Studies in Second Language Acquisition (2025), 1–36
doi:10.1017/S0272263125100867

METHODS FORUM

Modeling relationships between learning conditions, processes, and outcomes: An introduction to mediation analysis in SLA research

Ruirui Jia and Bronson Hui

University of Maryland, College Park, MD, USA
Corresponding author: Ruirui Jia; Email: rjia@umd.edu

(Received 11 September 2024; Revised 03 April 2025; Accepted 15 April 2025)

Abstract

In the past decade, researchers have been increasingly interested in understanding the process of language learning, in addition to the effect of instructional interventions on L2 performance gains (i.e., learning products). One goal of such investigations is to reveal the interplay between learning conditions, processes, and outcomes where, for example, certain conditions can promote attention to the learning targets, which in turn facilitates learning. However, the statistical modeling approach taken often does not align with the conceptualization of the complex relationships between these variables. Thus, in this paper, we introduce mediation analysis to SLA research. We offer a step-by-step, contextualized tutorial on the practical application of mediation analysis in three different research scenarios, each addressing a different research design using either simulated or open-source datasets. Our overall goal is to promote the use of statistical techniques that are consistent with the theorization of language learning processes as mediators.

Keywords: Mediation analysis; Direct effects; Indirect effects; Total effects; Learning conditions, processes, and outcomes

Introduction

In second language acquisition (SLA) research, instructed SLA in particular, researchers are often concerned about the effect of different instructional interventions on second language (L2) performance. These investigations encompass numerous topics and subareas within SLA, ranging from the impact of reading-while-listening on reading comprehension (e.g., Pellicer-Sánchez et al., 2020), to the effect of spacing on L2 collocation acquisition (e.g., Yamagata et al., 2022), and to the effect of different types of corrective feedback on L2 speech perception (e.g., Lee & Lyster, 2016), to name a few. However, many early instructional studies focused exclusively on L2 learning

outcomes (e.g., Leow, 2015; Godfroid, 2019). Exploring only the products but not the processes of learning could thus veil important questions in understanding L2 acquisition, such as what cognitive processes are involved in and conducive to L2 learning and how teachers and curriculum designers can promote these processes to better facilitate acquisition (Leow, 2015). Given that learning processes and products are indispensable to each other, investigating both can reveal the complex cognitive mechanisms underlying language development and offer more nuanced insights into language learning. Thus, bridging the connection between learning conditions, processes, and outcomes through understanding the causal relationships between these variables becomes essential in the study of language learning. It is in this context that we introduce mediation analysis to the field. This statistical method allows researchers to align their data modeling approach with the conceptualization of the relationships between learning conditions, processes, and outcomes, extending and refining the current practice of data analysis in SLA.

Measuring and linking learning processes and outcomes

In the past decade, studies that tap into both cognitive processes and language learning outcomes have been burgeoning, giving rise to what is called online methods. These methods range from online verbal reports in the form of, for example, think-aloud and stimulated recall (e.g., Leow, 2015), to the use of eye tracking (e.g., Godfroid, 2019), and the scrutiny of logs in computer-assisted language learning systems (e.g., Hui et al., 2023). In one of the early studies using verbal reports to capture L2 learners' noticing of the linguistic constructs in the input, Leow (1997) employed think-aloud protocols to elicit L2 Spanish learners' thought processes during a problem-solving task. Interested in the relationship between L2 learners' awareness on the target irregular morphological forms and the knowledge they further processed and eventually acquired, the researcher reported that learners who demonstrated a higher level of awareness on the target forms performed better on a subsequent recognition task, unveiling the important role of awareness in promoting L2 learning. As an important method to understand learning processes, stimulated recall and think-aloud have been widely used to tap into learners' noticing (e.g., Zalbidea, 2021), awareness (e.g., Woll, 2018), learning strategies (e.g., Gokturk & Chukharev-Hudilainen, 2023), and depth of processing (e.g., Leow et al., 2022), unpacking the link between learners' performance and what they received and processed during learning (Leow, 2015).

Another prominent method to capture learning processes is the use of eye tracking to reveal learners' attention and processing through gaze patterns (e.g., Conklin et al., 2018; Godfroid, 2020). Drawing on the eye-mind link (Rayner et al., 2012), researchers can analyze eye-tracking data gathered from different experimental conditions to reveal the relationship between learning conditions and the underlying cognitive processes. For example, Puimège et al. (2023) found that textual enhancement (i.e., underlining multiword units) could induce more visual attention to the target items. With longer reading time and less word skipping on the enhanced multiword units, learners were more likely to recall the word forms in the immediate posttest. Similarly, in Kang et al. (2022), participants were found to demonstrate different reading behavior when provided with different types of glosses (i.e., L1 and L2 glosses), with much longer time spent on L2 glosses than on L1 glosses. The relationship between lexical uptake and the time spent on processing the target words was also found to be different as a function of gloss types. Learners in the L2 gloss and no-gloss conditions demonstrated a significant, positive relationship, whereas learners in the L1 gloss condition did not. The

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

positive relationship suggests that learners who spent more time inferring the meaning from the context benefited more than those who spent less time on it, highlighting the important role of attention in L2 learning (Kang et al., 2022). As we can see, with eye-tracking measures, researchers can understand how learners direct their attention to different instructional manipulations and how such attention might or might not facilitate the encoding and retrieval of linguistic items.

Learning processes are by no means limited to cognitive processes. In computer-assisted language learning (CALL) research, system logs can also reveal how L2 learners engaged in learning under different treatment conditions delivered through digital platforms (i.e., the learning process) and its relationship to language learning outcomes (e.g., Hui et al., 2023; He & Loewen, 2022; Hwang et al., 2024; Révész et al., 2017). Using an interactive digital platform engineered to provide both general and specific corrective feedback on grammar exercises, Hui et al. (2023) extracted 19 learning process variables from the system logs (e.g., total time on task, task fields attempted, correct attempt, first correct) and reported that learning gains were significantly predicted by accuracy-focus (e.g., correctness in exercises) and finish time. Moreover, based on learners' learning process patterns, the researchers also revealed that the specific corrective feedback was beneficial only for people who demonstrated specific process patterns, namely those who submitted their work late. By illustrating the potential of these interaction logs, the authors encouraged researchers to unveil the black box of learning processes to understand how learning unfolds in the context of CALL.

In addition to cognitive processes and behavior, learning processes also include conative (e.g., motivation) and affective (e.g., enjoyment, boredom) variables. Advances in individual differences (ID) research in SLA have motivated researchers to develop and validate scales to index learner-internal factors (e.g., L2 grit: Teimouri et al., 2022; foreign language classroom anxiety: Botes et al., 2022). For example, in Li and Lu (2024), the researchers investigated not only the effect of task complexity on learners' L2 writing performance but also how task complexity influenced learners' cognitive–affective ID variables (i.e., working memory, trait enjoyment, task enjoyment, and task motivation) and the effect of these ID factors on their writing performance. With cognitive, motivational, and emotional ID variables functioning as part of the learning processes, the study highlighted the important role of task motivation in influencing the relationship between task complexity and L2 writing performance, finding that task complexity significantly enhanced task motivation and enjoyment with task motivation significantly predicting L2 writing performance consistently in both versions of the task. With ID measures, researchers are able to investigate how L2 learners, for example, enjoy instruction more or less as a function of cognitive demands in a speaking task (e.g., Chen, 2023) and unveil the relationships between learner emotion (anxiety, boredom, and enjoyment), engagement, and proficiency (e.g., Tsang & Dewaele, 2023). This is certainly a promising avenue to investigate how certain task features can promote favorable learning processes experienced by learners, leading to better acquisition outcomes.

With various methodological tools, such as eye tracking and validated scales, researchers are in a much better position to bridge the connection between learning conditions, processes, and outcomes. At the same time, researchers have often missed the opportunity to gain comprehensive insights into these relationships because many studies outline separate research questions to shed light on various parts of the bigger picture. Although it is not clear the extent to which the formulation of research questions has been influenced by modeling options accessible to the researchers, it is rather common in our field to separately address the independent associations between

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

(a) learning conditions and processes (e.g., condition predicting eye gaze, as an index of attention and processing), (b) learning processes and learning outcomes (e.g., eye gaze predicting offline measures of learning), and (c) learning conditions and outcomes. While these insights are very informative, researchers seldom focus on a system of relationships as a whole. In other words, current modeling practices often fail to take a holistic approach to capture the nuanced and comprehensive relationship between treatment, processes, and outcomes. Therefore, the analysis performed does not always align well with the theoretical conceptualization of these relationships. For instance, one might conceptualize that certain experimental conditions can promote attention to the target items, which in turn enhances learning. In this light, learning conditions can have not only a direct impact on learning but also an indirect influence on learning through promoting attention (see more discussion below). When research questions focus on isolated relationships, the broader learning mechanisms remain unclear. Thus, to capture learning conditions, processes, and outcomes as a whole, researchers need to take a holistic approach to identify the specific pathways through which interventions facilitate or hinder learning. Thus, conducting mediation analysis is important to capture this broader perspective. To lay the foundation for demonstrating the application of mediation analysis to the condition-process-outcome relationships, we first provide a brief introduction to mediation in the following section.

Mediation (which is not moderation)

Put simply, mediation involves a predictor impacting an outcome *through* a mediator. It is seldom covered by mainstream statistical textbooks in our field (e.g., Garcia, 2021; Larson-Hall, 2016; Loertes et al., 2020). There might also be confusion around this kind of relationship with what is known as moderation (or interaction) effects. A moderation analysis addresses, for example, how a treatment effect is moderated by a moderator (e.g., L1 background). It means that the treatment effect, namely its magnitude (or size) and/or direction (positive or negative), depends on or is based on, for example, the learner's native language. A significant interaction term, which mathematically is the product of the predictor and the moderator, often represents statistical evidence of such differential effects (Field, 2018). Depending on the nature of the interaction, researchers might find, for example, that treatment is generally helpful, but especially so for L1 Chinese learners, compared with Spanish learners (differs in magnitude, see Figure 1a). Another scenario could be that treatment is only helpful for some, but detrimental to others (differs in direction, see Figure 1b). All that is to say: a moderator variable changes (i.e., moderates) the relationship between a predictor and an outcome variable. In Figure 1c, we present how moderation is often visualized graphically in statistical texts.

Unlike moderators serving only on the predictor side (see Figure 1c above), a mediator is part of a causal sequence between the two variables, functioning as both a predictor and an outcome in the model (Baron & Kenny, 1986). Mediation represents the intermediary process where the effect of a predictor on the outcome variable is mediated by an intervening variable, known as the mediator (Field, 2018). Affecting the relationship between the independent and dependent variable, a mediator becomes part of the mechanism governing the causal relations (Field, 2018).

For better illustration, one can visualize the mediation relationships in a path diagram (see Figure 2; cf. Figure 1c for a moderation effect). In this diagram, X refers to the predictor variable (e.g., a learning condition variable), Y represents the outcome

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

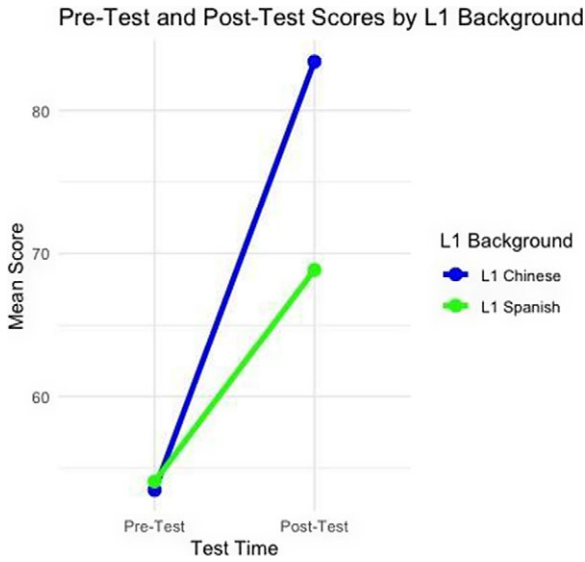


Figure 1a. Moderation example 1.

