A Bayesian approach to (re)examining learning effects of cognitive linguistics–inspired instruction: a close replication of Wong, Zhao, and MacWhinney (2018)

Man Ho Ivy Wong, Jakob Prange

Angaben zur Veröffentlichung / Publication details:

Wong, Man Ho Ivy, and Jakob Prange. 2024. "A Bayesian approach to (re)examining learning effects of cognitive linguistics–inspired instruction: a close replication of Wong, Zhao, and MacWhinney (2018)." Studies in Second Language Acquisition 46: 1493–513. https://doi.org/10.1017/s0272263124000603.

Studies in Second Language Acquisition (2024), **46**, 1493–1513
doi:10.1017/S0272263124000603

REPLICATION STUDY

A Bayesian approach to (re)examining learning effects of cognitive linguistics–inspired instruction*

A close replication of Wong, Zhao, and MacWhinney (2018)

Man Ho Ivy Wong[1,2] and Jakob Prange[3]

[1]Department of Education, University of York, York, United Kingdom; [2] Department of English, Hong Kong Shue Yan University, Hong Kong SAR, China and [3]Faculty of Applied Computer Science, University of Augsburg, Augsburg, Germany
Corresponding author: Man Ho Ivy Wong; Email: mhwong@hksyu.edu.

(Received 02 August 2023; Revised 11 August 2024; Accepted 03 September 2024)

Abstract

This study closely replicates Wong, Zhao, & MacWhinney (2018), who found that cognitive linguistics–inspired instruction (i.e., schematic diagram feedback) demonstrated a superiority effect over traditional instruction (i.e., rule and exemplar feedback or corrective feedback) on the translation test but not the cloze test. While the original study adopted the null hypothesis testing approach, the current study adopted Bayesian mixed effects logistic models to investigate how different variables might affect the learnability of prepositions among 81 Chinese-speaking learners of English. The research design, materials, and procedure are nearly identical to those of the original study except for an added delayed posttest. Our findings are generally consistent with the results reported in the original study, indicating that the cognitive linguistics–informed instruction demonstrates superiority effect. Furthermore, these positive learning outcomes persist over time, as evidenced by the results of the delayed posttest.

Introduction

Since the early 1980s, cognitive linguists have contended that prepositions carry multiple meanings, spanning both spatial and abstract semantic domains, yet are conceptually interconnected in a systematic manner (e.g., Brugman, 1988; Herskovits, 1986, 1988; Lakoff, 1987). This pioneering perspective presents significant advantages for pedagogical grammars over traditional methods like memorization and drilling. For example, demonstrating the spectrum of meanings associated with a single preposition in the form of a

*The online version of this article has been updated since original publication. A notice detailing the change has also been published

semantic network reduces arbitrariness, thereby diminishing the reliance on rote learning (Evans & Tyler, 2005). However, these semantic networks of meanings are often cognitively dense, which makes them difficult to translate for language teaching and learning purposes. With this goal in mind, Wong, Zhao, and MacWhinney (2018) translated cognitive semantic analyses of prepositions into instructional materials for second language (L2) learning using the English Preposition Tutor—a computerized tutorial system designed for preposition learning. The work of Wong et al. (2018) was selected for this replication study as it was the first to systematically combine two highly compatible usage-based theoretical frameworks, the competition model (CM) and cognitive linguistics (CL), to inform computer-assisted language learning (CALL), filling important research gaps in the fields of applied CM and CL. In addition, the original study has been cited over 60 times by various refereed publications in different areas of linguistics, second language acquisition (SLA), computational linguistics, and psycholinguistics as well as others. The robust methodological design of the paper has continued to influence research in SLA. Despite the important theoretical and pedagogical contributions of the original study, some limitations can be addressed through replication. For example, the null hypothesis testing (NHT) approach used on the original data can now be improved by using Bayesian mixed effects modeling (MEM), providing a more flexible approach to understanding both fixed and random effects contributions to preposition learning. Additionally, the imbalanced experimental group size in the original study raised concerns about the complex experimental design. Therefore, the Bayesian approach aims to enhance research validity by providing reliable estimates, even with small sample sizes. Another change made to the original study would be the inclusion of a delayed posttest to measure whether learning (if any) is retained over time. The major modification compared to the original study lies in our means of data analysis via a Bayesian statistical model. Thus, our goal is repeat as closely as possible Wong et al. (2018) in terms of all major aspects of the research methodology. Any other major modifications are likely to change the nature of our close replication (Porte & McManus, 2019).

Background literature

Cognitive linguistics–informed studies on L2 grammar instruction

Unlike the formal approaches to L2 learning, CL does not simply focus on syntax but the intersection between meaning and language, taking culture, anthropology, and psychology into consideration (Hajazo-Gascón & Llopis-García, 2019; Tyler, Huang, & Jan, 2018). Viewing language acquisition in relation to other general cognitive abilities, such as memory and attention, CL, as a theoretical framework, fits squarely with the functionalist perspective (e.g., usage-based linguistics). Therefore, CL has been widely adopted in many areas of applied linguistics, with SLA receiving the most attention (Boers & Lindstromberg, 2006; Robinson & Ellis, 2008; Tyler et al., 2018; Verspoor & Lowie, 2003). For example, CL has been widely applied to L2 grammar instruction (e.g., Archard & Niemeier, 2004; Rudzka-Ostyn, 2003; Tyler, 2012), bringing in plethora of evidence from experimental research to demonstrate the effectiveness of CL-informed methodologies (Jacobsen, 2018; Hwang, 2023; Qin, Wu, and Zhong, 2023; Wong et al., 2018; Wong, 2023).

Unlike the formal approaches to L2 learning, CL does not simply focus on syntax but the intersection between meaning and language, taking culture, anthropology, and psychology into consideration (Hajazo-Gascón & Llopis-García, 2019; Tyler, Huang, & Jan, 2018). Viewing language acquisition in relation to other general cognitive abilities, such as memory and attention, CL, as a theoretical framework, fits squarely

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

with the functionalist perspective (e.g., usage-based linguistics). Therefore, CL has been widely adopted in many areas of applied linguistics, with SLA receiving the most attention (Boers & Lindstromberg, 2006; Robinson & Ellis, 2008; Tyler et al., 2018; Verspoor & Lowie, 2003). For example, CL has been widely applied to L2 grammar instruction (e.g., Archard & Niemeier, 2004; Rudzka-Ostyn, 2003; Tyler, 2012), bringing in plethora of evidence from experimental research to demonstrate the effectiveness of CL-informed methodologies (Jacobsen, 2018; Hwang, 2023; Qin, Wu, and Zhong, 2023; Wong et al., 2018; Wong, 2023).

Jacobsen (2018) compared the efficacy of applying CL analysis of English conditionals to L2 instruction. The study found that the cognitive group performed significantly better than the task-supported group (without CL presentation of conditionals) in the production task but not in the comprehension task. This task variation effect was also observed in both Wong et al. (2018) and Wong (2023), where the superiority effect of CL-inspired preposition training was discernible exclusively in production tasks. The two instructional studies on preposition studies will be reviewed further in the upcoming section. In another study, Hwang (2023) explored pedagogical strategies to balance attention with form and meaning for enhanced efficiency in learning caused-motion constructions (e.g., "Sam put the apples into the box," "Jacky drove Peter to the beach"). Adopting a production task, the CL-inspired group, allowing simultaneous attention to both form and meaning, was found to be the most effective when compared to a form-oriented group and a meaning-oriented group. On the other hand, in their investigation of phrasal verb learning, Qin et al. (2023) found that CL-inspired instruction outperformed the traditional group, specifically in meaning recall tests, but not in a meaning recognition test. The authors attributed the lack of significance to inadequate training time and the novelty effect, stemming from the unconventional approach, affecting both the instructor and the learners. Nevertheless, these findings affirm the positive impact of CL-inspired instruction on L2 grammar instruction.

