EVALUATING DIFFERENT TRIAL AND TRAINING DESIGNS IN COMPUTERIZED FOREIGN-LANGUAGE VOCABULARY INSTRUCTION

by

JULIANA SEQUEIRA CESAR DE OLIVEIRA

Bachelor of Science, 2015
Universidade Federal do Para,
Para, Brazil

Master of Science, 2017
Universidade Federal do Para,
Para, Brazil

Master of Arts, 2020
Texas Christian University,
Texas, United States

Submitted to the Graduate Faculty of the
College of Science and Engineering
Texas Christian University
in fulfillment of the requirements
for the degree of

Doctor of Philosophy in Experimental Psychology

April, 2023

EVALUATING DIFFERENT TRIAL AND TRAINING DESIGNS IN COMPUTERIZED FOREIGN-LANGUAGE VOCABULARY INSTRUCTION

by

JULIANA SEQUEIRA CESAR DE OLIVEIRA

Dissertation approved:

Dr. Anna Ingeborg Petursdottir, Major Professor

Dr. Kenneth Leising

Dr. Sarah Uma Tauber

Dr. Mauricio Papini

Caio Miguel

Dr. Caio Miguel

Dr. Tim Barth, for the College of Science and Engineering

ACKNOWLEDGMENTS

I would like to begin by thanking my mentor, Dr. Anna Ingeborg Petursdottir. Dr. Petursdottir guidance and support throughout these five past years were crucial for my success. I started the program with a lot of insecurities and doubts about my skills and knowledge, but Dr. Petursdottir was able to create such a supportive and welcoming learning environment that my doubts simply became goals to work on. It has been a pleasure to learn from Dr. Petursdottir, and it I am beyond grateful for the mentor/mentee relationship that I have with her. I would also like to thank Dr. Caio Miguel, Dr. Uma Tauber, Dr. Kenneth Leising, and Dr. Mauricio Papini for serving on my dissertation committee. Their expertise and feedback have been essential to the development of this document and I am grateful for their support and time.

Next, I would like to thank my family (Luciano, Daniela, Daniel, Luciano, Marina, grandma Edna and uncle Alexandre) for their faith in me. Their love and constant support were essential to give me the strength to keep learning and growing in a place so far away from home. I would like to thank my best friends for life, Jessica, Maria, Isabela, Amanda, Biba, Vic, and Sara, and my forever roommates, Dr. Julie Swets and Dr. Addison Babineau. Their friendship was and will always be critical to my growth. Finally, I would like to thank my fiancée, Mitchell Gillen, and Zilda, for their daily love and support as I worked through my dissertation (and for celebrating every and each accomplishment together).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Evaluating Different Trial and Training Designs in Computerized Foreign-Language Vocabulary Instruction

People may seek to learn languages other than their native languages for different reasons, such as employability (Belpoliti & Pérez, 2019; Makumane & Ngcobo, 2018), increased cultural awareness, and communicative competence (for a review, see Fox et al., 2019). Additionally, research has shown that people who learn a foreign language have higher reading performance, perform better in verbal and nonverbal intelligence tests compared to those who do not, and can also show structural brain changes (Barac & Bialystok's, 2012; Hermanto et al., 2012; Weber, et al. 2016). Computer-assisted language-learning (CALL) applications for self-instruction in foreign languages have grown in popularity in recent years. For example, Duolingo® had up to 30 million active users around the world in 2020 (Mascarenhas, 2020), while sharing the market with other popular CALL applications, such as Rosetta Stone® and Babbel®.

Stimulus control over vocal foreign-language responses is an important outcome of CALL if the goal is for the learner to be able to communicate in the foreign language. Some CALL applications (e.g., Duolingo®; Rosetta Stone®) employ speech-recognition software that makes it possible for the learner to receive feedback on vocal responding as they repeat words or sentences modeled by the program. However, this mechanism is minimally used to either test or directly establish stimulus control by providing feedback on vocal naming of pictures or other representations of word referents. Instead, relationships between words and their referents are typically introduced in the absence of a requirement for the learner to vocalize. The purpose of the present study was to examine how stimulus control over vocal and other foreign-language responses is affected by the design and arrangement of teaching trials.

Learning a Foreign Language

For the purpose of the current research, I will briefly define terminologies in the field of second language acquisition (Gass & Selinker, 2001), that describe the language in which a person can already communicate or pursues to learn. A *native language* refers to the first language that a child learns, also known as the primary language, or L1. L1 is usually acquired in the process of growing up with the people who speak the same language in the first years of life. A *second language* (L2) is commonly described as any language that is learned after learning L1, regardless of whether it is a third or even a fourth language learned. People that are learning a L2 are usually older than children acquiring L1 and they commonly have acquired at least one language and prior knowledge in the L1 (Nor & Rashid, 2018). L2 is usually learned when the nonnative learner is living in an environment of native speakers (e.g., Brazilians learning Icelandic in Iceland). In contrast, a *foreign language* (FL) differs from L1 and L2 in that the nonnative student is learning the new language outside the environment where that second language is commonly spoken (e.g., Brazilians learning Icelandic in Brazil). FL acquisition is commonly accomplished within the context of the classroom or through CALL applications.

Acquiring a new language is a complex task that has many different components. For example, FL learners need to acquire the *syntax* of the new language (i.e., word order, sentence structure), and *morphology* (i.e., the form of the words), which may differ from their F1 language. The student will also need to learn the sounds system of the new language, which involves the *phonology* of the new language – learning to discriminate and articulate sounds that in some cases may not exist in the learner's F1 language. One important and often time-consuming component of FL learning is *vocabulary* acquisition. Acquiring the size of 3,000 or so high frequency words of a language is considered essential to everyday use (Nation & Waring, 1997), and the minimum of 5,000 words are required in order to read for pleasure (Hirsh & Nation, 1992). Because vocabulary is cornerstone of FL learning, most of the available CALL applications focus on teaching new words and simple sentence construction in the FL as their main early learning goals. Therefore, the present studies focused on vocabulary learning.

Computer-Assisted Language Learning (CALL)

The first known example of automated programmed instruction to teach a foreign language was described by Rocha e Silva and Ferster (1966), who developed a semi-automated machine to teach German vocabulary to American college students. After 14-18 hours of instruction, students were able to label, answer questions and maintain conversations in German. Since then, computers and the Internet have made FL self-study materials increasingly easy to access and use, and there are now many software applications available that facilitate self-study of languages. Some commercially available CALL applications (e.g., Duolingo®; Rosetta Stone®) share similar features of the program described by Rocha e Silva and Fester (1966); for example, they present learning material in discrete trials in which each presentation of a stimulus (e.g., a word or a sentence) is followed by an opportunity to respond (e.g., by selecting one of several options, or typing a response), feedback on the response, and stimuli gradually increase in complexity (e.g., from single words to sentences). In addition, some applications explicitly incorporate empirically based strategies from applied cognitive science to promote learning. For example, Duolingo® is informed by research on the effects of spaced repetition and incorporates user data into the design of its spaced-repetition algorithms (Settles & Meeder, 2016). Duolingo® and Rosetta Stone® have empirical support from outcome research that demonstrates their effectiveness in teaching foreign languages to adults (Garcia, 2013; Vesselinov, 2009). Nevertheless, little research exists on how basic elements of learning-trial design affect people's acquisition of FL vocabularies through interaction with instructional software, or assist with establishment of stimulus control over vocal responses.

Trial design for vocabulary learning varies both across and within popular CALL applications. One aspect that differs across programs is the extent to which students are required to make active responses to the presented material in order to advance through the program, as well as the nature of the required responses. For example, Rosetta Stone® teaches vocabulary largely through MTS trials in which a sample stimulus (e.g., a FL word) is presented, and the learner must select the correct comparison stimulus when presented in an array (e.g., a picture). By contrast, Duolingo® mixes trial types. For example, the learner might have to respond to a spoken FL word by typing or selecting either the FL word or the L1 equivalent, or might have to type or select a FL word that is missing in a sentence. Both of these CALL applications require an active student response in every trial. Other applications require less active student responding. For example, in Memrise®, many learning trials consist of words and pictures presented together across trials without requiring any typing or selection response from the student. The current study addressed, first, the effects of contingencies on active student responding (an issue related to the design of teaching trials), and second, the relative effects of using a single trial type versus mixed trial types during instruction (an issue related to the design of training sessions).

Trial Design: Passive Viewing and Active Student Responding

Some studies on computer-assisted programmed instruction in college student populations have found that active student responding (selecting, labeling or typing when required to do so) enhances learning outcomes compared to passive viewing of materials (Haggas & Hantula, 2002; Kumar et al., 1993; Kritch et al., 1995). Kritch et al. (1995), for example, compared a constructed-response interactive instruction to a click-to-continue and passive viewing format on posttest recall about AIDS information. The constructed-response group had a significantly higher score than the other two groups, while the posttest scores for the click-to-continue and passive-observation groups did not differ. In contrast, O'Grady et al. (2021) evaluated computer-assisted equivalence-based instruction to teach graph analysis to undergraduate students. The authors found that participants that were not required to perform an active response – they would look at the stimuli and advance to the next trial – performed similarly to participants who were required to respond to the material by selecting the correct response in each trial. Because a requirement for an active student response in each trial can slow the pace of instruction (Kritch et al., 1995), there is a need to further evaluate the role of active response requirements in computer-assisted programs, including CALL.

No published studies appear to have examined the effects of response contingencies in CALL vocabulary instruction. In an unpublished study, Smith, et al. (in preparation) compared the effects of a constructed-response condition with a contingency on typing foreign words in every trial with a passive viewing condition with intermittent contingencies on typing, while teaching Arabic words to college students in a CALL context. There were no differences in acquisition rate or constructed-response post-test performance between the two conditions, suggesting the greater density of active student responding in the constructed-response condition did not facilitate learning. This study, however, did not assess vocal production of the FL vocabulary words, nor were the learners exposed to the spoken forms of the FL words. In addition, because there were intermittent contingencies on active student responding in the passive viewing condition, it did not address the extent to which stimulus control would emerge in the absence of such contingencies.

The first two experiments (Experiment 1a and 1b) of the present study sought to further compare the effects of passive viewing and active student response methods in vocabulary learning. Contingencies on active student responding were implemented in the form of compound matching to sample (MTS), in which each trial presented a compound sample (i.e., an auditory and a textual FL stimulus), to which the learner was required to respond by selecting a matching picture from an array of choices. This approach is characteristic of vocabulary-learning trials in Rosetta Stone®. Passive viewing was represented by stimulus pairing (SP) trials, in which the auditory FL stimulus, the textual FL stimulus, and the matching picture were presented simultaneously, and no response was required from the learner other than clicking to advance to the next trial. This approach is characteristic of vocabulary-learning trials in Memrise®.

Theoretical Background

Skinner's (1957) analysis of verbal behavior provides a terminological framework for describing the different types of trials with active student responses that may be implemented in CALL programs to introduce vocabulary or assess learning outcomes. According to this framework, verbal behavior should be treated as operant behavior, which is reinforced and maintained by its consequences. Additionally, Skinner (1957) proposed that a functional analysis of language as a product of reinforcement contingencies provided a framework for effectively teaching new verbal skills. He also distinguished between the different roles of the speaker and the listener behavior during a verbal episode, in which the behavior of the speaker is reinforced through the mediation of the listener, who has been trained by the verbal community to reinforce the behavior of the speaker.

Skinner (1957) described several different types of verbal operants (i.e., three-term relations between discriminative stimuli, responses, and outcomes), distinguished by their specific antecedents and consequences. One of these operants is the tact. The tact is speaker behavior in which an object (an antecedent non-verbal stimulus) evokes a vocal or gestural response considered appropriate to the stimulus by the verbal community, as a result of a history of generalized conditioned reinforcement (Cruvinel & Hubner, 2013). For example, in the presence of a ball, the person may say “Ball!”. The tact is the primary verbal operant in Skinner’s analysis because it consists of relationships between words and referents. Another verbal operant described by Skinner is the intraverbal. The intraverbal is also a speaker behavior in which the antecedent event is a verbal stimulus (what someone said, signed or wrote) that evokes a verbal response (vocalizations, signs or written words), and there is no point-to-point correspondence between the antecedent stimulus and the response. For example, when presented with the question “How do you say ball in Portuguese?” and the other person answers “Bola”.

The tact and the intraverbal can be distinguished from listener behavior, as the listener orients to an object after previous exposure to a name (a verbal stimulus). For example, when the someone asks “Point to the bola”, a person may select the ball, when other objects are present at the time. Although the listener episode involves the same name and the same object as the tact, the stimuli and responses are different. In the first one, the antecedent is a verbal stimulus and the response is selecting or orienting to the stimulus (e.g., pointing or selecting the picture of the ball), while in the tact, the antecedent is a non-verbal stimulus (for example, a picture or an object) and the response is a vocal or gestural response (e.g., labeling the picture of the ball).

It follows from Skinner's (1957) analysis that speaker and listener repertoires are functionally independent of one another, as they participate in different three-term contingencies that may be established and maintained separately. Skinner acknowledged that in mature speakers, tacts and intraverbals nevertheless may emerge in the absence of reinforcement, sometimes simply as a result of seeing an object paired with its name, similar to an SP procedure. However, he did not provide a detailed explanation of emergent tact and intraverbal control; that is, control by objects or verbal stimuli over verbal responses that emerges in the absence of previous reinforcement.