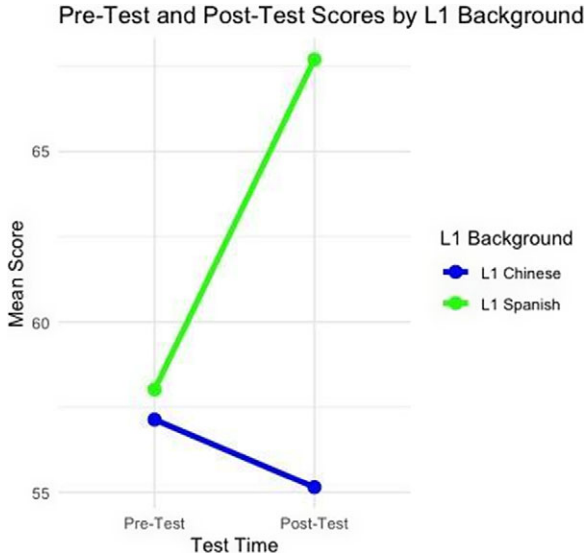


Figure 1b. Moderation example 2.

variable (e.g., a learning outcome variable), and M stands for the mediator (e.g., a learning process variable, as indexed, for example, by eye tracking). As can be seen, the path diagram is constituted by three causal relationships annotated as paths *a*, *b*, and *c'*, representing the causal relationship from X to M, from M to Y, and from X to Y, respectively. In this light, researchers can observe not only the individual effects indexed by the *a*, *b*, and *c'* paths, such as the effect of reading conditions on attention (the *a* path), the effect of attention on reading comprehension (the *b* path), and the

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

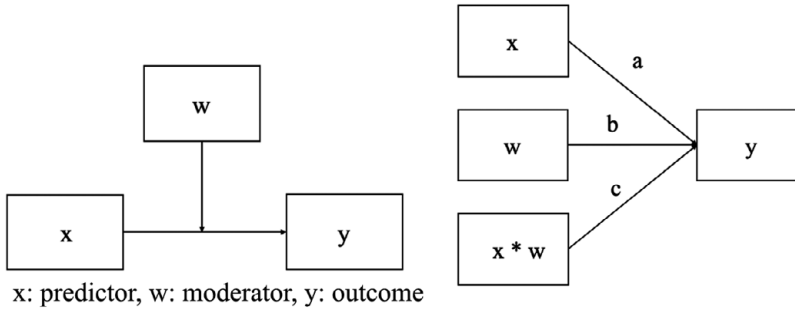


Figure 1c. Moderation diagrams (Field, 2018).

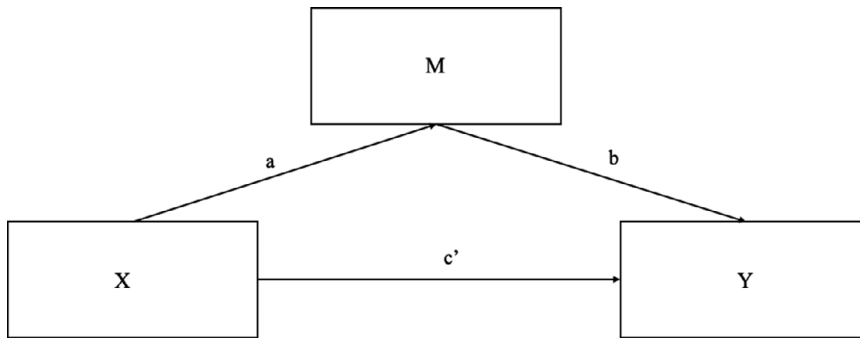


Figure 2. The mediation model (Field, 2018).

effect of reading conditions on reading comprehension while accounting for attention (the *c'* path), but also some broader relationships: first, the indirect effect of X on Y via M (e.g., the effect of reading conditions on reading comprehension via attention to pictures), calculated as the product of *a* and *b* paths: $a * b$, and second, the total effects of X on Y (e.g., the total effect of reading conditions on reading comprehension), determined by summing the indirect and direct effects: $a * b + c'$. For indirect effects (i.e., the effect of X on Y is mediated by M), two scenarios are possible: when the *c'* path is zero, there is complete mediation, meaning that the effect of X on Y is entirely mediated by the mediator M (Vuorre & Bolger, 2018). Conversely, if the *c'* path is not zero, the model indicates partial mediation where the effect of X on Y is influenced not only by the direct and indirect effects but also by other variables not considered in the model (Vuorre & Bolger, 2018). In this light, researchers are often interested in the extent of mediation (Shrout & Bolger, 2002). This magnitude helps enhance our understanding of how important the mediator is in the causal pathway from the predictor to the outcome variable. Using the eye-tracking example from above, the exact role of attention in this mediation relationship often reveals important information regarding the relationship between learning conditions and reading comprehension.

Mediation in SLA research

In SLA, mediation analysis is not yet very common, but not unprecedented, except in areas that call for the use of scales (e.g., measuring task motivation and enjoyment) with structural equation modeling (SEM). Technically, mediation analysis can be considered

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

a case of path analysis within the SEM framework. For example, exploring the mediation effect of individual variables, Sparks and Alamer (2022) have found that L2 aptitude and L2 achievement are significant mediators influencing the effect of L1 achievement on L2 anxiety. Moreover, motivation was found to not only influence L2 proficiency directly but also indirectly through self-confidence (Alrabai, 2022). As with many other studies (e.g., Li et al., 2022; Öztürk, 2023), these findings highlight the application of mediation analysis in uncovering complex relationships in individual differences research.

However, mediation analysis has seldom been seen in other SLA studies that have a cognitive focus. One exception is a study by Koval (2019) who was interested in the effect of spacing on attentional processing (i.e., the *a* path) and intentional vocabulary learning (i.e., the *c'* path) as well as how learners' attention mediates the spacing effect (i.e., $a * b$). Using the eye-tracking technique, the study revealed that learners' attention, captured by learners' total reading times, was a significant mediator of the spacing effect. Also, in another study, Hui and Godfroid (2021) explored how processing speed and automaticity affected L2 listening comprehension. Propositional comprehension was found to be a significant mediator in accounting for the effect of lexical processing on listening comprehension (i.e., $a * b$), which highlights the crucial role of lexical processing in facilitating L2 listening comprehension via comprehending propositional meaning.

So far, we have seen some examples of mediation in the literature that involve either an observational (e.g., Hui & Godfroid, 2021) or a between-participant experimental design (e.g., Koval, 2019). This is partly because these designs are common in SLA and social sciences in general. At the same time, a large portion of the literature has a within-participant design, especially in experimental contexts. This is because, oftentimes, researchers use within-participant designs to minimize confounding due to their inability to randomize participants. Such a design involves the same participant experiencing different learning or experimental conditions manipulated by the researchers, often in a counterbalanced manner. In that sense, learners are compared against themselves in the baseline condition, allowing the researchers to reduce any systematic bias introduced by the nonrandom assignment of conditions (Cook & Campbell, 1979). In addition, researchers have also been focusing on more fine-grained, item- or trial-level data. For example, looks at interest areas in eye tracking are often measured across different target items in different trials and often on different screens. In other words, each participant has multiple observations in the same experiment that are dependent on each other (because they are elicited from the same person). When analyzing such data, researchers typically adopt mixed-effects models to account for the nested data structure resulting from repeated measures (e.g., Linck & Cunnings, 2015). For example, in the study conducted by Tytko et al. (2024), the researchers investigated the impact of multimodal input on L2 reading comprehension using a within-participant design. This approach allowed every participant to experience all the reading conditions during the experiment, which included reading-only (RO), reading with image (RI), and reading with image and audio (RIA), presented in a counterbalanced order (see Figure 3 for the experiment flow chart). This means that six different lists were created, with participants assigned to different lists that specified the sequence of the three treatment conditions. Within each condition, each participant also encountered multiple areas of interest, and each area of interest was also read by multiple participants. Thus, building mixed-effects models with crossed random effects (i.e., by-participant and by-item random intercepts and slopes) is necessary to simultaneously account for the variability associated with both participants and items.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

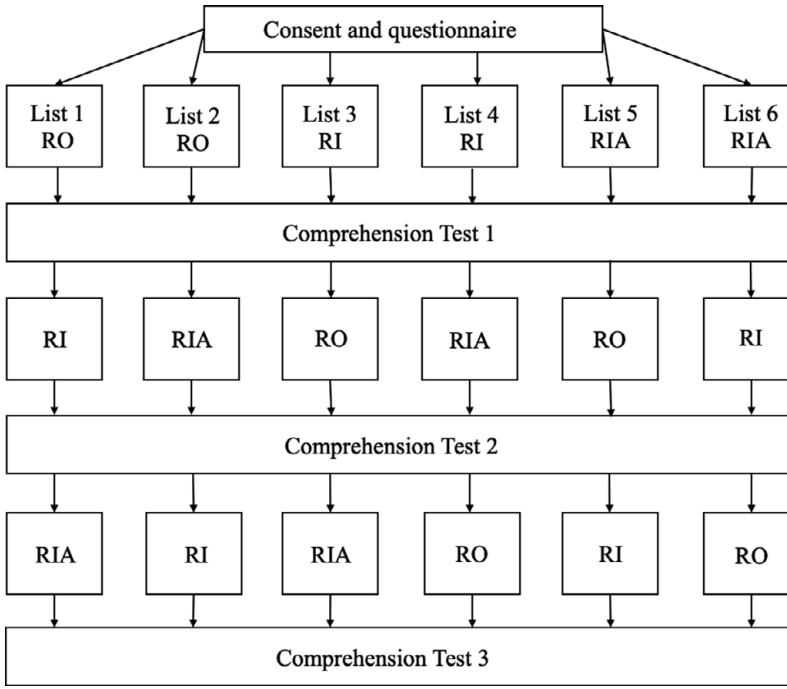


Figure 3. The experiment flow chart of Tytko et al. (2024).

Traditionally, mediation analysis has been discussed and applied less in studies with a within-participant design. This is because the more commonly employed mediation analysis, like ordinary regression analysis, has an assumption of independence, meaning that each observation should be independent of each other. Adding another layer of complexity in exploring mediation relationships is the need to account for the dependency in the trial-level data, which is typically achieved through incorporating both by-participant and by-item random effects. However, recent advances in quantitative methods have lifted the barriers for SLA researchers to model mediation with both between- and within-participant designs and with trial-level data. This would allow us to move beyond exploring the surface-level relationships to identifying the processes through which learning conditions influence learning outcomes. It is within this context that we offer a contextualized tutorial on mediation analysis in various types of research designs commonly seen in SLA research.

The tutorial

To demonstrate the application of mediation analysis in SLA, we present three working examples, each addressing a different study design: a between-participant design, a within-participant design, and a within-participant design with trial-level data. These examples highlight the need to adopt tailored statistical approaches to account for different data structures derived from these designs. For each example, we created either a hypothetical research scenario inspired by previous studies or adopted existing research to contextualize the analyses with either simulated data or open datasets

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

shared by the researchers. We acknowledge that longitudinal and mixed designs are also common in the field, especially in quasi-experimental ISLA research. However, the analysis can become very complex when mediation is involved, and dealing with this complexity falls beyond the introductory scope of the present tutorial.