Image schemas and preposition learning

CL posits that language is grounded in our cognitive experiences and conceptual systems. That is, our bodily experiences shape how we understand and use language to communicate meanings. As human beings, we continuously learn about objects not only as objective forms but also by conceptualizing how they can impact us (Holme, 2009; Talmy, 2005; Tyler, 2012). This body–mind awareness begins developing as early as infancy, preceding even language development (Mandler, 1992). Conceptual notions obtained through this body–mind awareness, such as animacy, inanimacy, agency, and containment, will be accumulated in the form of *image schemas*. These are structural contours that exist as recurring and analogue patterns beneath conscious awareness, prior to language acquisition (Hampe, 2005). Johnson (2005) emphasizes the critical role of image-schematic structures in deciphering concepts ranging from spatial relations to abstract notions of reason, mind, knowledge, justice, rights, and values. Importantly, Langacker (2008) argues that an imagistic approach is as capable as a propositional one in depicting complex structures. Consequently, the notion of image schema has been found to be particularly compatible with the teaching and learning of prepositions (Brugman, 1988; Lam, 2009; Tyler, 2012; Tyler & Evans, 2003; Wong et al., 2018), whose primary function is to convey spatial and temporal relationships.

Alongside compelling behavioral evidence, a recent event-related potential study confirmed the neurocognitive advantages of using diagram-based instruction for

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

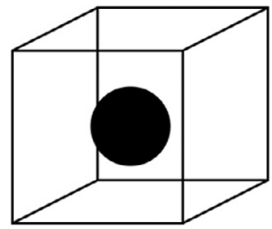

Figure 1. The CONTAINMENT schematization for the preposition *in*.

Figure 2. The POINT-ALONG-THE-ROUTE schematization for the preposition *at.*

acquiring English prepositions (Zhao, Huang, Zhou, & Wang, 2020). It is crucial to note that schematic diagrams and pictorial illustrations in CL-informed instruction serve a purpose beyond being visual aids. Their effectiveness in facilitating learning is maximized when they closely resemble the psychological gestalts stored in our mental lexicons. To demonstrate how schematization can enhance the learning of English prepositions, we will explore the CONTAINMENT schema (see Figure 1). The CONTAINMENT schema is one of the earliest acquired by infants, even before language development (Mandler, 1992). Figure 1 illustrates how the CONTAINMENT schema can convey both spatial and nonspatial senses of the preposition *in*. To grasp the schematic structure, one must identify two essential elements that realize an image schema: the Trajector (TR), spatially related to but more movable than the Landmark (LM). In the sentence "Mary is talking in her room," Mary is the TR, and her room is the LM. The TR, Mary, is spatially bounded by the exterior of the LM, such as the walls and the door of her room. Similarly, in the sentence "Joe is trapped in his marriage," Joe, the TR, is metaphorically bounded by the LM, his marriage, conceptualized metaphorically as a container limiting the movement of the TR. Hence, both spatial and nonspatial senses of the preposition *in* can be elucidated using the same schematization.

The difficulty in acquiring prepositions lies not in learners' inability to understand these schematizations but often in their uncertainty about which preposition to choose when faced with various options. For instance, the preposition *at* is commonly viewed as a typical English preposition not found in many other languages, such as Chinese or Italian. The vague definition of *at* as indicating a particular location provides minimal guidance when it competes with *in*. In contrast, the CL definition clarifies that the LM is conceptualized as a point along the route (see Figure 2), enabling learners to recognize cross-linguistic differences and the semantic distinctions conveyed by competing forms. Emphasizing conceptual understanding in this way directs learners' attention to how preposition forms convey various meanings. Thus, a meaning-oriented approach, supported by schematic diagrams in CL instruction, empowers learners to have greater control over their preposition knowledge and increased confidence in their preposition usage.

Despite the fruitful discussion of CL on preposition acquisition and successful applications of CL approaches to teaching prepositions, the diverse learning outcomes (Boieblan, 2023; Jacobsen, 2018; Lam, 2009; Tyler, 2012; Wong, 2023), as well as the limited number of experimental research studies, fail to offer robust empirical support for the pedagogical advantages associated with CL. Many CL-informed experimental studies conducted in the earlier phases, including some of the above-mentioned studies, primarily concentrated on the quality of pedagogical materials, such as CL

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

explanations. However, there was often a lack of emphasis on robust research design and appropriate statistical procedures, as pointed out by Jacobsen (2018), raising concerns about the generalizability of the corresponding findings. Wong et al. (2018) also raised their concern regarding the replicability of many CL-informed experimental studies due to (a) low transparency of adopted CL materials, (b) a lack of information regarding the selection of instructional paradigm, and (c) mostly researcher-fronted training.

The initial study

To address the multiple research gaps identified in applied CL research, Wong et al. (2018) developed an experimentalized CALL (eCALL) system called the English Preposition Tutor, to investigate the effectiveness of the CL approach in the CALL environment. While CL-informed studies focused on the pedagogical content of training tasks, studies of eCALL applications, motivated by the CM (MacWhinney, 1997), have examined other aspects of language teaching, including the role of cue focusing (i.e., presenting two competing cues only at a time), the types of feedback, and practice effects (Wong, 2023; Wong et al., 2018; Zhao & MacWhinney, 2018). Combining the two usage-based frameworks (i.e., CL and CM), the authors investigated whether the CL-informed approach, delivered via the eCALL tutor, had improved learnability of both spatial and nonspatial preposition polysemes.

Sixty-three Chinese-speaking learners of English, ages 15 to 16, were assigned to three treatment groups (schematic diagram feedback group, rule and exemplar feedback group, and corrective feedback group) and one test-only control. The three treatment groups received training on 12 prepositional polysemes selected from the prepositions *in*, *at*, and *over*. For each training trial, a picture stimulus and two contrasting sentences were provided to participants (see supplementary material S1 and S2 for details). The cue-focusing design, motivated by CM, aimed to draw participants' limited attentional resources to processing the two most competing cues, the target and the distractor prepositions. The major differences between the three treatment groups lay within the feedback that learners received: (a) the schematic diagram feedback was composed of a schematic diagram to reflect the spatial configuration of the target preposition, accompanied by a brief explanation on usage; (b) the rule and exemplar feedback consisted of a metalinguistic rule explanation paired with three example sentences to illustrate usage; and (c) in the corrective feedback group, learners would only receive feedback on whether their answer provided was correct. The test control group received online training on English articles instead.

A sentence-level cloze task and a translation task were adopted as measures. Pretest, training, and posttest were all delivered online (See Wong et al., 2018 for a full description of the instruction and testing). Repeated measures analysis of variance (ANOVA) confirmed that cue focusing was an effective instructional method. In addition, the gains of the three treatment groups were significant at the posttest. However, there was no interaction between time and types of instruction on the cloze task. The superiority effect of the CL approach was observed only on the translation task where the Bonferroni post hoc tests indicated the main source of variation was between the schematic feedback group and the corrective feedback group only. In other words, rich feedback (schematic feedback; rule and exemplar feedback) contributed to an increase in productive knowledge, as measured by the translation test, whereas minimal feedback (corrective feedback) would already be sufficient to promote receptive knowledge, as measured by the cloze test. For the learning of spatial and nonspatial polysemes,

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

learners across the three treatment groups demonstrated significantly better learning for spatial polysemes over nonspatial polysemes on the cloze test. Yet, the results from the translation test demonstrated significant yet comparable gains between the two types of polysemes.

Rationale for the replication study

Findings from the original study demonstrated a superiority effect of CL-informed instruction solely over one of the traditional groups (i.e., minimal feedback), specifically in the translation task. Further, no delayed posttest was administered. Therefore, the authors could not confirm whether the positive learning outcomes were sustained over time. Replicating the study would therefore enable us to elicit more evidence for interpretation.

