Peer-reviewed article

Repeat after me? Mapping The Relation of Overimitation to Foreign-Language Learning Aptitude

Dagmar Divjak

Department of Modern Languages, Department of Linguistics & Communication, University of Birmingham

d.divjak@bham.ac.uk

Justyna Mackiewicz

Department of Modern Languages, University of Birmingham

Petar Milin

Department of Modern Languages, University of Birmingham

Abstract

We investigated whether overimitation constitutes a previously unrecognised component of second language (L2) aptitude by examining its relationship with L2 proficiency, established aptitude measures, and domain-general cognitive abilities. Advanced L2 learners completed a novel language-based adaptation of the classic puzzle-box overimitation paradigm alongside measures of grammatical and collocational knowledge, implicit sequence learning, working memory, attentional control, and language aptitude. Graphical modelling revealed three distinct clusters. Linguistic proficiency was associated with grammatical sensitivity, collocational knowledge, and implicit sequence learning, whereas production-based language imitation grouped with explicit aptitude and executive resources. In contrast, procedural overimitation and self-reported mimicry formed an independent dimension, showing little association with either linguistic proficiency or established aptitude measures. These findings suggest that overimitation reflects a socio-motivational disposition that is functionally distinct from the mechanisms supporting L2 learning and linguistic knowledge.

Slovo 66:1 (2026), 177–194. DOI https://doi.org/10.33063/slovo.v66i1.1261
CC BY 4.0

1. Introduction

It has been claimed that the *homo sapiens* would have been more aptly named *homo imitans* as imitation lies at the heart of being human (Tennie et al., 2009). Humans imitate each other in unique ways: compared to other animals, humans are more willing to copy actions, including irrelevant ones, and this willingness increases with age (McGuigan et al., 2010). This propensity to imitate, and to overimitate more specifically, arguably forms the foundation of our culture and social life which includes many norms and rituals (Tomasello, 1998). Language, too, is a cultural artifact and one of the key artifacts that human beings transmit from one generation to the next. Despite the crucial role imitation plays in cultural transmission, the role of imitation in language learning remains virtually uncharted territory. After a brief introduction to the idea of imitation (Section 1.1), we outline our study (Section 1.2), in which we take a first step toward investigating a possible role of overimitation in second language (L2) learning and establish its position with respect to established measures of L2 learning aptitude.

1.1 (Over)imitation and language learning

Ever since the seminal study by Horner and Whiten (2005) researchers have been fascinated with the human tendency to imitate irrelevant actions: different from their close relative the chimpanzee, 3-4 year old children imitate actions that are visibly unnecessary in relation to the goal of an action sequence, e.g., to retrieve a reward. This is why Blackmore (2007, p. 1) opens her argument on "What makes us human?" succinctly with "To be human is to imitate.". And we *imitate at large*: person-to-person information copying has evolved as an independent social tool, going beyond (or above) the basic evolutionary system of copying information encoded in genes. The tendency to imitate causally unnecessary actions with great precision has been replicated in a number of studies (Hoehl et al., 2019) and has been termed "overimitation" (Lyons et al., 2011). Overimitation increases with age, with older children—and adults even more so—showing a strong tendency to faithfully reproduce irrelevant actions, especially when the action to be imitated is performed by an adult model (McGuigan et al., 2010). The tendency to overimitate also increases when the function of the imitated behaviour is opaque (Hoehl et al., 2019). It has been suggested that human children overimitate because they grow up in a jungle of artifacts and tools with complex underpinnings that are often inaccessible to direct inspection (Lyons et al., 2011). For this reason, automatically encoding *all* of an adult's actions as reliable indicators of the object's inner workings or causal structure may, in fact, be the most rational thing to do. Alternatively, and different from causal accounts, social accounts of overimitation stipulate that, given the high price that is paid for not fitting in, adopting a "copy-when-uncertain" social learning strategy is advantageous (for an overview of different accounts, see Farmer et al., 2018).

Despite the crucial role imitation plays in cultural transmission, the role of imitation in language learning remains virtually uncharted territory. Yet, an overimitation strategy could be advantageous too when learning the group's communication code.[1] The very few studies that have looked at the role imitation plays in first language development (Bannard et al., 2013; Klinger et al., 2016) do confirm that 3- to 10-year-old human children imitate faithfully, especially when the communicative function of a structure remains opaque (Klinger et al., 2016). That is, when children hear a linguistic form but do not yet understand what it accomplishes in communication, they are more likely to reproduce it with high fidelity. In such cases, imitation serves as a bridge: it allows the child to keep the unfamiliar form in circulation in their own speech and "fit in", which in turn provides further opportunities to observe how others use the form. Over time, these repeated exposures and reproductions can lead to a gradual grasp of the function and meaning of the structure.