Some more recent theories have addressed the development of emergent speaker control. The Naming Hypothesis (Horne & Lowe, 1996), and Relational Frame Theory (RFT; Hayes, et al., 2001), have specified early histories of operant reinforcement that may later enable tact control to emerge in the absence of reinforcement, whereas others have proposed that emergent tact control is a direct product of Pavlovian conditioning (Dugdale & Lowe, 1990; Tonneau, 2001) or a product of equivalence class formation as a direct outcome of operant reinforcement (Sidman, 2000). With respect to the goal of the current study, the Naming Hypothesis and RFT do not make any direct predictions regarding the role of response contingencies in promoting emergent stimulus control over vocal or other responses that are not themselves included in the response contingencies. Once the higher-order

(Naming Hypothesis) or generalized (RFT) operant is established, exposure to new name-object relations (e.g., a FL word and its referent object) should be sufficient for tact or intraverbal control to emerge. Similarly, the Pavlovian conditioning hypothesis makes no predictions regarding relative efficiency as long as contiguous and contingent presentations of names and objects occurs. All three theories nevertheless implicitly assume that the relevant stimuli must be observed or attended to, and it is possible that feedback contingencies involved on overt responding in learning trials serve to promote more reliable observation. By contrast, Sidman's (2000) theory proposes that emergent stimulus control is a product of overlapping reinforcement contingencies, and therefore, seems to more explicitly predict that contingencies on student responding should facilitate outcomes involving emergent stimulus control.

MTS and SP Procedures and Emergent Stimulus Control

Research on emergent stimulus control over FL speaker behavior has focused on understanding instructional and other variables that facilitate the emergence of derived relations, involving vocal responses. The subsections below will focus on research that has examined emergent stimulus control over vocal speaker behavior as a result of SP and MTS listener instruction. I will first describe how SP and MTS procedures are usually conducted. I will then summarize research on the emergence of FL and other forms of speaker behavior as a result of both procedures, and finally, I will review studies that have compared the effects of MTS and SP on derived relations that did not involve vocal or other topography-based responding.

MTS and SP Procedures. In research on emergent vocal responding following MTS listener instruction, each instructional trial usually consists of the presentation of an auditory sample stimulus (a spoken word) and three or more visual comparison stimuli (e.g., pictures). That is, they differ from MTS trials in the present study (and CALL applications such as Rosetta Stone®) in that the sample stimulus is solely auditory without a textual component. The trials may be delivered on a computer or in tabletop format using printed picture. The participant's target response usually is touching or pointing to a specific item or picture. The correct response (selecting the correct stimulus) is followed by the delivery of a consequence, such as the experimenter saying "That's right!", the written word "CORRECT" on the computer screen, token or point delivery, sounds or animated gifs. Incorrect responses may be followed by no consequence, error feedback (e.g., presentation of the correct response), or a correction procedure (e.g., repetition of the trial with a prompt to select the correct stimulus).

In SP instructional trials, there are no programmed contingencies of reinforcement and the participant is not required to emit any response. Instead, word-object relations are introduced via the experimenter's live modeling of a tact, or contiguous presentation of the auditory and visual stimuli via computerized instruction. If a response is required from the learner at all, it is only in the form of a non-differential response to advance from one trial to the next (e.g., pressing "next" on a computer screen).

MTS and SP: Emergence of Foreign-Language Speaker Behavior. A number of studies have examined the emergence of stimulus control over FL vocal responses (i.e., emergent FL tacts and intraverbals) as a result of either MTS or SP procedures. Most of these studies have examined the extent to which a particular procedure is sufficient to produce emergent tact and intraverbal control or compared different procedures that involved contingencies on active student responding, either for children of typical development (e.g., Cao & Greer, 2019; Cortez, et al., 2020, 2021; Haegele et al. 2011; May, et al., 2016; Petursdottir & Haflidadottir, 2009; Petursdottir, et al., 2008; Rosales, et al., 2011; Rosales et al., 2012) or adults (Rocha e Silva & Fester, 1966).

Out of these studies, none have directly compared MTS and SP procedures. Most of them have simply examined the efficacy of either MTS or SP procedures to produce emergent vocal responses, or compared MTS to vocal tact instruction (for a review, see Wooderson et al., 2022). Studies on the emergence of speaker behavior following MTS have shown mixed results. College students were reported to acquire emergent speaker relations (Rocha e Silva & Fester, 1966), while studies with children failed to consistently demonstrate emergence of tact or intraverbal relations (Cortez, et al., 2020, 2021; Petursdottir & Haflidadottir, 2009; Petursdottir et al., 2008; Rosales, et al., 2011). Only two studies to date have evaluated emergent FL tacts or intraverbals following SP (Cao & Greer, 2019; Rosales, et al., 2012). Cao and Greer (2019) conducted a SP procedure with preschoolers who were native English speakers, to evaluate the acquisition of tacts in Chinese. In the SP trials, the experimenter stated the name of the stimulus (e.g., “This is X”), and then asked the participants to match the stimulus with another identical stimulus in an array (e.g., “Match X with X”). Overall, the SP procedure was effective in establishing listener and tact responding to familiar and non-familiar visual stimuli in Chinese, but only after participants received training to echo Chinese speech sounds.

Combining the mixed results on the MTS studies with a scarce number of studies on SP procedure on deriving foreign-language speaker relations, the question of which procedure would more consistently produce emergence of speaker relations remains unanswered. However, additional studies have examined emergent speaker relations following MTS and SP outside of the context of foreign-language instruction.

MTS and SP: Emergence of Other Speaker Behavior. Studies that have examined effects of either MTS or SP procedures have sometimes found that speaker relations emerged, other times that they did not, or results were variable. Past research on speaker emergence after listener MTS instruction has often been conducted for the purpose of evaluating instructional sequences (i.e., receptive-before-expressive vs. expressive-before-receptive) in language instruction with children with neurodevelopmental disorders. In this population, listener MTS instruction has typically not produced high levels of speaker responding under stimulus control (for reviews, see Contreras et al., 2020; Petursdottir & Carr, 2011), but some studies have achieved more positive results (Keintz, et al., 2011; Kobari-Wright & Miguel, 2014). Mixed results have been obtained with children of typical development (Connell & McReynolds, 1981; Griffith et al., 2018; Horne, et al., 2004; Horne, et al., 2006; Miguel, et al., 2008), whereas adults generally demonstrate emergent speaker behavior after MTS instruction (Connell & McReynolds, 1981; Griffith et al., 2018).

In studies investigating the effects of SP on emergent stimulus control, speaker behavior has similarly emerged following SP for some, but not all participants with and without neurodevelopmental disorders (Boelens, et al., 2007; Byrne, et al., 2014; Omori & Yamamoto, 2013; Petursdottir et al., 2020; Ramirez, et al., 2009; Solares & Fryling, 2019; Takahashi, et al., 2011). Another set of studies have compared SP with direct tact or intraverbal instruction, in which tact and intraverbal relations are taught directly via prompting and prompt-delay procedures, while reinforcing correct responses and correcting incorrect responses. The primary purpose of these studies (Vladescu & Kodak, 2013;

Nottingham, et al., 2017) has been to evaluate the effects of “instructive feedback”, in which extra, non-target stimuli are presented as consequent events in instructional trials (e.g., during praise statements) and students are not required to respond to those additional stimuli – similar to an SP procedure. Most of the participants in these studies acquired the secondary targets without explicit instruction, suggesting SP had the same effects on vocal behavior as instruction that placed contingencies on vocal behavior,

In summary, both MTS and SP procedures have been found to produce emergent speaker behavior. However, results have varied across participants and studies, and only two studies have compared the effects of the two procedures on emergent speaker behavior. Perez-Gonzales, et al. (2014) compared MTS instruction to a SP procedure with children of typical development. Overall, the outcomes of tact probes for SP versus MTS were mixed. Four children performed with equal accuracy in both conditions. For two other participants, more instances of emergence occurred in the MTS condition, whereas the last participant showed more emergence in the SP condition. Vallinger-Brown and Rosales (2014) also compared SP and MTS procedures, however, the number of training trials was equated for both conditions. The participants were children with autism, and the dependent variable was intraverbal responses emitted following training in each procedure. The results revealed that MTS resulted in more emergent intraverbal responses compared to SP procedure for two participants, while SP and MTS conditions produced emergent intraverbal responses in similar accuracy for a third participant.

In summary, of the two studies that directly compared emergence of vocal speaker relations following MTS versus SP procedure, neither produced conclusive results. In addition, participants in both studies were children, as has also been the case in most of the literature on emergent speaker behavior following MTS and SP instructions. As a result, implications for trial design in CALLare difficult to determine.

MTS and SP: Effects on Derived non-Vocal Responding. Three studies have compared the effects MTS and SP procedures on emergent relations that did not involve vocal responding, with adults as participants. The primary dependent measure in these studies was equivalence test performance; that is, the emergence of novel relations among stimuli based on common nodes (e.g., an emergent relation between stimulus A and stimulus C after relations were trained between stimuli A and B and between stimuli A and C), as assessed on an MTS test. Leader and Barnes-Holmes (2001) conducted a series of experiments to evaluate the effectiveness of MTS and SP procedures on the emergence of equivalence responding. In the first two experiments, SP produced consistently superior equivalence test performance compared to MTS. Results of the third experiment suggested that these results could be attributed to the presence of negative comparisons in MTS trials. However, subsequent studies failed to replicate Leader and Barnes's (2001) results, finding either mixed results (Kinloch et al., 2013) or an advantage of MTS trials (Clayton & Hayes, 2004).

In summary, Leader and Barnes-Holmes' (2001) experiments suggested that exposure to stimulus relations was the crucial variable for derived relations to emerge. Participants exposed to SP performed better on equivalence compared to MTS participants only when the presence of negative comparisons made it possible to make incorrect responses in MTS trials, thereby preventing exposure to the target "correct" relation. The experiments by Clayton and Hayes (2004) suggested the opposite: That the presence of response contingencies in MTS trials might override the effects of having more exposure to the target stimulus relations in

SP. Finally, Kinloch et al. (2013) produced mixed findings. The discrepant results may be related to the particular stimuli used; for example, Leader and Barnes-Holmes (2001) used pronounceable printed pseudowords as stimuli, whereas Clayton and Hayes (2004) used written Chinese characters for non-Chinese speaking participants. Additional research is needed on conditions under which SP and MTS trials produce differential effects on learning.

Training Design: The Potential Efficiency of Teaching a Small Number of Relations

Another difference between CALL programs is that some tend to present predominantly one type of trial, interspersed by tests using other trial types, whereas other programs present different trial types through training, in a mixed arrangement. For example, in Rosetta Stone®, the relations between the auditory stimulus “sapo” (A), the textual stimulus “SAPO” (B), and a picture of a frog (C), are usually taught by presenting the auditory stimulus and the textual stimulus together as a sample, followed by the picture of the frog (AC-B). Alternatively, in other applications as Duolingo®, the stimulus relations might be presented in an intermixed fashion. For example, the learner might first be exposed to an MTS trial for the CB relation (i.e., see a picture of a frog and select the textual stimulus “SAPO” from several options), next to an AC trial (i.e., hear “sapo” and select a picture of a frog from several options), and after that a BD trial (i.e., see the word “SAPO” and select or type the English word, D, “FROG”), and so on. In the first example, only one type of stimulus relation is directly taught (i.e., the AB-C relation), whereas in the second example, multiple types of trials are mixed together to teach a single vocabulary target (i.e., the relation between the spoken foreign word, the written foreign word, and their visual referent). The former arrangement is similar to equivalence-based instruction (EBI), in which only the minimal number of relations needed to connect a network of stimuli is taught, whereas the latter arrangement is similar to the alternative of teaching all possible relations among the stimuli.

Stimulus Equivalence and Equivalence-Based Instruction

Basic research on stimulus equivalence (SE; cf. Sidman & Tailby, 1982) focuses on how new relationships between a set of three or more dissimilar stimuli emerge without reinforcement after teaching specific direct relations with stimulus subsets. For example, learning that A goes with B, B with C, and C with D will lead the learner to behave as if A goes with C and D goes with B, even though these stimuli have never been associated through direct experience. The learners come to treat physical dissimilar stimuli as functionally interchangeable as a result of acquiring overlapping conditional discriminations (e.g., in the presence of stimulus A1, select B1 but not B2 or B3, and in the presence of B1, select C1 but not C2 or C3).

EBI is an approach to instructional design that applies the principles of stimulus equivalence to teaching educationally relevant content. Over the years, EBI technology has been widely applied to teach socially relevant concepts and skills, to a variety of individuals (Brodsky & Fienup, 2018; Rehfeldt, 2011), as recycling skills to typical developing kids (Bolanos et al., 2020), piano skills to children with autism (Hill, et al., 2020), neuroanatomy (Pytte & Fienup, 2012), and statistics to undergraduate students (Sandoz & Hebert, 2016).