Moreover, in the original analysis, repeated measures ANOVAs were used, and the dependent variable was the aggregated mean scores. These traditional approaches to variance analysis continue to be widely adopted in applied CL research (Boieblan, 2023; Colasacco, 2019; Jacobsen, 2018; Qin et al., 2023). However, repeated measures ANOVA are only suitable for simple experimental designs, as several assumptions have to be met before conducting the analysis, including normally distributed variance, sphericity (e.g., constant variance across time points), no outliers in any of the repeated measurements, and balanced numbers across comparison groups (Jaeger, 2008). These assumptions are often unrealistic and thus pose great challenges on L2 interventional studies. Given the recent advances in computational software, we believe a *computational turn* is necessary and inevitable not only for CL (Divjak & Milin, 2023) but also for applied CL. Our focus of the reanalysis was to compare the outcomes derived from Bayesian MEM with those obtained through traditional ANOVA in the original study. To utilize the Bayesian mixed effects logistic models for the present study, the reanalysis was also a necessary step to facilitate prior distribution setting (Bürkner, Scholz, & Radev, 2023).

Unlike repeated measures ANOVA, MEM can address concerns, such as nonspherical error variance and heteroscestadicity, which are commonly observed in natural datasets. More importantly, MEM accounts for subjective variation by calculating several intercepts (means), one for each subject (patient), thereby teasing apart random effects from the fixed effects of parameters (e.g., predictors of an outcome variable), increasing the accuracy of interpretation on fixed effects. Because MEM can model variation explicitly (e.g., adjust estimates for imbalanced sampling or to study variation), preaveraging or averaging data to construct variable would not be necessary (McElreath, 2015). Also, the sigmoid function in logistic regression tapers the outliers and thus the presence of outlier data points does not exert great impact on the model performance compared to traditional ANOVA.

There is an increasing number of instructed L2 studies adopting MEM (Bovolenta & Williams, 2023; Hwang, 2023; Puimege, Perez, & Peters, 2023); however, most of these adopted a frequentist perspective, for example, NHT. Data analysis and interpretation using the NHT approach have begun to draw concerns in applied linguistics research (Norouzian, de Miranda, & Plonsky, 2018; Norris, 2015; Plonsky, 2015). A frequentist approach draws conclusions based on facts obtained from the observed data at hand only (i.e., excluding any prior information) and identifies probability with frequency, premising on imaginary data resampling (McElreath, 2015). This frequentist perspective continues to motivate researchers and reviewers to aim for large sample sizes. For example, if the coin toss experiment was performed 3,000 times or tossing continued

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

until 300 tails were observed, then the total probability, often denoted as the p value, could fall below a critical threshold (typically set at 0.05). In this context, $p < .05$ suggests that the likelihood of obtaining such results by random change is low, allowing a rejection to the null hypothesis. Hence, in the framework of NHT, a power analysis must be conducted to determine the optimal sample size of a study before data collection begins. In contrast, the Bayesian analysis yields a posterior distribution that updates its belief with every observation, making it less sensitive to sample size. Therefore, taking a Bayesian approach would be more appropriate and realistic for L2 intervention studies (Garcia, 2021, 2020; Wirtz & Pfenninger, 2023), considering the logistic and cost constraints inherent in the real-world scenarios.

From a frequentist perspective, the parameter used (e.g., $\theta = 0.5$ the two-sided coin is believed fair) to estimate the population is assumed to be fixed where there is only a single true parameter that is estimated and is not modeled as a probability distribution. When new data becomes available, it is used to perform statistical tests and predictions. On the other hand, Bayesian inference takes into the account initial estimates, the prior (i.e., prior information from the original study), which are both expressed in terms of probability distribution. In replication studies, incorporating prior information from the original research is crucial and is best achieved through Bayesian analysis. Neglecting such valuable information would be counterproductive, provided that the researcher acknowledges and appreciates the significance of the original study.

While it is our goal to replicate Wong et al. (2018) as closely as possible in most major aspects of their research methodology in order to either confirm or expand on their findings, we make two deliberate modifications to the setup as follows:

- Statistical data analysis: Rather than performing NHT for separate effects as in Wong et al. (2018), we use Bayesian MEM (Prange & Wong, 2023). One of the biggest advantages of the Bayesian approach is that we can incorporate the prior information from the original study into the model through prior distributions, leading to more robust and realistic inferences. While one can primarily obtain point estimates and confidence intervals (CIs) using frequentist MEM, the p values and CIs could be misleading and do not provide useful information for decision making in real life. Gelman and Hill (2007) pointed out that "all multilevel models are Bayesian in the sense of assigning probability distributions to the varying regression coefficients" (p. 276). Therefore, the Bayesian perspective fits squarely with multilevel modeling and its interpretation. Moreover, the Bayesian method allows us to understand not only the central tendency (e.g., mean) but also the uncertainty and shape of the distribution of each parameter. More importantly, one can directly state the probability that a parameter is greater than zero or that one condition is better than another. As a result, this is more intuitive and informative for decision making, which is particularly useful for L2 interventional research.
- Inclusion of a delayed posttest to examine whether the learning effect (if any) is sustainable over time.

Because the experimental manipulation of this replication is identical to that of the original study (as shown in Table 1), we addressed the research questions (RQs) as proposed in the original study but with greater precision and clarity:

> *RQ 1:* To what extent does cue focusing (i.e., minimal pair design) improve the learning of five target preposition forms (in, at, and over) immediately after instruction (at posttest) and three weeks later (at delayed posttest)?

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

Table 1. Comparison of methodologies in the original and replication studies

Methodology	The original study	The replication study
Participants	63 secondary 4 students (ages 15–16)	81 secondary 3 students (ages 14–15)
Target preposition*	In (1 spatial + 1 nonspatial) At (2 spatial + 2 nonspatial) Over (3 spatial + 3 nonspatial)	In (1 spatial + 1 nonspatial) At (2 spatial + 2 nonspatial) Over (3 spatial + 3 nonspatial)
Condition (i.e., different types of instruction)	Schematic diagram feedback (n = 17) Rule and exemplar feedback (n = 15) Corrective feedback (n = 13) Control (n = 18)	Schematic diagram feedback (n = 28) Rule and exemplar feedback (n = 20) Corrective feedback (n = 20) Control (n = 13)
Procedure	Pretest, training, and posttest	Pretest, training, posttest, and delayed posttest (after 3 weeks)
General proficiency	Intermediate†	Intermediate†
Measurement	1. Receptive task: a sentence-level multiple-choice cloze test 2. Productive task: a translation test	1. Receptive task: a sentence-level multiple-choice cloze test 2. Productive task: a translation test
Data analysis	The null hypothesis testing approach	Bayesian multilevel mixed effects model

*See supplementary material S1 for more details regarding each pair of spatial and nonspatial polysemes of the three target prepositions.
†The evaluation was provided by the English Panel Heads of the schools based on participants' internal English test performances.

RQ 2: To what extent do different types of feedback (CL-inspired feedback, rule and exemplar metalinguistic feedback, corrective feedback) improve learner accuracy on target preposition use immediately after instruction and three weeks later?

RQ 3: To what extent does the learning of spatial polysemes differ from that of nonspatial polysemes (e.g., in the kitchen vs. in time or love)?

Method

Participants

Identical to the original study, Chinese-speaking learners of English (n = 81) were recruited from two Band One secondary schools in Hong Kong that are recognized by the Education Bureau as "English-as-the-medium-of-instruction" (EMI) schools. Secondary schools in Hong Kong are categorized into three bands according to the students' performance in public examinations, where Band One is the best, followed by Band Two and Band Three. There are approximately 100 EMI schools in Hong Kong, accounting for roughly 30% of local secondary schools. EMI schools deliver instruction mainly in English for core subjects except for Chinese and Chinese history, Therefore, students from EMI schools are known to have higher English proficiency compared to their peers from non-EMI schools. Students are required to attend two English lessons per school day, making up a total of eight hours per week. The

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

participant characteristics closely resemble those described in Wong et al. (2018). However, the participants in the replication study were secondary 2 and 3 students (ages 13–15) instead of secondary 4 students (ages 16–17). After communicating with the English panels of both schools, we decided to label School A as the lower-performing group and School B as the higher-performing group.