Because a second language is often learned by older children and adults, who are even more strongly inclined to faithfully reproduce irrelevant aspects, imitation seems a valuable cognitive propensity to consider in L2 learning, too. After all, in many cases, the L2 is characterised by a large number of markers that are unfamiliar from the L1 point of view, not only in form but also in function. Interestingly, imitation has not been studied; yet, there is ample literature on the related concept of repetition and on repetition in the language learning classroom. The views of the 17th century Czech scholar Comenius who encouraged his students to "repeat after" him resurfaced in versions of the audio-lingual and oral-situational approaches that held sway in the early 20th century (Richards & Rodgers, 2014: 44-57) as they meshed well with the associative-behaviourist view that learning occurs as the result of repeated exposure to a stimulus. Heavily criticised during the cognitive revolution, behaviourist approaches were replaced with cognitive approaches in the second half of the 20th century. With behaviourism, repetition disappeared from the language learning stage, yielding the scene to measures of cognitive aptitude (Carroll & Sapon, 1959), aimed at identifying successful L2 classroom learners at the earliest stages of instruction (Robinson, 2005).

1.2 This study

In this study, we take a first step toward investigating a possible role of overimitation in second language learning and determining its position with respect to established measure of L2 learning aptitude.

To pursue this aim, we introduce a protocol based around the classic puzzle box imitation paradigm, adapting it to the domain of language. Specifically, we developed a set of tasks that translate the logic of the puzzle box into linguistically relevant activities:

[1] A voluminous literature on *vicarious learning*, of which imitation is essential part, has existed in cognitive psychology since Bandura's (1986) famous study. For a recent overview see Mayes (2015).

just as the puzzle box requires participants to reproduce actions whose necessity is not always transparent, our linguistic tasks require learners to reproduce linguistic forms whose communicative function is opaque. These tasks involve spoken sequence repetition and character drawing in Chinese, allowing us to assess the degree to which L2 learners engage in overimitation in language-specific contexts. Next, we tested learners' linguistic knowledge (of English) through targeted assessments of collocations and grammatical structures. We also administered a battery of established measures of L2 cognition that we want to position the overimitation tasks against. These include classical tasks such as Operation Span (OSpan) for working memory, the Serial Reaction Time (SRT) task for implicit sequence learning, the Flanker task for attentional control, and the Llama F task for explicit learning aptitude. The tasks are explained in more detail in Section 2.2.

Together, these measures provide a rich profile of both domain-general cognitive capacities and domain-specific linguistic competence, against which the role of overimitation in second language learning can be evaluated. We hypothesised that individuals with a greater propensity for overimitation would perform better on the assessments of collocations and grammatical structures, reflecting a positive association between overimitation and language learning. At the same time, for overimitation to qualify as a distinct aptitude component rather than a proxy for existing constructs, we expected it to show only limited overlap with established measures of language-learning aptitude and related cognitive abilities.

2. Methods[2]

2.1 Participants

Forty-two native speakers of Arabic, resident in the UK, took part in the study (29 females, M = 29.4, SD = 6.72; 13 males, M = 31.1, SD = 5.38). Participants came from Saudi Arabia (n = 20), Egypt (n = 12), Syria (n = 4), Jordan (n = 3), and one each from Algeria and Morocco. All were L2 learners of English who had arrived in the UK in adulthood, with no prior exposure to English during childhood and no knowledge of Chinese. Of the 42 participants, 36 (85.7%) reported learning English through formal education, while 6 (14.3%) were self-taught. Educational backgrounds varied: 7 had completed high school, 13 held an undergraduate or bachelor's degree, 15 a master's degree, and 7 a PhD or equivalent medical qualification. Participants were recruited via university campus leaflets and social media. Most were university students or members

[2] Data, including the full Materials file, are downloadable from the University of Birmingham Institutional Research Archive (UBIRA): https://edata.bham.ac.uk (DOI: https://doi.org/10.25500/edata.bham.00001739). Analysis code for the models reported in this paper is available from https://github.com/ooominds/Overimitation.

of the local community. The study was approved by the University of Birmingham Ethics Committee for Humanities and Social Sciences.

2.2 Task battery

The participants completed the task battery across two testing sessions. It consisted of imitation, language and cognitive tasks and questionnaires, which are described in turn below.

2.2.1 Imitation tasks

The imitation tasks tested general and linguistic imitation abilities. General imitation ability was assessed with the traditional 8-sequence puzzle box (McGuigan et al., 2010) while visual and auditory imitation abilities in the language domain were tested with a novel sentence repeating and character writing task. It took on average 15 minutes to complete all 3 imitation tasks. For all three tasks the same two types of scores, Fidelity and Overimitation, were calculated following McGuigan et al. (2010). Fidelity captures whether the demonstrated *order* of steps was retained (the steps needed to be performed but could be performed in different order); overimitation measures whether *unnecessary* steps are retained or skipped. The scoring protocols are available in Appendix.

Puzzle box. The general imitation task tested procedural imitation abilities where participants were presented with video displaying a sequence of 8 activities performed on a small box leading to obtaining a reward that was hidden inside.[3] Two versions of the video were created, with either a transparent or an opaque box (see Figures 1 and 2). The video was played three consecutive times; at the end of the video a voice instructed the participants that it was now their turn "to give it a go"; participants were handed a transparent or opaque box without further instructions. Participants were video recorded while performing this task.

[3] We would like to thank Dr Hui Sun for designing and testing the puzzle box while a postdoctoral research associate on the Leverhulme Trust Research Leadership Award (RL-2016-001 to D. Divjak) that funded this research.