Example 1: Mediation in between-participant designs

First, we introduce mediation in the context of a between-participant design. We based our research context on Rosa and Leow (2004), who explored how different feedback conditions influenced L2 learners' awareness levels and how learners' level of awareness affected their ability to generalize the target grammatical structure, Spanish contrary-to-fact conditional, to new exemplars. In Rosa and Leow (2004), participants were randomly assigned to one of the five learning conditions: explicit pretask + explicit feedback (EPEFE), explicit pretask + implicit feedback (EPIFE), explicit feedback (EFE), explicit pretask (EP), implicit feedback (IFE), or a control condition. Awareness was assessed using a think-aloud protocol during the treatment and a written questionnaire after the treatment. Based on their responses, participants were categorized into three awareness levels: No Report (NR), Noticing (N), and Understanding (U). Although the researchers collapsed the three levels into two for data analysis: Report (R) and No Report (NR), we kept the three levels in the original theorization in the present tutorial, making awareness an ordinal variable. Learning outcomes were measured through a multiple-choice recognition test and a controlled production test. For both tests, participants received 1 point for each correct selection/use of the target grammatical structure, with a maximum score of 10 points for both the old items and new exemplars. For demonstration purposes, we limited the assessment measure to the controlled production test of new items and focused only on three instructional conditions: explicit feedback (EFE), implicit feedback (IFE), and a control condition. Thus, a research question that can be addressed is "*To what extent does L2 learners' awareness mediate the effect of different feedback types on L2 grammar acquisition?*" Figure 4 visualizes the mediation model of this research scenario.

The simulated dataset

In this example, we used simulated data for demonstration purposes. Sixty participants were simulated for each instructional condition. Each participant, having been exposed

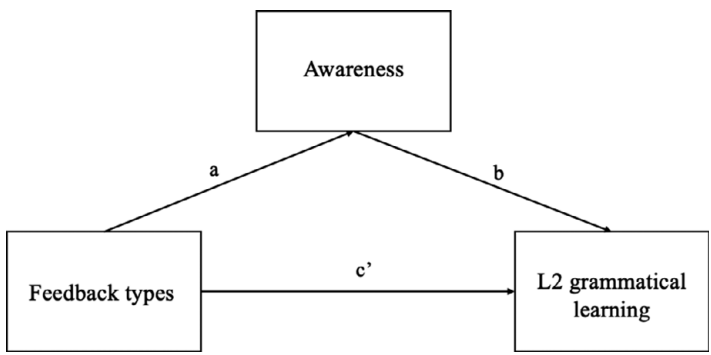


Figure 4. Visualization of the mediation model of Example 1.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

Table 1. Sample dataset for Example 1

participant_id	feedback_type	awareness	test_score
1	3	3	10
2	3	2	10
3	3	3	10
…	…	…	…
61	2	3	10
62	2	2	9
63	2	2	8
…	…	…	…
121	1	3	7
122	1	2	6
123	1	2	8

Note: For feedback types, the control condition was coded as 1, implicit feedback as 2, and explicit feedback as 3; For awareness level, NR was coded as 1, N as 2, and U as 3.

to only one instructional condition, is represented by a single row in the dataset, which records their assigned feedback condition (1 = Control, 2 = Implicit feedback, and 3 = Explicit feedback), awareness level (1 = NR, 2 = N, and 3 = U), and production test score. Table 1 presents the first several rows of the data in a long format. We have shared the R code for all the simulations and analyses conducted in this manuscript on the Open Science Framework (OSF) (https://osf.io/fzps6).

Mediation analysis and results

Firstly, we presented the descriptive statistics of the test scores and awareness by feedback types in Table 2 and visualized them in Figure 5. To conduct the between-group mediation analysis, we began by specifying separate models for the mediator and the outcome. In the outcome model, we applied a linear regression model using the lm() function, for which the dependent variable was "test_score" with "feedback_type" and "awareness" as categorical predictors (the *b* and *c'* paths in Figure 4). For the mediator model, "awareness" is treated as the ordinal outcome variable with "feedback_type" being the predictor. Because "awareness" is an ordered variable with three levels, we adopted the polr() function from the *MASS* package (Venables & Ripley, 2002) to build an ordinal regression model. The syntax of the polr() function is similar to that of the lm() function, with the addition to specify "Hess = TRUE" to estimate standard errors that are necessary to output inferential statistics in this function.

```
# Fit the outcome model
model1 <- lm(test_score ~ feedback_type + awareness, data = simulated_data)

# Fit the mediator model
model2 <- polr(awareness ~ feedback_type, data = simulated_data, Hess=TRUE)
```

To conduct mediation analysis, we combined the two regression models specified above using the mediate() function from the *mediation* package (Tingley et al., 2014). This function outputs a very simple and straightforward summary regarding the indirect effect (ACME), direct effect (ADE), total effect, and the proportion mediated.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

Table 2a. Descriptive statistics of test scores by feedback types

Feedback_Type	M (SD)	[95% CI]
Explicit	8.80 (1.29)	[8.46, 9.13]
Implicit	7.30 (1.71)	[6.86, 7.74]
Control	5.97 (1.90)	[5.48, 6.46]

Table 2b. Descriptive statistics of awareness by feedback types

Feedback_Type	Awareness	Count
Control	NR	13
Control	N	31
Control	U	16
Implicit	NR	9
Implicit	N	26
Implicit	U	25
Explicit	NR	4
Explicit	N	13
Explicit	U	43

Note: NR: No Report; N: Noticing; U: Understanding

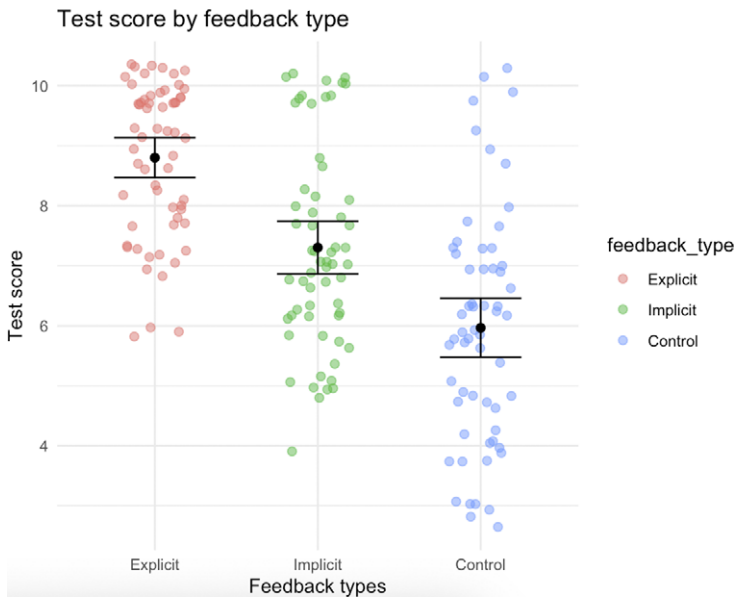


Figure 5a. Visualization of test scores by feedback types.

To build the model in the mediate() function, we first input the previously specified mediator and outcome models and then defined the predictor and the mediator after the arguments "treat" and "mediator," respectively.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

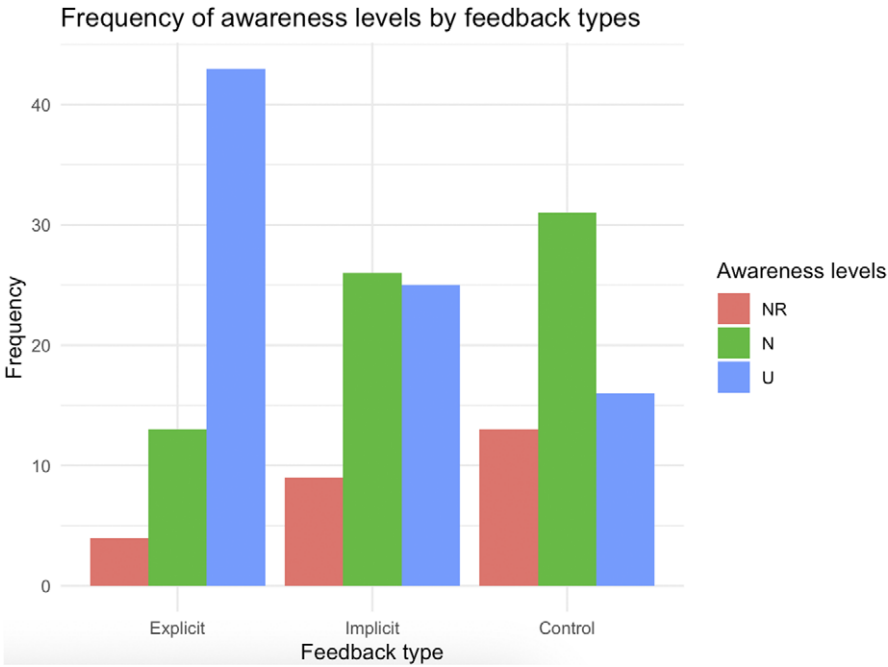


Figure 5b. Visualization of awareness by feedback types.
Note: NR: No Report; N: Noticing; U: Understanding

```r
# Fit the outcome model
model1 <- lm(test_score ~ feedback_type + awareness, data = simulated_data)

# Fit the mediator model
model2 <- polr(awareness ~ feedback_type, data = simulated_data, Hess=TRUE)

# Conduct the mediation analysis
mediation_result <- mediate(model2, model1, treat = "feedback_type",
                            mediator = "awareness",
                            control.value = 2,
                            treat.value = 3,
                            boot =TRUE,
                            sims = 1000)

summary(mediation_result)
```

Using the "control.value" and "treat.value" arguments, we further identified the two instructional conditions to contrast. Note that the mediate() function can only handle two treatment conditions at a time (Tingley et al., 2014). Thus, when there are more than two conditions, as in the present case, these parameters allow for a flexible

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

Table 3. Summary of the mediation model (Example 1)

	Estimate	2.5% CI	97.5% CI	*p*-value
ACME (a * b)	.222	.048	.490	.006 **
ADE (c')	1.260	.740	1.830	< .001 ***
Total Effect (a * b + c')	1.482	.975	2.060	< .001 ***
Prop. Mediated	.150	.033	.330	.006 **

Note: ACME refers to the indirect effect, and ADE refers to the direct effect. The parentheses were added for better interpretation. They are not part of the output.

specification by selecting any two conditions for comparison (Tingley et al., 2014). Here, we indicated "2" and "3" to represent the implicit and explicit feedback conditions, respectively. Researchers can choose the input values to make their hypothesized comparisons. Also, we need to set "boot = TRUE" to enable bootstrapping to calculate confidence intervals. Here, we used the default argument "sims = 1000" to indicate 1,000 resampling for bootstrapping (Tingley et al., 2014).

To output the mediation results, we used the summary() function. Table 3 displays the summary of the mediation model. The analysis revealed a significant indirect effect of feedback types on test scores via awareness (ACME: average causal mediation effect), $b = .222$, 95% CI [.048, .490], $p = .006$. This suggests that awareness played an important role in mediating the effect of feedback type on L2 grammar learning. In addition, with an estimate of 1.260 (95% CI [.740, 1.830], $p < .001$), the direct effect (ADE: average direct effect) of feedback types on test scores was also found significant. This indicates that explicit feedback led to greater learning gains than implicit feedback, independent of awareness. The total effect of feedback types on test scores ($b = 1.482$, 95% CI [.975, 2.060], $p < .001$) further emphasizes the importance of feedback types and awareness as two significant predictors of L2 grammar learning. Lastly, the proportion of the total effect mediated by awareness was .150 (95% CI [.033, .330], $p = .006$), meaning that approximately 15.0% of the total effect of feedback types on L2 grammar learning was mediated by awareness. The result highlights the role of awareness in promoting L2 grammar learning. At the same time, this role was not as critical as one might think because the proportion mediated was a relatively small number, which suggests that the current understanding of why and how feedback conditions have an effect on learning is far from a complete picture. At this point, we must remind readers that these results were based on our simulation for methodological demonstration purposes and remain silent on the substantive literature.