With the Bayesian approach, a power analysis was not necessary as Bayesian estimates are valid for any sample size (McElreath, 2015). While more data can increase the precision of the model, we focus on the credibility intervals of the effects of a range of variables on promoting preposition knowledge. Researchers who conduct intervention studies often had very little or even no control over the type and the number of participants they could recruit, as data collection involves significant logistic challenges in addition to high cost. This could also explain the biased sampling in many L2 intervention studies, as many samples are university based, leaving out a large population of young adolescents ages 14–17 (Mifka-Profozic, Behney, Gass, Macis, Chiuchiù, & Bovolenta, 2023; Peters, Puimège, Szudarski, 2023). Therefore, the Bayesian approach is particularly suitable for interventional studies with small sample sizes as it allows researchers to draw reliable inferences without compromising model validity (Garcia, 2023; Vasishth, Nicenboim, Beckman, Li, & Kong, 2018).

Design

The study design, timescale, procedure, and outcome measures were exactly as reported by Wong et al. (2018), except for the addition of a delayed posttest 1 week after the posttest. In line with the original study, the participants were randomly assigned into one of the four groups: the schematic diagram feedback group, the rule and exemplar feedback group, the corrective feedback group, and the control group. The training materials and testing stimuli also remained the same as in the original study. In Wong et al. (2018), the study was administered online in three sessions of data collection: a pretest (day 1), a training session (day 3), and a posttest (day 5), with each session lasting for approximately 60 minutes. The replication study did not vary each training and testing duration but added a delayed posttest three weeks later (day 26) to measure the long-term effect of any behavioral changes. The replica was also administered online but via a different platform, called Gorilla Experimental Builder (http://www.gorilla.sc). The Gorilla Experiment Builder serves as a more affordable alternative for researchers who intend to computerize their pedagogical materials.

Instructional treatment

The treatment components for each experimental condition were the same as those reported in the original study and is briefly explained below.

Following Wong et al. (2018), the three target prepositions selected were *in*, *at*, and *over*. The tutor delivered computerized training for six pairs of prepositional polysemes, including three pairs for *over*, two pairs for *at*, and one pair for *in*. See supplementary material S1 for detailed descriptions of all selected preposition usages and their respective schematic diagrams. The preposition training adopted a minimal-pair design, where two contrasting sentences would be presented to the learner on every trial. The minimal pair design was motivated by the CM (MacWhinney, 2012), aiming to draw learners limited attentional resources to processing highly competing polysemes only. The two sentences only differ in their choice of prepositions, one serving as

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

the target whereas the other serves as the distractor. Based on the contextual information provided in the pictorial illustration, the participant was asked to choose the best sentence to describe the picture. There were seven training items corresponding to each polyseme, resulting in a total of 84 training items (7 × 12). Each pair contained a spatial and a nonspatial polyseme, which were selected and paired up based on their shared schematic diagram. For instance, the spatial polyseme used in the sentence *The man is in his office* expresses the meaning of containment and such meaning is carried over to its nonspatial polyseme as in *the man is in love.* The shared schematic diagram (Figure 2) reflects the psychological gestalt that has been grounded in the embodied experience of human interacting with bounded landmarks (Grady, 1997; Tyler & Evans, 2003). While the conceptual link guided the selection of the preposition polysemes as well as the pictorial illustrations across all training conditions, only the schematic diagram feedback group was provided with the schematic diagrams to demonstrate the conceptual relationship between the spatial polyseme and its nonspatial counterpart. In addition to the schematic diagrams, the CL concepts were introduced to explain the conceptual meanings of the target preposition (Figure 3). The rule and exemplar feedback group and the corrective feedback group represent a more conventional approach to L2 preposition learning. The rule and exemplar feedback group received metalinguistic rule explanation (i.e., rule of thumb definition) together with three example sentences to illustrate usage whereas the corrective feedback group received feedback only on whether their selections were correct or incorrect.

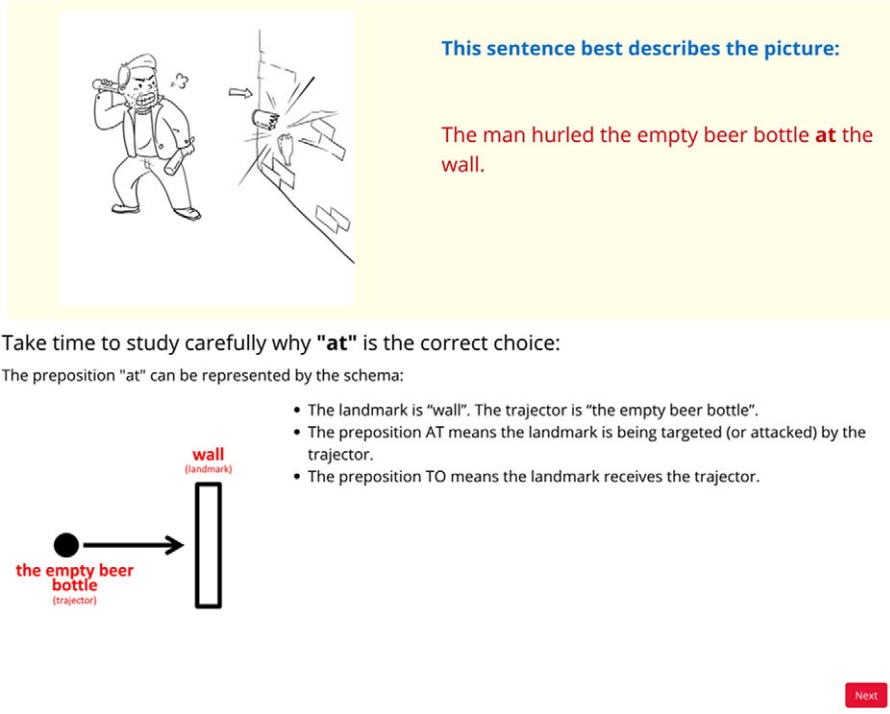

Figure 3. A sample training screen for the schematic diagram feedback group.

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

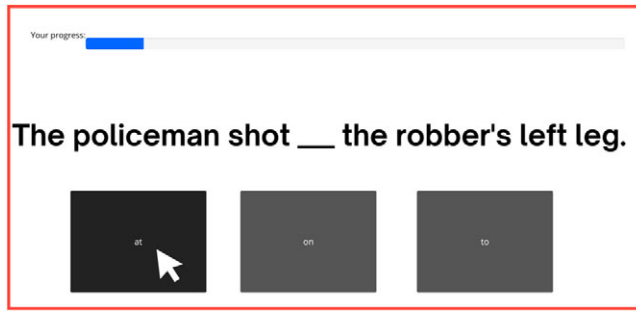

Figure 4. A sample screen of the sentence-level cloze task.

Figure 5. A sample screen of the translation task.

Outcome measures

All measures were identical to those used in Wong et al. (2018). All participants completed a sentence-level cloze task and a translation task for pretest, posttest, and delayed posttest accordingly. No feedback was provided to learners during testing. All test items in the pretest were recycled at the posttest and delayed posttest, but they were presented in a random order. The cloze test consisted of 65 sentences, 4 items for each polyseme (4 × 12) and 15 filler items. Participants were asked to click onto the most appropriate preposition that agrees with the context provided (Figure 4). During the translation test, participants were asked to translate a total of 50 sentences, 3 items for each polyseme (3 × 12) and 14 filler items. The main verb was provided to create an obligatory context for the target preposition (Figure 5). In addition, key noun phrases were provided to support their sentence formulation, as translation tasks were less common for junior level students. Adopting the same coding procedures as in the original investigation, we only coded the corrective for the preposition produced where other errors were not included in the analyses. Each test took approximately 90 minutes to complete (approximately 40–45 minutes/task).