The primary claimed benefit of EBI is that it is an efficient approach to instruction (e.g., Critchfield, 2018), as training one or a few relations can yield emergent responding of different relations that were not directly taught, thus potentially saving instructional time. Several studies have evaluated this claim by comparing the efficiency of EBI (i.e., teaching the minimal number of relations needed to connect a network of stimuli) to a control condition in which all possible relations among the stimuli, or a larger number of relations than necessary, are included in instruction (Fienup & Critchfield, 2011; Petursdottir & Oliveira, 2020; Oliveira, et al., 2021; Zinn et al., 2015). Across studies, results suggested that under some circumstances (e.g., depending on the training structure and mastery criteria used), EBI might require less training than CI, but yield similar performance on a later equivalence test.

Application to Foreign-Language Vocabulary Instruction

Matter et al. (2020) investigated the effects of teaching only one type of relations (tacts of visual stimuli) versus teaching multiple relations (tact, listener, native-foreign intraverbals and foreign-native intraverbals) when teaching Spanish words to four prekindergarten students. This study was similar to the aforementioned comparisons of EBI and control conditions that teach more than the minimal number of relations. However, EBI typically involves teaching two novel relations to connect three stimuli, whereas Matter et al. (2020) were able to connect foreign-language words, native-language words, and visual referents by teaching only relations between the foreign-language words and visual referents, as relations between native-language words and visual referents presumably already existed in the participants' repertoires.

Stimulus sets were either assigned to a Tact condition or to a Mixed condition in training sessions composed of 12 trials. During the Tact condition, participants were taught to label pictures in Spanish ("What is this in Spanish?"). In the Mixed condition, participants were taught to label pictures in Spanish, select pictures according to their name in Spanish ("Point to [Spanish word]"), translate words from English to Spanish ("What is [English word] in Spanish?"), and translate words from Spanish to English ("What does [Spanish word] mean?"). After reaching criterion of two consecutive sessions with 11 or 12 out of 12 trials correct for both Tact and Mixed conditions, posttest sessions assessed participants' performances for each relation (i.e., listener, tact, NF, and FN intraverbal relations). Overall, Tact training required fewer sessions to meet the mastery criterion than Mixed training. Additionally, Tact training was sufficient to produce criterion-level emergence of most untrained relations for all participants. Follow-up sessions conducted two and four months after the posttest session suggested that Tact training also resulted in better maintenance than Mixed training.

The last experiment (Experiment 2) of the current study built upon Matter et al. (2020) by comparing a training design consisting exclusively of AB-C compound MTS trials as in Experiment 1 to a mixed training design involving a greater number of relations.

Summary

In summary, the present study sought to evaluate trial designs and training designs that are commonly used in popular commercially available CALL programs. With respect to trial design, the goal was to evaluate emergent stimulus control over vocal and typed foreign-language responding following compound MTS and SP instruction. With respect to trial design, was to compare the efficiency with which participants acquired the same emergent stimulus relations as a result of an instruction that teaches a small number of relations (Compound MTS condition) versus an instruction that directly targets a greater number of relations separately (Mixed condition).

Experiment 1a

Experiment 1a used a between-subjects design to compare the effects of three computerized training conditions on emergent foreign-language vocal and typed responses.

The MTS condition was modeled after Rosetta Stone®. In each trial, the participant was exposed to a compound auditory-textual sample stimulus and learned to select the picture that corresponded with the specific compound stimuli. In the SP condition, the participant was exposed to the compound auditory-textual compound and a picture, and there was no response requirement; these trials were similar to those presented in Memrise®. The third condition was a Mixed condition, in which the participants received intermixed MTS and SP trials throughout training. Given the inconclusive results of previous comparisons of MTS procedures, neither procedure was predicted to have a consistent advantage over the other. However, Mixed was predicted to produce higher post-test performance compared to the other groups, as SP trials exposed them to all positive pairs of stimuli (Leader & Barnes-Holmes, 2001) and MTS trials included contingencies on active student responding throughout training (e.g., Kritch et al., 1995).

Method

Participants

Sixty undergraduate students (48 female; age range between 18 and 35 years old) were recruited from the psychology department's human subjects pool to participate for course credit, and were randomly assigned to three groups. The MTS group (N = 20) received MTS instruction, the SP group (N = 20) received SP instruction, and the Mixed group (N = 20) received SP and MTS training trials, in a random order. Demographic information for each group is shown in Table 1. The groups did not differ significantly on age (F = .97; p = .46) or reported gender identity (χ = 3.41; p = .49). All participants reported their native language as being English, except for one participant in the SP group reported their native language as being Hindi, one participant in the MTS reported Italian, and one participant in the Mixed condition reported their first language as being Chinese. There were no differences between groups on their ability to speak languages other than their native language, how fluent they considered themselves when speaking another language, and previous learning of a foreign language in a formal educational setting (*ps* ≥ .58). There were also no differences on the number of semesters participants learned a second language (*F* = .47, *p* = .63), and no differences were found on previous experience with CALL applications (χ = 12.48; *p* = .13).

Table 1.

Participants' Demographic information

Groups	Mean age (SD)	% female
SP	18.78 (1.67)	89.47
MTS	19.58 (1.43)	94.74
Mixed	20. 15 (3.77)	75

Setting and Materials

The experiment was completed in a single session, lasting two hours or less. The sessions occurred remotely, via the Zoom platform. SuperLab 6.0 software was used to present stimuli and record non-vocal responses.

Ten Portuguese words were selected as instructional targets. The target words were composed of two syllables each, with different consonants in each syllable. The visual stimuli consisted of ten 7 cm x 7 cm drawings (stimuli B), 10 text stimuli in Portuguese (stimuli C), and 10 text stimuli in English (stimuli C). All visual stimuli were displayed on white background (see Table 2). The auditory stimuli consisted of ten audio recordings of the spoken names of the pictures in Portuguese and ten audio recordings of the spoken names of the pictures in English.

Table 2.

Experimental stimuli

	Auditory stimuli (A)	Pictures (B)	Textual Stimuli F (C)	Textual stimuli N (D)	Auditory stimuli (E)
1	"Sapo"	Frog	SAPO	FROG	"Frog"
2	"Maca"	Stretcher	MACA	STRETCHER	"Stretcher"
3	"Bode"	Goat	BODE	GOAT	"Goat"
4	"Lixo"	Trash	LIXO	TRASH	"Trash"
5	"Loja"	Store	LOJA	STORE	"Store"
6	"Lata"	Can	LATA	CAN	"Can"
7	"Mato"	Grass	MATO	GRASS	"Grass"
8	"Fita"	Tape	FITA	TAPE	"Tape"
9	"Luva"	Gloves	LUVA	GLOVES	"Gloves"
10	"Foca"	Seal	FOCA	SEAL	"Seal"

Procedure

The first of every three participants scheduled was randomly assigned to one of the groups. The next two participants were randomly assigned to the remaining two groups, and the last participant was assigned to the last group available. Participants had access to the Zoom link via the Sona participant pool management system. Once the participant joined the session, the experimenter sent the link for the initial questionnaire using Qualtrics. The questionnaire included questions about participants' previous knowledge of Portuguese, access to a private space, followed by the informed consent, and a demographic questionnaire. Informed consent was obtained for each participant. In the informed consent, participants were told that they would work on a computerized task to learn some new words in Portuguese.

Then, the experimenter would share her screen and give remote control to the participant, followed by the experimenter's instruction: "I will now give you remote control of my screen and start the computer program. The program will let you know what to do and when you are done". After accepting the remote control, the participant was able to type and click on the stimuli presented on the experimenter's screen. Participants underwent a pre-test, training and post-test. At the end of the experiment, participants answered questions regarding the level of engagement, level of stress and previous experience learning a foreign language (see Appendix). Additionally, the experimenter vocally provided a debriefing, explaining that participants were also invited to another follow-up test two weeks later. In the follow-up test, participants retook the post-test.

Pre-, Post-, and Follow-Up Test. Participants underwent a pre-test, a post-test (administered after training) and a follow-up test (administered two weeks after the post-test). All tests were identical, except that the post-tests did not include echoic trials. Correct and incorrect responses were followed by a black screen for two seconds. During test, there were vocal trials (Echoic, Vocal Tacts, Vocal intraverbal NF and FN), Typing trials (included Typing Tact, Typing intraverbal NF and FN), and MTS trials. Each stimulus was presented once in each block, in a specific sequence (see Table 3).

Table 3.

Trial types and the number of trials in Pre-test and Post-test

Trial types	**Stimulus**	**Response**
Echoic (Pre-test only)	Auditory stimulus in Portuguese	Vocalizing in Portuguese
Vocal tact	Picture	Vocalizing in Portuguese
Typing tact	Picture	Typing in Portuguese
Vocal Intraverbal NF	Textual stimulus in English	Vocalizing in Portuguese
Typing Intraverbal NF	Textual stimulus in English	Typing in Portuguese
Vocal Intraverbal FN	Textual stimulus in Portuguese	Vocalizing in English
Typing Intraverbal FN	Textual stimulus in Portuguese	Typing in English
BC	Picture	Selecting textual stimulus in Portuguese
CB	Textual stimulus in Portuguese	Selecting picture
CD	Textual stimulus in Portuguese	Selecting textual stimulus in English
DC	Textual stimulus in English	Selecting textual stimulus in Portuguese

At the beginning of the test, participants saw the following instruction:

“Thank you for being part of this study. For this first part, please repeat out loud the words you hear.”. The first block consisted of echoic trials. They were only presented during Pre-test. When presented with stimuli A, participants were instructed to “please repeat out loud the words you hear”, or echo the word (e.g., repeat “mato”).

After that, participants saw the following instruction:

“Now, you will be asked to label some pictures and answer some questions. If you don't know the answer, you can say or type "I don't know".”

The second block presented vocal tact (VT) trials: When presented with B stimuli, the participant was asked “What is this in Portuguese?”. The third block included the typed tact (TT) trials, in which the B stimuli were presented, and the participant saw the instruction “Type the name of the picture in Portuguese”. The fourth block was the vocal intraverbal NF

(VNF), where when presented with D stimuli, the participant was instructed to say the word in Portuguese.

After that, the fifth block consisted of typed intraverbal Native-Foreign (TNF) trials, where the D stimuli was presented, and the participant would see the instruction “Type this word in Portuguese”. Then, participants were presented with a block of vocal intraverbal FN trials (VFN), in which when presented with A stimuli, the participant was instructed to provide the translation in English. Lastly, participants were presented with the typed intraverbal Foreign-Native (TFN) trials. Trials in this block presented the C stimuli, and the participant was instructed to type the translation in English. The last test block of the pre and post-test consisted of selection trials. In an MTS format, the participant was presented with the relations BC, CB, CD, DC in a randomized order. The first letter of the relations represents the sample stimuli and the second letter represents the comparison stimuli.

Training. Participants were randomly assigned to three different groups: MTS, SP, and Mixed condition. Training for the MTS group began with the following instruction:

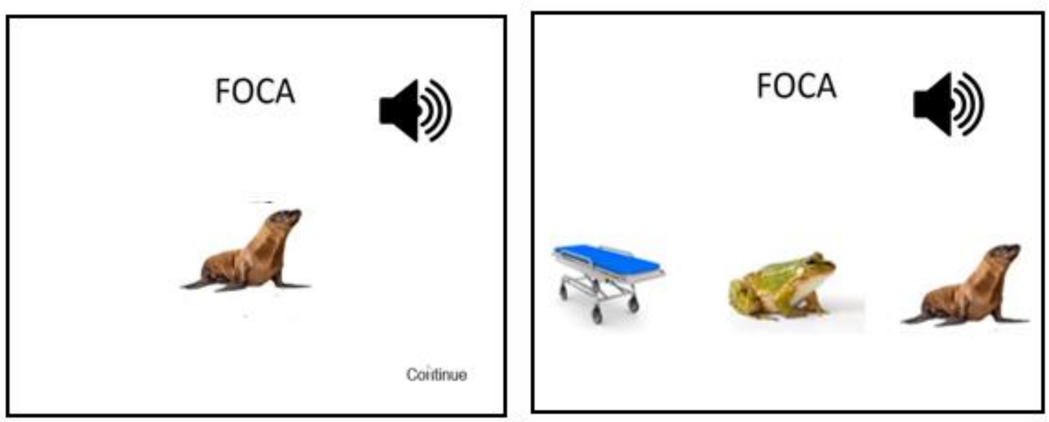
“Great job! Now, you will start the training phase.”

MTS Group. Participants in the MTS group underwent conditional discrimination training in MTS format, with the sample as an AC compound stimulus, and B as the comparison stimuli). Selection of the comparison stimulus was followed by the printed word “correct” in green letters in the center of the screen. Selection of an incorrect comparison was followed by the printed word “incorrect” in red letters, also in the center of the screen. The feedback was displayed for 1 s, followed by a 1-s intertrial interval. Training trials appeared once per block (10 trials in each block), in eighteen blocks. The number of training trials in each block was decided because in Smith et al. (in preparation), this number of training trials per target in SP produced variable outcomes with a mean of 79% correct in immediate tact post-tests. Therefore, a ceiling effect should not be an issue (i.e., MTS would have an opportunity to either outperform or underperform SP).