Figure 1. Opaque puzzle box



Figure 2. Transparent puzzle box

Spoken sentence repetition. Spoken Chinese is characterised not only by sounds but also by tones (rising, falling, flat). To the untrained ear, this poses a double problem: identifying and memorising the sounds as well as the tones. For consistency with the visual imitation task, we use 8-character sentences. The auditory stimuli were comprised of two Chinese sentences recorded by a native Chinese speaker: *I sincerely wish you a happy birthday* and *Welcome to this university*. They were of similar length and complexity. In the task, participants heard a Chinese sentence via headphones, and were asked to repeat it; their voice was recorded. As with the puzzle box task, the stimuli were played three consecutive times before participants were asked to repeat them. The productions were scored by a native speaker of Chinese for fidelity and, because it is impossible to judge whether components are unnecessary in spoken production when meaning is not available, productions were instead scored for pronunciation accuracy (average of all correctly repeated syllables), tone accuracy (average of all syllables produced with correct tone), and chunking (number of correctly segmented chunks).

Visual character drawing. Chinese characters are written using strokes that are executed in a particular sequence; however, to the uninitiated, characters could be formed in many different ways. Two characters of similar length and complexity were selected for this task: *honesty* and *depend on* (see Figures 3 and 4). Both characters consist of 8 strokes, some of which are simple (they consist of one line) and some of which are complex (they consist of 2 lines forming, e.g., an angle) and are executed in a specific order. A video of a Chinese character being written was presented three consecutive times and participants were asked to produce the same character. Participants were video recorded while completing the task and the recordings as well as the physical copies produced in these sessions were scored on a stroke-by-stroke basis for overimitation (flick of the hand and speed of writing), and fidelity (presence, sequence, accuracy and direction of a stroke). Further measures captured structure (spatial position of the character on the page), corrections (number of corrections made), and time to draw the character.



Figure 3. Chinese character: **honesty**



Figure 4 Chinese character: **depend on**

2.2.2 Language tasks

We ran two tasks that measure vocabulary size and grammar knowledge in English.

Collocations. The collocation task consisted of a 40-item *Words that go together* (WGT) questionnaire that tests receptive knowledge of collocations (Dąbrowska, 2014). Each item comprised a set of five phrases of varying frequency and collocation strength and the task was to choose one phrase that sounded the most natural or familiar. Half of the items consisted of a verb followed by an abstract noun (e.g., *raise standards*), and the other half of an adjective followed by an abstract noun (e.g., *bitter dispute*). Participants were encouraged to guess if they were not sure. The task was administered in Qualtrics and took 10 minutes to complete on average.

Grammaticality judgement task (GJT). The grammaticality judgement task consisted of 80 sentences half of which were grammatical and half ungrammatical. 60 sentences, half of which were grammatical, were selected from DeKeyser (2000): they contained Past tense, Plural, Third-Person Singular, Present Progressive, Determiners, Pronominalization, Particle Movement, Subcategorization, Yes-No Questions, Wh-Questions, Word Order. A further 20 sentences were created for the purpose of this study and contained examples assessing the knowledge of determiners and tenses (Present perfect, Future, Past Perfect and Present Progressive); see Appendix for the complete list of stimuli. The sentences appeared in a random order on a computer screen, and the participants were asked to press either M or Z on the keyboard to indicate sentence grammaticality. The experimental stimuli were preceded by 8 practice items. The task was created and administered in the PsychoPy software (Peirce et al., 2019) and took about 10 minutes to complete.

2.2.3. Language learning aptitude tasks

Language-learning aptitude was assessed using complementary measures of explicit language-learning aptitude (LLAMA F) and implicit sequence learning (Serial Reaction Time; SRT), capturing two mechanisms widely implicated in successful language acquisition.

Llama F task. Explicit language learning ability was tested with the online Llama F task which is available at www.lognostics.co.uk (Meara & Rogers, 2019).[4] In this task

[4] The online software was updated towards the end of the data collection which resulted in 6 participants taking the new version of Llama F. The new version has a timer on the screen and a revised scoring method. We

participants are presented with pictures and corresponding labels. They are given 4 minutes in the learning phase to figure out how the pictures relate to the labels. In the testing phase a new set of pictures is presented, and participants must select labels that describe the pictures. The Llama F score is represented on a scale from 0-20. The task took on average 10 minutes to complete.

Serial Reaction Time (SRT) task. The SRT task is a multi-choice, disjunctive task designed to evaluate improvements in procedural memory span for structured, statistically predictable sequences (Vakil et al., 2017). We conducted this task using the OpenSesame software (Mathôt et al., 2012). In the task, a dot appears on the screen in one of four positions (up, down, right, or left), and participants are required to press the corresponding button on a response pad as quickly as possible. We utilized second-order conditional sequences (SOC; Gabriel et al., 2013; Vakil et al., 2017; Wilkinson & Shanks, 2004), where the prediction of a target location depends on the two preceding locations. After completing the task, participants filled out a questionnaire designed to assess their awareness of the pattern. The entire task took approximately 5 minutes to complete.