Results

Model fitting

This study aimed to evaluate the effects of three different instructional conditions on students' learning of spatial and nonspatial uses of polysemes, considering the influence

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

of school. To address the three main research questions—*RQ 1: To what extent does cue focusing (i.e., minimal pair design) improve the learning of* three *target preposition forms (in, at, and over) immediately after instruction (at posttest) and* three *weeks later (at delayed posttest)?*; *RQ 2: To what extent do different types of feedback (CL-inspired feedback, rule and exemplar metalinguistic feedback, corrective feedback) improve learner accuracy on target preposition use immediately after instruction and three weeks later*; and *RQ 3: To what extent does the learning of spatial polysemes differ from that of nonspatial polysemes (e.g., in the kitchen vs. in time or love)?* We employed a Bayesian mixed effects logistic regression model using the **brms** package (Bürkner, 2017) to analyze the data, with **Condition** (schematic feedback, traditional feedback, and corrective feedback), **Test** (pretest, posttest, and delayed posttest), and **Idiomaticity** (spatial, nonspatial polysemes) as predictors, along with their interactions. The model also included random intercepts and slopes for participants and random intercepts for items. The response variable was binary (correct vs. incorrect), modeled with a Bernoulli distribution and a logit link function. The model formula was specified as follows:

Response ~ Condition.F * Test.F * Idiomaticity.F *School.F + (1 + Condition.F | Participant) + (1 | ItemNumber)

Because we recruited participants from two different EMI schools (see Participants section for more details), we included **School** as a fixed factor in our model. Using leave-one-out (LOO) cross-validation as the predictive model checking, our model demonstrated the best predictive performance for the cloze test data, with the highest expected log predictive density (**elpd_loo** = –5949.322) and the lowest leave-one-out information criterion (**LOOIC** = 11898.64). Similarly, for the translation data, our model was the best fit, with the highest **elpd_loo** (–7948.36) and the lowest **LOOIC** (15896.71). In other words, our model strikes the best balance between model fit and complexity. Markov chain Monte Carlo sampling was conducted with four chains, each running for 4,000 iterations for cloze test data and 5,000 iterations for translation test data. The adapt_delta parameter was set to 0.95 to enhance convergence due to the complexity of the full models. Computational resources were allocated with six cores. We report mean posterior point estimates and associated probabilities for each parameter, along with the 95% highest density interval, which is a type of credible interval (CrI). A CrI is a range of values within which an unknown parameter θ lies with a certain probability, given the observed data and prior distribution. A CI, originating from the frequentist perspective, is a range of values that, if we were to repeat the experiment many times, would contain the true parameter θ value in a certain proportion of those experiments (e.g., 95% of the time). In other words, the CrIs from the Bayesian approach provide a probability statement about the parameter itself, while the CIs only offer a long-run frequency statement about the procedure. For instance, a 95% CrI directly informs us of the probability that the mean lies within that interval, given the observed data and prior information. However, upon calculating the sample mean and standard error (SE) from the observed data, a 95% CI only tells that if we repeated this sampling process many times, 95% of the calculated intervals would contain the true mean. Therefore, the Bayesian approach gives us a more intuitive understanding of the observed data (McElreath, 2015).

Obtaining highly informative priors from Wong et al. (2018)

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

Table 2. Normal priors obtained for the cloze and translation test

	Cloze test	Translation test
Intercept (the control group)	Normal (0.38, 0.24)	Normal (–0.43, 0.60)
Condition.Fschematic	Normal (–0.07, 0.27)	Normal (–0.17, 0.33)
Condition.Ftraditional	Normal (–0.24, 0.26)	Normal (–0.17, 0.32)
Condition.Fcorrective	Normal (–0.32, 0.29)	Normal (–0.51, 0.34)
Test.Fposttest	Normal (0.04, 0.14)	Normal (0.17, 0.22)
Idiomaticity.F	Normal (0.44, 0.28)	Normal (0.87, 0.83)

One of the most important motivations of taking the Bayesian approach for this replication study is that it allows us to incorporate established knowledge from previous research into our Bayesian MEM, thereby improving the robustness and precision of the estimates (Kruschke, 2018; McElreath, 2015). To the best of our knowledge, this study is among the first to employ highly informative priors in a replication study within L2 intervention research, particularly in applied CL. In this replication study, we have selected a normal distribution for all our priors with means and standard deviations specified for each parameter, as indicated in Table 2. Additional normal priors were specified for interaction terms and variance components. The data and code necessary to reproduce the analyses reported in this article are available at Open Science Framework (OSF) https://osf.io/fxb89/.

In addition to the LOO analysis, posterior predictive checks were conducted to assess the fit of the model to our observed data. While both analyses indicated that the inclusion of informative priors improved model performance, the models without priors showed only a slight drop in performance that might not be practically significant. In other words, the priors were appropriately selected and were consistent with the data.

Cloze test data

```
summary(brm.cloze_fullmodel_prior <- brm(Response ~
                                           Condition.F*Test.F*Idiomaticity.F*School.F +
                                           (1+ Condition.F |Participant) +
                                           (1|ItemNumber),
                                           data=datCloze,
                                           family='bernoulli',
                                           prior = clozepriors,
                                           chains=4, iter=4,000, cores=6,
                                           control=list(adapt_delta=0.95)))
```

Interaction effects of predictors

The primary focus of this analysis was to examine the interaction effects between different instructional conditions, test points, idiomaticity, and schools (Figure 6). The interaction between **Condition** and **Test** was credible for both the posttest and the delayed posttest, with the schematic feedback group demonstrating the biggest

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

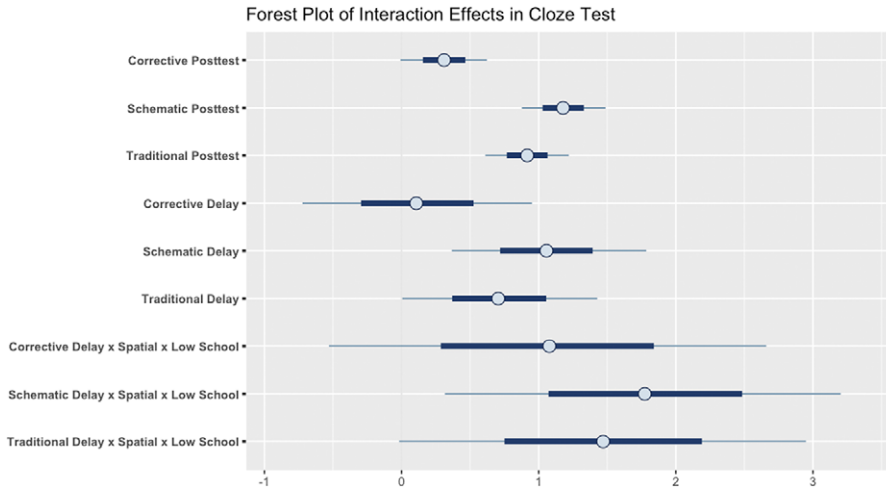

Figure 6. Forest plot of interaction effects in Cloze test. This plot displays the estimated interaction effects for different instructional conditions (Corrective, Schematic, Traditional) at the posttest and delayed posttest, including interactions with idiomaticity (Spatial, Non-spatial) and school (low performing vs. high performing). Each point represents the posterior median estimate, with the inner and outer intervals indicating the 66% and 95% credible intervals, respectively. The interactions illustrate how each instructional method impacts accuracy over time and in different contexts.

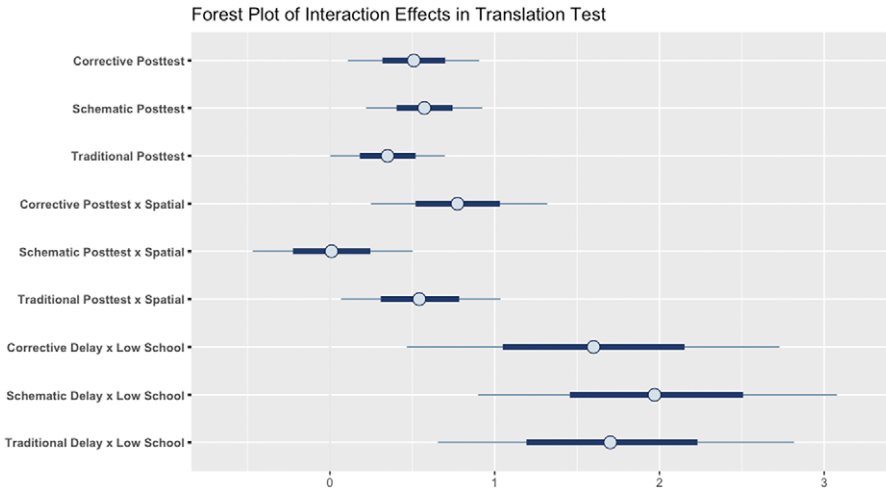

Figure 7. Forest plot of interaction effects in translation test. This plot displays the estimated interaction effects for different instructional conditions (Corrective, Schematic, Traditional) at the posttest, including interactions with idiomaticity (Spatial, Non-spatial) and school (Low performing, High performing). Each point represents the posterior median estimate, with the inner and outer intervals indicating the 66% and 95% credible intervals, respectively. The interactions illustrate how each instructional method impacts accuracy over time and in different contexts.

