifying key instructional variables that affect retention. As highlighted in the introduction, retention is a crucial training outcome for FL learners not immersed in a natural linguistic community and remains a significant barrier to acquiring the extensive vocabulary needed for fluency.

Chapter 6 investigated the fifth research sub-question: *e) Does repeated practice of foreign tact relations beyond initial mastery (i.e., accuracy criterion) using fluency-building procedures affect the retention accuracy of derived intraverbal relations during testing?*

This initial experimental study (Study 3) aimed to address this research question and tackle issues identified in previous studies, which had shown highly variable emergent learning outcomes across learners. It also provided a preliminary examination of the potential link between response strength and retention. In conducting this study, the standard discrete-trial FTT procedure was modified to enhance learners' acquisition and retention of emergent Korean vocabulary. These modifications were informed by literature on overlearning, which suggests that repeated practice beyond the initial accuracy criterion improves retention (Colman, 2015; Driskell et al., 1992). The study introduced fluency criteria and free-operant practice to the standard FTT protocol, making it the first to do so. The results were promising, with participants demonstrating high accuracy in emergent responding immediately following training. All five learners achieved substantial gains in emergent Korean vocabulary both immediately after and one month post training. Three participants showed improved retention with fluency-building at the six-month follow-up, while one participant performed marginally better with the standard FTT protocol. Although the exact reasons for this discrepancy were unclear initially, further investigations with the modified FTT protocol in subsequent experiments provided additional insights. Notably, none of the participants in the first experiment performed at criterion levels during the six-month follow-up tests. Whether maintaining criterion-level performance for five months without practice is a realistic goal is a difficult question to answer based on the existing research literature. However, understanding the extent of learning retention was central to this thesis's key research question.

Overall, these findings showed that retention accuracy of emergent NFI relations was higher for stimuli taught using fluency-based overlearning procedures, suggesting a functional relation between fluency-building and retention of emergent learning. Three out of the four participants who completed follow-up testing six months after training exhibited better retention in this condition. The results indicated that repeated practice beyond initial mastery criterion improved retention of emergent relations. These preliminary findings are important contributions to the field, as they add to our understanding of the types of training procedures that produce long-term learning outcomes.

Chapter 8 (Studies 3 and 4) examined the final two research sub-questions:

f) *Does an intertrial interval during training affect the retention accuracy and response rate of emergent and directly trained learning during post-testing?*

g) *What are the effects of weekly practice tests with and without feedback on retention of emergent and directly trained learning?*

Experiment 2 (Study 4) aimed to systematically replicate and extend the findings by modifying the research design to address potential confounds identified in the first experiment (Study 3). Study 3 did not control for differences in the quantity of trials and duration of training across the two comparison conditions, raising the possibility that the improved retention outcomes were due to extended practice rather than fluency-building (Doughty et al., 2004). However, if retention improvements were indeed a function of practice, it would be important to consider how to maximise practice opportunities during training. This could only be determined by incorporating controls for practice effects within the experimental design.

Previous research on fluency-building controlled for practice effects by equalising the number of trials across the comparison conditions, with mixed results (e.g., Darvell, 2006; Kong,

2009; Mathews, 2010; McGregor, 2006; Porritt et al., 2009; Wheatley, 2005). However, this approach may be less desirable, as total training time is likely a more salient factor than the number of practice trials. In educational settings, instructional time is often limited, making the efficiency with which learning is acquired just as important as the amount of learning achieved (Critchfield & Twyman, 2014).

Study 4 (Chapter 8) removed the accuracy training phase from the modified FTT protocol to improve efficiency (Coyle, 2005) and equalised the training duration across the comparison conditions to control for differences in the amount of instruction. The modified FTT protocol was then compared with a restricted-operant protocol, which was similar in design to the standard FTT protocol. The restricted-operant FTT procedure included a programmed three-second intertrial delay, characteristic of discrete-trial presentation sequences. It was expected that the modified free-operant FTT condition would provide more practice trial opportunities than the restricted-operant FTT condition when total training time was held constant, and that more trials would increase response strength (i.e., higher rate of responding) and, subsequently, improve retention.

Contrary to the findings of the first experiment, results from the second experiment (Study 4, Chapter 8) failed to demonstrate a functional relationship between training type and the retention of emergent or directly taught FL vocabulary. Equivocal retention levels were observed across post-tests following FTT, regardless of whether free-operant or restricted-operant procedures were used. These findings were unexpected. First, the intertrial delay implemented in restricted-operant FTT did not systematically affect learners' response rates. One participant's terminal training response rate was similar across both conditions, suggesting that rate was not a determining factor for retention. Second, both FTT conditions in Study 4 produced high rates of

accurate trained tact and emergent NFI responding immediately following training, but monthly retention probes conducted over six months post-training demonstrated equivocal differences between the two conditions. The outcomes were undifferentiated despite there being four times as many practice trials in the modified FTT condition.

It became apparent that other variables affecting retention had not been accounted for in the earlier procedural modifications to FTT. Of particular interest were the weekly post-testing results and individual retention curves from Study 3, which showed the retention patterns for the first month post-training. The retention levels from Study 3 appeared to be much higher and more stable compared to the data from Study 4. Notably, the results from Study 3 indicated that two of the five participants' response accuracy with emergent NFI relations was as high as 95% (mean = 83.75%) one month post training following free-operant FTT and weekly testing without feedback. In contrast, in Study 4, P1 and P2's highest NFI response accuracy one month after training was 50% (mean = 50%). While the exact causes remain unclear, the main difference was that Study 3 included weekly post-tests, whereas Study 4 did not.

While devising the research questions and methodology for Study 5, the retention data from Study 3 were reviewed in light of the conflicting findings from Study 4. This was possible because the participants in Study 4 also took part in Study 3. These findings prompted a more systematic examination of spacing in the third experiment, which directly compared spaced weekly post-testing with delayed post-testing one month after training. It was hypothesised that introducing weekly post-tests would improve retention.

The third experiment's results (Study 5, Chapter 8) systematically reintroduced the spaced weekly testing used in the first experiment's post-testing phase (Study 3) and showed improved retention compared to the delayed testing employed in Study 4. Study 5 aimed to

address the poor retention results found in Study 4 and test the hypothesis that spaced weekly practice improves retention, as observed in Study 3.

Study 5 commenced one month after the final post-tests of Study 4. The same two individuals from Study 4 relearned 30 words that they had failed to retain in Study 3. Key procedural changes were implemented during the first month of post-training, comparing free-operant fluency-building FTT conditions with and without weekly practice. The findings indicated a spacing effect, where weekly practice testing improved the retention of emergent NFI relations. This outcome was consistent with the results seen in Study 3 (weekly post-testing), but not with those in Study 4 (monthly post-testing).

Additionally, Study 5 compared conditions with and without feedback, as well as a delayed testing condition, whereas Study 3 and 4 withheld corrective feedback during all post-testing sessions. The results following weekly practice tests with and without feedback demonstrated increasingly higher levels of performance across repeated one-week retention intervals; however, the weekly practice with feedback condition produced the highest performance levels.

The finding that spaced practice testing improves retention is far from new (Carpenter et al., 2022; Ebbinghaus, 2013). In fact, the spacing effect is among the most consistent and well-supported principles in learning research and has important implications for the field of emergent learning. It is surprising, then, that it has not received more attention in the FL emergent learning literature or been discussed more often in behaviour analysis in general (but see Johnson & Layng, 1996). Study 5 is the first to explore this effect in the context of FL emergent learning, offering numerous opportunities for future studies.

Taken together, the findings of this study offer valuable insights into the factors that influence emergent learning and retention. FTT is more efficient and effective than other emergent learning procedures, and the use of fluency-building and spaced practice strategies demonstrably improves FTT retention. However, to fully appreciate the practical implications of these findings, it is important to consider both the potential benefits and drawbacks of the implemented strategies. The following sections consider the specific findings and recommendations from the current line of research, as well as related issues that impacted emergent learning outcomes.

9.2 Implications for emergent learning programming

Emergent learning is efficient because learners can acquire vocabulary both explicitly and implicitly within a single set of instructional procedures. This results in 'free' learning, where training in one skill set can lead to proficiency in derived untaught skills (Critchfield, 2018), thereby optimising and enhancing learning outcomes (Critchfield & Twyman, 2014). The FTT procedures examined in this thesis reliably increased both receptive and productive language without the need for direct training in all skills, and all participants consistently emitted a range of emergent relations.

These results align with the behaviour analytic account of emergence, where learners are able to emit responses not directly trained due to the development of the basic units that serve as prerequisites for this emergent behaviour, rather than relying on cognitions, internal representations, or insight to explain the emergence of behaviour that has no history of reinforcement (Sidman, 1986). This is because FTT sets the occasion for learners to develop and derive equivalence relations between key verbal and non-verbal stimuli used in vocabulary learning: native-language words, foreign-language words, and related pictures (May et al., 2013).

According to stimulus equivalence theory, associating a picture with its native-language referent and its foreign-language counterpart facilitates the emergence of multiple untrained relations (Sidman, 2018), including NFI, FNI, mands and listener behaviour. All participants demonstrated prior knowledge of the native word during pre-assessment tests and were trained to relate the picture cards to the foreign words during tact training. They then demonstrated mastery of implicitly derived equivalence relations.

9.21 Recommended FTT protocol

This thesis determined that FTT is the most efficient and effective emergent learning procedure for producing emergent foreign vocabulary. Despite FTT's superiority over other methods, such as FNI, NFI, and listener training, it can be further improved. This was evident in the experiments using modified FTT procedures, which showed increased retention of both emergent and trained vocabulary following the addition of fluency-building and spaced practice.

Several modifications to the standard FTT protocol were trialled within Studies 3, 4, and 5. The modified FTT procedure that produced the highest level of retention of emergent Korean vocabulary (Chapter 9, Study 5) comprised the following features:

1. Daily training sessions with six to eight timings (60 seconds per timing) per day until the learner achieves the acquisition criteria—100-80 foreign tacts per minute with no errors across two consecutive timings.
2. Ten training stimuli presented at once in a randomised order, no feedback or instructor interaction during timings, and the learner controls the pace of trials to limit constraints on the rate of responding.
3. Corrective feedback on incorrect or unknown words following each timing, including modelling of correct pronunciation.

4. Spaced weekly post-acquisition sessions with six to eight NFI and foreign tact timings (60 seconds per timing) per session until the learner reacquires or exceeds the tact acquisition criteria in the absence of practice, following a one-week retention interval.