2.2.4 Memory and Attention tasks

In addition, we ran several tasks that are standardly used in L2 research and tested working memory and attention.

The Operation Span Task (OSPAN). OSPAN is a complex task designed to assess working memory capacity. The version used in this study incorporates an arithmetic component and was administered using the PsychoPy software (Peirce et al., 2019). During the task, participants first see a letter on the screen, followed by an arithmetic equation (e.g., 3 x 7 = 21). They must then determine whether the equation is correct by pressing the appropriate key. At the end of each trial, participants are required to recall and enter the letter that was shown at the beginning. This concludes the trial, and a new letter appears to start the next trial. The task progresses by increasing the number of letters in each trial, with the longest sequence containing up to 7 letters. The participant's score reflects the number of letters correctly recalled in their original order. The entire task typically took around 15 minutes to complete.

Flanker task. To assess executive control mechanisms, we administered the Flanker task, which measures the ability to suppress interference from distracting stimuli while focusing attention on a task-relevant target. Participants saw a row of 5 arrows on the screen and had to press a key corresponding to the direction of the arrow at the centre. We used congruent conditions in which all arrows pointed the same direction (no interference from any distractors), and incongruent conditions where the target stimulus

successfully covered the timer on the screen to minimise the distraction; however, the new scores were calculated in a different way that cannot be translated into the old scores.

is flanked by distractors (arrows pointing in another direction) that may cause response interference. The main task was preceded by some practice trials. It took on average 3 minutes to complete the task.

2.2.5 Demographic and Mimicry Questionnaires

In addition to a **language background questionnaire** which elicited information on participants' demographic and educational profiles, language learning history, and cultural orientation we administered a **mimicry questionnaire** which assessed the extent to which individuals engage in mimicry—the behaviour of imitating or copying others' actions, expressions, or mannerisms. This instrument was adapted from the Iowa Mimicry Questionnaire (IMQ; Ash et al., 2024) to focus specifically on language-related imitation. A brief six-item self-report questionnaire assessed participants' inclination to mimic linguistic input. Items targeted behaviours such as talking to oneself in the target language, memorising songs or poems, mimicking native-like accents, and reproducing phrases encountered in speech or text. Participants also indicated whether they mimic random environmental sounds and, if so, whether such mimicry occurs spontaneously or is consciously controlled. Each item was rated on a Likert-type scale reflecting frequency or agreement.

2.3 Testing sessions: order of tasks

Participants completed two testing sessions scheduled one week apart, during which they completed our battery of imitation, language, and cognitive tasks as well as questionnaires. At the beginning of each session, participants completed one version of the imitation tasks (Chinese sentence repetition and character writing) and the puzzle box task. Each participant was assigned to only one puzzle-box condition throughout the study (transparent or opaque),

This was followed by the Serial Reaction Time task, which was also identical across both sessions, for the same reasons. In the first session, participants additionally completed the Grammaticality Judgement Task, a collocation task, and a language background questionnaire. In the second session, they completed measures of memory, attention, LLAMA, and the Mimicry Questionnaire. Each testing session lasted approximately one hour. Participants received a £25 Amazon voucher upon completion of both sessions.

3. Results

3.1 Consistency of performance

Because some of our tasks are novel, we started out by assessing the reliability of our instruments. Performance across the box imitation task showed limited variance, with participants scoring near ceiling. Note that the imitation-box task is not intended to differentiate participants by difficulty, but rather to ensure universal recall of the steps so that any skipped or re-ordered actions reflect choice rather than memory limitations. We used paired-samples t-tests to compare performance across the two sessions. In addition to statistical significance testing, Cohen's d was calculated as a measure of effect size. Following conventional benchmarks, values of approximately 0.20, 0.50, and 0.80 were interpreted as small, medium, and large effects, respectively (Cohen, 1988). The results showed consistent performance across both versions of the box task (for details see Table 1).

Measure	***t-test* (df)**	***p-value*[5]**	**Cohen's *d***
Over-imitation	0.40 (41)	.688	-0.06
Fidelity	1.42 (41)	.164	-0.22
Fidelity-specific	1.30 (41)	.200	-0.20

Table 1. Performance differences on the box imitation task across sessions

To examine the effect of box type, participants were divided into two independent groups according to the condition they completed (opaque vs. transparent). Group differences were analysed using independent-samples t-tests followed by Cohen's effect sizes. Box type (transparent vs opaque) had no effect on the results for over-imitation in Session 1 or Session 2, see Table (2). The only significant difference was found in the time taken to perform the task with the second session taking less time, as expected (t(42) = 3.23, p = .002, Cohen's d = -0.49).

Measure	**Session**	***t-test* (df)[6]**	***p-value***	**Cohen's *d***
Over-imitation	1	-0.45 (39.07)	.652	-0.14
Over-imitation	2	0.00 (38.64)	1.00	0.00
Fidelity	1	0.28 (36.34)	.784	0.09
Fidelity	2	0.63 (36.97)	.535	0.19
Fidelity-specific	1	0.63 (34.69)	.536	0.19
Fidelity-specific	2	0.00 (39.80)	1.00	0.00

[5] The p-values reported in all tables are uncorrected for multiple comparisons.
[6] Degrees of freedom are non-integer because Welch's t-tests for independent samples were used, which do not assume equal variances between groups.