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

improvement in accuracy from pretest to posttest (β = 1.18, *SE* = 0.16, 95% CrI [0.88, 1.49]). This corresponds to a probability of approximately 0.76, suggesting that the schematic feedback group benefitted more than the traditional and corrective groups. Specifically, the improvement for the schematic feedback group was higher (β = 1.18) compared to the traditional (β = 0.92, *SE* = 0.16, 95% CrI [0.61, 1.22], probability ≈ 0.72) and corrective (β = 0.31, *SE* = 0.16, 95% CrI [–0.01, 0.62], probability ≈ 0.58). Similarly, the schematic feedback group demonstrated the most learning at the delayed posttest (β = 1.06, *SE* = 0.36, 95% CrI [0.37, 1.78], probability ≈ 0.74), suggesting the benefits of CL-informed instruction were sustained over time. While the traditional feedback group showed a sustained effect at the delayed posttest (β = 0.71, *SE* = 0.36, 95% CrI [0.01, 1.43]), the probability has dropped from ≈ 0.72 to ≈ 0.67 at the delayed posttest. Although there was no credible interaction found between **Condition, Test,** and **Idiomaticity** (β = 0.84, *SE* = 0.47, 95% CrI [–1.77, 0.07]), strong evidence was observed for the schematic feedback group at the delayed posttest (β = 1.78, *SE* = 0.74, 95% CrI [0.32, 3.20]) in a four-way interaction between **Condition, Test, Idiomaticity**, and **School**. For students in the lower-performing school, the schematic feedback was particularly effective in facilitating the learning of spatial polysemes (probability ≈ 0.86). However, the benefits of the schematic feedback took longer to unfold as such evidence was not observed at the posttest (β = 0.47, *SE* = 0.44, 95% CrI [–0.39, 1.34], probability ≈ 0.62). In summary, the interaction effects indicate that the schematic feedback group was the most effective overall, particularly in lower-performing schools and for spatial items in the long run. For detailed estimates of all parameters and their associated probabilities, please refer to supplementary material S3.

The random effects

The standard deviation of the intercepts for items was estimated to be 0.76 (*SE* = 0.08, 95% CrI [0.62, 0.95]), indicating variability in baseline responses across items. Similarly, the standard deviation of the intercepts for participants was 0.66 (*SE* = 0.09, 95% CrI [0.48, 0.84]), reflecting variability in baseline responses across participants. Nevertheless, the correlations among the random effects were generally weak, indicating that individual differences in baseline responses were relatively independent of condition-specific effects. Such independence matters as the weak correlation implies that the instructional conditions have similar effect across participants, regardless of their starting point.

Translation test data

```
summary(brm.trans_fullmodel_prior <- brm(Response ~
                                         Condition.F*Test.F*Idiomaticity.F*School.F +
                                         (1+ Condition.F |Participant) +
                                         (1|ItemNumber),
                                         data=datCloze,
                                         family='bernoulli',
                                         prior = transpriors,
                                         chains=4, iter=5,000, cores=6,
                                         control=list(adapt_delta=0.95)))
```

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

Interaction effects of predictors

The primary focus of this analysis was to examine the interaction effects between different instructional conditions, test points, idiomaticity, and schools. The interaction between **Condition** and **Test** was credible at the posttest, with the schematic feedback group demonstrating the biggest improvement in accuracy from pretest to posttest ($\beta = 0.57$, $SE = 0.18$, 95% CrI [0.22, 0.92], probability ≈ 0.78). Nonetheless, both the traditional feedback and corrective feedback groups also improved substantially, with an estimate of 0.35 ($SE = 0.51$, 95% CrI [0.00, 0.70], probability ≈ 0.63) and 0.51 ($SE = 0.51$, 95% CrI [0.11, 0.91], probability ≈ 0.74) respectively. However, the learning effects gained were not sustainable across the three instructional groups as no evidence was found at the delayed posttest. As for the interaction between **Condition, Test,** and **Idiomaticity,** the traditional feedback group showed the biggest improvement on spatial polysemes at the posttest ($\beta = 0.78$, $SE = 0.27$, 95% CrI [0.25, 1.32], probability ≈ 0.76). Similarly, the corrective feedback group showed substantial learning on spatial polysemes ($\beta = 0.55$, $SE = 0.25$, 95% CrI [0.07, 1.04], probability ≈ 0.73). However, such improvement on spatial polysemes did not sustain over time. The schematic feedback group showed similar learning outcomes for both spatial and nonspatial polysemes at posttest and delayed posttest. Finally, an interaction was found between **Condition, Test,** and **School** across the three instructional groups. The traditional feedback group from the lower-performing school demonstrated credible learning at both the posttest ($\beta = 0.63$, $SE = 0.31$, 95% CrI [0.02, 1.24], probability ≈ 0.64) and the delayed posttest($\beta = 1.71$, $SE = 0.55$, 95% CrI [0.66, 2.82], probability ≈ 0.85) while the schematic condition only showed a potential positive effect at the posttest ($\beta = 0.59$, $SE = 0.30$, 95% CrI [0.02, 1.24], probability ≈ 0.64), although the CrI included zero. Nevertheless, both schematic feedback and corrective feedback groups of the lower-performing school also showed substantial learning at the delayed posttest, with an estimate of 1.98 ($SE = 0.55$, 95% CrI [0.90, 3.08], probability ≈ 0.88) and 1.60 ($SE = 0.58$, 95% CrI [0.47, 2.73], probability ≈ 0.83) respectively. Four ways interaction was not found in the translation data. Overall, the findings from the translation data highlight the effectiveness of different instructional conditions at different test phases, particularly in low-performing schools. The schematic feedback group showed the most substantial learning at the delayed posttest, suggesting the long-lasting benefits of CL-informed instruction. The traditional feedback group also demonstrated consistent positive effects across both posttest and delayed posttest. However, the benefits of the corrective feedback were more pronounced at the delayed posttest than at the posttest. These insights cast light on future instructional strategies to enhance learning outcomes, especially in low-performing educational settings.

The random effects

The standard deviation of the intercepts for items was estimated to be 1.06 ($SE = 0.08$, 95% CrI [0.83, 1.37]), showing substantial variability in baseline responses across items. Similarly, the standard deviation for participant-level intercepts was 0.99 ($SE = 0.09$, 95% CrI [0.79, 1.20]), indicating credible variability in baseline responses across participants. Nevertheless, the correlations among the random effects were generally weak, indicating that individual differences in baseline responses were relatively independent of condition-specific effects. Such independence matters as the weak correlation implies that the instructional conditions have similar effect across participants, regardless of their starting point.

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

Discussion and conclusions

This replication study addressed the first research question regarding the role of cue focusing on preposition learning. Previous eCALL studies have suggested that practice and cue focusing promote L2 grammar learning (Presson, Davy, & MacWhinney, 2013; Presson, MacWhinney, & Tokowicz, 2014), but none of those studies has examined the learning of prepositions. Similar to the original study, we did not find testing effect, and all treatment groups improved substantially upon training with the eCALL tutor. More importantly, with the added delayed posttest, we were able to observe internalization of accurate form–function mappings. As learners were repeatedly exposed to the polysemic contrasts, they were able to differentiate between the target and its highly competing cue.