To summarize, when conducting mediation analysis in between-participant designs, we first specified the mediator and the outcome models separately, before testing the mediation effects using the mediate() function. The mediate() function can accommodate not only the ordinal data as in the present example, but also many other types, including continuous, binary, and quantile outcomes (Tingley et al., 2014). Researchers can apply different mediator and outcome models (e.g., lm(), glm(), polr()) depending on the data type they are working with (Tingley et al., 2014). Moreover, as noted earlier, the mediate() function can only handle two treatment conditions at a time, so researchers need to manually specify which two conditions they are interested in comparing (Tingley et al., 2014). As a technical note, it is important to convert different level names into numbers for the mediate() function to run properly. Lastly, although the mediate() function works perfectly for regression models of various kinds, it becomes limited when applied to mixed-effects models. Specifically, it cannot accommodate some well-established and widely used mixed-effects modeling packages such

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

as *nlme* (Pinheiro & Bates, 2024) and *glmmTMB* (Brooks et al., 2017). Moreover, it falls short when complex data structures are involved, such as residual correlation and cross-random structures. Thus, in the following sections, we demonstrate how mediation analysis can be applied to more complex data structures.

Example 2: Mediation in within-participant designs

In this example, we demonstrated mediation analysis for a hypothetical study with a within-participant design (i.e., each participant experienced all treatment conditions in a counterbalanced fashion). This hypothetical study was conceptualized within the task-based language teaching (TBLT) literature. More specifically, we showcased the modeling of the relationships between task complexity, cognitive load, and linguistic complexity. The mediation relationships were visualized in Figure 6.

The conceptualization of this mediation model was supported by empirical and/or theoretical evidence. The *a* path, namely the effect of task complexity on cognitive load, has been investigated by studies such as Xu et al. (2022), Lee (2019), and Révész et al. (2016). The *c'* path, which deals with the direct effect of task complexity on linguistic diversity, has been explored by studies such as Lee (2019), Révész et al. (2017), and Vasylets et al. (2017). Although the *b* path, concerning the effect of cognitive load on linguistic complexity, has seldom been formally investigated in the context of TBLT, studies such as Christodoilides (2016) and Lively et al. (1993) have reported the effect of cognitive load on speech production, and Li et al. (2024) has studied the influence of working memory on L2 writing production. The logic of these investigations can be somewhat applied to the *b* path. Moreover, theories such as Skehan's (2014) Limited Attentional Capacity (LAC) model also support the investigation of the *b* path. The LAC model proposes that people have limited attentional capacity, and task complexity can impose challenges to learners' limited cognitive resources, resulting in a tradeoff among complexity, accuracy, and fluency in learners' production. Given that the LAC model implies a causal relationship between cognitive load and linguistic complexity in production, the investigation of the *b* path is also theoretically valid and informative. Taken together, the three paths constituting the hypothetical model are either empirically or theoretically supported for investigation.

On the one hand, most of these effects (i.e., particularly paths *a* and *c*), in isolation, are rather well investigated. On the other hand, the whole mechanism governing these causal relationships has received less, if any, attention. For example, what is the indirect

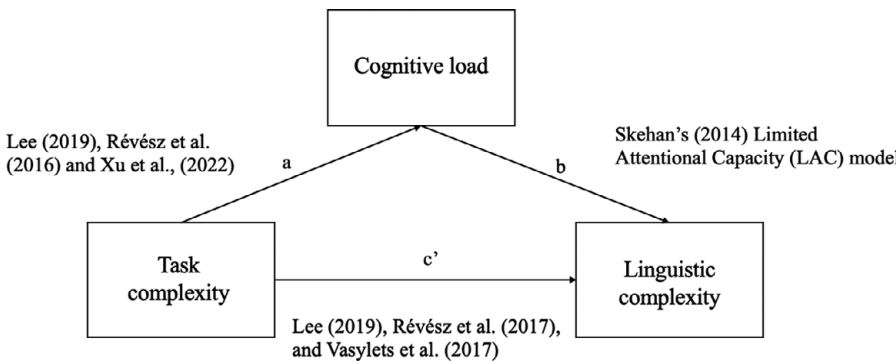


Figure 6. Visualization of the mediation model of Example 2.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

effect of task complexity on linguistic complexity, via cognitive load? What is the total effect, including both direct and indirect effects? What exactly is the role of cognitive load in this system (i.e., the mediation effects)? Might other factors play a role in understanding linguistic complexity? The answer to these questions can further our nuanced understanding of the implications of manipulating task complexity and the exact role of cognitive load.

Thus, for the purpose of demonstration, we conceptualized a within-participant study. The research question that could be addressed is "*To what extent is the effect of task complexity on the linguistic complexity of L2 speaking mediated by learners' cognitive load?*" Data from 60 participants were simulated for both the simple and complex versions of a speaking task. The tasks were administered at two time points, spaced one week apart (see Figure 7 for the procedure). We operationalized cognitive load as learners' perceived difficulty of the tasks (e.g., Lee, 2019; Révész et al., 2016; Révész et al., 2017; Vasylets et al., 2017), using a nine-point Likert scale. The linguistic complexity of learners' oral production was operationalized as lexical diversity indexed by VOCD or the D measure (e.g., Lee, 2019; Révész et al., 2017), drawn from their recorded speaking performance. Thus, in this design, each participant provided two difficulty ratings (one after each complexity level) and had two VOCD measure scores (again, one for each complexity level).

The simulated dataset

A dataset was simulated to represent this hypothetical research scenario. Given the within-participant design, we accounted for the correlation between cognitive load ratings and VOCD scores across the two complexity levels of the spoken task. Thanks to a reviewer's feedback, we also incorporated residual heterogeneity to reflect realistic differences in variance between different task conditions, as variability at each complexity level is unlikely to be identical. We assume that a simple task might yield a smaller variance because participants tend to produce similar responses, whereas a

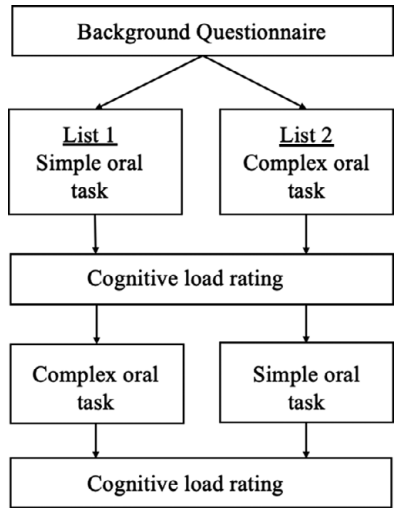


Figure 7. Procedure of the hypothetical experiment in example 2.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

Table 4. Descriptive statistics of cognitive load rating and VOCD by task complexity

	M (SD) [95% CI]	
	Simple	Complex
Cognitive load rating	4.55 (1.43) [4.18, 4.92]	6.68 (1.44) [6.31, 7.06]
VOCD	68.84 (9.85) [66.30, 71.38]	77.37 (1.17) [74.74, 80.00]

complex task may result in a larger variance due to unexpected difficulties. In sum, the simulation accommodated both data dependency and variance differences across tasks within the multilevel modeling framework. In the final dataset, again, each participant has two values for cognitive load rating: one for the simple version and one for the complex version. The same is true for the outcome variable, VOCD. Below, we present and visualize the descriptive statistics in Table 4 and Figure 8. In addition, we include the scatterplot to illustrate the relationship between cognitive load and VOCD by task complexity in Figure 9.

Mediation analysis and results

Given the design, two modeling approaches can be considered: First, one can adopt (Bayesian) multilevel modeling (MLM), which accounts for data dependency and residual heterogeneity; second, one can also engage mediation analysis within a path-analytic framework (e.g., Montoya & Hayes, 2017). Since the third example (see more below) will employ the first modeling approach to handle trial-level data, we focus on the path-analytic framework to approach this dataset. Specifically, we used the MEMORE (Mediation and Moderation for Repeated Measures) macro (Montoya & Hayes, 2017) as an alternative to the multilevel modeling approach for the present research scenario.

The MEMORE macro, developed specifically for two-condition within-participant designs, operates within a path-analytic framework using ordinary least squares (OLS) regression (Montoya & Hayes, 2017). It models mediation (and moderation) relationships by focusing on the differences between two conditions for each participant. Rather than treating the mediator as a single variable, MEMORE calculates the difference in the mediator across the two conditions as the mediator itself and the difference in the outcome across the two conditions as the outcome itself (Montoya & Hayes, 2017). By subtracting the mediator and the outcome across the two conditions within each participant, the difference scores allow us to not only cancel out individual differences but also isolate the effect of the predictor. Again, in this research example, the same participant rated their cognitive load for both the simple and complex versions of the task. Their ratings may depend on each person's individual characteristics, such as their cognitive capacity. These individual characteristics can create correlations between the two treatment conditions, making it difficult to isolate the specific effect of task complexity on cognitive load rating and lexical diversity. However, when calculating difference scores, we remove these individual characteristics from each condition by focusing only on the changes in cognitive load rating and lexical diversity as a function of task complexity, teasing out any individual-related influence from the model. This also means that we no longer need to account for data dependency by specifying correlations in the model. It simplifies model specifications and makes it

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

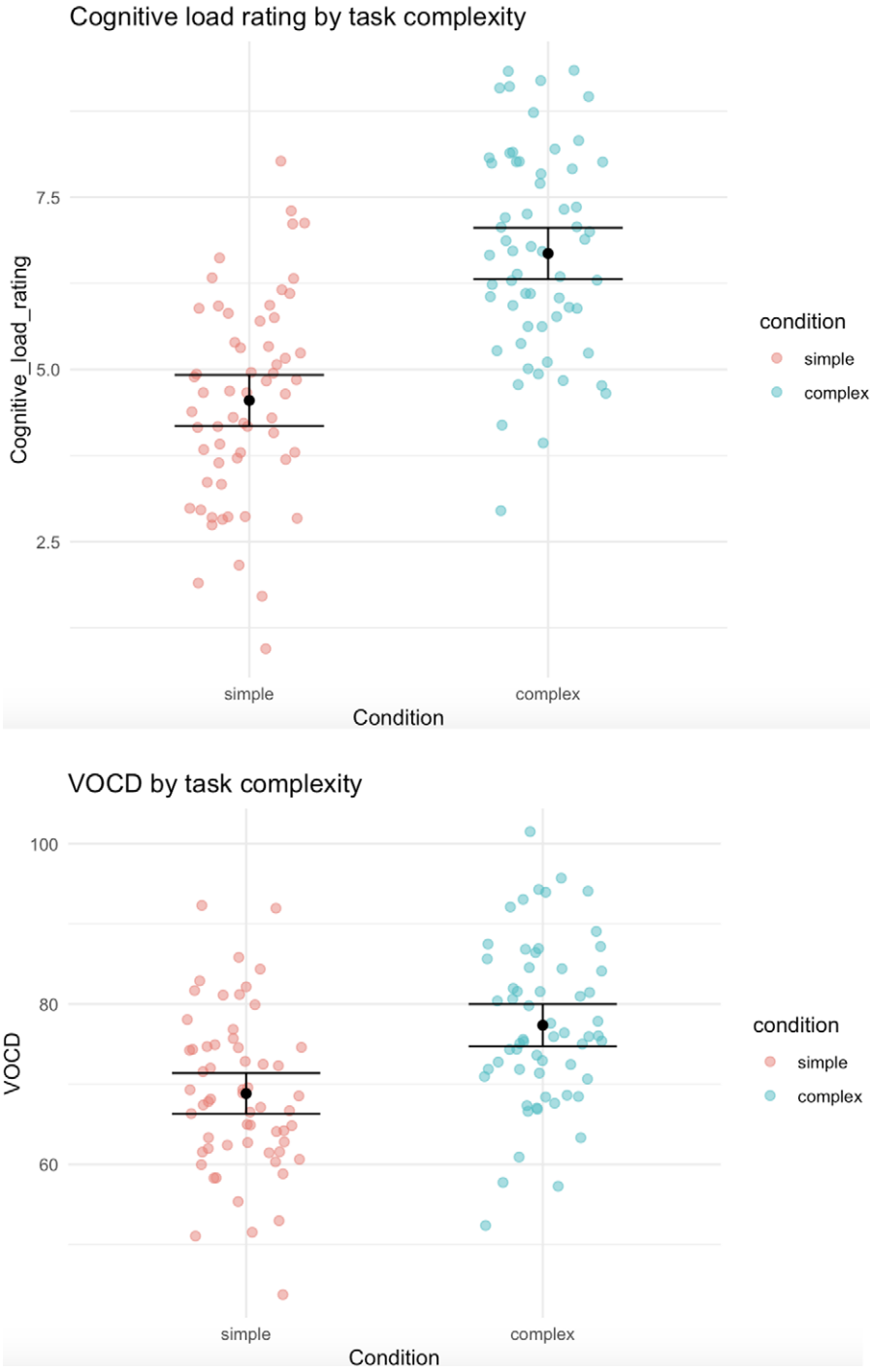


Figure 8. Visualizing the descriptive statistics of cognitive load rating and VOCD.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