Duration	1	-0.04 (39.50)	.971	-0.01
Duration	2	0.43 (33.54)	.670	0.13

Table 2. Performance differences on the box imitation task by condition

Performance in the language imitation tasks was stable across stimuli, with no major differences between the two types of either spoken or written items, indicating that both the materials and scoring procedure were reliable. No significant differences were found in the speaking task between the sentences for pronunciation, tone or chunking; for details see Table (3).

Measure	***t-test* (df)**	***p-value***	**Cohen's *d***
Pronunciation	1.47 (39)	.151	-0.23
Tone	0.60 (39)	.551	-0.10
Chunking	1.97 (38)	.056	-0.32

Table 3. Performance differences across speaking stimuli

In the character writing task, no differences emerged between the two characters for overimitation, correction or fidelity but structure did yield significant results; for details see Table (4).

Measure	***t-test* (df)**	***p-value***	**Cohen's *d***
Over-imitation	0.70 (41)	.487	-0.11
Time	1.35 (41)	.184	-0.21
Correction	1.03 (41)	.309	-0.16
Fidelity	2.05 (41)	.047	-0.32
Structure	3.11 (41)	.003 (0.015 after correction)	-0.48

Table 4. Performance differences across writing stimuli

Among the cognitive tasks, the SRT task was performed twice because concerns have been raised about its reliability as an individual differences measure (Oliveira et al., 2023; Oliveira et al., 2024). Using paired-samples t-tests, we found no significant differences between the participant-specific Intercepts (t-test values: $t(40) = -0.39$, $p = .696$, $d = 0.06$) or Slopes (t-test values: $t(40) = 0.53$, $p = .601$, $d = -0.08$) estimated from the SRT task across the two testing sessions. This demonstrated that our (short) version of SRT yields a reliable measure of implicit learning ability, revealing a stable individual trait.

For completeness we also report accuracy in the processing component of the OSPAN task which was high with 95% of participants achieving at least the cut-off score of 60, indicating good task compliance. In the Flanker task, participants' mean

reaction times were 544.57 ms (SD=83.9) for congruent trials and 619.3 ms (SD=73.85) for incongruent trials, showing the expected effect, with slower responses to incongruent stimuli.

3.2 Data preparation

We applied Principal Component Analysis (PCA) to responses on the newly developed Language Mimicry Questionnaire for two reasons. First, because the questionnaire was adapted from the Iowa Mimicry Questionnaire (Ash et al., 2024) to a language-specific context, PCA allowed us to examine whether the items clustered into interpretable dimensions consistent with the intended construct. Although Exploratory Factor Analysis is often used to investigate latent constructs, our primary goal was data reduction rather than the development of a formal measurement model. Second, PCA reduced the number of predictors entering subsequent analyses by summarising responses across multiple items into a smaller set of component scores, thereby simplifying the modelling procedure and reducing the risk of multicollinearity. All PCA calculations, including psychometric characteristics (α, λ_6, ω), were obtained using the psych package (v. 2.4.6.26) in R (v. 4.4.2). The questionnaire demonstrated acceptable internal consistency and construct validity (Cronbach's $\alpha = 0.69$; Guttman's $\lambda_6 = 0.74$; McDonald's $\omega_{total} = 0.83$). Three principal components were retained based on inspection of the scree plot. To aid interpretability, varimax rotation was applied, yielding a solution with three orthogonal (i.e., uncorrelated) dimensions of self-reported mimicry. The first component (External Mimicry) reflected imitation of observable linguistic behaviour such as accents, phrases, and speech sounds. The second component (Internal Mimicry) captured internalised rehearsal behaviours, including silently repeating language, self-directed speech, and memorisation of verbal material such as lyrics or poems. The third component (Compulsive Mimicry) reflected a tendency toward involuntary or compulsive imitation of environmental or linguistic sounds. Component scores were extracted for each participant and used as summary indices of these three dimensions in subsequent analyses.

We also applied principal component analysis (PCA) separately within each imitation modality to reduce the number of behavioural imitation measures and derive a single summary score. The manual imitation PCA included four measures of fidelity and overimitation obtained from the puzzle-box task; the written imitation PCA included four measures of fidelity and overimitation from the reproduction of two Chinese characters; and the spoken imitation PCA included six measures assessing pronunciation, tone, and chunking in the imitation of two Chinese sentences. For each modality, only the first principal component was retained because it captured the variance shared across the individual measures. The resulting component scores served

as summary indices of Manual Imitation, Written Imitation, and Spoken Imitation, respectively, and were used in all subsequent analyses.

3.3 Graphical modelling

The data analysis proceeded in two stages. First, we tested whether the relationships between the predictors and GJT differed across the two experimental conditions (BoxType: Transparent vs. Opaque) using the InvariantCausalPrediction package (v. 0.8) in R. Invariant Causal Prediction evaluates whether predictor–outcome relationships remain stable across different environments or conditions. We specified GJT as the dependent variable, the remaining 12 measures as predictors, and BoxType as the environmental/condition variable. All p-values exceeded .10, indicating no evidence that the predictor–outcome relationships varied between the Transparent and Opaque conditions.[7] We therefore proceeded with a single graphical model pooling observations across both conditions.