Through combining the two usage-based frameworks to address the second research question, Wong et al. (2018) demonstrated synergetic effect of CL and CM on L2 preposition learning. Interestingly, our findings aligned with the reanalysis of Wong et al.'s (2018) using Bayesian MEM, in which we observed a superiority effect of the CL-informed instruction in the cloze task at the posttest ($\beta = 1.00$, $SE = 0.22$, 95% CrI [0.57, 1.43], probability ≈ 0.73). While the authors of the original study observed a trend of the schematic feedback group demonstrating the most learning, the difference between conditions was not significant, as $p > .05$. This concurs with the reanalysis using the Bayesian approach as the traditional feedback group and the corrective feedback also demonstrated a comparable probability of ≈ 0.70 and ≈ 0.66 to improve respectively. Nevertheless, this replication study provides further evidence that the superiority effect of CL-informed instruction sustained over time whereas the improvement of the traditional instruction at the posttest did not. We believe that the participants were attuned to the conventional approach, in which rule-of-thumb definitions and examples were provided, prompting them to adopt the memorization strategy. However, without deeper conceptual stimulation, learners might fail to consolidate and internalize the form-meaning mappings in long run (Negueruela, 2003). This might also explain why many L2 learners, even those with advanced proficiency, fail to fully acquire the English preposition system (Tyler, 2012).The translation data of the present study demonstrated a similar trend to that of the original study, in which the schematic feedback group demonstrated the highest probability (≈ 0.66) of making improvement, but it was not substantially different from those receiving traditional (≈ 0.59) or corrective feedback (≈ 0.62).

Unlike the cloze test, the positive instructional effects of the CL-informed instruction did not sustain over time. This finding did not concur with the findings of Wong (2023) where learners who received CL-informed instruction significantly outperformed the corrective feedback group at the delayed posttest. One possible reason for the substantial regression could be age related. The participants were junior secondary students; thus, the CL-inspired pedagogy might take more time to consolidate. Further, unlike sentence-level cloze task, translation tasks are uncommon for junior L2 learners, which might also have posed an additional cognitive challenge. Although we were unable to show sustainable learning effect at the production task as seen in previous research, varying learning outcomes were repeatedly found in different assessment tasks, thereby urging L2 researchers to incorporate multiple task types to better assess instructional effects and students' learning trajectories.

For the learning of spatial and nonspatial polysemes, as in the original study, we observed no interaction effects between condition, time, and idiomaticity in the cloze task. However, with the newly added factor (**School**), the replication study found that

https://doi.org/10.1017/S0272263124000603 Published online by Cambridge University Press

CL-informed instruction benefits the learning of spatial polysemes among low-performing students. As for the translation task, learners from the traditional and corrective feedback groups did substantially better for spatial polysemes, whereas learners from the schematic feedback group showed the poorest performance. This observation is not consistent with previous findings, as the CL-informed instruction often shows better learning. Further, the findings on spatial and nonspatial polysemes did not replicate Wong's (2023) findings where learners obtained higher accuracy for nonspatial polysemes in the receptive task (e.g., grammaticality judgement test). We observed a general trend that the three types of instruction benefited the learning of spatial polysemes, especially among lower-performing learners. Although many might think nonspatial polysemes pose a greater challenge due to their abstractness, spatial polysemes are much more versatile, increasing the difficulty for learners to choose in various contexts. For instance, the prepositions *over*, *on*, *onto* are all valid options for the sentence *James put the blanket __ his bed.* Therefore, when provided with options, learners might find nonspatial polysemes easier to recall than spatial polysemes as they are less abstract (Langacker, 2008). The substantial learning of spatial polysemes might indirectly suggest learners had not received sufficient help previously; therefore, the different instructions, despite variation, were useful in training spatial uses. Overall, the replication study provided further support for incorporating CL concepts into L2 preposition teaching and learning. With the Bayesian perspective, we gain a more comprehensive view of learning trajectories with increased flexibility, particularly on variable identification.

Future replication research

Our replication study marked an effort to apply Bayesian analysis to experimental research within applied CL. Subsequent replication endeavors should consider integrating prior information derived from the original research through the Bayesian lens, contrasting with the frequentist approach that tends to overlook insights from initial studies. In essence, researchers should prioritize the incorporation of informative priors in replication research to optimize the advantages inherent in adopting the Bayesian method. Moreover, posterior predictive checks should be performed for Bayesian analysis, as they are useful in assessing the validity of models, pinpointing any model inadequacy, quantifying lack of fit, and informing model improvement. In addition, when the original dataset is available, a replication study is advised to reanalyze the original data using the Bayesian approach, facilitating direct comparisons with new findings. Crucially, to foster robust replication research, it is essential for future studies to embrace open science principles, e.g., sharing complete sets of pedagogical intervention materials and collected data as well as the source codes for built models. Finally, future replication studies could explore how different proficiencies may lead to different degrees of uptake from various types of instruction (i.e., feedback types).

Acknowledgment. This work was supported by the Hong Kong Standing Committee on Language Education and Research (SCOLAR) grant (EDB(LE)/P&R/EL/203), awarded to the first author, and the Hong Kong PolyU grant 1-YWBW, awarded to the second author. We extend our sincere thanks to the learners who participated in this study and to Michaela Wong for her assistance with recruitment. We also wish to express our gratitude to Kevin McManus for his valuable work as Editor, and to the reviewers for their insightful comments. Finally, we are deeply grateful to Petar Milin for his generous guidance with the Bayesian analysis. We acknowledge that all errors that remain are our own.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263124000603.

Data availability statement. The experiment in this article earned Open Data and Materials badges for transparent practices. The data and materials are available at https://osf.io/fxb89/ and https://app.gorilla.sc/openmaterials/649408 respectively.

References

Achard, M., & Niemeier, S. (2004). Introduction: Cognitive Linguistics, Language Acquisition, and Pedagogy. In *Cognitive Linguistics, Second Language Acquisition, and Foreign Language Teaching* (p. 1–12). Mouton de Gruyter. https://doi.org/10.1515/9783110199857.1

Boers, F., & Lindstromberg, S. (2006). Cognitive linguistic applications in second or foreign language instruction: rationale, proposals, and evaluation. In *Cognitive Linguistics: Current Applications and Future Perspectives* (p. 305–358). Mouton de Gruyter. https://doi.org/10.1515/9783110197761.4.305

Boieblan, M. (2023). Enhancing English spatial prepositions acquisition among Spanish learners of English as L2 through an embodied approach. *International Review of Applied Linguistics in Language Teaching*, *61*, 1391–1420. https//https://doi.org/10.1515/iral-2021-0151

Bovolenta, G. & Williams, J. N. (2023). Implicit learning in production: Productive generalization of new form–meaning connections in the absence of awareness. *Language Learning*, *73*, 723–758.

Brugman, C. (1988). *The story of "over": Polysemy, semantics and the structure of the lexicon.* Garland.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Bürkner, P. C., Scholz, M., & Radev, S. T. (2023). Some models are useful, but how do we know which ones? Towards a unified Bayesian model taxonomy. *Statistics Surveys.* doi:10.1214/23-SS145

Colasacco, M.A. (2019). A cognitive approach to teaching deictic motion verbs to German and Italian students of Spanish. *International Review of Applied Linguistics in Language Teaching*, *57*, 71–95. https://doi.org/10.1515/iral-2018-2007

Divjak, D. & Milin, P. (2023). Using computational cognitive modeling in usage-based linguistics. In M. Díaz-Campos & S. Balasch (Eds.), *The handbook of usage-based linguistics* (pp. 307–324). Wiley Blackwell.

Evans, V., & Tyler, A. (2005). Applying cognitive linguistics to pedagogical grammar: The English prepositions of verticality. *Revista Brasileira De Linguística Aplicada*, *5*, 11–42. https://doi.org/10.1590/S1984-63982005000200002

Garcia, G. D. (2023). Statistical modeling in L3/Ln acquisition. In J. Cabrelli, A. Chaouch-Orozco, J. G. Alonso, S. M. P. Soares, E. Puig-Mayenco, & J. Rothman (Eds.), *The Cambridge handbook of third language acquisition* (pp. 744–770). Cambridge University Press. https://doi.org/10.1017/9781108957823.030

Garcia, G. D. (2021). *Data visualization and analysis in second language research*. Routledge.