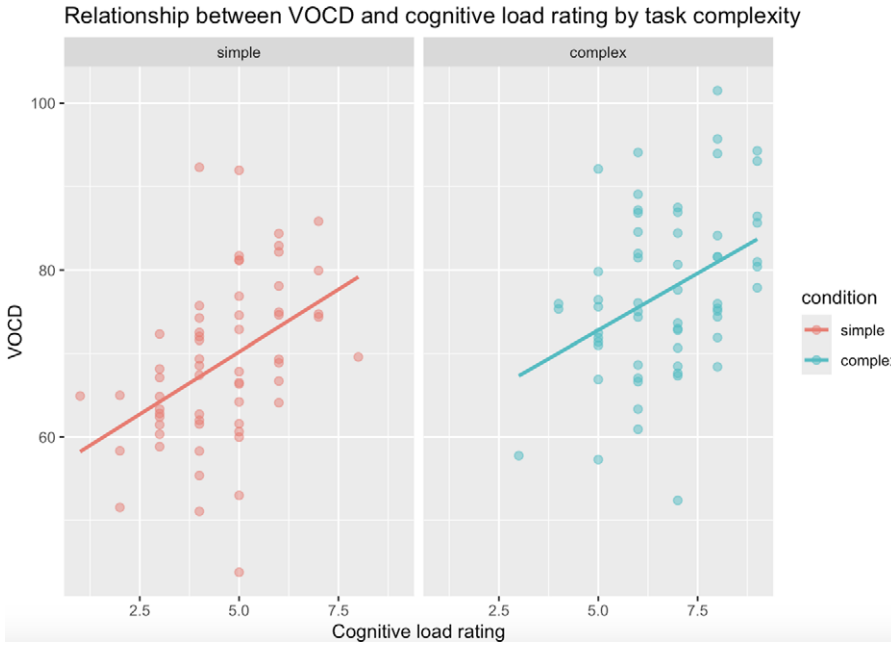


Figure 9. Visualizing the relationship between cognitive load and VOCD by task complexity.

easier to interpret the effect of the treatment. Additionally, we also included the mean of the mediator (i.e., the average of the two cognitive ratings by the same person between the conditions) in the model to control for the baseline differences across participants. In this way, we separate the participant-specific effect from the task-specific effect without the interference of the baseline-level variability among participants. Figure 10 illustrates the statistical diagram of this approach.

Currently, the MEMORE macro is only available for SPSS and SAS. However, for researchers who prefer a point-and-click interface over extensive code-writing, MEMORE offers a highly user-friendly solution. Here, we provide a step-by-step, contextualized tutorial on how to use the MEMORE macro in SPSS to conduct a two-condition within-participant mediation analysis with the hypothetical research scenario described above.

To get started, we need to download MEMORE from the following website: https://www.akmontoya.com/spss-and-sas-macros. The macro named "memore_V2.1.spd" was made publicly available and free to download by Amanda Montoya, who also provided detailed instructions for its use. After downloading, the macro needs to be installed in SPSS. Under the "Extensions" menu, we need to select "Install custom dialog" from the dropdown menu of "Utilities." When prompted, we need to locate the downloaded file "memore_V2.1.spd" in our saved directory and complete the installation. Once the macro is successfully installed, it can be accessed under the "Regression" category within the "Analyze" menu.

To ensure the data format is compatible with MEMORE, the first step is to convert the dataset from a long format to a wide format before opening it in SPSS. This step is necessary because MEMORE automatically calculates the difference scores between conditions for both the mediator and the outcome. Therefore, we must specify the

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

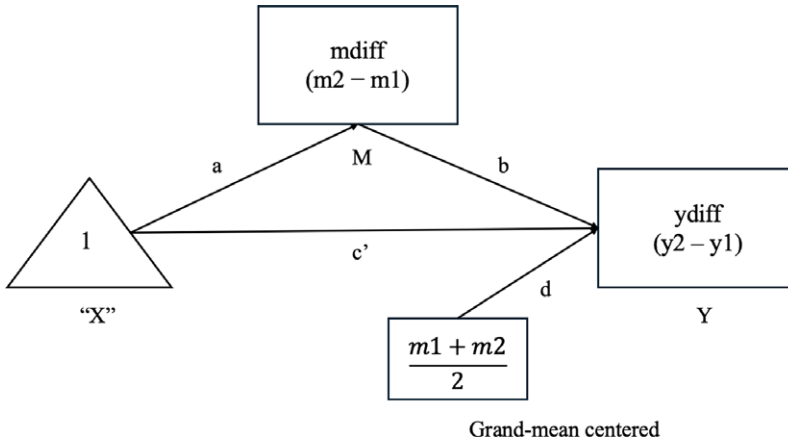


Figure 10. The statistical diagram of the mediation model (Montoya & Hayes, 2017).
Note: In this model, the predictor X, which refers to task complexity (simple vs. complex), is modeled as a fixed variable. This is because in a within-subject design, participants experience both treatment conditions, meaning that it does not vary across participants. "mdiff" represents the differences in cognitive load rating between the simple and complex tasks, indicating how cognitive load rating changes due to task complexity. "ydiff" refers to the differences in lexical diversity scores indexed by VOCD, representing how lexical diversity scores change due to the changes in cognitive load rating and task complexity. The inclusion of the grand-mean centered covariate is to account for the absolute level of cognitive load rating. In other words, it controls for individual differences in cognitive load rating, as participants with higher levels of cognitive load rating might exhibit different levels of lexical diversity compared to those with lower levels of cognitive load rating. Difference scores, which reflect only within-participant differences, cannot capture between-participant variations. Thus, to account for individual differences, absolute values must be controlled for.

Table 5. Sample dataset for Example 2

sub_id	m1	m2	y1	y2
1	4	5	72.02	76.42
2	4	7	61.99	72.95
3	7	9	74.35	77.86
4	4	7	58.30	68.47
5	4	7	55.37	67.63

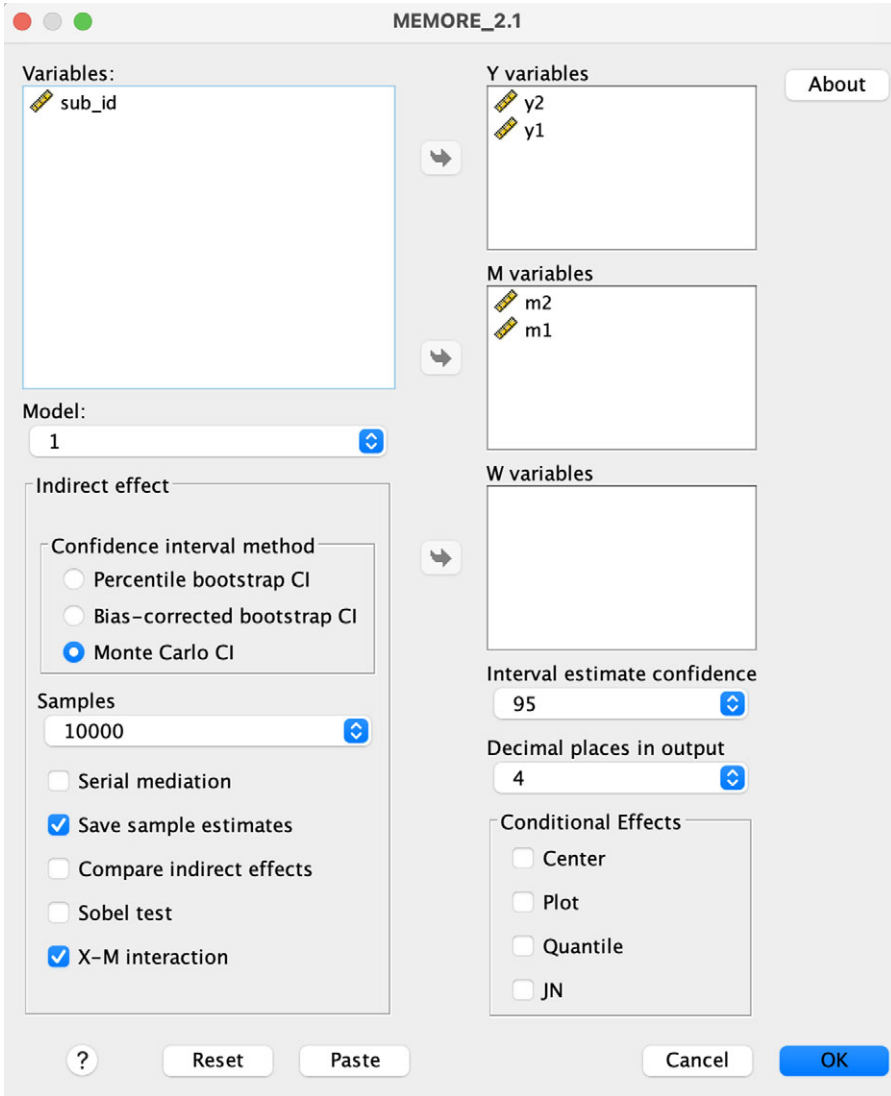
Note: m1 refers to the cognitive load ratings for the simple task; m2 represents the cognitive load ratings for the complex task; y1 refers to the VOCD scores for the simple task, and y2 represents the VOCD scores for the complex task.

mediator and the outcome for each condition so that MEMORE can correctly identify and process these columns. Table 5 presents a sample of the dataset that was uploaded for analysis.

To start the analysis, we need to first ensure that the mediator and the outcome variables are properly set as "scales" in the "Variable View" interface. Then, we can click on the "Analysis" menu and select "Regression," where we can find the MEMORE function just installed. In the MEMORE window, it is important to classify the mediator variables (i.e., m1 and m2) into the "M variables" box and the outcome variables (i.e., y1 and y2) into the "Y variables" box. Additionally, the variables from each condition should be entered in the correct order, as MEMORE will perform subtraction based on this sequence. Thus, in our case, we first selected "y2" and then "y1" to the "Y variables" box and "m2" and "m1" in sequence to the "M variables" box. This setup tells MEMORE to subtract "y1" from "y2" for the outcome and "m1" from "m2" for the

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

mediator to obtain the difference scores. Moving forward, we set the model as Model 1, which specifies a within-participant repeated measures design. Additionally, we



selected the Monte Carlo option to estimate confidence intervals and set the number of resampling iterations to 10,000. Researchers can adjust the value for their needs. Moreover, it is also essential to check the option to save sample estimates and specify the confidence interval as 95% to ensure accurate interval estimates for the analysis. Once all parameters are specified and the analysis is conducted, the results (see Figure 11) are presented as follows:

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

```
Model:
  1

Variables:
Y = y2       y1
M = m2       m1

Computed Variables:
Ydiff =          y2        -       y1
Mdiff =          m2        -       m1
Mavg =  (        m2        +       m1       )        /2       Centered

Sample Size:
  60

**************************************************************************************
Outcome: Ydiff =  y2        -       y1

Model
        Effect         SE          t          p       LLCI       ULCI
'X'     8.5245      .4712    18.0897      .0000     7.5816     9.4675

Degrees of freedom for all regression coefficient estimates:
  59

**************************************************************************************
Outcome: Mdiff =  m2        -       m1

Model
        Effect         SE          t          p       LLCI       ULCI
'X'     2.1333      .0965    22.1178      .0000     1.9403     2.3263

Degrees of freedom for all regression coefficient estimates:
  59

**************************************************************************************
Outcome: Ydiff =  y2        -       y1

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .3491      .1219    12.1108     3.9548     2.0000    57.0000      .0246

Model
           coeff         SE          t          p       LLCI       ULCI
'X'       5.3254     1.3697     3.8881      .0003     2.5826     8.0681
Mdiff     1.4996      .6065     2.4725      .0164      .2851     2.7141
Mavg      -.4515      .3264    -1.3831      .1720    -1.1052      .2022

Degrees of freedom for all regression coefficient estimates:
  57
```

Figure 11. The MEMORE output for Example 2.