We analysed the conditional independence structure of the 13 variables using Graphical Modelling. Partial correlations, computed using the ppcor package (v. 1.1) for R, served as sufficient statistics for model estimation. These were transformed via Fisher's Z to obtain p-values, implemented through a custom-made R function. The resulting graphical structure was inferred using the pcalg package (v. 2.7-12) and visualised with the help of the igraph package (v. 2.1.1) for R. The resulting graph (presented in Figure 5, R output manually adapted for accessibility) reveals a clear and interpretable network of relationships: vertices of different shapes represent different classes of variables, whereas edges denote statistically significant conditional associations between them.

[7] As the Collocation scores can also be considered a dependent variable, we tested the invariance for Collocation as well (excluding only GJT). Again, all p-values exceeded 0.1, thus, confirming the invariance of effects.

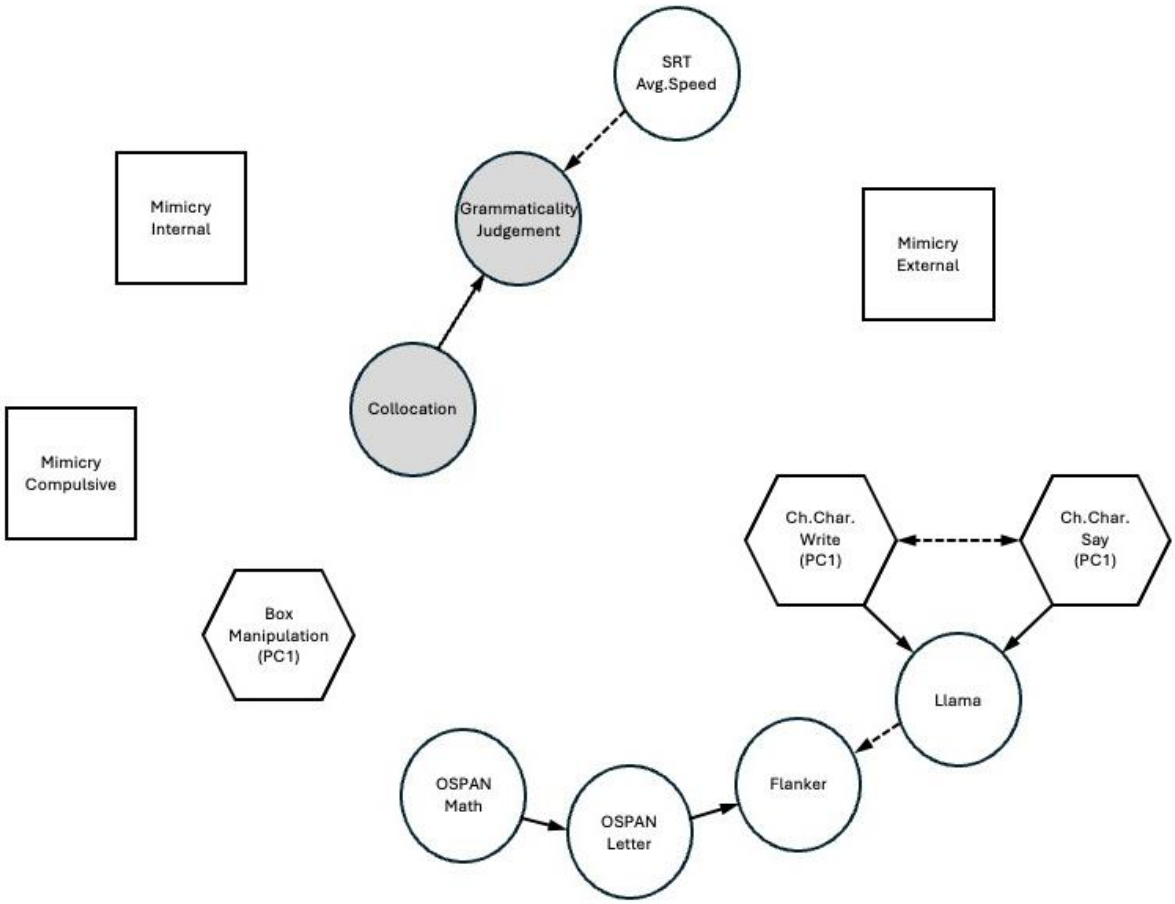


Figure 1. Graphical model. Gray-shaded circles = language knowledge measures; squares = principal components of mimicry test; hexagons = principal components of behavioural imitation tests; white circles = various cognitive tests, including Llama F. Dashed-line edges represent negative conditional associations, whereas solid-line edges represent positive conditional associations.