SP Group. Participants in the SP group were presented with stimuli A, B and C together. The participant was not required to perform any selection response, except for clicking on the arrow. The click on the arrow was followed by a white screen during a 1-s intertrial interval, and the next trial began. Participants had the same number of training trials as the participants in the MTS group. See Figure 1.

Figure 1.

Sample of an MTS trial and an SP Trial



Mixed Group. Participants in the Mixed Group also received 180 training trials (90 MTS trials and 90 SP trials). The two trial types were presented in a random order throughout the training phase.

Data Collection and Interobserver Agreement

Selection and typed responses were scored by the software. Typed responses were scored as correct only if spelled correctly. The vocal trials were video recorded and participants' vocal responses were scored as correct or incorrect by two independent observers. A second observer independently collected data on 30% of the sessions (six participants for each group) and the observers' records were compared to assess interobserver agreement. Correct responses were scored as correct only if they occur within 5 s of the presentation of the stimuli and the corresponding vocal or non-vocal instruction (e.g., "What is this?", or when presented with a picture of a frog). Responses such as "I think it's sapo" or "maybe sapo?" were counted as correct responses, as long as the participant did not include any other experiment word (e.g., "it's maybe mato or sapo"). Incorrect responses were defined as: (a) saying any of the other names used in the experiment, (b) saying any other words not used in the experiment but that were not correct, and (c) not saying any of the experimental names within 5 s. Approximations that consisted of minor articulation errors were accepted and were defined for each participant based on articulation in the echoic trials.

Interobserver agreement for each Tact, Vocal Intraverbal N-F and Vocal Intraverbal F-N test trials was calculated by dividing the number of agreements by the total number of trials, and converting to a percentage. For Tact test trials, mean agreement was 90% (range, 80% - 100%) for MTS group, 90% (range, 80% - 100%) for SP group and 90% (range, 80% - 100%) for Mixed group. For Vocal Intraverbal N-F, mean agreement was 90% (range, 80% - 100%) for MTS group, 88.33% (range, 70% - 100%) for SP group and 90% (range, 80% - 100%) for Mixed group. Finally, for Vocal Intraverbal F-N, mean agreement was 96.67% (range, 90% - 100%) for MTS group, 95% (range, 80% - 100%) for SP group and 98.33% (range, 90% - 100%) for Mixed group.

Data Analysis

Across the three experiments, the primary dependent measure was performance accuracy on the final test after training, broken down by trial type. Other measures included the time each group spent during training in each condition and their performance during the follow-up test.

Inspection of the data revealed the distribution of some dependent variables departed at least mildly from normality ($p < .05$ on the Shapiro Wilk test). Based on data on the robustness of ANOVA and ANOVA alternatives to normality violations (Blanca et al., 2017; Lix et al., 1996), the following decision rules were used to tests for between-group differences: The Kruskal-Wallis test was used in case of extreme departures from distribution symmetry as determined by skewness values exceeding ±2.0. For less extreme departures from normality, Fisher's ANOVA was used if variances were homogeneous ($p \geq .05$ on Levene's test) and Welch's ANOVA for heterogeneous variances. Outliers were removed, when applicable, first by visually identifying potential outliers on boxplots. After that, participants that performed two standard deviations below or above the mean were removed from the analysis.

Results

Pre-test

During Pre-test (see Table 4), one participant in the SP group responded correctly for 2 trials in the Vocal Intraverbal and Typing Intraverbal F-N trials, and some correct selection responses were made for MTS trials (range from 0 to 65% correct responses). For MTS group, four participants responded correctly for 1 or 3 stimuli across the trial types. The range of correct responses for MTS trials was from 0 to 62.5%. In the Mixed group, occasional correct trials occurred for MTS trials (range 0 – 67.5%).

Table 4.

Mean (SD) Percent Correct during Pre-test

	SP	MTS	Mixed
Echoic	97.5 (5.5)	98.95 (3.15)	99.44 (2.36)
Vocal Tact	0	0	0
Typing Tact	0	2.5 (5.5)	0
Vocal Intraverbal N-F	0	0.5 2.24)	0
Typing Intraverbal N-F	0	2 (5.23)	0
Vocal Intraverbal F-N	1.05 (4.59)	1.5 (6.71)	0
Typing Intraverbal F-N	1 (4.47)	2 (6.96)	0
MTS	22.25 (18.79)	31 (12.44)	28 (17.71)

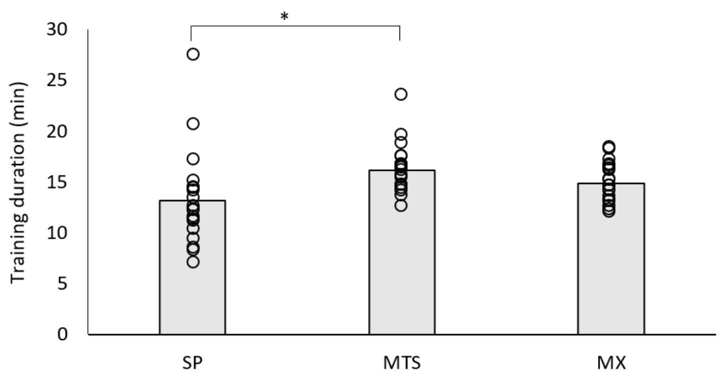
Performance during Training

Participants in all groups were exposed to 180 training trials total. During training, MTS participants required a mean of 6.15 blocks (range 0-16) before achieving 100% of correct responding in two consecutive training blocks, while Mixed group required only 3.2 blocks (range 0-12) until reaching the same criterion in MTS trials. This difference was not statistically significant (t = 1.69, p = 0.20). All participants, in both groups, reached criterion.

Training condition had an effect on the duration of the training phase, F = 4.46, p = 0.02 (see Figure 2). The SP group required statistically less time to complete training (M = 11.83 minutes, SD = 2.59) when compared to the MTS group (M = 15, SD = 3.32), p = 0.012, but there were no differences in training duration between the SP and Mixed groups (M = 14.42, SD = 1.97), or the MTS and Mixed groups ($ps \geq 0.24$).

Figure 2.

Mean and Individual Training Duration in Minutes



Note. Bars represent group means, and open circles represent individual participants data.

Post-test results

Although SP numerically outperformed MTS and Mixed groups for most of the variables during post-test, there were no significant statistical differences between group means ($ps \geq .29$). See Table 5 for post-test results.

Table 5.

Mean (SD) Percent Correct during Post-test

	SP	**MTS**	**Mixed**
Vocal Tact	70 (15)	71.56 (21.50)	72 (25.73)
Typing Tact	70 (6.67)	65 (35.05)	63 (23.59)
Vocal Intraverbal N-F	71.11 (16.91)	58.75 (24.16)	70 (23.09)
Typing Intraverbal N-F	68.89 (18.33)	65 (30.24)	53 (30.93)
Vocal Intraverbal F-N	96.67 (7.07)	86.23 (25.05)	96 (8.43)
Typing Intraverbal F-N	97.78 (4.01)	90 (16.04)	85 (12.69)
MTS	99.4 (1.10)	94.38 (6.65)	98.5 (2.11)

The criterion to pass the test for each measure was defined as 90% correct responses or more. Overall, numerically more participants in the Mixed group passed each measure during Post-test when compared to the other groups, for almost all the measures assessed. The difference seemed to be more pronounced in one of the main dependent variables, Vocal Tact trials, with a difference of 20% for SP group and 30% for MTS group (see Table 6). However, there were no statistical differences between groups ($ps \geq 0.07$). There was a marginal difference between groups in MTS trials ($\chi = 5.86$; $p = .053$). All participants in the Mixed and SP groups passed MTS trials test, while only 17 out of 20 participants in the MTS group passed the test.

Table 6.

Percentage (%) of Participants that Met Passing Criterion in Post-test

	SP	**MTS**	**Mixed**
Vocal Tact	30	22.22	52.94
Typing Tact	14.29	20	20
Vocal Intraverbal N-F	20	22.22	33.33
Typing Intraverbal N-F	14.29	40	30
Vocal Intraverbal F-N	70	77.78	88.89
Typing Intraverbal F-N	60	80	70
MTS	100	85	100

Follow-up Test

Thirty-two participants (11 participants from SP group, 10 from MTS group and 11 from Mixed group) received a Follow-up test two weeks after the experimental session.

Overall, SP group continued to do numerically better at Follow-up test on all measures, including the main dependent variables, except during MTS trial types. The Mixed group had numerically the poorest retention on all measures except for MTS trial types (see Table 7).

Table 7.

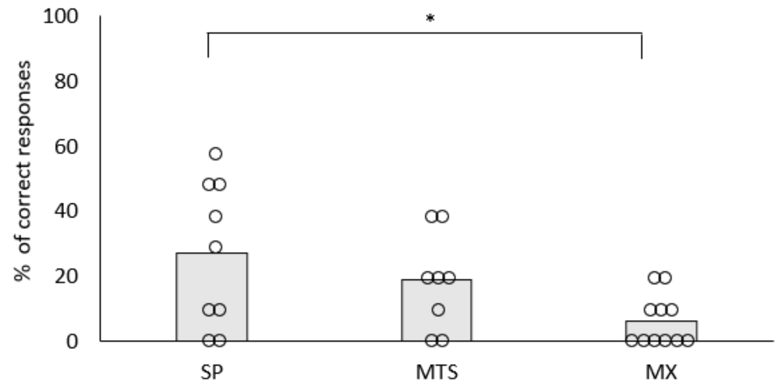
Mean (SD) Percent Correct during Follow-up Test

	SP	MTS	Mixed
Vocal Tact	26.67 (23.45)	18.75 (15.53)	5 (7.07)
Typing Tact	37.78 (29.62)	25 (23.30)	15 (15.09)
Vocal Intraverbal N-F	30 (25)	20 (18.52)	10 (8.16)
Typing Intraverbal N-F	42.22 (31.53)	22.5 (19.82)	18 (16.87)
Vocal Intraverbal F-N	68.89 (23.69)	56.25 (28.25)	49 (14.5)
Typing Intraverbal F-N	71.11 (26.19)	58.75 (24.16)	58 (20.98)
MTS	81.82 (18.51)	88.13 (13.87)	92.95 (5.68)

Training had a statistically significant effect only in Vocal Tact trials ($W = 4.37$, $p = 0.02$), in which SP group performed significantly better than the Mixed group ($p = 0.04$). See Figure 3. There were no differences between conditions for the other variables analyzed during the Follow-test ($ps \geq 0.22$).

Figure 3.

Mean Percent Correct for Vocal Tact trials during Follow-up Test



Note. Bars represent group means, and open circles represent individual participants data.

In general, numerically, more participants in the SP group passed the Typing Intraverbal N-F, Vocal and Typing Intraverbal F-N trial types than MTS and Mixed groups (see Table 8). In MTS trials types, numerically, more participants in the Mixed group passed the Follow-up test (72.72%), followed by MTS group (60%) and SP group (27.27%). However, these differences were not statistically significant ($ps \geq 0.09$).

Table 8.

Percentage (%) of Participants that Passed each Trial Type during Follow-up Test

	SP	**MTS**	**Mixed**
Vocal Tact	0	0	0
Typing Tact	0	0	0
Vocal Intraverbal N-F	0	0	0
Typing Intraverbal N-F	11.11	0	0
Vocal Intraverbal F-N	22.22	12.5	9.09
Typing Intraverbal F-N	44.44	20	0
MTS	27.27	60	72.72

Post-Questionnaire

No differences between groups were found for measures of interest, engagement, or level of stress during the task (*ps* ≥ 0.68).

Discussion

Overall, there were no differences between conditions for most of the measures assessed during Post and Follow-up tests, except for differences in training duration and Vocal Tact trials during Follow-up test. Overall, it was expected that the SP group would require less training time, as participants were not required to engage in selection responses followed by feedbacks. It was also expected the Mixed condition would perform better during Post-test and Follow-up test measures, but results were inconsistent throughout the experiment. For example, even though participants in the Mixed group numerically outperformed participants in the SP group when looking at the percentage of participants that passed each measure on Post-test, SP group statistically performed better than Mixed condition during Vocal Tact trials in the Follow-up test.

A potential explanation to why the Mixed condition did not consistently outperform the other conditions could be that because trials in the Mixed condition were intermixed in a random fashion. As a result, participants could be exposed to MTS trials for certain targets before being exposed to any SP trials for those targets. Therefore, Experiment 1b repeated Experiment 1a with a change in the Mixed condition.

Experiment 1b

The main goal of Experiment 1b was to evaluate if the effects of the Mixed condition would be enhanced by guaranteeing that participants assigned to this condition would be exposed to positive pairing for all AB-C relations before being required to respond to any of them in MTS trials. Meanwhile, MTS and SP conditions received the same training as before.

Method

Participants

Sixty-one undergraduate students participated. The groups did not differ significantly on age (F = 1.77; p = 0.18) or reported gender identity distribution (χ = 0.84; p = 0.66). See Table 9. All participants reported their native language as being English, except for two participants in the SP group reported their native language as being Swahilli and Burmese, one participant in the MTS reported Spanish, and one participant in the Mixed condition reported their first language as being Spanish as well. There were no differences between groups on their ability to speak languages other than their native language, and previous learning of a foreign language in a formal educational setting ($ps \geq .24$). There were also no differences on the number of semesters participants learned a second language, $F = .71$, $p = .50$, and no differences were found on previous experience with CALL applications ($F = .22$; $p = .80$). Six participants in the SP group considered themselves fluent when speaking another language, compared to only one participant in each MTS and Mixed groups ($\chi = 6.72$, $p = .04$).