Based on the findings of the current studies, fluency-building is a beneficial component for improving learner's retention of emergent learning, but may be best used in combination with spaced practice. The spacing effect, which suggests that retention improves when learning trials are distributed rather than massed together, is one of the most consistent findings in educational research (Carpenter, et al., 2022). During spaced practice, learning trials are distributed over time with periods of no practice between sessions. In contrast, massed practice sessions are concentrated in a short period with little or no time between sessions. For example, a student engaging in spaced practice might learn 10 new vocabulary words for 10 minutes on Monday, practice them on Wednesday for another 10 minutes, and have a final 10-minute review on Friday (Bloom & Shuell, 1981). A student employing massed practice, on the other hand, might complete a single, extended learning session lasting 30 minutes in one day. Recent research highlights the effectiveness of spaced practice for improving both immediate recall and retention in FL learning (Kim & Webb, 2022). The key implication from these findings is that fluency-building without the appropriate spacing of practice sessions may fail to produce meaningful, persistent improvements in learning (Johnson & Layng 1996). However, these results should be interpreted cautiously given that a direct comparison of massed versus spaced practice was not conducted.

9.22 Costs and benefits of fluency-building

Despite the positive results produced by fluency-building in Studies 3 and 5, the overall findings call into question its benefits regarding emergent FL learning. The implementation of fluency-building without careful consideration and planning of practice sessions may negatively impact retention. In particular, the results of the current studies identified both costs and benefits associated with the use of fluency-building, notably, the practice protocols employing free-operant practice and targeting high rates of responding. These issues are not apparent in the precision-teaching literature but are more clearly noted in the broader literature on retention and spacing.

During free-operant FTT, participants' rates of responding increased sharply, and the resulting high-density tests were characterised by shorter spacing between individual word items, which may have negatively impacted retention (Karpicke & Bauernschmidt, 2011; Zigterman et al., 2015). Zigterman et al., for example, showed that the larger the within-session spacing between the repetition of individual word items, the greater the improvement in retention.

On the other hand, Study 3 showed that free-operant extended practice positively impacted retention. Notably, the highest levels of retention across all three experiments were observed in Study 5, where free-operant practice was implemented in combination with weekly spacing and corrective feedback. Additionally, free-operant FTT generally resulted in higher response accuracy and rates of responding during the initial acquisition phase.

Finally, the two participants who took part in all three experiments commented that they preferred free-operant FTT over the restricted-operant procedure.

Taken as a whole, the findings suggest that the potential benefits of fluency-building retrieval practice outweigh the costs when FTT is combined with appropriate spacing and free-operant practice trials (Higham et al., 2022).

To further understand the factors affecting retention, it is also important to consider the influence of learning materials, such as the pronunciation guides, on the learning process.

9.23 The pronunciation guide's effect on spacing

The pronunciation guide introduced in Study 4 was implemented to improve initial acquisition. The guide included pictures of each word along with text modelling their correct pronunciation. It was presented prior to the first training timing of each day and again following every timing for exactly 30 seconds. This also ensured that the total training duration, including study and test sessions, was equal across all comparison conditions.

In Study 5, the pronunciation guide was not shown before the first timing during the weekly post-testing phase to allow retention levels to be tested after one week without practice. Retention improved in Study 5, but it is not possible to determine if this improvement was solely due to changes in the use of the pronunciation guide.

Presenting the guide before the first timing in Study 4 may have had an unforeseen negative effect on retention, as it meant that the time between study (presentation of the guide) and practice (timings) was typically less than one minute. Karpicke and Roediger (2007) found that delaying the time between study and retrieval practice positively impacts retention. According to the authors, when testing occurs immediately or following a brief delay after the presentation of the learning material, it is easier for the learner to respond correctly than if the testing is delayed. This may improve initial acquisition but negatively impact retention.

Training sessions in Study 4 were scheduled daily, as this is a common practice in precision teaching: short, intensive sessions repeated daily to encourage learners to achieve the fluency aim (Johnson & Street, 2013). However, these sessions were always closely preceded by exposure to the pronunciation guide. Study 4's free-operant FTT might have produced better retention had the pronunciation guide not been shown before the first timing each day, but following each timing.

Additional research is recommended to explore the impact of response rate on retention and to gain further insights into the effectiveness of fluency-building techniques.

9.24 Response rate and retention

It was anticipated that by keeping total training time constant in Study 4, greater differences would arise between the terminal training response rates achieved in free-operant and restricted-operant practice. However, what was found instead was that as participants became more proficient, they responded faster with the stimuli taught in both free-operant and restricted-operant conditions (Kong, 2009; Darvell, 2006; Wheatley, 2005), and there was no clear functional relationship between response rate and retention.

In Study 3, some participants who failed to achieve the fluency criterion nonetheless demonstrated improved retention from fluency-building regardless. This suggests that fluency may not be a necessary predictor of retention, as initially hypothesised.

None of the current studies were successful in systematically isolating the effects of response rate on retention. This is a longstanding issue that previous studies also failed to address (e.g., Darvell, 2006; Kong, 2009; Mathews, 2010; McGregor, 2006; Wheatley, 2005), and it is recommended that future studies shift their focus to variables associated with the frequency and spacing of practice testing.

This is not to say that fluent, free-flowing performance is an unimportant outcome of effective instruction; rather, the current findings suggest that response rate should be viewed as characteristic, and not a causal factor of performance that is more likely to be retained over the long-term.

9.3 General limitations and recommendations for future research

There are several limitations to the findings of this thesis that could be addressed by future studies. First, Study 4 was limited by the lack of a control condition with the same total training time as the two spaced weekly practice conditions. The delayed-testing condition had far less trials, and training duration was less than half that of the other two conditions. The impact of spacing and practice testing on the emergent relations could be evaluated more systematically in a future study by increasing the duration of training in the control condition during the initial post-test so that it matches the spaced practice conditions. Bloom and Shuell (1981) tested this type of arrangement by allocating an equal amount of time across spaced and massed practice conditions. Their spaced condition was divided into three sessions lasting 10-minutes each across three consecutive days, and the massed condition was a single 30-minute session.

The way the pronunciation guide was implemented in Study 4 may have negatively impacted retention by reducing the delay between study and testing activities. A follow-up study is recommended where the guide is withheld prior to the first practice test, or a delay and distraction task is implemented between study and testing trials. This could help isolate the variables that were responsible for the low levels of retention observed during Study 4.

Although the current line of studies showed FTT was effective at producing emergent NFI, FNI, and listener relations, vocabulary learning typically has much broader applications beyond the simple tests for paired associations conducted in these studies (Nation, 2022). Future research

should extend these testing procedures to include more complex repertoires indicative of fluent language learners. For example, a suggested line of inquiry could examine the combination of FTT with simple sentence construction to assess whether fluent vocabulary performance increases fluency in related sentence construction. This is a common approach to instructional design in precision teaching: train component skills to fluency, then test for the emergence of higher-level composite skills. Kapoor et al. (2023), for example, recently taught two prerequisite (component) maths skills (reading and counting numbers between 0-20) to fluency criterion with four schoolchildren. Following the initial acquisition of the required component skills, the experimenter taught and tested the composite skill (mixed subtraction and addition facts) with the four participants and a control group comprising same-age children. The students who were taught the component skills using fluency-building procedures demonstrated greater mastery of the composite skill.

Finally, the results of the current study are not generalisable beyond the current participants and settings. Replication with participants from different demographic populations is recommended. Additionally, it is unlikely that the results are specific to the study of Korean vocabulary; however, replication with other FLs would help extend the generalisability of the findings. Future research could also be conducted using a randomised control trial (RCT) to assess the effectiveness of the modified FTT procedure at the group level and in a range of settings. In an RCT, participants would be randomly assigned either to the modified FTT intervention or to an alternative or standard teaching approach, allowing for a more robust comparison of outcomes at the group level and helping to rule out potential confounding variables. Ideally, such procedures could be evaluated in an FL classroom, where learners are already enrolled in a structured instructional program, and thus provide further insights. An RCT

design would additionally enable the collection of data on a range of learner variables, such as language proficiency, motivation, and prior vocabulary knowledge.

9.4 Conclusion

The aim of all educational programs should be to optimise learning by making instruction as effective and efficient as possible (Gilbert & Gilbert, 1992). To this end, it is recommended that learning programs are based on systematic, empirically derived instructional technologies, and behaviour analysis has much to offer in this regard (Binder & Watkins 1990). This thesis focussed on behaviour analytic approaches to FL learning with a series of technical and applied studies aimed at improving vocabulary acquisition by combining components of tact training, emergent learning, overlearning, and fluency-building procedures.

The current studies are among the first empirical investigations of this nature, and the findings provide important considerations for the design of FL vocabulary instruction. Moreover, there are broad implications that extend beyond verbal behaviour learning. Instructional programs typically view learning as complete when learners achieve an accuracy criterion. However, it is evident from the findings in the current study that learning is occurring as the learner approaches the accuracy criterion and continues as they progress towards the fluency criterion and beyond.

While the present research focused on FL vocabulary acquisition, the principles and methodologies explored here have potential applications for a range of educational contexts. Foreign language researchers should continue to investigate how these behaviour-analytic approaches can be refined and adapted to other domains of learning. Additionally, further exploration into retention of emergent learning and the role of fluency within a variety of

contexts could provide valuable insight into the types of training arrangements that optimise learning outcomes.

Although not examined in the present studies, it should be acknowledged that learning continues to occur as the acquired behaviour is applied to various contexts and in combination with various other behaviours. Traditional educational approaches typically take a much narrower view of learning. There could be much to gain from revising these longstanding and outdated views about learning and acquisition.

As we move forward, it is essential that educators and instructional designers embrace these insights and consider how behaviour-analytic approaches can be integrated into modern educational practices. By doing so, we can engineer more efficient, effective, and enduring learning experiences for all students.

References

- Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*. <https://doi.org/10.1007/s10648-021-09595-9>
- Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*: Bloomsbury Academic.
- Antoniou, M., Gunasekera, G. M., & Wong, P. C. M. (2013). Foreign language training as cognitive therapy for age-related cognitive decline: A hypothesis for future research. *Neuroscience & Biobehavioral Reviews*, 37(10), 2689-2698
<https://doi.org/10.1016/j.neubiorev.2013.09.004>
- Ash, I. K., Jee, B. D., & Wiley, J. (2012). Investigating insight as sudden learning. *The Journal of Problem Solving*, 4(2), 2. <https://doi.org/10.7771/1932-6246.1123>
- Aydin, O., & Yassikaya, M. Y. (2022). Validity and reliability analysis of the Plotdigitizer software program for data extraction from single-case graphs. *Perspectives on Behavior Science*, 45(1), 239–257. <https://doi.org/10.1007/s40614-021-00284-0>
- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal Of Applied Behavior Analysis*, 1(1), 91–97.
<https://doi.org/10.1901/jaba.1968.1-91>
- Bak, M. S., Dueñas, A. D., Avendaño, S. M., Graham, A. C., & Stanley, T. (2021). Tact instruction for children with autism spectrum disorder: A review. *Autism & Developmental Language Impairments*, 6, 239694152199901. <https://doi.org/10.1177/2396941521999010>
- Barnes-Holmes, D., Finn, M., McEntegart, C., & Barnes-Holmes, Y. (2018). Derived stimulus relations and their role in a behavior-analytic account of human language and cognition.

Perspectives on Behavior Science, 41(1), 155-173. <https://doi.org/10.1007/s40614-017-0124-7>

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, 27, 223-247.