3.4 Final results

The network shown in Figure 5 splits into two largely distinct clusters of variables: (1) a linguistic proficiency cluster, linking the Grammaticality Judgement Task, Collocation, and Serial Reaction Time task, and (2) a production-based imitation cluster centred on Flanker with two wings leading into it: OSPAN (Letter and Math) directly to Flanker and Speaking (Chinese Character Say) and Writing (Chinese Character Write) via LLAMA F to Flanker. Self-reported mimicry measures (Mimic Internal, Mimic External and Mimic Compulsively) as well as Box Manipulation were largely isolated from the main language network, showing no direct dependency on any other variable in the set.

The linguistic proficiency cluster captures relationships among measures of explicit and implicit linguistic aptitude and knowledge. Grammaticality judgments were affected positively by collocational knowledge (p = .002), indicating shared lexical–

grammatical proficiency. Grammaticality judgments were also negatively dependent on the SRT intercept (p = .002), suggesting that participants with higher grammatical sensitivity demonstrated faster implicit sequence learning. Together, these links point to an integrated linguistic competence combining explicit grammatical awareness with implicit learning efficiency.

In the production-based imitation cluster, Llama F revealed a positive dependency on Speaking (p = .003) and Writing (p = .013), indicating that participants with stronger explicit learning aptitude performed better in both modalities. However, the relationship between Speaking and Writing was negative and bidirectional (p = .044). Flanker was negatively affected by Llama F (p = .015) and positively by OSPAN Letter (p = .048): participants with stronger explicit learning aptitude performed better (i.e., faster or more accurately) on the Flanker task, while better working-memory performance was associated with less efficient Flanker performance. OSPAN Letter was affected by OSPAN Math (p = .000), which is explained by the fact that both measures were part of the same task, tapping into the same memory construct.

4. Discussion

We set out to investigate a possible role of overimitation in second language learning and establish its position with respect to established measure of L2 learning aptitude. More specifically, we examined how overimitation and associated imitation behaviours relate to established measures of language learning, of aptitude for language and domain-general cognitive abilities such as memory and attention. We hypothesised that individuals with a greater propensity for overimitation would perform better on the Grammaticality Judgement and Collocations Tasks, reflecting a positive association between overimitation and language learning. At the same time, for overimitation to qualify as a distinct aptitude component rather than a proxy for existing constructs, we expected it to show only limited overlap with established measures of language-learning aptitude and related cognitive abilities.

Across tasks, a consistent picture emerges for our group of advanced L2 learners: while linguistic, language aptitude and general cognitive abilities form coherent, interpretable structures aligned with existing theoretical models, overimitation behaves as a largely independent construct. Crucially, overimitation does not covary with language knowledge or language aptitude, suggesting functional independence between the constructs.

A key outcome of the graphical modelling is the emergence of a linguistic proficiency cluster that combines grammatical sensitivity, collocational knowledge, and implicit sequence learning. The tight association between grammaticality judgments and collocational knowledge reinforces a long-standing view in usage-based and constructionist approaches: that grammatical knowledge and lexical knowledge are

deeply intertwined expressions of a shared underlying sensitivity to distributional patterns (Langacker, 1987). More surprising is the way this integrated linguistic competence is also connected to performance on an implicit sequence-learning task. This suggests that individuals who are better attuned to lexical–grammatical structure may also possess a more general sensitivity to patterned input, extending beyond the linguistic domain. Such a link resonates with theoretical proposals that grammar and lexicon emerge from the same (statistical) learning mechanisms, but it also raises questions about how domain-general sequential processing supports higher-order linguistic judgments.

In contrast, production-based language imitation (spoken and written Chinese reproduction) grouped with explicit learning aptitude and executive resources. This pattern suggests that short-term linguistic imitation, despite its surface resemblance to procedural imitation, is supported primarily by controlled processes: encoding novel forms, maintaining them in working memory, and reproducing them accurately. The separation of production-based imitation from the linguistic proficiency cluster underscores that immediate reproduction of novel linguistic material does not necessarily tap into the same mechanisms that support long-term grammatical or lexical knowledge. This division is consistent with models positing the existence of a "phonological loop" (Baddeley et al., 2017) which supports initial encoding but is only indirectly related to the consolidation of abstract linguistic structure.

Most strikingly, procedural overimitation and self-reported mimicry were largely unconnected to either of these linguistic or cognitive clusters. This theoretically important dissociation suggests that propensity for overimitation—at least as elicited by a non-social, object-based paradigm as the puzzle box—may not reflect a language aptitude-like cognitive capacity. Instead, it may represent a socio-motivational disposition, consistent with accounts that frame overimitation as driven by affiliative goals born out of a desire for social connectedness (often with a specific individual or group, Over & Carpenter, 2013) or norm-tracking born out of a desire to conform to how things ought to be done within a specific social or cultural context (Keupp et al., 2013). The lack of integration with explicit aptitude, implicit learning, or working memory constraints indicates that overimitation propensity may operate along dimensions orthogonal to those traditionally associated with language learning.

Taken together, these findings support a multidimensional model, in which linguistic proficiency, controlled linguistic imitation, and social imitation reflect separable systems with distinct cognitive underpinnings. For theories linking imitation to language learning, the results imply that only certain forms of imitation—those activating the language network and relying on controlled, explicitly supported mechanisms—align with known language-learning aptitudes. Procedural overimitation, in contrast, may occupy a different functional niche, one that is socially rather than cognitively grounded. Future work should test whether different levels of proficiency yield different results;

whether more socially embedded overimitation paradigms produce tighter links to language learning; or whether overimitation should indeed be treated as a domain-general socio-cognitive trait independent of aptitude. Longitudinal and predictive modelling approaches will be essential to clarifying whether these clusters map onto real-world learning trajectories.