Table 9.

Participants' Demographic Information

Groups	Mean age (SD)	% female
SP	19.28 (1.58)	80.95
MTS	20.05 (2.89)	73.68
Mixed	18.90 (0.85)	68.42

Materials and Procedure

Experiment 1b was identical to Experiment 1a, except that instead of presenting MTS and SP trials intermixed, participants assigned to the Mixed group received one block of SP trials (10 trials consisting of one presentation of each experimental stimulus), followed by one block of MTS trials. The experimental stimuli, the sequence of trial types presentation on tests, and the number of training trials were the same as Experiment 1

Data Collection and Analysis

Data were analyzed and interobserver agreement was assessed in the same way as in Experiment 1a. For Tact test trials, mean agreement was 92.5% (range, 80% - 100%) for MTS group, 88% (range, 80% - 90%) for SP group and 93.33% (range, 70% - 100%) for

Mixed group. For Vocal Intraverbal N-F, mean agreement was 95% (range, 90% - 100%) for MTS group, 88% (range, 80% - 90%) for SP group and 86.67% (range, 70% - 100%) for Mixed group. Finally, for Vocal Intraverbal F-N, mean agreement was 95% (range, 90% - 100%) for MTS group, 98% (range, 90% - 100%) for SP group and 93.33% (range, 80% - 100%) for Mixed group.

Results

Pre-test

During pre-test, one participant in the SP condition responded correctly in two trials Typing Tact trials, three Typing Intraverbal N-F trials and eight Typing Intraverbal F-N trials. Three other participants made one correct response in the Typing Intraverbal F-N. The SP group's correct selection responses in MTS trials ranged from 12.5% to 97.5%. In the MTS group, two participants made one to three correct responses in Tact, Typing Intraverbal N-F, and Typing Intraverbal F-N trials. Correct selection responses for MTS trials ranged from 0 to 60%. Finally, in the Mixed condition, some participants responded one to three correct trials in Typing Intraverbal N-F and Typing Intraverbal F-N trials. Correct responses in the MTS block ranged from 17.5 to 85%. See Table 10.

Table 10.

Mean (SD) Percent Correct during Pre-Test

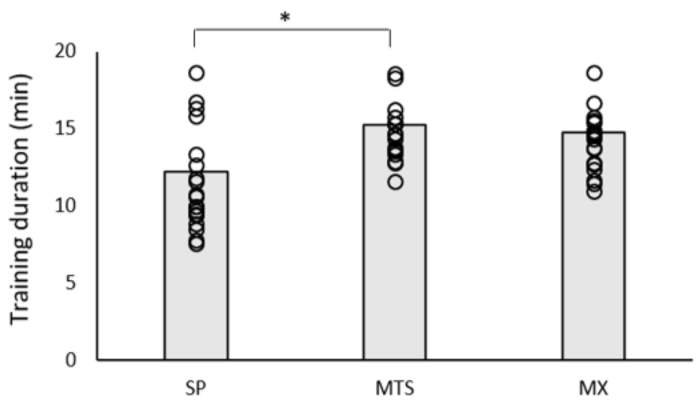
	SP	**MTS**	**Mixed**
Echoic	95.83 (14.43)	95.83 (7.93)	98.13 (5.44)
Vocal Tact	0	0.83 (2.89)	0.63 (2.5)
Typing Tact	0	3.33 (7.78)	0
Vocal Intraverbal N-F	0	0	0.63 (2.5)
Typing Intraverbal N-F	0	3.33 (8.88)	0.63 (2.5)
Vocal Intraverbal F-N	0	0	1.25 (5.0)
Typing Intraverbal F-N	1.67 (3.89)	3.33 (7.78)	3.13 (8.73)
MTS	33.96 (22.35)	28.79 (22.75)	37.97 (23.65)

Performance during Training

During training, MTS participants required a mean of 7.0 blocks (range 3-16) before achieving 100% correct responding in two consecutive training blocks, while the Mixed group required only 2.9 blocks in MTS trials (range 2-8) until reaching the same criterion. Two participants (one in MTS and another in the Mixed group) did not reach criterion during training. Training condition had an effect on the duration the training phase, $F = 4.91$, $p = 0.011$ (see Figure 4). The SP group required statistically less time to complete training ($M = 13.07$ minutes, $SD = 4.50$) when compared to the MTS group ($M = 16.17$, $SD = 2.34$), $p = 0.003$.

Figure 4.

Mean and Individual Training Duration in Minutes



Note. Bars represent group means, and open circles represent individual participants data.

Post-test results

Overall, MTS and Mixed tended to numerically outperformed SP group in most of the measures assessed (see Table 11). However, the only significant effect of condition was on MTS trials ($K = 6.92$, $p = 0.03$), in which the Mixed group performed significantly better than SP group, $p = 0.009$ (see Figure 5).

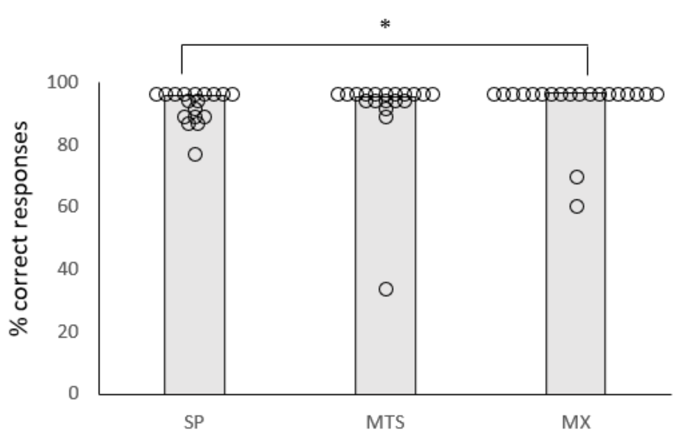
Table 11.

Mean (SD) Percent Correct during Post-Test

	SP	MTS	Mixed
Vocal Tact	70.08 (24.54)	68.87 (19.22)	75.83 (21.96)
Typing Tact	60 (28.84)	70 (21.76)	69 (21.74)
Vocal Intraverbal N-F	64.16 (24.73)	64.70 (22.67)	67.78 (25.10)
Typing Intraverbal N-F	64.5 (31.20)	67.5 (24.03)	67 (22.50)
Vocal Intraverbal F-N	82.38 (17.86)	92.94 (13.58)	91.11 (19.37)
Typing Intraverbal F-N	82.38 (18.41)	93 (17.50)	91 (16.83)
MTS	95.66 (10.62)	95.26 (14.74)	96.75 (10.13)

Figure 5.

Mean Percent Correct for MTS Trials during Post-Test



Note. Bars represent group means, and open circles represent individual participants data.

Overall, there were no differences between groups on other variables assessed (*ps* ≥ 0.13). However, when outliers were removed from the analysis, there was an effect of condition on Vocal Intraverbal F-N (W = 7.58, p = 0.001) and Typing Intraverbal F-N (W = 5.51, p = 0.009). In Vocal Intraverbal F-N, SP performed statistically worse than MTS (p = 0.019) and Mixed groups (p = 0.003). In typing intraverbal F-N, participants assigned to the Mixed group performed significantly higher than participants in SP group (p = 0.01). See Figures 6 and 7.

Figure 6.

Mean Percent Correct for Vocal Intraverbal F-N during Post-test

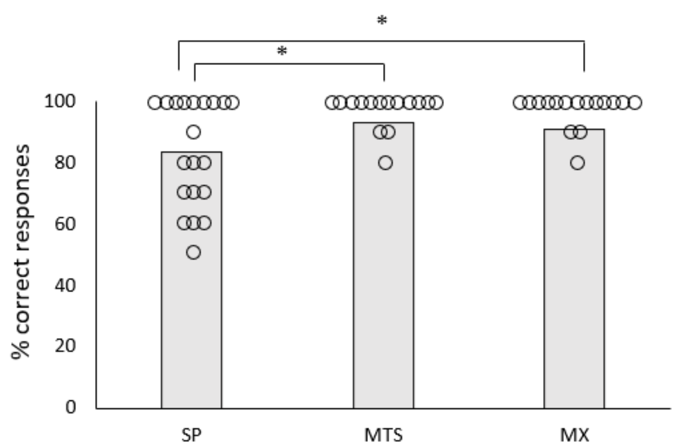
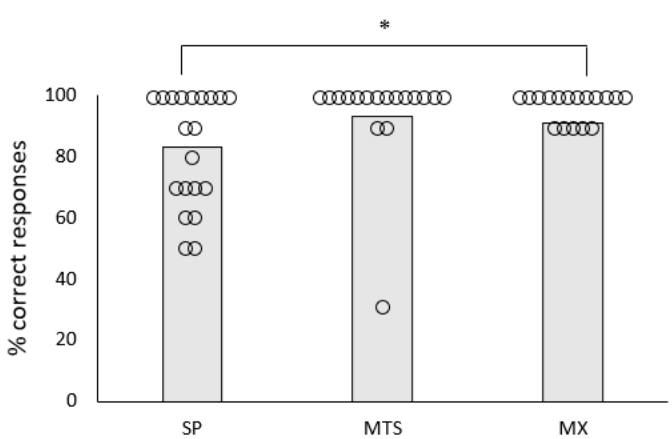


Figure 7.

Mean Percent Correct for Typing Intraverbal F-N trials during Post-test



Note. Bars represent group means, and open circles represent individual participants data.

In general, numerically, more participants in the MTS and Mixed groups passed most of the assessed measures when compared to SP group. The difference between the proportion of participants that passed each measure in Mixed and SP groups is evident for example, in Vocal Intraverbal N-F (32% difference) and Typing Intraverbal F-N (38% difference). However, these differences were not statistically significant (*ps* ≥ .13), except for Typing Intraverbal F-N trials (χ = 9.29, *p* = .01). More participants in the MTS and Mixed groups (18 and 17 respectively) passed Typing Intraverbal F-N test trials, while 11 participants in the SP group passed the test.

Table 12.

Percentage (%) of Participants that Passed each Trial Type during Post-Test

	SP	**MTS**	**Mixed**
Vocal Tact	20	21.05	31.58
Typing Tact	20	30	30
Vocal Intraverbal N-F	15.79	22.22	33.33
Typing Intraverbal N-F	35	25	30
Vocal Intraverbal F-N	52.38	77.78	77.78
Typing Intraverbal F-N	47.62	90	85
MTS	94.74	95	90

Follow-up Test results

The last experimental session occurred in one of the last days of the school semester, and the psychology department's human subjects pool closing for the semester. Thus, it was not possible to conduct a Follow-up session in this experiment.

Post-Questionnaire

There were no differences between groups on measures of interest, engagement, or level of stress during the task (*ps* $\geq$ 0.10).

Discussion

The results of Experiment 1b were similar to the results of Experiment 1a in that there were no consistent differences between groups. The change in the Mixed condition from Experiment 1b did not produce the predicted advantage of this condition over the MTS and SP conditions. A difference between the two experiments is that in Experiment 1a, the only significant difference occurred at Follow-up and was in favor of the SP group over the Mixed group, and in general, the SP group tended to numerically outperform the other groups at Post-test and Follow-up. In Experiment 1b, by contrast, the SP group performed worse than at least one of the other groups on two post-test measures. The reasons for this are unclear.

One possibility was that participants in the SP group clicked very fast through the SP trials in the second experiment. To assess this possibility, we conducted correlations between training time and the measures assessed during test. In both experiments, we found no correlation between training time and performance during Post-test for SP group (*ps* $\geq$ 0.054). We also analyzed the median reaction time for each participant then calculated the mean of medians across the two experiments. Overall, SP group in Experiment 1b responded numerically slower (M = 2556.6 ms, SD = 979.89) than participants in SP group in Experiment 1a (M = 2245.28 ms, SD = 1039.69), but no statistical differences were found between groups (p = 0.33).

Overall, the two experiments did not suggest an advantage of one trial design over the other, nor did it suggest an advantage of mixing MTS and SP trials in terms of facilitating the emergence of tact or intraverbal relations.

Experiment 2

As previously mentioned, some CALL applications available (e.g., Rosetta Stone®) employ the strategy that is similar to the MTS condition in the first two previous experiments. In this training design, participants were presented with the auditory stimulus (A) together with the written word in Portuguese (C). Participants engaged in a selection response (B) followed by feedback. In this compound MTS condition, participants learn AC-B relations across the experiment. Overall, participants assigned to this condition performed similarly to participants in the Mixed condition, for most of the measures assessed on Post and Follow-up Tests.

Other CALL applications (e.g., Duolingo®) present the different trial types in a mixed arrangement. For example, in some trials, the individual practices an BA relation, followed by AC, BC or CB relations. The stimulus relations are usually presented intermixed throughout training. The goal of the present study was to compare the compound MTS trial arrangement alone (similar to the MTS condition in the previous experiments) to intermixed trials that address different types of relations related to the same word.