[https://doi.org/10.1016/S0346-251X\(99\)00018-4](https://doi.org/10.1016/S0346-251X(99)00018-4)

Blair, B. J., & Shawler, L. A. (2019). Developing and implementing emergent responding training systems with available and low-cost computer-based learning tools: Some best practices and a tutorial. *Behavior Analysis in Practice*, 13(2), 509–520.

<https://doi.org/10.1007/s40617-019-00405-x>

Binder, C. (1979). *Response rate measurement in a mediated transfer paradigm: Teaching severely retarded students to read* [Paper presentation]. Meeting of the Association for Behavior Analysis, Dearborn, MI.

Binder C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst*, 19(2), 163–197. <https://doi.org/10.1007/BF03393163>

Binder C. (2004). A refocus on response-rate measurement: comment on Doughty, Chase, and O'Shields (2004). *The Behavior Analyst*, 27(2), 281–286.

<https://doi.org/10.1007/BF03393186>

Binder, C., & Watkins, C. L. (1990). Precision teaching and direct instruction: Measurably superior instructional technology in schools. *Performance Improvement Quarterly*, 26(2), 73-115. <https://doi.org/10.1002/piq.21145>

Biosoft. (2004). Ungraph (Version 5.0.1). Retrieved from

<http://www.biosoft.com/w/ungraph.htm>

- Bloom, K.C. & Shuell, T.J (1981) Effects of massed and distributed practice on the learning and retention of second-language vocabulary, *The Journal of Educational Research*, 74:4, 245-248, <https://doi.org/10.1080/00220671.1981.10885317>
- Bormann, I. (2020). DigitizeIt (Version 2.5.3): Bormisoft. Retrieved from <http://www.digitizeit.xyz/>
- Bortoloti, R., & de Rose, J. C. (2009). Assessment of the relatedness of equivalent stimuli through a semantic differential. *The Psychological Record*, 59(4), 563-590. <https://doi.org/10.1007/BF03395682>
- Bortoloti, R., & de Rose, J. C. (2011). An "Orwellian" account of stimulus equivalence. Are some stimuli "more equivalent" than others? *European Journal of Behavior Analysis*, 12(1), 121-134. <https://doi.org/10.1080/15021149.2011.11434359>
- Bortoloti, R., Rodrigues, N. C., Cortez, M. D., Pimentel, N., & De Rose, J. C. (2013). Overtraining increases the strength of equivalence relations. *Psychology & Neuroscience*, 6(3), 357-364. <https://doi.org/10.3922/j.psns.2013.3.13>
- Bucklin, B. R., Dickinson, A. M., & Brethower, D. M. (2000). A comparison of the effects of fluency training and accuracy training on application and retention. *Performance Improvement Quarterly*, 13(3), 140-163. <https://doi.org/10.1111/j.1937-8327.2000.tb00180.x>
- Cao, Y., & Greer, R. D. (2018). Mastery of echoics in Chinese establishes bidirectional naming in Chinese for preschoolers with naming in English. *The Analysis of Verbal Behavior*, 34(1-2), 79-99. <https://doi.org/10.1007/s40616-018-0106-1>

- Cariveau, T., Batchelder, S., Ball, S., & La Cruz Montilla, A. (2021). Review of methods to equate target sets in the adapted alternating treatments design. *Behavior Modification*, 45(5), 695–714. <https://doi.org/10.1177/0145445520903049>
- Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, 1(9), 496-511. <https://doi.org/10.1038/s44159-022-00089-1>
- Cheng, K., Deng, Y., Li, M., & Yan, H. M. (2015). The impact of L2 learning on cognitive aging. *ADMET & DMPK*, 3(3). <https://doi.org/10.5599/admet.3.3.206>
- Chomsky, N. (1957). *Syntactic Structures*: Mouton.
- Colman, A. (2015). Overlearning. In *A Dictionary of Psychology*.: Oxford University Press. Retrieved 12 Mar. 2023, from <https://www-oxfordreference-com.ezproxy.lib.uts.edu.au/view/10.1093/acref/9780199657681.001.0001/acref-9780199657681-e-5918>
- Contreras, B. P., Cooper, A. J., & Kahng, S. (2020). Recent research on the relative efficiency of speaker and listener instruction for children with autism spectrum disorder. *Journal of Applied Behavior Analysis*, 53(1), 584-589. <https://doi.org/10.1002/jaba.543>
- Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied Behavior Analysis* (2nd ed. ed.). Upper Saddle River, N.J: Pearson/Merrill-Prentice Hall.
- Cortez, M. D., da Silva, L. F., Cengher, M., Mazzoca, R. H., & Miguel, C. F. (2021). Teaching a small foreign language vocabulary to children using tact and listener instruction with a prompt delay. *Journal of Applied Behavior Analysis* (Advance online publication). <https://doi.org/10.1002/jaba.885>

- Cortez, M. D., da Silva, L. F., Cengher, M., Mazzoca, R. H., & Miguel, C. F. (2022). Teaching a small foreign language vocabulary to children using tact and listener instruction with a prompt delay. *Journal of Applied Behavior Analysis*, 55(1), 249–263.
<https://doi.org/10.1002/jaba.885>
- Cortez, M. D., dos Santos, L., Quintal, A. E., Silveira, M. V., & Rose, J. C. (2020). Learning a foreign language: Effects of tact and listener instruction on the emergence of bidirectional intraverbals. *Journal of Applied Behavior Analysis*, 53(1), 484-492.
<https://doi.org/10.1002/jaba.559>
- Council for Exceptional Children (2014). *Standards for evidence-based practices in special education*. Retrieved from <https://www.cec.sped.org>.
- Cowley, B. J., Green, G., & Braunling-McMorrow, D. (1992). Using stimulus equivalence procedures to teach name-face matching to adults with brain injuries. *Journal of Applied Behavior Analysis*, 25(2), 461-475. <https://doi.org/10.1901/jaba.1992.25-461>
- Coyle, C. (2005). *An investigation of the fluency paradigm: The effects of accuracy training before rate-building and incremental increases in response rates on skill retention, endurance, stability, application and adduction* (Doctoral Thesis). Murdoch University. Retrieved from <https://researchportal.murdoch.edu.au/esploro/outputs/doctoral/An-investigation-of-the-fluency-paradigm/991005542174307891#file-0>
- Critchfield, T. (2018). Efficiency is everything: Promoting efficient practice by harnessing derived stimulus relations. *Behavior Analysis in Practice*, 11(3), 206-210.
<https://doi.org/10.1007/s40617-018-0262-8>

- Critchfield, T., Barnes-Holmes, D., & Dougher, M. J. (2018). Editorial: What Sidman did -- historical and contemporary significance of research on derived stimulus relations. *Perspectives on Behavior Science*, 41(1), 9-32. <https://doi.org/10.1007/s40614-018-0154-9>
- Critchfield, T., & Twyman, J. S. (2014). Prospective instructional design: Establishing conditions for emergent learning. *Journal of Cognitive Education and Psychology*, 13(2), 201-217. <https://doi.org/10.1891/1945-8959.13.2.201>
- Cummins, J. (2009) Bilingual and immersion programs. In M. Long, & C. Doughty, (Eds.) *The handbook of language teaching*. Malden, MA: Wiley-Blackwell.
<https://doi.org/10.1002/9781444315783.ch10>
- Daly, D., & Dounavi, K. (2020). A comparison of tact training and bidirectional intraverbal training in teaching a foreign language: A refined replication. *The Psychological Record*, 70(2), 243-255. <https://doi.org/10.1007/s40732-020-00396-0>
- Darvell, A. (2006). *Fast and fluent or careful and correct: an empirical study of rate-building methods* (Unpublished Master's Thesis). The University of Auckland, Auckland, New Zealand.
- Delfs, C. H., Conine, D. E., Frampton, S. E., Shillingsburg, M. A., & Robinson, H. C. (2014). Evaluation of the efficiency of listener and tact instruction for children with autism. *Journal of Applied Behavior Analysis*, 47(4), 793-809. <https://doi.org/10.1002/jaba.166>
- Dixon, M. R., & Stanley, C. R. (2020). Relational frame theory: Basic relational operants. In M. Fryling, R. A. Rehfeldt, J. Tarbox, & L. J. Hayes (Eds.), *Applied behavior analysis of language and cognition: Core concepts and principles for practitioners*: New Harbinger Publications.

- Dougher, M., Twohig, M. P., & Madden, G. J. (2014). Editorial: Basic and translational research on stimulus–stimulus relations. *Journal of the Experimental Analysis of Behavior*, 101(1), 1-9. <https://doi.org/10.1002/jeab.69>
- Dougherty, K. M., & Johnston, J. M. (1996). Overlearning, fluency, and automaticity. *The Behavior Analyst*, 19(2), 289-292. <https://doi.org/10.1007/BF03393171>
- Doughty, S. S., Chase, P. N., & O'Shields, E. M. (2004). Effects of rate building on fluent performance: A review and commentary. *The Behavior Analyst*, 27(1), 7–23. <https://doi.org/10.1007/BF03392086>
- Dounavi, K. (2011). A comparison between tact and intraverbal training in the acquisition of a foreign language. *European Journal of Behavior Analysis*, 12(1), 239-248. <https://doi.org/10.1080/15021149.2011.11434367>
- Dounavi, K. (2014). Tact training versus bidirectional intraverbal training in teaching a foreign language. *Journal of Applied Behavior Analysis*, 47(1), 165-170. <https://doi.org/10.1002/jaba.86>
- Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. *Behavior Modification*, 41(2), 323-339. <https://doi.org/10.1177/0145445516673998>
- Driskell, Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology*, 77(5), 615–622. <https://doi.org/10.1037/0021-9010.77.5.615>
- Ebbinghaus, H. (2013). Memory: A Contribution to Experimental Psychology. *Annals of Neurosciences*, 20(4). <https://doi.org/10.5214/ans.0972.7531.200408>

Eguz, E. (2019). Learning a second language in late adulthood: Benefits and challenges.

Educational Gerontology, 45(12), 701-707.

<https://doi.org/10.1080/03601277.2019.1690273>

Eilifsen, C., & Arntzen, E. (2017). Effects of immediate tests on the long-term maintenance of stimulus equivalence classes. *The Psychological Record*, 67(4), 447-461.

<https://doi.org/10.1007/s40732-017-0247-y>

Engku Ibrahim, E. H., Othman, K., Sarudin, I., & Jariah Muhamad, A. (2013). Measuring the vocabulary size of Muslim pre-university students. *World Applied Sciences Journal*, 21, 44-49. <https://doi.org/10.5829/idosi.wasj.2013.21.s1l.2136>

Epstein, R. (1991). Skinner, creativity, and the problem of spontaneous behavior. *Psychological Science*, 2(6), 362-370. <https://doi.org/10.1111/j.1467-9280.1991.tb00168.x>

Epstein, R., Kirshnit, C. E., Lanza, R. P., & Rubin, L. C. (1984). 'Insight' in the pigeon: Antecedents and determinants of an intelligent performance. *Nature*, 308(5954), 61.