The first section of the output indicates the model type executed, the variables included in the dataset, and all the variable manipulation behind the screen. MEMORE automatically calculated the difference in the mediator (*Mdiff*) and the outcome (*Ydiff*) across task conditions, as well as the mean of the mediator (*Mavg*). This step allows the software to prepare the dataset for subsequent analyses.

Starting from the second section, the output reports the estimates of the total effect of task complexity on the changes in lexical diversity indexed by the VOCD scores. This is followed by the third section, which presents the effect of task complexity on the changes in learners' cognitive load captured by their perceived difficulty of the tasks

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

```
************************** TOTAL, DIRECT, AND INDIRECT EFFECTS **************************

Total effect of X on Y
     Effect         SE          t         df          p       LLCI       ULCI
     8.5245      .4712    18.0897    59.0000      .0000     7.5816     9.4675

Direct effect of X on Y
     Effect         SE          t         df          p       LLCI       ULCI
     5.3254     1.3697     3.8881    57.0000      .0003     2.5826     8.0681

Indirect Effect of X on Y through M
         Effect       MCSE     MCLLCI     MCULCI
Ind1     3.1992     1.2997      .6541     5.7053

Indirect Key
Ind1  'X'      ->       Mdiff    ->       Ydiff

******************************** ANALYSIS NOTES AND WARNINGS ********************************

Number of samples for Monte Carlo condifidence intervals:
  10000

The following variables were mean centered prior to analysis:
 (        m2        +       m1       )        /2

Level of confidence for all confidence intervals in output:
      95.00

------ END MATRIX -----
```

Figure 11. Continued.

(referred to as the *a* path). In the fourth section, three relationships are provided: (a) the direct effect of task complexity on the changes in lexical diversity (the *c'* path), (b) the effect of the changes in cognitive load rating on the changes in lexical diversity (the *b* path), and (c) the effect of the average cognitive load rating on the changes in lexical diversity. These sections together provide a detailed breakdown of the relationships between task complexity, cognitive load, and lexical diversity.

Although the first part of the output provides all the necessary information for interpretation, a concise summary of the estimates for the total, direct, and indirect effects can be found in the second part of the output. It can be seen that the total effect of task complexity on the changes in VOCD was significant, $b = 8.525$, SE = .471, 95% CI [7.582, 9.468], $p < .001$. This indicates that engaging in the more complex task resulted in an average increase of 8.525 units in VOCD compared to the simpler task, suggesting that higher task complexity was associated with greater lexical diversity in their oral performance. Moreover, the direct effect of task complexity on the changes in lexical diversity remained significant after controlling for the changes in cognitive load rating, $b = 5.325$, SE = 1.370, 95% CI [2.583, 8.068], $p < .001$. This finding demonstrates that task complexity continued to predict lexical diversity even after accounting for learners' perceived difficulty of the tasks, highlighting its independent contribution to the changes in lexical diversity. Furthermore, the indirect effect of task complexity on the changes in lexical diversity, via the changes in cognitive load rating, was also significant, $b = 3.199$, SE = 1.300, Monte Carlo 95% CI of [.654, 5.705]. This indicates that learners' cognitive load, as indexed by their perceived difficulty of the tasks, served as a meaningful mediator, influencing the effect of task complexity on lexical diversity. Specifically, approximately 37.52% of the total effect (3.199/8.525) was accounted for by learners' perceived difficulty of the tasks, underscoring the role of cognitive perception

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

in bridging task complexity and lexical diversity. Again, we remind readers that these results and interpretations are based on our simulated dataset for the present demonstration.

As demonstrated, MEMORE provides a straightforward output that clearly outlines the estimates for each path, as well as the indirect and total effects of interest. Additionally, MEMORE offers the flexibility to model multiple mediators (see Montoya & Hayes, 2017 for more details), which enhances researchers' ability to explore more complex relationships between learning conditions, processes, and outcomes. While there is no direct equivalent of MEMORE in R, the point-and-click interface in SPSS largely reduces the need for extensive coding, making it a powerful tool for conducting mediation analysis with repeated measures.

For researchers who are familiar with Mplus, Montoya and Hayes (2017) also provide the Mplus code for carrying out within-participant mediation analysis. However, we acknowledge that SPSS, SAS, and Mplus are behind paywalls. Thus, we translated the Mplus code provided by Montoya and Hayes (2017) into R using the *lavaan* package (Rosseel, 2012) to enable researchers to conduct this type of analysis with open-source statistical tools. Both the R and Mplus codes can be found in our shared OSF link.

Example 3: Mediation in within-participant designs with cross-random structures

Unlike Example 2, where each participant only has one score for each condition, researchers often analyze trial-level data where one participant has multiple values (i.e., multiple rows in the dataset) for each condition. Adding to the picture is the fact that each item elicits data from multiple participants. Given this cross-nested data structure, which is common in, for example, eye-tracking research (e.g., Godfroid & Hui, 2025), the two statistical methods introduced above fall short in accounting for such complexity. Thus, in this example, we introduce another statistical method to accommodate mediation with cross-random data structures in within-participant designs.

To do this, we engaged in mixed-effects modeling within the Bayesian framework using the *brms* package (Bürkner, 2017) in R. The primary reason to introduce the Bayesian (as opposed to frequentist) approach is that there are currently readily available functions and packages in R that can accommodate mediation with model objects commonly used to analyze such data without manual calculations of the indirect and total effects with bootstrapping. In addition to practicality, the Bayesian approach is powerful in its ability to offer more accurate estimates based on posterior distribution, more straightforward output, much simpler code for implementation, and greater flexibility to accommodate complex data structures when mediation is involved (Bürkner, 2017; Vuorre & Bolger, 2018). We acknowledge that the Bayesian approach can seem intimidating to many applied researchers who might not have much relevant training, and indeed, it can take a lot of computational resources. But we wanted to stress that many will be able to draw on their knowledge in constructing mixed-effects models, using, for example, the *lme4* package (Bates et al., 2015).

We adopted the dataset shared by Tytko et al. (2024), available on Open Science Framework (https://osf.io/hq57n/?view_only=ed878de940e44cadade885f21efeea45). It was obtained for an eye-tracking study that investigated the effectiveness of various types of multimodal input in scaffolding reading comprehension. Sixty-six adult L2

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

learners were exposed to different parts of a text under three experimental conditions: (a) reading-only (RO), (b) reading with an image (RI), and (c) reading with both an image and audio (RIA). During exposure, participants read or read and listened to the text materials on 21 screens for each condition, with their eye movements being captured. Several eye-tracking indices were used to index their eye movement behavior, including the percentage of their dwell times on images. Moreover, comprehension questions were administered to measure the participants' understanding of the text. Some screens have only one comprehension question associated with them, while some have more. The researchers examined not only the effect of different input conditions on reading comprehension (i.e., the *c'* path) but also the relationship between image processing and comprehension (i.e., the *b* path). In this context, a pertinent mediation question that can be addressed through multilevel mediation is: "*What is the role of L2 learners' attention to images in influencing the effect of different multimodal inputs on reading comprehension*?" Figure 12 demonstrates the visualization of this mediation model.

For demonstration purposes, we only included two conditions: RI and RIA. Participants' attention to images was quantified as the percentage of their dwell times on images relative to their total dwell times on both images and text. Participants' responses to comprehension questions were scored as 0 or 1, with 0 being incorrect and 1 being correct. Thus, the accuracy data were binary.

It is important to emphasize again that, unlike the former example where responses vary only at the participant level, the current dataset is cross-classified. This means that the participants' responses or behaviors vary not only at the participant level but also at the item level. As reflected in this study, each participant had 21 proportion values (one for each screen) for each of the two conditions, resulting in a total of 42 eye-tracking data points per participant. Additionally, each screen elicited eye-movement data from participants going through the two conditions and each comprehension question was also answered by participants experiencing both conditions. This cross-classified data structure requires statistical models to account for both the by-participant and by-item variability. Table 6 presents the data format of this study. Applied psycholinguistics, including eye-tracking researchers, should be fairly familiar with the structure of this kind of trial-level data.

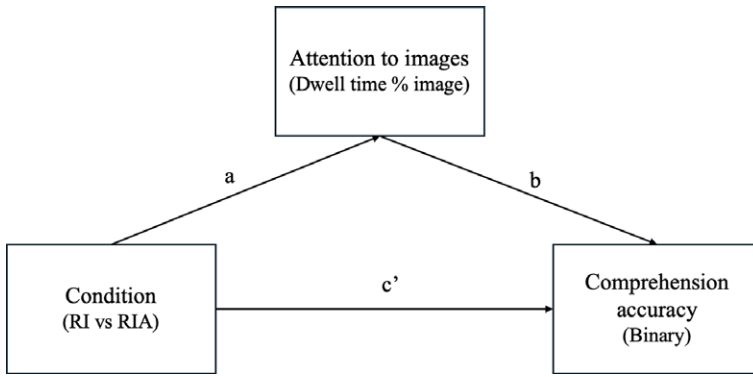


Figure 12. Path diagram of the eye-tracking mediation model (binary outcome).

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

Table 6. Sample eye-tracking dataset

ID	TRIAL_INDEX	con	pic_percent	Q	acc	List
P101	22	0	.156	Q32	1	List1
P101	22	0	.156	Q31	1	List1
P101	23	0	.120	Q33	1	List1
P101	23	0	.120	Q34	1	List1
P101	24	0	.068	Q35	1	List1
P101	25	0	.109	Q36	1	List1
…						
P101	43	1	.045	Q61	1	List1
P101	44	1	.038	Q62	1	List1
…						
P102	1	0	.137	Q1	0	List2
P102	1	0	.137	Q2	0	List2
…						
P102	22	1	.082	Q31	1	List2
P102	22	1	.082	Q32	1	List2

Note: "TRIAL_INDEX" refers to the screen participants read texts on. "con" refers to the two reading conditions: The reading + image condition was coded as 0, and the reading + image + audio condition was coded as 1. "pic_percent" stands for the percentage of participants' dwell times on images. "Q" refers to the comprehension questions. "acc" refers to the participants' accuracy of the comprehension questions, and "List" stands for the experimental list that each participant was assigned to.

Table 7. Descriptive statistics of dwell time percentage on images and comprehension accuracy by reading conditions

	RI		RIA
		M (SD) [95% CI]	
pic_percent	.160 (.095) [.155, .164]		.191 (.115) [.186, .196]
		M (SD) [95% CI]	
Comprehension accuracy	.666 (.472) [.645, .687]		.662 (.473) [.641, .683]

Note: RI = Reading + Image; RIA = Reading + Image + Audio

We first present the descriptive statistics of learners' dwell time percentage on images by different reading conditions and the proportion of accurate comprehension responses by conditions in Table 7 with visualization in Figures 13 and 14, respectively. To conduct mediation analysis, we need to first install and load the *brms* package (Bürkner, 2017) and build our Bayesian models. We need to first specify two mixed-effects models, one for the mediator (i.e., the *a* path) and the other for the outcome (i.e., the *b* and *c'* paths). The syntax of the mixed-effects models mirrors that of the lmer () function in the *lme4* package (Bates et al., 2015), with the key difference being the replacement of "lmer" with "bf" to accommodate the Bayesian framework.