A recent study (Colasurdo et al., 2023) demonstrated that adults participants exposed to a compound arbitrary stimulus in an MTS task required significantly fewer trials and training blocks to demonstrate emergence of new and untrained relations when compared to participants assigned to a single compound group. Overall, participants assigned to compound and single stimuli group performed similarly during post-test, suggesting a potential benefit of efficiency when incorporating compound stimuli during MTS task. It is important to note that Colasurdo et al. (2023) trained a minimal number of relations in both conditions, so they both could be considered EBI conditions. In the current study, the mixed MTS trials were equivalent to a complete instruction (Petursdottir & Oliveira, 2020; Oliveira et al., 2021), except that the trials with auditory comparisons (e.g., spoken words in English and in Portuguese) were omitted from training.

In the current study, the first prediction was that participants in the Compound MTS condition would require less training trials than Mixed MTS, as the Mixed condition includes redundant relations (e.g., BC and CB). Previous studies comparing EBI vs. CI have found that at least under some circumstances, a condition similar to CI might take more trials to complete training compared to training that addresses a minimal number of relations (Fienup & Critchfield, 2011; Oliveira et al., 2021; Petursdottir & Oliveira, 2020; Zinn et al., 2015). A second prediction was that both groups would perform similarly during Post-test. It is possible that Mixed MTS group could have an advantage over the Compound MTS group at test, because stimulus control would have been established by all elements over relevant selections. By contrast, in the Compound MTS group, it is possible that only the auditory stimulus or only the textual stimulus would gain control over the relevant picture selection. However, Colasurdo et al. (2023) found that participants who received Compound training did as well at test as those who received training with elements in isolation.

Method

Participants

Forty-one students were randomly assigned to two different groups: Compound MTS ($N = 20$) and Mixed MTS ($N = 21$). The groups did differ significantly on age ($t = 2.19$; $p =$

0.03), or reported gender identity (χ = 0.00; p = 1.00). See Table 13. Three participants in the Compound MTS group reported their native language as being Spanish and another participant reported their first language as being Hindi. Two participants in the Mixed MTS reported their first language as Spanish, and one participant reported their first language as being Vietnamese. There were no differences between groups on their ability and fluency to speak languages other than their native language, previous learning of a foreign, the number of semesters participants learned a second, and on previous experience with CALL applications ($ps \geq .14$).

Materials and Procedure

The software, stimuli, and procedures were the same as the previous experiments. Experiment 2 statistical decision rules were the same as Experiments 1a and 1b, with the exception that Mann-Whitney U test was used instead of Kruskal-Wallis.

Table 13.

Participants' Demographic information

Groups	**Mean age (SD)**	**% female**
Compound MTS	20.37 (1.26)	89.47
Mixed MTS	19.47 (1.26)	88.89

Participants in both groups received feedback for correct and incorrect response in the same way as previous experiments. Participants were first exposed to a fixed number of 70 training trials, followed by the presentation of more training trials, in which were repeated until the participant reached the mastery criterion of 20 consecutive correct trials. The mastery criterion was the same for both groups. Additionally, pre and post-tests were the same as the previous experiments, except that MTS test trials also included the relations AB, AD, and EC.

The training phase began with the following instruction:

“Great job! Now, you will start the training phase. The more correct responses you make, the sooner you will be done.”

Mixed MTS group. Participants received an MTS training that included the relations AB, AD, BC, CB, CD, DC and EC. Each of the ten target words were presented once for each relation type, resulting in a total of 70 different trial types. This decision was made to guarantee that participants were exposed to each experimental stimulus at least once, for each relation type. After receiving the first 70 trials, participants received more training trials, until reaching the criterion of 20 consecutive correct responses.

Compound MTS group. Participants received MTS training in which A and C stimuli were presented together as the sample, while B was the comparison (i.e., AC-B). Participants were first exposed to seven training blocks (total of 70 training trials), to equate the number of training trials that participants in the Mixed MTS group had to receive in order to be exposed to at least one experimental stimulus for each relation type. Then, training trials were presented and repeated, until the participant achieved the mastery criterion of 20 consecutive correct trials.

Table 14.

Trial Types and Number of Trials in Pre and Post-tests

Trial types	Number of trials
Echoic (only Pre-test)	10
Vocal tact	10
Typing tact	10
Vocal Intraverbal NF	10
Typing Intraverbal NF	10
Vocal Intraverbal FN	10
Typing Intraverbal FN	10
AB	10
AC	10
AE	10
BC	10
CB	10
CD	10
DC	10

Interobserver agreement was assessed in the same way as Experiments 1a and 1b. For Tact test trials, mean agreement was 92% (range, 90% - 100%) for Compound MTS group, and 90% (range, 80% - 100%) for Mixed MTS group. For Vocal Intraverbal N-F, mean agreement was 90% (range, 80% - 100%) for Compound MTS group, and 98% (range, 90% - 100%) for Mixed MTS group. Finally, for Vocal Intraverbal F-N, mean agreement was 94% (range, 90% - 100%) for Compound MTS group, and 100% for Mixed MTS group.

Results

Pre-test

During Pre-test (see Table 15), few participants responded correctly across the assessed measures. The higher number of correct responses on Pre-test occurred on MTS test trials, with variable range of responding for Compound MTS (range 0 – 62.85%) and for Mixed MTS group (0 – 77.14%).

Table 15.

Mean (SD) Percent Correct during Pre-Test

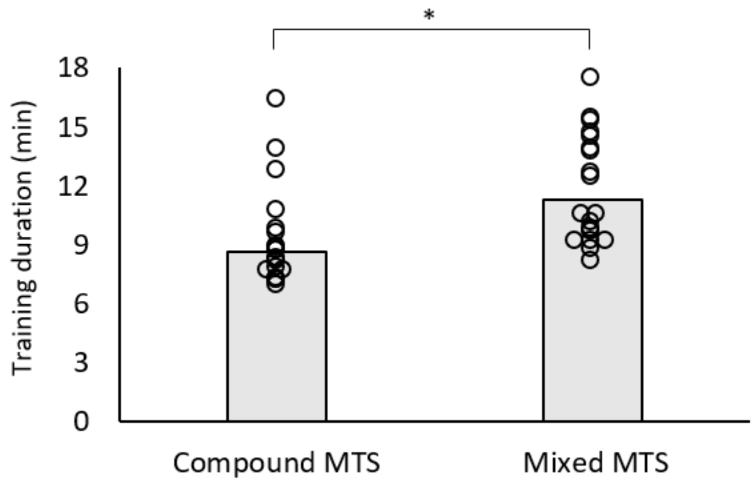
	Compound MTS	**Mixed MTS**
Echoic	97.33 (5.94)	98.13 (4.03)
Vocal Tact	0.5 (2.23)	1.43 (6.55)
Typing Tact	0.5 (2.24)	1.9 (6.80)
Vocal Intraverbal N-F	1 (4.47)	1.43 (6.55)
Typing Intraverbal N-F	0	1.43 (6.55)
Vocal Intraverbal F-N	1 (4.47)	1.43 (6.55)
Typing Intraverbal F-N	2 (6.96)	1.9 (6.02)
MTS	32.93 (15.88)	34.54 (19.43)

Performance during Training

There was an effect of condition on training time ($U = 321$, $p = 0.001$). Participants assigned to the Compound MTS group required less time to reach criterion in minutes (M = 8.50, SD = 2.73) when compared to participants in the Mixed MTS group (M = 11.52, SD = 2.98). See Figure 8.

Figure 8.

Mean and Individual Training Duration in Minutes

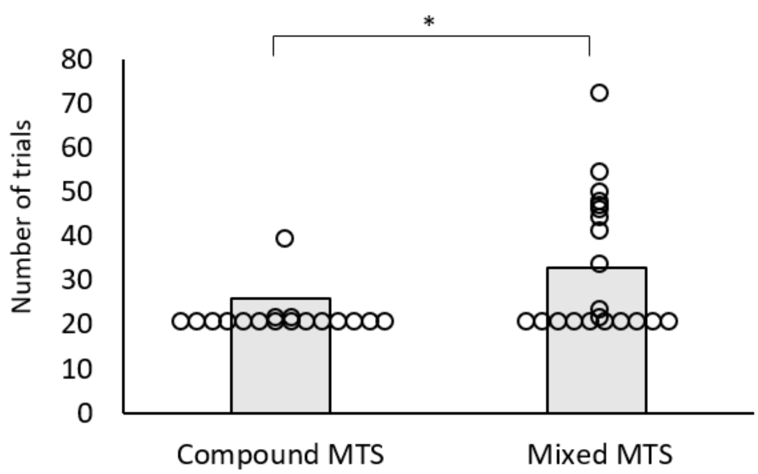


Note. Bars represent group means, and open circles represent individual participants data.

There was also an effect of condition on the number of trials to reach criterion during training when outliers were removed from the analysis (U = 253, p = 0.03). Participants in the Compound MTS group required fewer training trials (M = 25.39, SD = 15.93) than the Mixed MTS group (M = 33.50, SD = 16.95). See Figure 9. The first 70 trials that both groups were exposed to were excluded from the analysis.

Figure 9.

Trials to Reach Criterion during Training



Post-test results

Overall, there was no effect of condition on any of the variables assessed during Post-test (*ps* ≥ 0.08). See Table 16.

Table 16.

Mean (SD) Percent Correct during Post-Test

	Compound MTS	**Mixed MTS**
Vocal Tact	75.15 (17.08)	67.72 (22.88)
Typing Tact	75.55 (22.55)	74.21 (21.94)
Vocal Intraverbal N-F	70 (20)	67.37 (22.81)
Typing Intraverbal N-F	73.89 (21.18)	74.21 (21.16)
Vocal Intraverbal F-N	91.11 (16.05)	93.16 (11.08)
Typing Intraverbal F-N	90.55 (18.41)	94.76 (7.72)
MTS	95.71 (9.92)	97.67 (6.67)

Overall, numerically, it looks like more participants in the Mixed MTS group passed the test for most of the other measures (achieving at least 90% correct responses during test) than Compound MTS group. However, there were no differences between groups (*ps* ≥ 0.07).

Table 17.

Percentage (%) of Participants that Passed each Trial Type Test during Post-Test

	Compound MTS	**Mixed MTS**
Vocal Tact	33.33	9.52
Typing Tact	40	38.09
Vocal Intraverbal N-F	22.22	28.57
Typing Intraverbal N-F	40	40
Vocal Intraverbal F-N	72.22	75
Typing Intraverbal F-N	65	84.21
MTS	90	94.74

Post-Questionnaire

There were no differences between groups on measures of interest, engagement, or level of stress during the task (*ps* ≥ 0.11).

Follow-up results

Thirty participants (15 participants from Compound MTS group and 15 from Mixed MTS group) received a Follow-up test two weeks after the experimental session. There were no statistically significant differences between groups for any of the measures assessed (*ps* ≥ 0.26). See table below.

Table 18.

Mean (SD) Percent Correct during Follow-up Test

	Compound MTS	**Mixed MTS**
Vocal Tact	20.67 (21.87)	20.71 (14.92)
Typing Tact	22.67 (24.04)	20 (21.38)
Vocal Intraverbal N-F	26 (24.14)	23.57 (20.23)
Typing Intraverbal N-F	23.33 (24.69)	23.33 (21.27)
Vocal Intraverbal F-N	62 (32.34)	72.86 (15.41)
Typing Intraverbal F-N	66.67 (30.63)	69.33 (24.63)
MTS	84.28 (19.97)	84.81 (20.45)

None or few participants passed the first four measures assessed during Post-test. Numerically, more participants in the Compound MTS group passed Vocal and Typing Intraverbal N-F compared to Mixed MTS group. Eight participants in the Compound MTS group compared to nine participants in the Mixed MTS group passed the MTS test trials. However, there was no statistical difference between groups (*ps* ≥ 0.31).

Table 19.

Percentage (%) of Participants that Passed each Trial Type Test during Follow-up Test

	Compound MTS	**Mixed MTS**
Vocal Tact	0	0
Typing Tact	6.67	0
Vocal Intraverbal N-F	0	0
Typing Intraverbal N-F	6.67	0
Vocal Intraverbal F-N	40	21.42
Typing Intraverbal F-N	40	33.33
MTS	53.33	60

Discussion

Consistent with previous studies comparing EBI and CI conditions (e.g., Fienup & Critchfield, 2011; Oliveira et al., 2021; Petursdottir & Oliveira, 2020; Zinn et al., 2015), the Compound MTS condition was more efficient than Mixed MTS condition. The Compound MTS group required less training time, less training trials, and achieved similar performance to the Mixed MTS condition during Post-test and Follow-up measures. Overall, participants in the first group were exposed to only three stimuli at each trial, while participants in the Mixed condition received training on different stimulus relations in each trial. Directly learning a small number of relations and acquiring a larger number of new relations "for free" (Critchfield, 2018) is an important benefit for CALL applications, as users might benefit from a more efficient instruction.

General discussion

In the present study, the first set of experiments compared different trial designs across three different groups (SP, MTS, and Mixed condition), while the last experiment evaluated two different training designs, comparing a Compound MTS condition versus a Mixed MTS condition. The prediction for Experiments 1a and 1b was that the Mixed group would have an advantage over the two other conditions, as participants assigned to this condition received both response contingencies and reliable presentation of each picture along with its associated foreign auditory-visual compound. This prediction was not confirmed. By contrast, Experiment 2 confirmed the prediction that the compound MTS condition would require fewer trials to complete to mastery, while resulting in similar post- and follow-up test performance.