<https://doi.org/10.1038/308061a0>

European Commission. (2020). Pupils by education level and number of modern foreign languages studied - absolute numbers and % of pupils by number of languages studied.

Retrieved from

https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=educ_uoe_lang02&lang=en

Evans, A. L., Bulla, A. J., & Kieta, A. R. (2021). The precision teaching system: A synthesized definition, concept analysis, and process. *Behavior Analysis in Practice*, 14(3), 559-576.

<https://doi.org/10.1007/s40617-020-00502-2>

Fienup, D. M., & Brodsky, J. (2017). Effects of mastery criterion on the emergence of derived equivalence relations. *Journal of Applied Behavior Analysis*, 50(4), 843-848.

<https://doi.org/10.1002/jaba.416>

Fienup, D. M., & Carr, J. E. (2021). The use of performance criteria for determining “mastery” in discrete-trial instruction: A call for research. *Behavioral Interventions*, 36(4), 756–763.

<https://doi.org/10.1002/bin.1827>

Flower, A., Mckenna, J. W., & Upreti, G. (2016). Validity and reliability of GraphClick and DataThief III for data extraction. *Behavior Modification*, 40(3), 396-413.

<https://doi.org/10.1177/0145445515616105>

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*, 109(8), 1067–

1083. <https://doi.org/10.1037/edu0000197>

Fuller, J. L., & Fienup, D. M. (2018). A preliminary analysis of mastery criterion level: Effects on response maintenance. *Behavior Analysis in Practice*, 11(1), 1–8.

<https://doi.org/10.1007/s40617-017-0201-0>

Geomatix. (2021). XYit (Version 3.1.10). Retrieved from <https://www.geomatix.net/xyit/>

Gilbert, T. F., & Gilbert, M. B. (1992). Potential contributions of performance science to education. *Journal of Applied Behavior Analysis*, 25(1), 43–49.

Gist, C., Bulla, A.J. (2020). A systematic review of frequency building and precision teaching with school-aged children. *Journal of Behavioral Education*, 31, 43–68.

<https://doi.org/10.1007/s10864-020-09404-3>

- Greer, R. D., & Speckman, J. (2009). The integration of speaker and listener responses: A theory of verbal development. *The Psychological Record*, 59(3), 449-488.
<https://doi.org/10.1007/bf03395674>
- Ha, H. T. (2021). Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension. *Language Testing in Asia*, 11(1). <https://doi.org/10.1186/s40468-021-00131-8>
- Haegele, K. M., McComas, J. J., Dixon, M., & Burns, M. K. (2011). Using a stimulus equivalence paradigm to teach numerals, English words, and Native American words to preschool-age children. *Journal of Behavioral Education*, 20(4), 283-296.
<https://doi.org/10.1007/s10864-011-9134-9>
- Hartshorne, J. K., Tenebaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition* (177), 263-277.
- Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post Skinnerian account of human language and cognition*. New York: Plenum Publishers.
- Higham, P. A., Zengel, B., Bartlett, L. K., & Hadwin, J. A. (2022). The benefits of successive relearning on multiple learning outcomes. *Journal of Educational Psychology*, 114(5), 928–944. <https://doi.org/10.1037/edu0000693>
- Horne, P. J., & Lowe, C. F. (1996). On the origins of naming and other symbolic behavior. *Journal of the Experimental Analysis of Behavior*, 65(1), 185-241.
<https://doi.org/10.1901/jeab.1996.65-185>
- Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-*

case intervention research: Methodological and statistical advances. (pp. 91-125).

Washington, DC, US: American Psychological Association.

Hull, C. L. (1939). The problem of stimulus equivalence in behavior theory. *Psychological Review*, 46(1), 9-30. <https://doi.org/10.1037/h0054032>

Johnson, K. R., & Layng, T. V. J. (1996). On terms and procedures: Fluency. *The Behavior Analyst*, 19(2), 281-288. <https://doi.org/10.1007/bf03393170>

Johnson, K. R., & Street, E. M. (2004). *The Morningside Model of Generative Instruction: What It Means to Leave No Child Behind*. Concord, MA: Cambridge Center for Behavioral Studies.

Johnson, K.R., & Street, E. M. (2013). *Response to Intervention and Precision Teaching: Creating Synergy in The Classroom*. [Guilford Press](#)

Johnston, J. M., & Pennypacker, H. S. (2010). *Strategies and Tactics of Behavioral Research*: Taylor & Francis.

Jonathans, P. M., Widiati, U., Astutik, I., & Ratri, D. P. (2021). The practices of intentional vocabulary acquisition for Asian EFL learners: A systematic review. *Journal of English Education*, 9(2). <https://doi.org/10.25134/erjee.v9i2.4350>

Joyce, B. G., & Joyce, J. H. (1993). Using stimulus equivalence procedures to teach relationships between English and Spanish words. *Education and Treatment of Children*, 16(1), 48-65. Retrieved from <http://www.jstor.org/stable/42899293>

Kapoor, G., Vostanis, A., Mejía-Buenaño, S., & Langdon, P. E. (2023). Using precision teaching to improve typically developing student's mathematical skills via

- teleconferencing. *Journal of Behavioral Education*. <https://doi.org/10.1007/s10864-023-09520-w>
- Karpicke, J.D. & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 37. <https://doi.org/10.1037/a0023436>
- Kazdin, A. E. (2021). Single-case experimental designs: Characteristics, changes, and challenges. *Journal of the Experimental Analysis of Behavior*, 115(1), 56-85.
<https://doi.org/10.1002/jeab.638>
- Kim, S.K, & Webb, S. (2022), The effects of spaced practice on second language learning: A meta-analysis. *Language learning*, 72: 269-319. <https://doi.org/10.1111/lang.12479>
- Klimova, B., Pikhart, M., Cierniak-Emerych, A., Dziuba, S., & Firlej, K. (2021). A comparative psycholinguistic study on the subjective feelings of well-being outcomes of foreign language learning in older adults from the Czech Republic and Poland. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.606083>
- Kodak, T., Halbur, M., Bergmann, S., Costello, D. R., Benitez, B., Olsen, M., Gorgan, E., & Cliett, T. (2019). A comparison of stimulus set size on tact training for children with autism spectrum disorder. *Journal of Applied Behavior Analysis*.
<https://doi.org/10.1002/jaba.553>
- Köhler, W. (1925). *The Mentality of Apes*. New York: Harcourt Brace.
- Kong, X. (2009). *Precision teaching: Fast practice or merely more practice results in better learning?* (Master's Thesis). The University of Waikato, Hamilton, New Zealand.
Retrieved from <https://hdl.handle.net/10289/3940>

- Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26-38. <https://doi.org/10.1177/0741932512452794>
- Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews. Neuroscience, 5*(11), 831-843. <https://doi.org/http://dx.doi.org/10.1038/nrn1533>
- Lafrance, D. L., & Tarbox, J. (2020). The importance of multiple exemplar instruction in the establishment of novel verbal behavior. *Journal of Applied Behavior Analysis, 53*(1), 10-24. <https://doi.org/10.1002/jaba.611>
- Lalonde, K. B., Dueñas, A. D., Neil, N., Wawrzonek, A., & Plavnick, J. B. (2020). An evaluation of two tact-training procedures on acquired tacts and tacting during play. *The Analysis of Verbal Behavior, 36*(2), 180-192. <https://doi.org/10.1007/s40616-020-00131-4>
- Laufer, B. (2001). Quantitative evaluation of vocabulary: How it can be done and what it is good for. In C. Elder, Hill, K., Brown, A., Iwashita, N., Grove, L., Lumley, T., and McNamara, T. (Ed.), *Experimenting with Uncertainty*. Cambridge: Cambridge University Press.
- Ledford, J. R., Lane, J. D., & Severini, K. E. (2018). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment, 19*(1), 4-17. <https://doi.org/10.1017/brimp.2017.16>
- Lee, G. T., & Singer-Dudek, J. (2012). Effects of fluency versus accuracy training on endurance and retention of assembly tasks by four adolescents with developmental disabilities. *Journal of Behavioral Education, 21*(1), 1-17. <https://doi.org/10.1007/s10864-011-9142-9>
- Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*(3), 210-212. https://doi.org/10.1207/S15328023TOP2903_06

- Lindsley O. R. (1996). The four free-operant freedoms. *The Behavior Analyst*, 19(2), 199–210.
<https://doi.org/10.1007/BF03393164>
- Lobo, M. A., Moeyaert, M., Baraldi Cunha, A., & Babik, I. (2017). Single-case design, analysis, and quality assessment for intervention research. *Journal of Neurologic Physical Therapy*, 41(3), 187-197. <https://doi.org/10.1097/npt.0000000000000187>
- Longino, E., Richling, S. M., McDougale, C. B., & Palmier, J. M. (2021). The effects of mastery criteria on maintenance: A replication with most-to-least prompting. *Behavior Analysis in Practice*, 15(2), 397–405. <https://doi.org/10.1007/s40617-021-00562-y>
- Maggin, D. M., Lane, K. L., & Pustejovsky, J. E. (2017). Introduction to the special issue on single-case systematic reviews and meta-analyses. *Remedial and Special Education*, 38(6), 323-330. <https://doi.org/10.1177/0741932517717043>
- Martin, S. E. (1966). Lexical evidence relating Korean to Japanese. *Language*, 42(2), 185–251.
<https://doi.org/10.2307/411687>
- Martinho, M. T., Booth, N., Attard, N., & Dillenburger, K. (2021). A systematic review of the impact of precision teaching and fluency-building on teaching children diagnosed with autism. *International Journal of Educational Research*, 116, 102076.
<https://doi.org/10.1016/j.ijer.2022.102076>
- Mathews, S. T. (2010). *Teaching time telling and examining the relative effects of rate-building and rate-controlled practice on the retention and generalization of the time cues* (Master's Thesis). University of Waikato, Hamilton, New Zealand. Retrieved from
<https://hdl.handle.net/10289/4993>