References

Ash, P. A., Benzaquén, E., Gander, P. E., Berger, J. I. & S. Kumar. 2024. Mimicry in misophonia: A large-scale survey of prevalence and relationship with trigger sounds. *Journal of clinical psychology* 80(1). 186–197. https://doi.org/10.1002/jclp.23605.

Baddeley, A. D., Gathercole, S. E. & C. Papagno. 2017. The phonological loop as a language learning device. *Exploring working memory*, 164–198.

Bannard, C., Klinger, J. & M. Tomasello. (2013). How Selective Are 3-Year-Olds in Imitating Novel Linguistic Material? *Developmental Psychology* 49. 2344–2356. https://doi.org/DOI: 10.1037/a0032062.

Blackmore, S. 2007. Imitation makes us human. In C. Pasternak (ed.), *What makes us human?,* 1-16. Oxford: Oneworld.

Carroll, J. B. & S. M. Sapon. 1959. *Modern Language Aptitude Test*. Psychological Corporation.

Dąbrowska, E. 2014. Words that go together: Measuring individual differences in native speakers' knowledge of collocations. *The Mental Lexicon* 9(3). 401–418. https://doi.org/10.1075/ml.9.3.02dab.

DeKeyser, R. M. 2000. The robustness of critical period effects in second languageacquisition. *Studies in Second Language Acquisition* 22(4). 499–533. https://doi.org/10.1017/S0272263100004022.

Farmer, H., Ciaunica, A. & A. F. d. C. Hamilton. 2018. The functions of imitative behaviour in humans. *Mind & language* 33(4). 378–396. https://doi.org/10.1111/mila.12189.

Hoehl, S., Keupp, S., Schleihauf, H., McGuigan, N., Buttelmann, D. & A. Whiten. 2019. 'Over-imitation': A review and appraisal of a decade of research. *Developmental Review* 51. 90–108. https://doi.org/10.1016/j.dr.2018.12.002.

Horner, V. & A. Whiten. 2005. Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal Cognition* 8. 164–181. https://doi.org/https://doi.org/10.1007/s10071-004-0239-6.

Keupp, S., Behne, T. & H. Rakoczy. 2013. Why do children overimitate? Normativity is crucial. *Journal of Experimental Child Psychology* 116(2). 392–406. https://doi.org/10.1016/j.jecp.2013.07.002.

Klinger, J., Mayor, J. & C. Bannard. 2016. Children's Faithfulness in Imitating Language Use Varies Cross-Culturally, Contingent on Prior Experience. *Child Development* 87(3). 820–833. https://doi.org/10.1111/cdev.12503.

Langacker, R. W. 1987. *Foundations of cognitive grammar.* Vol. I. *Theoretical prerequisites.* Stanford, CA: Stanford University Press.

Lyons, D. E., Damrosch, D. H., Lin, J. K., Macris, D. M. & F. C. Keil. 2011. The scope and limits of overimitation in the transmission of artefact culture. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 366(1567). 1158–1167. https://doi.org/https://doi.org/10.1098/rstb.2010.0335.

McGuigan, N., Makinson, J. & A. Whiten. 2010. From over-imitation to super-copying: Adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. *British Journal of Psychology* 102. 1-18. https://doi.org/10.1348/000712610X493115.

Oliveira, C. M., Hayiou-Thomas, M. E. & L. M. Henderson. 2023. The reliability of the serial reaction time task: meta-analysis of test–retest correlations. *Royal Society Open Science* 10(7). https://doi.org/10.1098/rsos.221542.

Oliveira, C. M., Hayiou-Thomas, M. E. & L. M. Henderson. 2024. Reliability of the serial reaction time task: If at first you don't succeed, try, try, try again. *Quarterly Journal of Experimental Psychology* 77(11). 2256–2282. https://doi.org/10.1177/17470218241232347.

Over, H. & M. Carpenter. 2013. The social side of imitation. *Child development perspectives* 7(1). 6-11. https://doi.org/10.1111/cdep.12006.

Richards, J. C. & T. S. Rodgers. 2014. *Approaches and Methods in Language Teaching.* 3rd edn. Cambridge: Cambridge University Press.

Robinson, P. 2005. Aptitude and second language acquisition. *Annual review of applied linguistics* 25. 46-73. https://doi.org/10.1017/S0267190505000036.

Tennie, C., Call, J. & M. Tomasello. 2009. Ratcheting up the ratchet: on the evolution of cumulative culture. *Philos Trans R Soc Lond B Biol Sci*. 364(1528). 2405–2415. https://doi.org/doi: 10.1098/rstb.2009.0052.

Tomasello, M. 1998. Social Cognition and the Evolution of Culture. In J. Langer & M. Killen (eds.), *Piaget, evolution, and development*, 221-245. Hove: Psychology Press.