We first specified the full mixed-effects models (e.g., Barr et al., 2013). In the mediator model (i.e., the *a* path), participants' percentage of dwell time on images ("pic_percent") was treated as the outcome variable, with "condition" dummy coded (RI: 0; RIA: 1) as the predictor (or the fixed effect). Both by-item (i.e., TRIAL_INDEX) and by-participant (i.e., ID) random intercepts are specified, along with condition (i.e., con) as the random slopes. Given that the percentage of dwell time on images (i.e., "pic_percent") is proportional, the data was beta transformed to better handle the limited range. Also, we identified eight cases of zero for "pic_percent" in our dataset, so

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

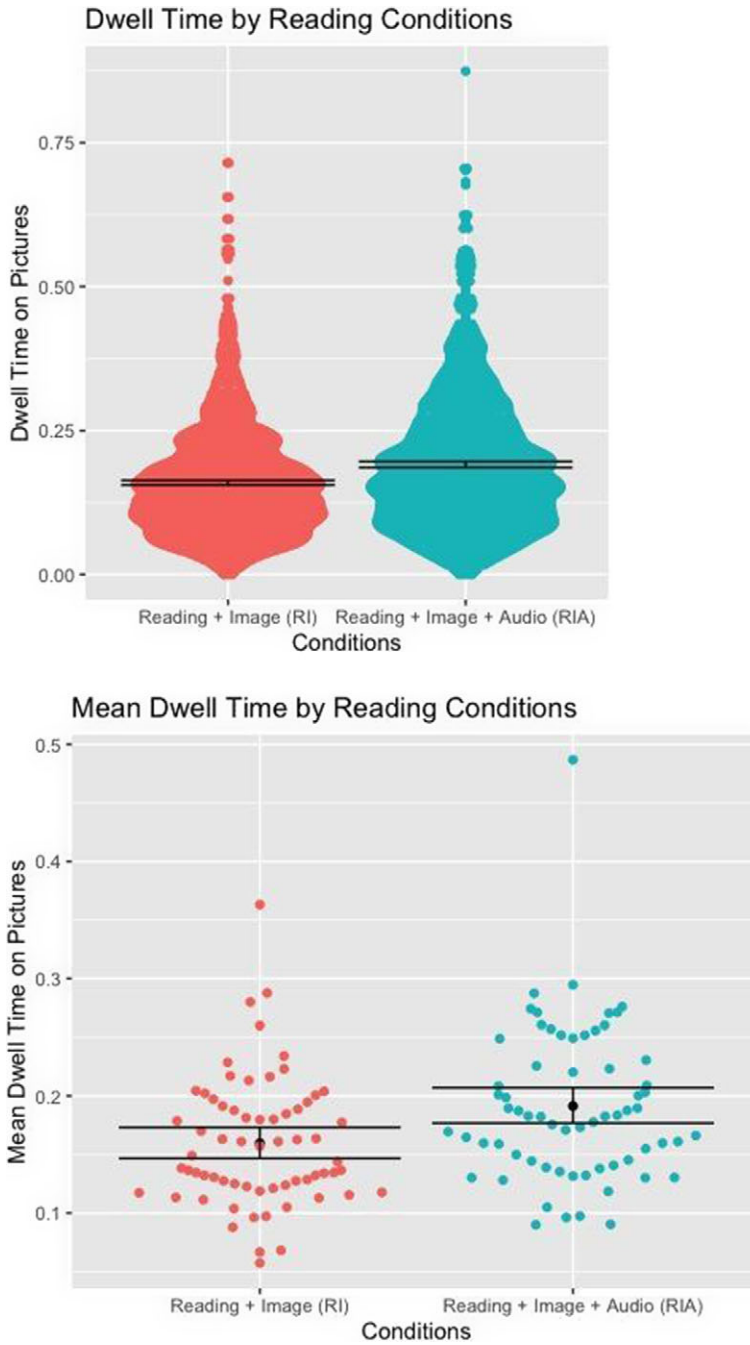


Figure 13. Visualization of raw and mean dwell time percentage to images by reading conditions.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

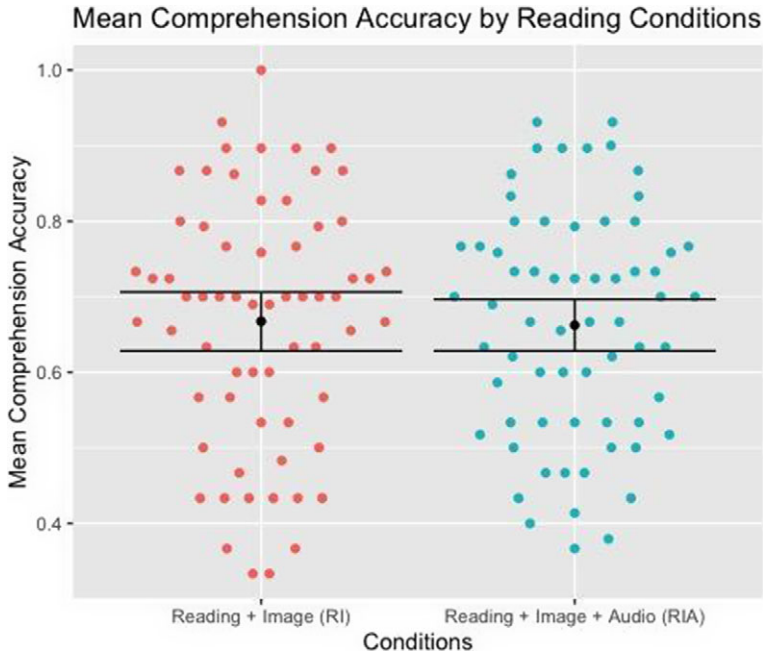


Figure 14. Visualization of the proportion of accuracy by reading conditions.

we chose the zero-inflated beta model to account for those exact zeros. As part of the zero-inflated beta model, we specified zi = ~1 to indicate that the probability of the occurrence of zeros does not relate to any predictors in the model, meaning that the occurrence of zeros may be due to some unknown processes unrelated to the reading conditions such as participants not paying attention to the images[1].

```
# Full model
# Define the mediator model
mediator_formula <- bf(
  pic_percent ~ con + (1 + con | ID) + (1 + con | TRIAL_INDEX),
  zi = ~ 1,
  family = zero_inflated_beta())

# Define the the outcome model
outcome_formula <- bf(
  acc ~ con + pic_percent + (1 + con + pic_percent | ID) +
    (1 + con + pic_percent | Q), family = bernoulli())
```

[1]We thank an anonymous reviewer who provided very useful feedback on our modeling.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

For the outcome model (i.e., the *b* and *c'* paths), comprehension accuracy at the item level (i.e., acc: 0 or 1) served as the outcome variable, with both condition and dwell time as predictors. Given that both questions and participants exhibit variability based on reading conditions and dwell time, "con" and "pic_percent" were treated as random slopes. Similar to the generalized linear mixed-effects models, identifying the distribution of the outcome variable is crucial. In this case, we specify family = bernoulli() to indicate the binary outcome. The bernoulli distribution models the probability of a binary event where a single trial has only two possible outcomes (Gelman et al., 2013), which aligns with the characteristics of the outcome variable (i.e., comprehension accuracy) in the present example. It is a special case of the binomial distribution often specified in generalized linear mixed-effects models using glmer().

Once the mixed-effects models are defined, we can proceed to fit them within the Bayesian mediation framework using the brm() function. The first argument in the brm () function contains the defined mixed-effects models connected by the "+" sign. The second parameter requires the identification of the long-format dataset. After that, we need to specify the number of chains and iterations for model execution (Bürkner, 2017). Optionally, a starting point can be set using the "seed" parameter for reproducibility. This model requires more computational demands and thus takes longer to run, and the running time is contingent upon model complexity, number of observations, number of iterations, etc. After successfully sampling from the posterior distribution of model parameters, we can use the summary() function to produce a summary of the result, where information such as random effects, fixed effects, and model convergence details is listed.

```
# Fit the Bayesian model with both mediator and outcome models
model1 <- brm(
  mediator_formula + outcome_formula,
  data = df,
  seed = 123,
  chains = 4,
  iter = 10000
)

summary(model1)
```

After the full model is built, the next step is to identify the best-fitting model by engaging in model comparison. This process is similar to how researchers usually do model comparison in mixed-effects models. Recommended by Barr et al. (2013), we initially adopted the maximal random effects structures by including all random components in the model. Then, we took a backward model selection approach to evaluate the random effects (Matuschek et al., 2017). This process involves sequentially removing one random component (starting from random slopes and then random intercepts) at a time to test whether deleting one random component with the smallest standard deviation would significantly decrease the model fit. However, for model comparison, instead of relying on traditional criteria such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or chi-square statistics as is

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

Table 8. Random effects of model 1

	Estimate	Est. error	2.5%	97.5%	Rhat	Bulk_ESS	Tail_ESS
~ID							
sd(picpercent_Intercept)	.37	.04	.31	.45	1.00	4227	9275
sd(picpercent_con1)	.37	.04	.30	.45	1.00	3565	8107
sd(acc_Intercept)	.81	.14	.55	1.09	1.00	10300	14718
sd(acc_con1)	.17	.12	.01	.44	1.00	4440	9428
sd(acc_picpercent)	2.00	.84	.37	3.67	1.00	2271	5139
~TRIAL_INDEX							
sd(picpercent_Intercept)	.29	.03	.24	.35	1.00	6421	11211
sd(picpercent_con1)	.16	.03	.11	.22	1.00	8298	12612
~Q							
sd(acc_Intercept)	1.30	.18	.98	1.67	1.00	7526	12488
sd(acc_con1)	.19	.14	.01	.50	1.00	4036	9007
sd(acc_picpercent)	3.25	.95	1.51	5.28	1.00	886	376

Note: The correlation between each random effect is omitted in the table.

```
# Reduced model 1
mediator_formula2 <- bf(
  pic_percent ~ con + (1 + con | ID) + (1 | TRIAL_INDEX),
  zi = ~ 1,
  family = zero_inflated_beta())

outcome_formula2 <- bf(
  acc ~ con + pic_percent + (1 + con + pic_percent | ID) + (1 + con + pic_percent | Q),
  family = bernoulli()
)

model2 <- brm(mediator_formula2 + outcome_formula2,
              data = df,
              seed = 123,
              chains = 4,
              iter = 10000)

summary(model2)
```

typically done in mixed-effects modeling, here, we need to perform Leave-One-Out Cross-Validation (LOO-CV) for both models and then compare their respective expected log predictive density (ELPD) value to make informed decisions (see details below). We prefer the use of LOO-CV over other methods, such as Widely Applicable Information Criterion (WAIC) for model comparison because LOO-CV offers less biased estimates of out-of-sample predictive performance (i.e., how well the model could generalize the pattern/relationship to new data) and it is more reliable for datasets with small sample sizes (Vehtari et al., 2017), as is the case in the present study.

Table 8 presents the random effects of model 1. As can be seen, the by-trial random slope of "con" (i.e., condition) demonstrates the smallest standard deviation among all the random slopes (*Estimate* = .16). Therefore, we eliminated this component by fitting a reduced model, referred to as model 2. Then, we conducted a LOO-CV test (Bürkner, 2017) on both models using the loo() function from the *loo* package (Vehtari et al., 2024). LOO-CV estimates the model's predictive accuracy by systematically leaving out

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

one observation at a time from the dataset, fitting the model to the remaining data, and then predicting the left-out observation (Vehtari et al., 2017). Next, the predictive accuracy of the two models is compared using the loo_compare() function. This function calculates the difference in expected log predictive density (ELPD) between the models, along with its standard error (Vehtari et al., 2017). A higher ELPD indicates a superior predictive accuracy.

```
## Model comparison
loo_fit <- loo(model1)
loo_fit1 <- loo(model2)
comparison1 <- loo_compare (loo_fit, loo_fit1)
print(comparison1)
```

The comparison table (see Table 9) demonstrates that both the ELPD difference and its standard error for model 1 are zero. This indicates that model 1 serves as the reference model for assessing the predictive accuracy of the reduced model. Notably, the ELPD difference for model 2 is -23.5, which is smaller than zero, indicating that model 2 exhibits much lower predictive accuracy compared to model 1. Moreover, given that the absolute value of *elpd_diff* for model 2 (23.5) is larger than the *se_diff* value multiplied by 1.96 (i.e., *se_diff* * 1.96), which is 15.876 (1.96 corresponding to $p < .05$), the two models differ significantly from each other in terms of predictive accuracy. Because model 1 is more complex and demonstrates significantly better predictive accuracy, it is considered our best-fitting model.