- Matter, A. L., Wiskow, K. M., & Donaldson, J. M. (2020). A comparison of methods to teach foreign-language targets to young children. *Journal of Applied Behavior Analysis*, 53(1), 147-166. <https://doi.org/10.1002/jaba.545>
- May, R. J., Chick, J., Manuel, S., & Jones, R. (2019). Examining the effects of group-based instruction on emergent second-language skills in young children. *Journal of Applied Behavior Analysis*, 52(3), 667-681. <https://doi.org/10.1002/jaba.563>
- May, R. J., Downs, R., Marchant, A., & Dymond, S. (2016). Emergent verbal behavior in preschool children learning a second language. *Journal of Applied Behavior Analysis*, 49(3), 711-716. <https://doi.org/10.1002/jaba.301>
- May, R. J., Hawkins, E., & Dymond, S. (2013). Brief report: Effects of tact training on emergent intraverbal vocal responses in adolescents with autism. *Journal of Autism and Developmental Disorders*, 43(4), 996-1004. <https://doi.org/10.1007/s10803-012-1632-7>
- McDougale, C. B., Richling, S. M., Longino, E. B., & O'Rourke, S. A. (2020). Mastery criteria and maintenance: A descriptive analysis of applied research procedures. *Behavior Analysis in Practice*, 13(2), 402-410. <https://doi.org/10.1007/s40617-019-00365-2>
- McGregor, S. J. (2006). *Practice makes the difference: The effect of rate-building and rate-controlled practice on retention* (Master's Thesis). The University of Waikato, Hamilton, New Zealand. Retrieved from <https://hdl.handle.net/10289/2515>
- Melvin-Brown, R., Garcia, Y., Rosales, R., Mahoney, A., & Fuller, J. (2022). A review of second language acquisition in verbal behavior analysis. *Journal of Behavioral Education*, 1-24. <https://doi.org/10.1007/s10864-022-09471-8>

- Michael, J., Palmer, D. C., & Sundberg, M. L. (2011). The multiple control of verbal behavior. *The Analysis of Verbal Behavior*, 27(1), 3-22. <https://doi.org/10.1007/bf03393089>
- Miguel, C. F. (2016). Common and intraverbal bidirectional naming. *The Analysis of Verbal Behavior*, 32(2), 125-138. <https://doi.org/10.1007/s40616-016-0066-2>
- Moeyaert, M., Maggin, D., & Verkuilen, J. (2016). Reliability, validity, and usability of data extraction programs for single-case research designs. *Behavior Modification*, 40(6), 874-900. <https://doi.org/10.1177/0145445516645763>
- Morgan-Short, K., & van Hell, J.G. (Eds.). (2023). *The Routledge Handbook of Second Language Acquisition and Neurolinguistics* (1st ed.). Routledge.
<https://doi.org/10.4324/9781003190912>
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59-82. <https://doi.org/10.3138/cmlr.63.1.59>
- Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781009093873>
- New American Economy. (2017). Not Lost in Translation: The Growing Importance of Foreign Language Skills in the U.S. Job Market. Retrieved from
http://www.newamericaneconomy.org/wp-content/uploads/2017/03/NAE_Bilingual_V9.pdf
- Norris, J. & Ortega, L. (2005). Does type of instruction make a difference? Substantive findings from a meta-analytic review. *Language Learning*. 51, 157 - 213.
<https://doi.org/10.1111/j.1467-1770.2001.tb00017.x>

- Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, 18(2), 161-175. [https://doi.org/10.1016/S0889-4906\(98\)00005-2](https://doi.org/10.1016/S0889-4906(98)00005-2)
- Page, M. J., Mckenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., Mcguinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1). <https://doi.org/10.1186/s13643-021-01626-4>
- Paribakht, T. S., & Wesche, M. (1999). Reading and “incidental” L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language Acquisition*, 21(2), 195-224. <https://doi.org/10.1017/S027226319900203X>
- Pavlik, P. I., Jr., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29, 559–586. https://doi.org/10.1207/s15516709cog0000_14
- Peladeau, N., Forget, J., & Gagne, F. (2003). Effect of paced and unpaced practice on skill application and retention: How much is enough? *American Educational Research Journal*, 40(3), 769-801, <https://doi.org/10.3102/00028312040003769>
- Pennypacker, H.S., Gutierrez, A, Jr., Lindsley, O.R. (2003). *Handbook of the standard celeration chart*. Cambridge Center for Behavioral Studies.

- Petursdottir, A. I., & Haflíðadóttir, R. D. (2009). A comparison of four strategies for teaching a small foreign-language vocabulary. *Journal of Applied Behavior Analysis*, 42(3), 685-690. <https://doi.org/10.1901/jaba.2009.42-685>
- Petursdottir, A. I., Lepper, T., & Peterson, S. (2014). Effects of collateral response requirements and exemplar training on listener training outcomes in children. *Psychological Record*, 64(4), 703-717. <https://doi.org/10.1007/s40732-014-0051-x>
- Petursdottir, A. I., Olafsdottir, A. R., & Aradóttir, B. (2008). The effects of tact and listener training on the emergence of bidirectional intraverbal relations. *Journal of Applied Behavior Analysis*, 41(3), 411-415. <https://doi.org/10.1901/jaba.2008.41-411>
- Petursdottir, A. I., & Oliveira, J. S. (2023). Teaching foreign language. In Matson, J. L. (Ed) *Handbook of applied behavior analysis: Integrating research into practice* (pp. 1059–1076). Springer International Publishing.
- Picker Wheel (n.d.) [Computer software]. Retrieved January 21, 2020, from <https://pickerwheel.com/tools/random-team-generator/>
- Piller, I. (2016). *Linguistic Diversity and Social Justice: An Introduction to Applied Sociolinguistics*. Oxford University Press.
- Pitts, L., & Hoerger, M. L. (2021). Mastery criteria and the maintenance of skills in children with developmental disabilities. *Behavioral Interventions*, 36(2), 522–531. <https://doi.org/10.1002/bin.1778>
- Polack, C. W., & Miller, R. R. (2022). Testing improves performance as well as assesses learning: A review of the testing effect with implications for models of learning. *Journal of Experimental Psychology. Animal Learning and Cognition*, 48(3), 222–241. <https://doi.org/10.1037/xan0000323>

- Polson, D. A. D., Grabavac, D. M., & Parsons, J. A. (1997). Intraverbal stimulus–response reversibility: Fluency, familiarity effects, and implications for stimulus equivalence. *Analysis of Verbal Behavior*, 14(1), 19-40. <https://doi.org/10.1007/BF03392914>
- Polson, D. A. D., & Parsons, J. A. (2000). Selection-based versus topography-based responding: An important distinction for stimulus equivalence? *Analysis of Verbal Behavior*, 17(1), 105-128. <https://doi.org/10.1007/BF03392959>
- Porritt, M., Van Wagner, K., & Poling, A. (2009). Effects of response spacing on acquisition and retention of conditional discriminations *Journal of Applied Behavior Analysis*, 42(2), 295–307. <https://doi.org/10.1901/jaba.2009.42-295>
- Quigley, S. P., Peterson, S. M., Frieder, J. E., & Peck, K. M. (2018). A review of SAFMEDS: Evidence for procedures, outcomes and directions for future research. *Perspectives on Behavior Science*, 41(1), 283–301. [10.1007/s40614-017-0087-8](https://doi.org/10.1007/s40614-017-0087-8)
- Rakap, S., Rakap, S., Evran, D., & Cig, O. (2016). Comparative evaluation of the reliability and validity of three data extraction programs: UnGraph, GraphClick, and DigitizeIt. *Computers in Human Behavior*, 55, 159-166. <https://doi.org/10.1016/j.chb.2015.09.008>
- Ramirez, J., Rehfeldt, R. A., & Ninness, C. (2009). Observational learning and the emergence of symmetry relations in teaching Spanish vocabulary words to typically developing children. *Journal of Applied Behavior Analysis*, 42(4), 801-805. <https://doi.org/10.1901/jaba.2009.42-801>
- Regaço, A., Zapparoli, H. R., Aggio, N. M., Silveira, M. V., & Arntzen, E. (2023). Maintenance of stimulus equivalence classes: A bibliographic review. *The Psychological Record*, 73(1), 1–11. <https://doi.org/10.1007/s40732-023-00535-3>

- Rehfeldt, R. A. (2011). Toward a technology of derived stimulus relations: An analysis of articles published in the Journal of Applied Behavior Analysis, 1992-2009. *Journal of Applied Behavior Analysis*, 44(1), 109-119. <https://doi.org/10.1901/jaba.2011.44-109>
- Reynolds, G. S. (1961). Attention in the pigeon. *Journal of the Experimental Analysis of Behavior*, 4(3), 203–208. <https://doi.org/10.1901/jeab.1961.4-203>
- Richling, S. M., Williams, W. L., & Carr, J. E., (2019). The effects of different mastery criteria on the skill maintenance of children with developmental disabilities. *Journal of Applied Behavior Analysis*, 52(3), 701–717. <https://doi.org/10.1002/jaba.580>
- Robson, S. G., Baum, M. A., Beaudry, J. L., Beitner, J., Brohmer, H., Chin, J., Jasko, K., Kouros, C.D., Laukkonen, R.D., Moreau, D., Searston, R.A., Slagter, H.A., Steffens, N.K., Tangen, J.M., Thomas, A. (2021). Promoting Open Science: A holistic approach to changing behavior. <https://doi.org/10.31234/osf.io/zn7vt>
- Rocha e Silva, M. I., & Ferster, C. B. (1966). An experiment in teaching a second language. *IRAL - International Review of Applied Linguistics in Language Teaching*, 4(1-4), 85-114. <https://doi.org/10.1515/iral.1966.4.1-4.85>
- Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249-255. <https://doi:10.1111/j.1467-9280.2006.01693.x>
- Roediger, H. L., III, & Nestojko, J. F. (2015). The relative benefits of studying and testing on long-term retention. In J. G. W. Raaijmakers, A. H. Criss, R. L. Goldstone, R. M. Nosofsky, & M. Steyvers (Eds.), *Cognitive modeling in perception and memory: A festschrift for Richard M. Shiffrin* (pp. 99-111). New York, NY, USA: Psychology Press.

- Rohatgi, A. (2020). Webplotdigitizer (Version 4.4). Retrieved from <https://automeris.io/WebPlotDigitizer>
- Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education*, 40(1), 4–17. <https://doi.org/10.5951/jresmetheduc.40.1.0004>
- Rohrer, D., Taylor, K., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology*, 19(3), 361–374. <https://doi.org/10.1002/acp.1083>
- Romagnoli, C., & Conti, S. (2019). How and how much? Vocabulary learning strategies and vocabulary size in CFL. *Chinese as a Second Language*, 54(2), 93-121. <https://doi.org/10.1075/csl.18006.rom>
- Rosales, R., Rehfeldt, R. A., & Huffman, N. (2012). Examining the utility of the stimulus pairing observation procedure with preschool children learning a second language. *Journal of Applied Behavior Analysis*, 45(1), 173-177. <https://doi.org/10.1901/jaba.2012.45-173>
- Rosales, R., Rehfeldt, R. A., & Lovett, S. (2011). Effects of multiple exemplar training on the emergence of derived relations in preschool children learning a second language. *Analysis of Verbal Behavior*, 27(1), 61-74. <https://doi.org/10.1007/BF03393092>
- Saunders, R. R. (1996). From review to commentary on Roche and Barnes: Toward a better understanding of equivalence in the context of relational frame theory. *The Psychological Record*, 46(3), 477. <https://doi.org/10.1007/BF03395178>
- Schlosser, R., & Sigafoos, J. (2007). Evidence-based communication assessment and intervention-purpose and procedures. *Evidence-Based Communication Assessment & Intervention*, 1(1), 52-54. <https://doi.org/10.1080/17489530701269748>

- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*: Palgrave Macmillan UK.
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26-43.
<https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- Seel, N. M. (2012). Learning and thinking. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1797-1799). Boston, MA: Springer US.
- Shepley, C., Ault, M. J., Ortiz, K., Vogler, J. C., & McGee, M. (2020). An exploratory analysis of quality indicators in adapted alternating treatments designs. *Topics in Early Childhood Special Education*, 39(4), 226–237. <https://doi.org/10.1177/0271121418820429>
- Shibata, K., Sasaki, Y., Bang, J. W., Walsh, E. G., Machizawa, M. G., Tamaki, M., Chang, L.-H., & Watanabe, T. (2017). Overlearning hyperstabilizes a skill by rapidly making neurochemical processing inhibitory-dominant. *Nature Neuroscience*, 20(3), 470–475.
<https://doi.org/10.1038/nn.4490>
- Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech & Hearing Research*, 14(1), 5-13. <https://doi.org/10.1044/jshr.1401.05>
- Sidman, M. (1986). Analysis and Integration of Behavioral Units. In M. D. Zeiler & T. Thompson (Eds.), *Analysis and Integration of Behavioral Units*.: Routledge.
- Sidman, M. (2009). Equivalence relations and behavior: An introductory tutorial. *The Analysis of Verbal Behavior*, 25(1), 5-17. <https://doi.org/10.1007/bf03393066>

- Sidman, M. (2011). Can an understanding of basic research facilitate the effectiveness of practitioners? Reflections and personal perspectives. *Journal of Applied Behavior Analysis*, 44: 973-991. <https://doi.org/10.1901/jaba.2011.44-973>
- Sidman, M. (2018). What is interesting about equivalence relations and behavior? *Perspectives on Behavior Science*, 41(1), 33-43. <https://doi.org/10.1007/s40614-018-0147-8>
- Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, 37(1), 5-22. <https://doi.org/10.1901/jeab.1982.37-5>
- Sindelar, P. T., Rosenberg, M. S., & Wilson, R. J. (1985). An adapted alternating treatments design for instructional research. *Education and Treatment of Children*, 8, 67-76. <https://www.jstor.org/stable/42898888>
- Singer-Dudek, J., Greer, R.D. (2005). A long-term analysis of the relationship between fluency and the training and maintenance of complex math skills. *Psychological Record*, 55, 361–376. <https://doi.org/10.1007/BF03395516>
- Skinner, B. F. (1957). *Verbal Behavior*: Appleton-Century-Crofts.
- Stewart, I. (2017). RFT as a functional analytic approach to understanding the complexities of human behavior: A Reply to Killeen and Jacobs. *The Behavior analyst*, 40(1), 65. <https://doi.org/10.1007/s40614-017-0099-4>
- Stewart, I. (2018). Derived relational responding and relational frame theory: A fruitful behavior analytic paradigm for the investigation of human language. *Behavior Analysis: Research and Practice*, 18(4), 398-415. <https://doi.org/10.1037/bar0000129>
- Sundberg, M. L. (2015). The most important verbal operant. *VB News*, 14(2), 3-5.

- Susanto, A. (2017). The teaching of vocabulary: A perspective. *Jurnal KATA*, 1(2), 182.
<https://doi.org/http://doi.org/10.22216/jk.v1i2.2136>
- The jamovi project. (2020). jamovi (Version 1.6). Retrieved from <https://www.jamovi.org>.
- Torgerson, C. J. (2006). Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies*, 54(1), 89-102. <https://doi.org/10.1111/j.1467-8527.2006.00332.x>
- Tummers, B. (2015). DataThief III (Version 1.7). Retrieved from <https://datathief.org/>
- U.S. Department of State. (n.d.). Foreign Language Training. Retrieved from <https://www.state.gov/foreign-language-training/>
- Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, 47(1), 64-75. <https://doi.org/10.1080/09571736.2016.1191527>
- Van Der Zee, T., & Reich, J. (2018). Open Education Science. *AERA Open*, 4(3), 233285841878746. <https://doi.org/10.1177/2332858418787466>
- Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, 34(4), 457-479. <https://doi.org/10.1093/applin/ams074>
- Vargas, J. S. (2020). *Behavior analysis for effective teaching* (3rd ed.): Routledge.
- Vargas, J.S. and McLaughlin, T.F. (1977), *Behavioral Psychology for Teachers*. Nonprofit Management Leadership, 16: 14-14. <https://doi.org/10.1002/pfi.4180161009>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

- Wallace, M. D., Iwata, B. A., & Hanley, G. P. (2006). Establishment of mands following tact training as a function of reinforcer strength. *Journal of Applied Behavior Analysis*, 39(1), 17-24. <https://doi.org/10.1901/jaba.2006.119-04>
- Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*. <https://doi.org/10.1111/lang.12424>
- Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*, 104(4), 715-738. <https://doi.org/10.1111/modl.12671>
- Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, 3, 240-245. doi:10.1111/j.1467-9280.1992.tb00036.x
- Wheetley, B. (2005). *The effects of rate of responding on retention, endurance, stability, and application of performance on a match-to-sample task*. (Doctoral dissertation) Retrieved from University of North Texas Libraries, UNT Digital Library, <https://digital.library.unt.edu>.
- Wolfe, K., Barton, E. E., & Meadan, H. (2019). Systematic protocols for the visual analysis of single-case research data. *Behavior Analysis in Practice*, 12(2), 491-502. <https://doi.org/10.1007/s40617-019-00336-7>
- Wooderson, J. R., Bizo, L. A., & Young, K. (2022). A systematic review of emergent learning outcomes produced by foreign language tact training. *The Analysis of Verbal Behavior*, 38, 157-178. <https://doi.org/10.1007/s40616-022-00170-z>

- Wooderson, J. R., Bizo, L. A., & Young, K. (2024). Extracting published graphical data: A guide for researchers wanting to synthesize single-case data. *Experimental Analysis of Human Behavior Bulletin*, 34, 1-8. <https://doi.org/10.17605/osf.io/rfhjx>
- Wooderson, J. R., Bizo, L. A., & Young, K. (2024?). under review - Retention of emergent Korean vocabulary following foreign tact training and overlearning
- Wu, W. L., Lechago, S. A., & Rettig, L. A. (2019). Comparing mand training and other instructional methods to teach a foreign language. *Journal of Applied Behavior Analysis*, 52(3), 652-666. <https://doi.org/10.1002/jaba.564>
- Yousefi, M. H., & Biria, R. (2018). The effectiveness of L2 vocabulary instruction: a meta-analysis. *Asian-Pacific Journal of Foreign language Education*, 3(1). <https://doi.org/https://doi.org/10.1186/s40862-018-0062-2>
- Zigterman, J. R., Simone, P. M., & Bell, M. C. (2015). Within-session spacing improves delayed recall in children. *Memory*, 23(4), 625–632. <https://doi.org/10.1080/09658211.2014.915975>
- Zhang, P., & Graham, S. (2020). Vocabulary learning through listening: Comparing L2 explanations, teacher codeswitching, contrastive focus-on-form and incidental learning. *Language Teaching Research*, 24(6), 765-784. <https://doi.org/10.1177/1362168819829022>
- Zittoun, T., & Brinkmann, S. (2012). Learning as Meaning Making. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1809-1811). Boston, MA: Springer US.

Appendix 1

Ethics information and consent forms

Study 3

PARTICIPANT INFORMATION SHEET

Retention of Korean vocabulary learning following two different training procedures

Approval Number: UTS HREC REF NO. ETH20-5630

WHO IS DOING THE RESEARCH?

My name is John Wooderson, and I am a doctoral student at the University of Technology Sydney (UTS). My supervisor is Dr Kirsty Young (Kirsty.Young@uts.edu.au)

WHAT IS THIS RESEARCH ABOUT?

This research is to find out what teaching methods lead to better retention of foreign vocabulary. Specifically, we want to examine how well you remember Korean words for up to one month after training has ended.

WHY HAVE I BEEN ASKED?

You have been invited to participate in this study because you expressed an interest in foreign-language learning. Also,

1. You are not currently enrolled in, nor have you previously participated in any formal Korean language instructional programs; and
2. You do not live with a native-Korean speaker.

IF I SAY YES, WHAT WILL IT INVOLVE?

If you decide to participate, I will invite you to join an online Korean vocabulary training and follow-up program. This will involve up to 7 weeks (7 hours) of participation and one follow up

test 6 months following training. The program is delivered 1:1 with the instructor (not group sessions). During this program:

- I will test your existing knowledge of Korean words before training begins by asking you to pronounce and translate some words. Altogether, these pre-training tests will take up to 1 hour to complete over three separate days.
- I will then provide individualized training in two sets of 10 Korean words. The training sessions will take 30 minutes per day and continue for up to 10 days.
- After training has ended, I will test you once per week for four weeks during post-tests and once again up to 6 months later during a follow-up test to see how many words you remember. Each post-training test will take about 10 minutes to complete.
- I will video-record every training and testing session for data analysis via Microsoft Teams.

ARE THERE ANY RISKS/INCONVENIENCE?

Yes, there are some risks/inconveniences. Specifically, you may feel some discomfort or fatigue during the training program because I will ask you to continue to practice the learning material after you have learnt it. This is because we want to evaluate the retention levels that result from overlearning compared with the levels of retention produced by more traditional learning approaches. You might also experience embarrassment if you fail to learn the vocabulary.

DO I HAVE TO SAY YES?

Participation in this study is voluntary. It is entirely up to you whether you decide to take part.

WHAT WILL HAPPEN IF I SAY NO?

If you decide not to participate, it will not affect your relationship with the researchers or the University of Technology Sydney. If you wish to withdraw from the study once it has started, you can do so at any time without having to give a reason, by contacting John Wooderson (John.R.Wooderson@student.uts.edu.au)

If you withdraw from the study, we will destroy all information about you that we collected.

CONFIDENTIALITY

By signing the consent form you consent to the research team collecting and using personal information about you for the research project. All this information will be treated confidentially. Any identifying data will be kept separate from your responses and will only be accessible by members of the research team. The researchers will store session video recordings and data on learner responses during teaching and testing trials on a password-secured computer only accessible to the researchers. We will use participant's initials only on records so you will not be identifiable to others.

Copies of records to be used for reliability coding purposes will be placed in a secure Microsoft Team's repository with access available only to the research team. The researchers will keep digital copies for up to 5 years following the study's completion, after which they will be destroyed. In all instances your information will be treated confidentially.

We plan to publish the results within a research journal, to add to the existing body of scientific knowledge on effective instructional procedures. In any publication, information will be provided in such a way that you cannot be identified.

WHAT IF I HAVE CONCERNS OR A COMPLAINT?

If you have concerns about the research that you think I or my supervisor can help you with, please feel free to contact me on [REDACTED]

You will be given a copy of this form to keep.

NOTE:

This study has been approved in line with the University of Technology Sydney Human Research Ethics Committee [UTS HREC] guidelines. If you have any concerns or complaints about any aspect of the conduct of this research, please contact the Ethics Secretariat on ph.: +61 2 9514 2478 or email: Research.Ethics@uts.edu.au and quote the UTS HREC reference number. Any matter raised will be treated confidentially, investigated and you will be informed of the outcome.