Garcia, G. D. (2020). Language transfer and positional bias in English stress. *Second Language Research*, *36*, 445–74. https://doi.org/10.1177/0267658319882457

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge.

Grady, J. (1997). *Foundations of meaning: Primary metaphors and primary scenes*. (Unpublished doctoral dissertation). University of California at Berkeley.

Hampe, B. (2005). Image schemas in Cognitive Linguistics: Introduction. In *From Perception to Meaning* (p. 1–14). Mouton de Gruyter. https://doi.org/10.1515/9783110197532.0.1

Herskovits, A. (1986). *Language and spatial cognition: An interdisciplinary study of the prepositions in English*. Cambridge University Press.

Herskovits, A. (1988). Spatial expressions and the plasticity of meaning. In B. Rudzka-Ostyn (Ed.), *Topics in cognitive linguistics* (pp. 271–298). John Benjamins. https://doi.org/10.1075/cilt.50.11her

Hajazo-Gascón, A. & Llopis-García, R. (2019) Applied cognitive linguistics and foreign language learning. Introduction to the special issues. *International Review of Applied Linguistics in Language Teaching*, *57*, 1–20. https://doi.org/10.1515/iral-2018-2004

Holme, R. (2009). *Cognitive Linguistics and Language Teaching*. Palgrave Macmillan.

Hwang, H. (2023). Image-schema-based-instruction enhanced L2 construction learning with the optimal balance between attention to form and meaning. *International Review of Applied Linguistics in Language Teaching.* https://doi.org/10.1515/iral-2023-0015

Jacobsen, N. D. (2018). The best of both worlds: Combining cognitive linguistics and pedagogic tasks to teach English conditionals. *Applied Linguistics*, *39*, 668–693.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Johnson, M. H. (2005). *Developmental Cognitive Neuroscience*. (2nd Ed) Oxford: Blackwell.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Sceince*, *1*, 270–280.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago.

Lam, Y. (2009). Applying cognitive linguistics to teaching the Spanish prepositions por and para. *Language Awareness*, 18(1), 2–*18*. https://doi.org/10.1080/09658410802147345

Langacker, R.W. (2008). *Cognitive Grammar: A Basic Introduction*. OxfordUniversity Press.

MacWhinney, B. (1997). Second language acquisition and the Competition Model. In A. M. B. de Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 113–142). Lawrence Erlbaum Associates Publishers.

MacWhinney, B. (2012). The logic of the unified model. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 211–227). Routledge. https://doi.org/10.4324/9780203808184.ch13

Mandler, J. M. (1992). How to build a baby: II. *Conceptual primitives. Psychological Review*, *99*(4), 587–604. https://doi.org/10.1037/0033-295x.99.4.587

McElreath, R. (2015). *Statistical rethinking. A Bayesian course with examples in R and Stan*. Chapman Hall.

Mifka-Profozic, N., Behney, J., Gass, S.M., Macis, M., Chiuchiù, & Bovolenta, G. (2023). Effects of Form-Focused Practice and Feedback: A Multisite Replication Study of Yang and Lyster (2010). *Language Learning*, *73*, 1164–1210. https://doi.org/10.1111/lang.12623

Negueruela, E. (2003). Systemic-theoretical instruction and L2 development. UnpublishedPhD dissertation, Pennsylvania State University. State College.

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, *68*, 1032–1075. https://doi.org/10.1111/lang.12310

Norris, J.M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65* (Suppl. 1), 97–126.

Peters, E., Puimège, E., & Szudarski, P. (2023). Repetition and Incidental Learning of Multiword Units: A Conceptual Multisite Replication Study of Webb, Newton, and Chang (2013). *Language Learning*, *73*(4), 1211–1251. Portico. https://doi.org/10.1111/lang.12621

Plonsky, L. (2015). Statistical power, p values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23–45). Routledge.

Porte, G., & McManus, K. (2019). *Doing replication research in applied linguistics*. Routledge.

Prange, J., & Wong, M. H. I. (2023). Reanalyzing L2 preposition learning with Bayesian mixed effects and a pretrained language model. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12722–12736). Association for Computational Linguistics. 10.18653/v1/2023.acl-long.712

Presson, N., Davy, C., & MacWhinney, B. (2013). Experimentalized CALL for adult second language learners. In J. Schwieter (Ed.), *Innovative research and practices in second language acquisition and bilingualism* (pp. 139–164). John Benjamins. https://doi.org/10.1075/lllt.38.10pre

Presson, N., MacWhinney, B., & Tokowicz, N. (2014). Learning grammatical gender: The use of rules by novice learners. *Applied Psycholinguistics*, *35*, 709–737. https://doi.org/10.1017/S0142716412000550

Puimege, E., Perez, M. M., & Peters, E. (2023). The effects of typographic enhancement on L2 collocation processing and learning from reading: An eye-tracking study. *Applied Linguistics*, *45*, 88–110. https://doi.org/10.1093/applin/amad003

Qin, J., Wu, Z. & Zhong, S. (2023). When concept-based language instruction meets cognitive linguistics: teaching English phrasal verbs with *up* and *out*. *International Review of Applied Linguistics in Language Teaching*, *61*, 1455–1480. https://doi.org/10.1515/iral-2021-0164

Robinson, P., & Ellis, N. C. (Éds.). (2008). Handbook of Cognitive Linguistics and Second Language Acquisition. Routledge. https://doi.org/10.4324/9780203938560

Rudzka-Ostyn, B. (2003). *Word Power: Phrasal Verbs and Compounds: A Cognitive Approach*. Mouton de Gruyter. https://doi.org/10.1515/9783110197235

Talmy, L. (2005). The fundamental system of spatial schemas in language. In B. Hampe (ed.), *From perception to meaning: Image schemas in cognitive linguistics*, pp. 199234. Mouton de Gruyter.

Tyler, A. (2012). *Cognitive linguistics and second language learning: Theoretical basics and experimental evidence*. Routledge.

Tyler, A., & Evans, V. (2003). *The semantics of English prepositions: Spatial scenes, embodied meanings and cognition*. Cambridge University Press. https://doi.org/10.1017/cbo9780511486517

Tyler, A., Huang, L., & Jan, H. (Eds.). (2018). *What is applied cognitive linguistics: Answers from current SLA research* [Kindle version]. Mouton de Gruyter.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Verspoor, M., & Lowie, W. (2003). Making Sense of Polysemous Words. *Language Learning*, *53*(3), 547–586. Portico. https://doi.org/10.1111/1467-9922.00234

Wirtz, M. A., & Pfenninger, S. E. (2023). Variability and individual differences in L2 sociolinguistic evaluations: The GROUP, the INDIVIDUAL, and the HOMOGENOUS ENSEMBLE. *Studies in Second Language Acquisition*, *45*, 1186–209. doi:10.1017/S0272263123000177

Wong, M. H. I. (2023). Fostering conceptual understanding through computer-based animated schematic diagrams and cue contrast. *TESOL Quarterly*. https://doi.org/10.1002/tesq.3195

Wong, M. H. I., Zhao, H., & MacWhinney, B. (2018). A cognitive linguistics application for second language pedagogy: The English preposition tutor. *Language Learning*, *68*, 438–468. https://doi.org/10.1111/lang.12278

Zhao, H., & MacWhinney B. (2018). the instructed learning of form-function mappings in the English article system. *The Modern Language Journal*, *102*, 99–119. https://doi.org/10.1111/modl.12449

Zhao, H., Huang, S., Zhou, Y., & Wang, R. (2020). SCHEMATIC DIAGRAMS IN SECOND LANGUAGE LEARNING OF ENGLISH PREPOSITIONS: A BEHAVIORAL AND EVENT-RELATED POTENTIAL STUDY. *Studies in Second Language Acquisition*, *42*(4), 721–748. https://doi.org/10.1017/s027226311900069x

Cite this article: Wong, M. H. I., & Prange, J. (2024). A Bayesian approach to (re)examining learning effects of cognitive linguistics–inspired instruction: A close replication of Wong, Zhao, and MacWhinney (2018). *Studies in Second Language Acquisition,* 46: 1493–1513. https://doi.org/10.1017/S0272263124000603