Previous studies that compared MTS and SP have not provided conclusive evidence of an advantage of one condition over the other (e.g., Clayton & Hayes, 2004; Kinloch, et al., 2013; Leader & Barnes-Holmes, 2001). In the first two experiments, there was also not any reliable difference between MTS and SP conditions. In addition, the Mixed condition did not reliably outperform the MTS or the SP conditions. It is possible that the experiments may have been underpowered to detect an effect; however, numerical differences between group were not consistent across experiments. Another possibility is that trial design (i.e., SP vs. MTS) is simply not an important determinant of acquisition in adults' foreign-language vocabulary learning. To date, only one study has found a consistent difference between the outcomes of SP and MTS training, and this difference was in favor of SP (Leader & Barnes-Holmes, 2021). The failure to replicate this finding in later studies that used different kinds of stimuli, and also in the present study, could suggest that the effect is limited to the use of particular stimuli. Specifically, all stimuli in Leader and Barnes-Holmes were textual pseudoword stimuli. It is possible that learning relations between pairs of these stimuli required verbal rehearsal, which in turn may have been promoted in the SP condition due to the consistent presentation of positive pairs.,

In Experiment 2, it was expected that the Compound MTS condition would require less training trials than the Mixed condition, but both groups would perform similarly during Post-test trials. In line with the initial prediction, the Compound MTS condition was more efficient than the Mixed MTS condition, as participants in the first group required less training trials and instructional time, and yet performed similarly to the second group in all Post-test measures. These results are consistent with prior comparisons of EBI versus CI (Fienup & Critchfield, 2011; Oliveira et al., 2021; Petursdottir & Oliveira, 2020; Zinn et al., 2015), and with studies that compared the potential advantage of using compounds compared to single stimuli (Colasurdo et al., 2023).

The relative efficiency of EBI and CI has previously been found to depend on training structure (Oliveira et al., 2021). Training structure (Saunders & Green, 1999) refers to the way stimuli are linked and the specific order in which conditional discriminations are established during training. Three typically used training structures are one-to-many (OTM), many-to-one (MTO) and linear series (LS). In OTM (e.g., AB, AC), one set of stimuli act as the sample (A) with different sets of comparison stimuli (B and C) across trials. In MTO (.e.g, BA, CA), the samples are different (B and C), but the same set of stimuli act as comparisons (A). In LS (e.g., AB, BC), one set of stimuli (B) serves as comparison stimuli in some trials (AB) and as samples in other trials (AB). Previous research suggests that OTM and MTO procedures increase the probability of class formation and may require less training trials to criterion compared to LS training structure (Arntzen, et al., 2010; Zaring-Hinkle et al., 2016).

Training structure is not directly applicable to the current study, as only one relation was taught in the Compound condition that was presumed to overlap with existing English-language relations. However, the Compound condition could be thought of as being similar to both MTO (participants could likely already select pictures given English-language auditory-textual compounds) and LS training structure (participants could likely already say or select English-language words given the pictures). Taken together, the present findings suggest that there is an efficiency benefit to include just one type of relation for each target word, regardless of whether that relation is auditory-textual compound to picture as in the current study, or picture to spoken word as in Matter et al. (2020).

Some limitations should be noted. First, feedbacks employed throughout the current experiments consisted of the written words “Correct” and “Incorrect”. Previous literature discusses the potential positive impacts of representing a trial following an error. Staropoli, et al. (2022), for example, compared a corrective feedback condition, in which a trial was repeated after an incorrect response, against a condition in which incorrect trial were not repeated. Results revealed that equivalence classes were learned with similar accuracy in both conditions, but the corrective feedback condition resulted in higher maintenance of acquired stimulus relations. The current study did not represent a trial followed by an error, and maybe this would have produced different outcomes, as an enhanced performance for MTS and Mixed groups during Post-tests or Follow-up tests.

A second limitation is that (a) no Follow-up tests were conducted in Experiment 1b, and (b) Follow-up tests conducted for Experiments 1a and 2 yielded low levels of correct responding. For some participants, the tests occurred more than two weeks later (e.g., 17 days later). For example, in Experiment 1a, the mean percentage of correct responses on Vocal Tact trials during the Follow-up test was 26.67% in SP group, 18.75% in MTS group, and 5% in Mixed group. In Experiment 2, the performance on Vocal Tact trials was only 20.67% for Compound MTS group and 20.71% for Mixed MTS group. Potential modifications that could have improved maintenance performance include conducting multiple training session instead of just one, and scheduling Follow-up tests one week after the experiment.

Future research could investigate how different types of stimuli (e.g., textual/auditory vs. images/shapes/patterns, and novel vs. previously familiar/meaningful) affect relative effects of SP vs. MTS trial design. Future studies could also examine if it is crucial for efficiency to have the taught relation include a picture/object, or if the same effect would be seen when teaching, for example, NF intraverbals.

References

Arntzen, E., Grondahl, T. & Eilifsen, C. (2010). The effects of different training structures in the establishment of conditional discriminations and subsequent performance on tests for stimulus equivalence. *Psychology Record*, *60*, 437-461. https://doi.org/10.1007/BF03395720

Bao, S., Sweatt, K. T., Lechago, S. A., & Antal, S. (2017). The effects of receptive and expressive instructional sequences on varied conditional discriminations. *Journal of applied behavior analysis*, *50*(4), 775-788. doi: https://doi.org/10.1002/jaba.404

Barac, R., & Bialystok, E. (2012). Bilingual effects on cognitive and linguistic development: Role of language, cultural background, and education. *Child development*, *83*(2), 413-422. doi: https://doi.org/10.1111/j.1467-8624.2011.01707.x

Belpoliti, F., & Pérez, M. E. (2019). Service learning in Spanish for the health professions: Heritage language learners' competence in action. *Foreign Language Annals*, *52*(3), 529-550. https://doi.org/10.1111/flan.12413

Boelens, H., Hofman, B., Tamaddoni, T., & Eenink, K. (2007). Specific effect of modeling on young children's word production. *The Psychological Record*, *57*(1), 145–166. https://doi.org/10.1007/BF03395569

Bolanos, J. E., Reeve, K. F., Reeve, S. A., Sidener, T. M., Jennings, A. M., & Ostrosky, B. D. (2020). Using stimulus equivalence-based instruction to teach young children to sort recycling, trash, and compost items. *Behavior and Social Issues*, *29*, 78–99. https://doi.org/10.1007/s42822-020-00028-w

Brodsky, J., & Fienup, D. M. (2018). Sidman goes to college: A meta-analysis of equivalence-based instruction in higher education. *Perspectives on Behavior Science, 41*, 95-119. https://doi.org/10.1007/s40614-018-0150-0

Byrne, B. L., Rehfeldt, R. A., & Aguirre, A. A. (2014). Evaluating the effectiveness of the stimulus pairing observation procedure on tact and listener responses in children with autism. *The Analysis of Verbal Behavior, 15*(2), 160–169. https://doi.org/10.1007/s40616-014-0020-0

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), 79-99. doi: https://doi.org/10.1007/s40616-018-0106-1

Clayton, M. C., & Hayes, L. J. (2004). A comparison of match-to-sample and respondent-type training of equivalence classes. *The Psychological Record*, *54*(4), 579-602. doi: https://doi.org/10.1007/BF03395493

Colasurdo, C. R., Reeve, K. F., Jennings, A. M., Vladescu, J. C., Albright, L. K., & Reeve, S. A. (2023). Comparing coumpond pairs and single stimuli during match-to-sample to establish arbitrary stimulus classes with adults of typical development. *European Journal of Behavior Analysis*. https://doi.org/10.1080/15021149.2022.2164828

Connell, P. J., & McReynolds, L. V. (1981). An experimental analysis of children's generalization during lexical learning: Comprehension or production. *Applied Psycholinguistics*, *2*(4), 309-332. https://doi.org/10.1017/S0142716400009760

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

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*. https://doi.org/10.1002/jaba.885

Cortez, M. D., Dos Santos, L., Quintal, A. E., Silveira, M. V., & de 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

Critchfield, T. S. (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

Cruvinel, A. C., & Hübner, M. M. C. (2013). Analysis of the acquisition of verbal operants in a child from 17 months to 2 years of age. *The Psychological Record*, *63*(4), 735-750. https://doi.org/10.11133/j.tpr.2013.63.4.003

Dugdale, N. A., & Lowe, C. F. (1990). Naming and stimulus equivalence. In D. E. Blackman & H. Lejeune (Eds.), *Behaviour analysis In Theory and Practice: Contributions and Controversies*. 115–138. Hove, England: Erlbaum.

Fienup, D. M., & Critchfield, T. S. (2011). Transportability of equivalence-based programmed instruction: Efficacy and efficiency in a college classroom. *Journal of Applied Behavior Analysis, 44*(3), 435-450. https://doi. org/10.1901/jaba.2011. 44-435

Fox, R., Corretjer, O., Webb, K., & Tian, J. (2019). Benefits of foreign language learning and bilingualism: An analysis of published empirical research 2005–2011. *Foreign Language Annals*, *52*(3), 470-490. https://doi.org/10.1111/flan.12418

Garcia, I. (2013). Learning a language for free while translating the web. Does duolingo work?. *International Journal of English Linguistics*, *3*(1), 19. https://doi.org/10.5539/ijel.v3n1p19

Gass, S. M., & Selinker, L. (2001). Second Language Acquisition, an Introductory Course (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Griffith, K. R., Ramos, A. L., Hill, K. E., & Miguel, C. F. (2018). Using equivalence-based instruction to teach piano skills to college students. *Journal of Applied Behavior Analysis*, *51*(2), 207-219. https://doi.org/10.1002/jaba.438

Hill, K. E., Griffith, K. R., & Miguel, C. F. (2016). Using equivalence-based instruction to teach piano skills to children. *Journal of Applied Behavior Analysis,* 1–21. https://doi.org/10.1002/jaba.547

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

Haggas, A. M., & Hantula, D. A. (2002). Think or click? Student preference for overt vs. covert responding in web-based instruction. *Computers in Human Behavior*, *18*(2), 165-172. https://doi.org/10.1016/S0747-5632(01)00041-3

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). Relational frame theory: A post-Skinnerian account of human language and cognition. New York: Plenum.

Hermanto, N., Moreno, S., & Bialystok, E. (2012). Linguistic and metalinguistic outcomes of intense immersion education: How bilingual?. *International journal of bilingual education and bilingualism*, *15*(2), 131-145. https://doi.org/10.1080/13670050.2011.652591

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2), 689-696.

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

Horne, P. J., Hughes, J. C., & Lowe, C. F. (2006). Naming and categorization in young children: IV: Listener behavior training and transfer of function. *Journal of the Experimental Analysis of Behavior*, *85*(2), 247-273. https://doi.org/10.1901/jeab.2006.125-04

Horne, P. J., Lowe, C. F., & Randle, V. R. (2004). Naming and categorization in young children: II. Listener behavior training. *Journal of the Experimental Analysis of Behavior*, *81*(3), 267-288. https://doi.org/10.1901/jeab.2004.81-267

Keintz, K. S., Miguel, C. F., Kao, B., & Finn, H. E. (2011). Using conditional discrimination training to produce emergent relations between coins and their values in children with autism. *Journal of Applied Behavior Analysis*, *44*(4), 909-913. https://doi.org/10.1901/jaba.2011.44-909

Kinloch, J. M., McEwan, J. S. A., & Foster, T. M. (2013). Matching-to-sample and stimulus-pairing-observation procedures in stimulus equivalence: The effects of number of trials and stimulus arrangement. *The Psychological Record*, *63*(1), 157-174. https://doi.org/10.11133/j.tpr.2013.63.1.012

Kobari-Wright, V. V., & Miguel, C. F. (2014). The effects of listener training on the emergence of categorization and speaker behavior in children with autism. *Journal of Applied Behavior Analysis*, *47*(2), 431-436. https://doi.org/10.1002/jaba.115

Kritch, K. M., Bostow, D. E., & Dedrick, R. F. (1995). Level of interactivity of videodisc instruction on college students' recall of aids information. *Journal of Applied Behavior Analysis*, *28*(1), 85-86. https://doi.org/10.1901/jaba.1995.28-85

Lakshmanan, U., & Selinker, L. (2001). Analyzing interlanguage: how do we know what learners know?. *Second Language Research*, *17*(4), 393-420. https://doi.org/10.1177%2F026765830101700406

Leader, G., & Barnes-Holmes, D. (2001). Matching-to-sample and respondent-type training as methods for producing equivalence relations: Isolating the critical variable. *The Psychological Record*, *51*(3), 429-444. https://doi.org/10.1007/BF03395407

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test. *Review of Educational Research*, *66*(4), 579–619. https://doi.org/10.3102/00346543066004579

Makumane, M. A., & Ngcobo, S. (2018). The socio-economic value of French language education in Lesotho: The learners' voices. *South African Journal of African Languages*, *38*(2), 167-175. https://doi.org/10.1080/02572117.2018.1463705

Mascarenhas, N. (2020). Duolingo can't teach you how to speak a language, but now it wants to try. TechCrunch+. https://techcrunch.com/2021/05/03/duolingo-ec1-future/

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*, 147-166. https://doi.org/10.1002/jaba.545

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, 711-716. https://doi.org/10.1002/jaba.301.