CONSENT FORM

Comparing the impact of mastery and fluency training procedures on the retention of derived intraverbal relations (ETH20-5630)

Approval Number: UTS HREC REF NO. ETH20-5630

I _____ agree to participate in the research project “Retention of Korean vocabulary learning following two different training procedures (ETH20-5630)” being conducted by John Wooderson, John.R.Wooderson@student.uts.edu.au.

I have read the Participant Information Sheet, or someone has read it to me in a language that I understand.

I understand the purposes, procedures and risks of the research as described in the Participant Information Sheet.

I have had an opportunity to ask questions, and I am satisfied with the answers I have received.

I freely agree to participate in this research project as described and understand that I am free to withdraw at any time without affecting my relationship with the researchers or the University of Technology Sydney.

I understand that I will be given a signed copy of this document to keep.

I agree to be:

☐ Audio recorded

☐ Video recorded

I agree that the research data gathered from this project may be published in a form that:

☐ Does not identify me in any way

☐ May be used for future research purposes

I am aware that I can contact John Wooderson if I have any concerns about the research.

_____	____/____/____
Name and Signature [participant]	Date

_____	____/____/____
Name and Signature [researcher or delegate]	Date

Study 4 and 5

PARTICIPANT INFORMATION SHEET

ETH22-7405 - "The effects of response rate on retention of emergent Korean vocabulary"

WHO IS CONDUCTING THIS RESEARCH?

My name is John Wooderson, and I am a student at UTS. My supervisor is Dr Kirsty Young (Kirsty.Young@uts.edu.au)

WHAT IS THE RESEARCH ABOUT?

The purpose of this research is to find out what teaching methods lead to better retention of foreign vocabulary. Specifically, we want to examine how well you remember Korean words for up to six months after training.

WHY HAVE I BEEN INVITED?

You have been invited to participate because you expressed an interest in foreign-language learning.

Before you decide to participate in this research study, please check the selection criteria:

1. You are not currently enrolled in any formal Korean language instructional programs; and
2. You do not live with a native-Korean speaker.

WHAT DOES MY PARTICIPATION INVOLVE?

If you decide to participate, I will invite you to join an online Korean vocabulary training and follow-up program. This will involve up to 10 months (12 hours) of participation. The program is delivered 1:1 with the instructor (not group sessions). During this program:

Retention of emergent foreign vocabulary

- I will test your existing knowledge of Korean words before training begins by asking you to pronounce and translate some words. Altogether, these pre-training tests will take up to 1 hour to complete over three separate days.
- I will then provide individualized training in six sets of 10 Korean words. The training sessions will take 30 minutes per day and continue for up to 1 month.
- After training has ended, I will test you once per month for six to eight months to see how many words you remember. Each post-training test will take about 5 minutes to complete.
- After the post-training testing has completed, I will retrain you in the words that you can't recall and retest your retention four weeks later.
- I will video-record every training and testing session via Microsoft Teams for data analysis.

ARE THERE ANY RISKS/INCONVENIENCE?

Yes, there are some risks/inconveniences. Specifically, you may feel some discomfort or fatigue during the training program because I will ask you to continue to practice the learning material after you have learnt it. This is because we want to evaluate the retention levels that result from overlearning compared with the levels of retention produced by more traditional learning approaches. You might also experience embarrassment if you fail to learn the vocabulary

DO I HAVE TO TAKE PART IN THIS RESEARCH PROJECT?

Participation in this study is voluntary. It is completely up to you whether or not you decide to take part.

If you decide not to participate, or to withdraw from the study, it will not affect your relationship with the researchers.

WHAT IF I WITHDRAW FROM THIS RESEARCH PROJECT?

If you wish to withdraw from the study once it has started, you can do so at any time without having to give a reason, by contacting John Wooderson (John.R.Wooderson@student.uts.edu.au).

If you decide to leave the research project, we will not collect additional personal information from you (e.g. name, address, date of birth etc.), although personal information already collected will be retained to ensure that the results of the research project can be measured properly and to comply with law. You should be aware that data collected up to the time you withdraw will form part of the research project results. If you do not want me to do this, you must tell me before you join the research project.

WHAT WILL HAPPEN TO INFORMATION ABOUT ME?

By signing the consent form you consent to the research team collecting and using personal information about you for the research project. All this information will be treated confidentially.

Your information will only be used for the purpose of this research project, and it will only be disclosed with your permission, except as required by law.

It is anticipated that the results of this research project will be published and/or presented in a variety of forums. In any publication and/or presentation, information will be provided in such a way that you cannot be identified, except with your permission. Any identifying data will be kept separate from your responses and will only be accessible by members of the research team. The researchers will store session video recordings and data on learner responses during teaching and

testing trials on a password-secured computer only accessible to the researchers. We will use participant's initials only on records so you will not be identifiable to others.

Copies of records to be used for reliability coding purposes will be placed in a secure Microsoft Team's repository with access available only to the research team. The researchers will keep digital copies for up to 5 years following the study's completion, after which they will be destroyed. In all instances your information will be treated confidentially.

In accordance with relevant Australian and/or NSW Privacy laws, you have the right to request access to the information about you that is collected and stored by the research team. You also have the right to request that any information with which you disagree be corrected. Please inform the research team member named at the end of this document if you would like to access your information.

The results of this research may also be shared through open access (public) scientific databases, including internet databases. This will enable other researchers to use the data to investigate other important research questions. Results shared in this way will always be de-identified by removing all personal information (e.g. name, address, date of birth etc.).

WHAT IF I HAVE ANY QUERIES OR CONCERNS?

If you have queries or concerns about the research that you think I or my supervisor can help you with, please feel free to contact me on [REDACTED]

You will be given a copy of this form to keep.

NOTE:

This study has been approved in line with the University of Technology Sydney Human Research Ethics Committee [UTS HREC] guidelines. If you have any concerns or complaints about any aspect of the conduct of this research that you wish to raise independently of the research team, please contact the Ethics Secretariat on ph.: +61 2 9514 2478 or email: Research.Ethics@uts.edu.au] and quote the UTS HREC reference number. Any matter raised will be treated confidentially, investigated and you will be informed of the outcome.

CONSENT FORM

ETH22-7405 - "The effects of response rate on retention of emergent Korean vocabulary"

I _____ agree to participate in the research project being conducted by John Wooderson, John.R.Wooderson@student.uts.edu.au, _____

I have read the Participant Information Sheet, or someone has read it to me in language that I understand.

I understand the purposes, procedures and risks of the research as described in the Participant Information Sheet.

I have had an opportunity to ask questions, and I am satisfied with the answers I have received.

I freely agree to participate in this research project as described and understand that I am free to withdraw at any time without affecting my relationship with the researchers.

I understand that I will be given a signed copy of this document to keep.

I am aware that I can contact John Wooderson if I have any concerns about the research.

Name and Signature [participant]

____/____/____

Date

Name and Signature [researcher or delegate]

____/____/____

Date

Appendix 2

Research protocol for Study 3

Purpose

To determine if foreign tact training results in greater response retention of derived intraverbal (native-to-foreign) relations following instruction under a) accuracy or b) accuracy+rate-building conditions.

Participants

Five English speaking adults

Setting and Materials

Sessions will take place on-line via Microsoft Teams®. Materials will include datasheets, computer and webcam, and Microsoft PowerPoint® slides containing English target nouns or pictures of objects.

Retention of emergent foreign vocabulary

Set 1			Set 2		
Picture	Korean noun	English equivalent	Picture	Korean noun	English equivalent
	sajin	photo		janggap	gloves
	gawi	scissors		naembi	pot
	oi	cucumber		gamja	potato
	chima	skirt		chimdae	bed
	usan	umbrella		begae	pillow
	baji	pants		jido	map
	ageo	crocodile		subak	watermelon
	chamgmun	window		namu	tree
	sagwa	apple		sangeo	shark
	yeonpil	pencil		chiyak	toothpaste

Dependent variables

Correct answers

Tact training trials: the participant vocally labels a picture using the Korean referent

Derived native-to-foreign intraverbal (NFI) probes: the participant vocalizes the Korean referent after being presented with its English written equivalent.

Incorrect answers

"I don't know."

Target word in the incorrect language

Incorrect word

A nonsense word

Design

Adapted alternating treatments design with pre- and post-tests (counterbalanced across participants):

Set 1: 10 nouns assigned to tact instruction under accuracy conditions

Set 2: 10 nouns assigned to tact instruction under accuracy+rate-building conditions

Pre-assessment procedures

Assigning nouns to tact accuracy+rate-building training, and tact accuracy training conditions:

Assign ten common nouns - Set 1 and Set 2

Try to control for the complexity of words that are in each set of 10 nouns by making sure the number of syllables in each word is the same (e.g., ten words with two syllables for each 10-noun set)

Retention of emergent foreign vocabulary

Do not include similar sounding words in a set of 10 nouns (e.g., ori and oi) or nouns that sound similar in Korean and English (e.g., sajin and sergeant)

Try to keep the sets of 10 nouns the same (e.g., always keep clock, desk, and plum together), but change the conditions across participants (e.g., if you assign ten nouns to the accuracy condition set for Participant 1, assign the same nouns to the accuracy+rate-building condition set for Participant 2)

Conduct each of the following two assessments on separate days:

Pronunciation pre-test:

Ask the participant to repeat each noun in Sets 1 and 2 in Korean

Provide praise on an FR-1 schedule for correct responses

If the participant closely approximates the word, ask the participant to repeat the word to make sure he or she can pronounce the word correctly

If the participant cannot correctly pronounce a word, replace the noun with another one

Native tact pre-test:

Ask the participant to label pictures of each noun in Set 1 and 2 in English by asking "What is this?". Present one slide up at a time.

Provide praise on an FR-1 schedule for correct responses

If the participant's response matches the expected English equivalent noun, state, "Yes, that's (expected English equivalent)." If the participant provides a response that is synonymous with the expected noun (e.g. 'quilt' instead of 'blanket'), ask the participant to repeat the word and then acknowledge his response by stating, "Okay, you called that (noun). Would you prefer to call that (noun) or (expected English equivalent)?" Note the participant's preference, and adjust the learning materials accordingly. If the participant provides a non-synonymous response, replace

the noun with another one, and conduct pronunciation and native tact tests again with the new noun.

Assessment procedures

Pre-tests and post-tests

Conduct pre-test and post-test sessions for Training Set 1 and Set 2 using the native-to-foreign relations only.

Post-testing schedule:

Post-tests will be conducted once per week for one month

Session procedure

Before each session, tell the participant "*I am going to ask you what an English word is in Korean. I'm not going to tell you if you're right or wrong, but I want you to try your best*".

Probes will comprise 10 trials, delivered twice for each stimulus set, with sets alternated for each session. During each probe, test all ten nouns once in a randomized order.