Miguel, C. F., Petursdottir, A. I., Carr, J. E., & Michael, J. (2008). The role of naming in stimulus categorization by preschool children. *Journal of the Experimental Analysis of Behavior*, *89*(3), 383-405. https://doi.org/10.1901/jeab.2008-89-383

Nation, P. and Waring, R. (1997): Vocabulary size, text coverage, and word lists. In Schmitt, N. and McCarthy, M. Vocabulary: description, acquisition, and pedagogy. Cambridge: Cambridge University Press.

Nor, N. M., & Ab Rashid, R. (2018). A review of theoretical perspectives on language learning and acquisition. *Kasetsart Journal of Social Sciences*, *39*(1), 161-167. https://doi.org/10.1016/j.kjss.2017.12.012

Nottingham, C. L., Vladescu, J. C., Kodak, T., & Kisamore, A. N. (2017). Incorporating multiple secondary targets into learning trials for individuals with autism spectrum disorder. *Journal of Applied Behavior Analysis*, *50*(3), 653-661. https://doi.org/10.1002/jaba.396

O'Grady, A. C., Reeve, S. A., Reeve, K. F., Vladescu, J. C., & Deshais, M. (2021). Comparing computer-based training and lecture formats to teach visual analysis of baseline-treatment graphs. *Behavioral Interventions*, *36*(1), 67-92. https://doi.org/10.1002/bin.1752

Oliveira, J. S., & Petursdottir, A. I. (In preparation). Evaluating class reorganization in equivalence-based instruction.

Oliveira, J. S., Freitas, L., Tomlinson, G. M., & Petursdottir, A. I. (2021). Translational evaluation of training structures in equivalence-based instruction. *Journal of the Experimental Analysis of Behavior*, *115*(1), 393-404. https://doi.org/10.1002/jeab.657

Omori, M., & Yamamoto, J. I. (2013). Stimulus pairing training for Kanji reading skills in students with developmental disabilities. *Research in Developmental Disabilities*, *34*(4), 1109-1118. https://doi.org/10.1016/j.ridd.2012.12.016

Petursdottir, A. I., & Carr, J. E. (2011). A review of recommendations for sequencing receptive and expressive language instruction. *Journal of applied behavior analysis*, *44*(4), 859-876. https://doi.org/10.1901/jaba.2011.44-859

Petursdottir, A. I., & Hafliðadóttir, R. D. (2009). A comparison of four strategies for teaching a small foreign language vocabulary. *Journal of Applied Behavior Analysis*, *42*, 685-690. https://doi.org/10.1901/jaba.2009.42-685.

Petursdottir, A. I., & Oliveira, J. S. (2020). Efficiency of equivalence-based instruction: A laboratory evaluation. *Journal of the Experimental Analysis of Behavior*, *114*(1), 87-105. https://doi.org/10.1002/jeab.617

Petursdottir, A. I., Ólafsdóttir, R., & Aradóttir, B. (2008). The effects of tact and listener training on the emergence of bidirectional intraverbal relations. Journal of Applied Behavior Analysis, 41, 411-415. https://doi.org/10.1901/jaba.2008.41-411

Pytte, C. L., & Fienup, D. M. (2012). Using equivalence-based instruction to increase efficiency in teaching neuroanatomy. *The Journal of Undergraduate Neuroscience Education*, *10*, A125–A131.

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

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

Rocha e Silva, N. I., & Ferster, C. B. (1966). An experiment in teaching a second language. *International Review of Applied Linguistics in Language Teaching*, *4*, 85-113. https://doi.org/10.1515/iral.1966.4.1-4.85

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. *The Analysis of verbal behavior*, *27*(1), 61-74. https://doi.org/10.1007/BF03393092

Sandoz, E. K., & Hebert, E. R. (2017). Using derived relational responding to model statistics learning across participants with varying degrees of statistics anxiety. *European Journal of Behavior Analysis*, *18*, 113–131.

Saunders, R. R., & Green, G. (1999). A discrimination analysis of training-structure effects on stimulus equivalence outcomes. *Journal of the Experimental Analysis of Behavior, 72*(1), 117–137. https://doi.org/10.1901/ jeab.1999.72-117

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. *Proceedings of the 54th annual meeting of the association for computational linguistics 1*, 1848-1858.

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of behavior, 74*(1), 127-146. https://doi.org/10.1901/jeab.2000.74-127

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*, 5–22. https://doi.org/10.1901/jeab. 1982.37-5

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Smith, C., Perry, H., Oliveira, J. S., Cox, R. E., & Petursdottir, A. I. (in preparation). Comparison of dense and lean response contingencies in computer-assisted language learning.

Solares, L., & Fryling, M. J. (2019). Further evaluation of the stimulus pairing observation procedure with children with autism spectrum disorder. *The Analysis of verbal behavior*, *35*(1), 85-93. https://doi.org/10.1007/s40616-018-0101-6

Staropoli, D. M., Colasurdo, C. R., Reeve, K. F., Jennings, A. M., Reeve, S. A., & Deshais, M. A. (2022). Effects of repeating or not repeating trials immediately when errors occur during training of equivalence classes with college students. *European Journal of Behavior Analysis.* https://doi.org/10.1080/15021149.2023.2170954

Takahashi, K., Yamamoto, J. I., & Noro, F. (2011). Stimulus pairing training in children with autism spectrum disorder. *Research in Autism Spectrum Disorders*, *5*(1), 547-553. https://doi.org/10.1016/j.rasd.2010.06.021

Tonneau, F. (2001). Equivalence relations: A critical analysis. *European Journal of Behavior Analysis*, *2*(1), 1-33. https://doi.org/10.1080/15021149.2001.11434165

Vallinger-Brown, M., Rosales, R. (2014). An Investigation of Stimulus Pairing and Listener Training to Establish Emergent Intraverbals in Children with Autism. *The Analysis of Verbal Behavior*, *30*, 148–159. https://doi.org/10.1007/s40616-014-0014-y

Vesselinov, R. (2009) Measuring the effectiveness of Rosetta Stone: Final report. Queens College, City University of New York.

Vladescu, J. C., & Kodak, T. M. (2013). Increasing instructional efficiency by presenting additional stimuli in learning trials for children with autism spectrum disorders. *Journal of Applied Behavior Analysis*, *46*(4), 805-816. https://doi.org/10.1002/jaba.70

Weber, K., Christiansen, M. H., Petersson, K. M., Indefrey, P., & Hagoort, P. (2016). fMRI syntactic and lexical repetition effects reveal the initial stages of learning a new language. *Journal of Neuroscience*, *36*(26), 6872-6880. https://doi.org/10.1523/JNEUROSCI.3180-15.2016

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

Zaring-Hinkle, B., Carp, C. L., & Lepper, T. L. (2016). An evaluation of two stimulus equivalence training sequences on the emergence of novel intraverbals. *The Analysis of Verbal Behavior*, *32*, 171-193. https://doi.org/10.1007/s40616-016-0072-4

Zinn, T. E., Newland, M. C., & Ritchie, K. E. (2015). The efficiency and efficacy of equivalence-based learning: A randomized controlled trial. *Journal of Applied Behavior Analysis, 48*(4), 865-882. https://doi.org/10.1002/ jaba.258

Appendix

Final questionnaire

1) How would you rate your interest in the learning task you just completed?
    - Extremely interested
    - Very interested
    - Slightly interested
    - Not interested
2) How would you rate your engagement in the learning task you just completed?
    - Not engaged
    - Slightly engaged
    - Very engaged
    - Extremely engaged
3) How would you rate your stress level while you were completing the learning task?
    - Not stressed
    - Slightly stressed
    - Very stressed
    - Extremely stressed
4) What strategies did you use to help you remember the words you had to?
5) What is your native language or language(s)? A native language is a language you have spoken since early childhood and you can speak very well.
6) How would you describe your ability to speak languages other than your native language?

- I can string together sentences in one language other than my native language(s)
- I can string together sentences in two or more languages other than my native language(s)
- I cannot string together sentences in any language other than my native language(s)

7) Do you consider yourself a fluent speaker of any language other than your native language(s)?
    - Yes, I am fluent in one or more language(s) other than my native language(s)
    - No, I am not a fluent speaker of any language(s) other than my native language(s)
8) Have you learned a foreign language in a formal education setting (e.g., in a college class or in high school)?
    - No
    - Yes. Number of semesters?
9) What was the foreign language or languages you learned in a formal educational setting?
10) Have you ever tried to learn a foreign language through a computerized self-instruction program like Rosetta Stone, Duolingo, or Babbel?
    - Yes, but I did not learn very much
    - Yes, and I learned a lot
    - No

VITA

Personal Information Juliana Sequeira Cesar de Oliveira

j.oliveira@tcu.edu

Education

Master of Arts, Experimental Psychology, emphasis: Human Learning and Verbal Behavior
Texas Christian University, 2020. Mentor: Dr. Anna Petursdottir

Master of Science, Experimental Behavior Analysis
Federal University of Para, Brazil, 2017. Mentor: Dr. Romariz Barros

Bachelor of Psychology
Federal University of Para, Brazil, 2015

Publications

Oliveira, J. S. C. D., & Petursdottir, A. I. (Under review). Summation in Convergent Multiple Control Over Verbal Behavior.

Petursdottir, A. I., & Oliveira, J. S. C. D. (2022). Teaching foreign language. In J. L. Matson (Ed.), *Applied behavior analysis: A comprehensive handbook*. Springer.

Miller, A. C., Cox, R. E., Swensson, R. M., Oliveira, J. S. C. D., & Petursdottir, A. I. (2021). Effects of blocking echoic responses on tact emergence following stimulus pairing. *European Journal of Behavior Analysis*. Advance online publication, https://doi.org/10.1080/15021149.2021.1896070

Oliveira, J. S. C. D., Freitas, L., Tomlinson, G.M., & Petursdottir, A. I., (2021). Translational evaluation of training structures in equivalence-based instruction. *Journal of the Experimental Analysis of Behavior, 115*(1), 393-404. https://doi: 10.1002/jeab.657.

Petursdottir, A. I., & Oliveira, J. S. C. D. (2020). Efficiency of equivalence-based instruction: A laboratory evaluation. *Journal of the Experimental Analysis of Behavior, 114*, 87-105. https://doi.org/10.1002/jeab.617

Silva, A. J. M., Keuffer, S. I. C., Oliveira, J. S. C. D., & Barros, R. S. (2018). Acquisition of intraverbal repertoire via equivalence-based instruction in children with autism spectrum disorder. *Trends in Psychology*, *26*(3), 1155-1171. https://dx.doi.org/10.9788/tp2018.3-02pt

Pereira, F. S., Lobato, S. N. S., Oliveira, J. S. C., Yamaguchi, C. T., Cordeiro, J. C. B., & Galvão, O. F. (2017). Auditory-visual discrimination in adults with post lingual hearing impairment and cochlear implants. *Temas em Psicologia*, *25*(3), 1373-1384. https://dx.doi.org/10.9788/TP2017.3-20Pt

Experience

Teaching Assistantship, Texas Christian University

2018-2022

ABSTRACT

EVALUATING DIFFERENT TRIAL AND TRAINING DESIGNS IN COMPUTERIZED FOREIGN-LANGUAGE VOCABULARY INSTRUCTION

By Juliana Sequeira Cesar de Oliveira, Ph.D., 2023
Department of Psychology
Texas Christian University

Dissertation Advisor: Dr. Anna Ingeborg Petursdottir, Associate Professor of Psychology

The present study sought to evaluate trial designs and training designs that are commonly used in popular commercially available computer-assisted language-learning (CALL) programs. The first two experiments (Experiment 1a and 1b) compared the effects of passive viewing and active student response methods in vocabulary learning. Contingencies on active student responding were implemented in the form of compound matching to sample (MTS), in which each trial presented a compound sample (i.e., an auditory and a textual FL stimulus), to which the learner was required to respond by selecting a matching picture from an array of choices. This approach is characteristic of vocabulary-learning trials in Rosetta Stone®. Passive viewing was represented by stimulus pairing (SP) trials, in which the auditory FL stimulus, the textual FL stimulus, and the matching picture were presented simultaneously, and no response was required from the learner other than clicking to advance to the next trial. This approach is characteristic of vocabulary-learning trials in Memrise®. With respect to training design, the goal of Experiment 2 was to compare the efficiency with which participants acquired the same emergent stimulus relations as a result of an instruction that teaches a small number of relations (Compound MTS condition) versus an instruction that directly targets a greater number of relations separately (Mixed condition). The prediction for Experiments 1a and 1b was that the Mixed group would have an advantage

over the two other conditions, as participants assigned to this condition received both response contingencies and reliable presentation of each picture along with its associated foreign auditory-visual compound. This prediction was not confirmed. By contrast, Experiment 2 confirmed the prediction that the compound MTS condition would require fewer trials to complete to mastery, while resulting in similar post- and follow-up test performance.