Provide neutral comments (e.g., "Okay") for correct (within 5 seconds) and incorrect responding ('I don't not know', the target word in the wrong language, an incorrect word, or a nonsense word).

If the participant consistently responds correctly to any of the nouns during the pre-test sessions, replace the nouns and repeat the pre-assessment procedures with new nouns

When to begin post-tests:

Accuracy training condition - After twenty-five trials or the participant has met the accuracy condition training criterion (2 consecutive trials at 100% correct), whichever comes first.

Conduct post-test sessions for the accuracy training stimulus set only. Conduct post-test sessions using the procedure described above, but do not replace any of the nouns.

Accuracy+rate-building – After the participant completes ten training days or attains the rate criterion (80 corrects per minute) , whichever comes first.

Training Procedures

Session procedure

Conduct baseline and instructional sessions for both training conditions (10 sessions per day) and their assigned stimulus sets using the tact relations only and an alternating treatments design.

Initially, implement the same instructional procedures (i.e., accuracy training) for both training conditions until participants achieve 100% correct responding in two consecutive 10-trial blocks or the participant completes 25 10-trial blocks, whichever occurs first. After attaining the accuracy criterion or the end of the final training trial in the stimulus set assigned to the accuracy only condition, implement the post-testing phase for that assigned stimulus set only.

Immediately after the participant achieves accuracy criterion in the stimulus set assigned to the accuracy+ rate-building condition, implement the rate-building procedures using that stimulus set.

Baseline

Conduct baseline sessions of the assigned training conditions of Set 1 and Set 2 stimuli.

Conduct at least two baseline sessions before commencing training in each set. The important thing is to demonstrate experimental control.

Before each session, tell the participant "*I am going to show you slides with pictures and ask you to label each picture you see in Korean. After you label a picture, I'm not going to tell you if you're right or wrong, but I want you to try your best*".

Wait for 5 seconds for the participant to respond before providing a neutral comment and moving to the next slide.

Accuracy training

Tell the participant "I am going to show you the slides with pictures again and ask you to label as many pictures in Korean as you can. This time, after you label a picture, I will tell you if you were right or tell you the correct answer if you were wrong. If you are unsure about any of the pictures, you can say 'I don't know'

For each trial, wait 5 seconds before providing feedback (i.e., "You're right. That's _____", or "No, that is _____")

At the end of 10-trial session, score the number of correct responses and errors. Give feedback to the participants on their performance and encourage them to increase their accuracy if they have not yet attained the accuracy criterion. For example, "You scored higher than the previous session, well done" or "You are trying so hard, let's see if you can improve your score on the next session"

Randomise the order of slides before commencing each new trial.

Rate-building training

Only conduct this training with the stimulus set assigned to the accuracy+rate-building condition after the participant has achieved 100% correct responding in two consecutive 10-trial blocks during the accuracy training phase or the participant has completed 25 10-trial blocks, whichever occurs first. Do not conduct this training with the stimulus set assigned to the accuracy-only training condition.

Tell the participant "I am going to show you slides with pictures of the words you have learnt, and you will have one minute to label as many pictures in Korean as you can. When the timer starts, and I say 'please begin', label each picture left to right, starting with the top row, and moving to the bottom row. Then, we will move to the next slide. If you are unsure about any of

the pictures, you can say 'Don't know' or guess. Do not skip any pictures. When the timer ends, I will tell you how well you did.”

At the end of each one-minute trial, score the number of correct responses and errors. Give feedback to the participants on their performance and encourage them to increase their responding rate if they have not yet attained the target aim rate. For example, "You scored 60 corrects per minute during the last trial. Your target is 80 corrects per minute. Try to increase your rate of correct responses during the next trial".

Interobserver Agreement

A second observer will watch videos of 30% of the Microsoft Teams session recordings across all pre-test, instruction and post-test phases for each participant.

The second observer will independently collect data on participant responses during trial presentations. We will then compare the data collected by both observers on a trial-by-trial basis. If both observers record an incorrect or correct response for the same trial, we will this as an agreement; otherwise, we will score the trial as a disagreement.

We will calculate interobserver agreement for each session in which both observers collected data by dividing the number of agreements by the total number of agreements and disagreements and multiplying by 100.

Retention of emergent foreign vocabulary

Condition: Baseline/ Accuracy training /Rate-building training/Pre-test/Post-test			Date:
			Participant:
			Session:
Trial	Stimulus	Target	Response
1	apple	sagwa	Correct / Incorrect
2	pants	baji	Correct / Incorrect
3	skirt	chima	Correct / Incorrect
4	umbrella	usan	Correct / Incorrect
5	cucumber	oi	Correct / Incorrect
6	window	changmun	Correct / Incorrect
7	scissors	gawi	Correct / Incorrect
8	pencil	yeonpil	Correct / Incorrect
9	crocodile	ageo	Correct / Incorrect
10	photo	sajin	Correct / Incorrect

Treatment Integrity

During baseline, pre-, and post-test phases, the second observer will score whether the experimenter: 1) presents the discriminative stimulus, 2) waits up to 5 seconds for the participant to respond, and 3) provides neutral feedback for all learner responses.











During the accuracy phase the observer will also check whether the experimenter provides acknowledgment of a correct response or error correction rather than neutral feedback.

Finally, during the fluency phase, the observer will score whether the experimenter provides feedback to the participant on the number of correct responses only after each one-minute trial.

Instructions for participants in Study 4




Pronunciation guides

Set 1

				
ohm-jee	ee-barl	parl-goom- chee	byarm	moo-rup
				
mork	ee-mar	sorn	kor	beh











Retention of emergent foreign vocabulary

Set 2

				
dar-ree	ohk-geh	sorn-gar-rark	gwee	ip-sool
				
noon	moh-ree	barl	tohk	hyoh











Retention of emergent foreign vocabulary

Set 3

				
nuk-deh	geh	dark	dweh-jee	mool-gor-gee
				
goh-wee	nark-tar	kor-ki-ree	yarng	geh-mee

Retention of emergent foreign vocabulary

Set 4

				
sar-sum	sor	marl	yohm-sor	gor-yarng-ee
				
tar-jor	kool-bohl	geh-goo-ree	behm	sar-jar

Retention of emergent foreign vocabulary

Set 5

				
gorng	gee-char	mor-jar	jar-dorng-char	wee-jar
				
york-jor	jip	ar-gee	kar-barng	marng-chee

Set 6

				
chehk	johp-shee	shin-barl	jar-john-goh	yarng-marl
				
darm-yor	gort	yohl-seh	goh-ool	pehn-chee

Experimenter scripts and guidance

Record all sessions!

Native tact preassessment

Experimenter: "I am going to show you cards with pictures and ask you to label each picture you see in English"

Ask the participant to label pictures of each noun in Set 1-6 in English by asking "What is this?"

Present one picture card up at a time.

If the participant's response matches the expected English equivalent noun, state, "Yes, that's (expected English equivalent)."

If the participant provides a response that is synonymous with the expected noun (e.g. 'quilt' instead of 'blanket'), ask the participant to repeat the word and then acknowledge his response by stating, "Okay, you called that (noun). Would you prefer to call that (noun) or (expected English equivalent)?" Note the participant's preference and adjust the learning materials accordingly. If the participant provides a non-synonymous response, replace the noun with another one, and conduct pronunciation and native tact tests again with the new noun.

NFI sets (pre-test)

Before the first NFI timing of each day's session, tell the participant,

"Press the 'start timer' and 'next card' buttons when you are ready to begin. You will see a flashcard appear on the screen. Say the Korean word that matches the card. You can say 'don't know' if you are unsure, but don't skip any words. You will have one minute to respond to as many cards as possible."

Do not give feedback!

NFI and Tact sets (post-test)

“Press the ‘start timer’ and ‘next card’ buttons when you are ready to begin. You will see ten flashcards appear on the screen. From left to right, top to bottom, say the Korean word that matches each card. You can say ‘don’t know’ if you are unsure, but don’t skip any words. You will have one minute to respond to as many cards as possible.”

Do not give feedback!

FNI sets (post-test)

“Press the ‘start timer’ and ‘next card’ buttons when you are ready to begin. You will see ten flashcards appear on the screen. From left to right, top to bottom, say the English word that matches each card. You can say ‘don’t know’ if you are unsure, but don’t skip any words. You will have one minute to respond to as many cards as possible.”

Do not give feedback!

Lis sets (post-test)

“Press the ‘start timer’ and ‘next card’ buttons when you are ready to begin. You will see four flashcards appear on the screen. The card in the top row is a Korean word. Select the English word that matches it from one of the three cards in the second row. After you select the card, it will then have a red border. You can guess if you are unsure but don’t skip any words. You will have one minute to respond to as many cards as possible.”

Do not give feedback!

FTT sets

Model set

Before the first FTT training timing of the day for each stimulus set, vocalise each picture/word and have the participant repeat each word once; tell the participant,

“I will show you all ten pictures in the next training set and their Korean names for 30 seconds.

Repeat each word once after I say it.”

FTT no-delay (10 cards)

Before the first FTT no-delay timing of each day’s session, tell the participant,

“Press the ‘start timer’ and ‘next card’ buttons when you are ready to begin. You will see ten flashcards appear on the screen. From left to right, top to bottom, say the Korean word that matches each card. You can say ‘don’t know’ if you are unsure, but don’t skip any words. You will have one minute to respond to as many cards as possible. Your aim is to score 80 or more corrects with no errors”

FTT no-delay (10 cards)

Before the first FTT no-delay timing of each day’s session, tell the participant,

“Press the ‘start timer’ and ‘next card’ buttons when you are ready to begin. You will see a set of flashcards appear on the screen. Say the Korean word that matches each card from left to right, starting with the top row and moving to the bottom. You can say ‘don’t know’ if you are unsure, but don’t skip any words. You will have one minute to respond to as many cards as possible. Your aim is to score 80 or more corrects with no errors”

FTT delay

Before the first FTT delay timing of each day’s session, tell the participant,

“Press the ‘start timer’ and ‘next card’ buttons when you are ready to begin. You will see a flashcard appear on the screen. Say the Korean word that matches the card. You can say ‘don’t know’ if you are unsure, but don’t skip any words. You will have one minute to respond to as many cards as possible. Your aim is 100% accuracy”

Feedback set

FTT no-delay

At the end of each one-minute FTT no-delay trial, score the correct responses and errors. Give feedback to the participants on their performance.

"You scored 60 corrects per minute and 2 errors during the last trial. Try to improve your rate on the next trial"

FTT delay

At the end of each one-minute FTT delay trial, score the correct responses, calculate the percentage of correct responses and provide feedback on the participant's accuracy. Encourage participants to improve their accuracy during the next timing if they had not achieved 100% correct.

"You scored 85% correct during the last trial. Try to increase your accuracy on the next trial."

Review

Review the training set for 30 seconds. Model the words that the participant got wrong"

"I will show you all ten pictures again and their Korean names for 30 seconds. We will practice any words that you got wrong."