Table 10a presents the model summary of the effects of paths *a*, *b*, and *c'* for our best-fitting model. For each parameter, the output table presents the estimate (i.e., the posterior mean of the parameter being estimated), standard error (SE), 95% credible interval at the lower and upper bounds, the potential scale reduction factor (*Rhat*) calculated based on the posterior distribution, bulk effective sample size (Bulk_ESS), and tail effective sample size (Tail_ESS). Note that *Rhat*, which is a convergence diagnostic index, is recommended to be 1 or at least very close to 1 to indicate good

Table 9. Model comparison results

models	elpd_diff	se_diff
model1	.0	.0
model2	−23.5	8.1

Table 10a. Mediation model summary output from the brm() function

	Estimate	SE	2.5%	97.5%	Rhat	Bulk_ESS	Tail_ESS
picpercent_Intercept	−1.71	.06	−1.83	−1.59	1.00	3889	6179
zi_picpercent_Intercept	−6.11	.34	−6.83	−5.51	1.00	16249	10851
acc_Intercept	.84	.21	.43	1.25	1.00	6905	12753
picpercent_con1 (a)	.20	.05	.10	.30	1.00	5996	9439
acc_con1 (c')	−.06	.10	−.26	.13	1.00	20917	14819
acc_pic_percent (b)	1.66	.83	.10	3.38	1.00	11530	13003
phi_picpercent	26.03	.62	24.82	27.26	1.00	26097	13390

Note: (a), (c'), and (b) are not in the output. The denotation is added here for better understanding and illustration. The table only includes fixed effects. The whole output also includes by-item, by-participant, and by-question random effects (see Table 7).

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

Table 10b. Mediation model summary output from the mediation() function

Effect	Estimate	95% ETI
Direct effect (ADE)	-.063	[-.261, .131]
Indirect effect (ACME)	.306	[.017, .762]
Mediator effect	1.631	[.096, 3.376]
Total effect	.249	[-.098, .717]

Note: Proportion mediated: 123.25% [-350.25%, 596.74%]. Direct and indirect effects have opposite directions. The proportion mediated is not meaningful.

model convergence (Bürkner, 2017). If the *Rhat* value exceeded 1.1, increasing the number of iterations would help resolve the convergence issue (Bürkner, 2017). Additionally, serving as diagnostics for sampling efficiency in the bulk and tail of the posterior distribution, respectively, Bulk_ESS and Tail_ESS are expected to be at least 100 per Markov Chain to indicate reliable sampling (Vehtari et al., 2021). To output the indirect and total effect, we adopted the mediation() function from the *bayestestR* package (Makowski et al., 2019) with the Bayesian model just fitted. This function facilitates the estimation of the indirect and total effects, returning a summary table easy and straightforward for interpretation (see Table 10b).

```
# Mediation
mediation(model1)
```

As we can see from Table 10a, the model demonstrates a good convergence as all the *Rhat* values are 1.00. The first and third parameters, "picpercent_Intercept" and "acc_Intercept," refer to the intercept of dwell time percentage on pictures and accuracy, respectively. The second parameter "zi_picpercent_Intercept" represents the intercept of the zero-inflated part of the mediator model, which models the probability of the dwell time percentage ("pic_percent") being zero regardless of the reading conditions. As can be seen, the estimate of the intercept of the zero-inflated model is negative, b = -6.11, SE = .34, 95% Credible Interval [-6.83, -5.51], suggesting that the log-odds of the mediator being zero was extremely low. Starting from the fourth parameter, variables are connected with the underscore symbol (i.e., _). This could be seen as equivalent to the "~" in the regression model. Thus, "picpercent_con1" refers to the regression coefficient of the predictor "con" (i.e., condition) on the outcome variable (i.e., dwell time percentage on images), namely the *a* path in the mediation model. The result indicates a positive effect of reading conditions on learners' attention to images, b = .20, SE = .05, 95% Credible Interval [.10, .30], suggesting that learners were more likely to pay more attention to images in the RIA condition than in the RI condition. The relatively narrow credible interval indicates a high degree of certainty in this estimate. Regarding the *b* path (i.e., "acc_pic_percent"), the result also reveals a positive effect of attention to images on reading comprehension, with an estimated effect of b = 1.66, SE = .83, 95% Credible Interval [.10, 3.38] on the log-odds scale. Although the relatively wide credible interval suggests some uncertainty in the precision of this estimate, the fact that it does not include zero indicates a high probability that there was an effect. In addition, "acc_con1," which refers to the *c'* path, failed to demonstrate a credible effect of reading conditions on reading comprehension, while accounting for dwell time percentage on images, b = -.06, SE = .10, 95% Credible Interval [-.26, .13]. This result indicates that learners tended to perform similarly across the RI and RIA conditions when controlling for their time spent on images.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

Supplementing the results from the brm() function, Table 10b, derived from the mediation() function, directly showcased the relations that would otherwise be manually calculated: the indirect effect, the total effect, and the proportion of the effect mediated, each with 95% credible interval. As can be seen, learners' attention to images was not only a credible predictor of reading comprehension, b = 1.63, 95% Credible Interval [.10, 3.38] but also a credible mediator influencing the effect of reading conditions on reading comprehension, b = .31, 95% Credible Interval [.02, .76]. However, the relatively wide credible interval of the b path suggests a relatively high level of uncertainty regarding the role of attention to images in predicting reading comprehension. In contrast, the relatively narrower credible interval of the indirect effect suggests a greater degree of certainty in the role of attention to images as a mediator between reading conditions and reading comprehension. In addition, the total effect of reading conditions on reading comprehension was found to be uncertain, b = .25, 95% Credible Interval [-.10, .72]. This implies that the indirect effect might be balanced out by the noncredible direct effect, b = -.06, SE = .10, 95% Credible Interval [-.26, .13]. By outputting the effect of each pathway as well as the indirect and total effects, the results reveal the important role of attention in not only predicting reading comprehension but also mediating the relationship between reading conditions and comprehension. Without such information, we would miss out on the important pathways through which reading conditions could make a difference in reading comprehension. Lastly, when the direct and indirect effects have opposite signs, the proportion mediated becomes difficult to interpret (Vuorre & Bolger, 2018). Here, the presence of a positive, credible indirect effect, along with a noncredible direct effect, indicates that compared to the RI condition, the addition of audio in the RIA condition enhanced learners' attention to images. This increased attention, in turn, positively influenced reading comprehension. Importantly, these results suggest that the value of audio emerged only when it drew learners' attention to the visual information (i.e., the images). Without the boosted attention, the RIA condition provided almost no added value, compared to the RI condition. This interpretation would have been masked if researchers were to carry out separate analyses with the data.

In sum, the brm() function offers a flexible and advanced tool to manage complex data structures commonly encountered in eye-tracking and psycholinguistic research. Its capability expands the range of research where researchers can apply the Bayesian approach to address multilevel mediation questions. As a result, L2 researchers can manipulate this sophisticated modeling technique to uncover new insights into the complexities of second language learning. However, for practical consideration, it should be noted that compared to the frequentist approach, the Bayesian approach typically requires more time for models to run due to its sampling procedures, and the time needed is contingent upon sample size, model complexity, and the number of iterations executed. Moreover, when the brm() function is used, it will take even longer if multiple models have to run sequentially to identify the best-fitting model. However, despite these challenges, the strengths of the Bayesian approach make it a powerful tool for SLA researchers to handle complex data structures and go for more nuanced research questions with within-participant designs, which could facilitate our understanding of more complex processes underlying L2 learning and development.

Strengths, considerations, and limitations

To sum up, this tutorial demonstrates how mediation analysis can be applied to various research designs in SLA research to bridge the connection between learning conditions,

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

processes, and outcomes. It offers a comprehensive understanding of how specific learning conditions influence learning outcomes through intermediary learning processes. This nuanced analysis can uncover the pathways through which instructional interventions exert their effects, offering deeper insights into the mechanisms of L2 learning and informing more effective pedagogical methods.

Nevertheless, it is important to address some considerations when conducting mediation analysis. Since the mediator functions as both a predictor and outcome variable, two issues need to be considered: Firstly, it is important to ensure that the *b* path (i.e., M → Y) reflects a causal relationship. That is to say, for the mediation model to function properly, the mediator M must exert a causal influence on the outcome variable Y (Vuorre & Bolger, 2018) and the model assumes that the direction of the causality flows from M to Y, and it does not support the reverse causality from Y to M or merely a correlational relationship between M and Y (Vuorre & Bolger, 2018). Contextually speaking, for example, in a hypothetical study that examined the effect of task complexity on syntactic complexity in L2 writing, time on task is often used to capture learners' engagement. Although the more complex task would likely result in learners spending more time on the task than the simpler one (i.e., the *a* path), it would be inappropriate to assume a causal relationship between time on task and syntactic complexity (i.e., the *b* path). This is because time on task does not necessarily predict syntactic complexity in L2 writing. Simply spending more time on a task does not guarantee the production of more complex grammatical structures. Thus, treating time on task as a mediator in this context would violate the causality assumption.

Secondly, as a predictor of the *b* path, the mediator being measured is assumed to be free from measurement errors (Vuorre & Bolger, 2018). If a mediator contained a lot of measurement errors or, technically speaking, is not reliable in terms of measures, the prediction of the outcome variable would be less precise and could potentially mask the causal relations that should have existed. Thus, when researchers try to capture a construct of interest, it is essential to ensure that the measures adopted are reliable enough to comply with this "free from measurement errors" assumption. One example of introducing noise to a mediator would be the use of reaction time differences measures to index cognitive individual differences (Tan & Yan, 2016).

Conclusion

To conclude, the tutorial paper demonstrated the process of conducting mediation analysis within the context of SLA research. Three working examples were presented, each corresponding to a different type of research design and data structure commonly encountered in SLA research. By providing detailed, step-by-step guidance on data preparation, statistics selection, code writing, and result interpretation, we hope that the current tutorial on mediation analysis could help equip SLA researchers with the tools to uncover complex causal relationships in second language learning, explore more complex hypotheses with regard to learning conditions, learning processes, and learning outcomes, and promote the development of more comprehensive and robust theoretical frameworks in the field.

Acknowledgments. We would like to extend our thanks to the following researchers who generously shared their data and analysis code in online repositories. Their commitment to open science has made the current secondary analysis possible.

Competing interest. We have no known conflict of interest to disclose.

https://doi.org/10.1017/S0272263125100867 Published online by Cambridge University Press

References

Alrabai, F. (2022). The predictive role of anxiety and motivation in L2 proficiency: An empirical causal model. *Language Teaching Research*, Advance online publication. https://doi.org/10.1177/13621688221136247

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Barr D, Levy R, Scheepers C, & Tily, H. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://doi.org/10.18637/jss.v067.i01

Botes, E., Van Der Westhuizen, L., Dewaele, J. M., MacIntyre, P., & Greiff, S. (2022). Validating the short-form foreign language classroom anxiety scale. *Applied Linguistics*, *43*(5), 1006–1033. https://doi.org/10.1093/applin/amac018

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.3929/ethz-b-000240890

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chen, T. H. (2023). Dynamic fluctuations in foreign language enjoyment during cognitively simple and complex interactive speaking tasks. *Studies in Second Language Learning and Teaching*, *13*(3), 627–661. https://doi.org/10.14746/ssllt.31194

Christodoulides, G. (2016). *Effects of cognitive load on speech production and perception* [Doctoral dissertation, UCL-Université Catholique de Louvain].

Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues in field settings*. Houghton Mifflin.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th *edition*). Sage Publications Limited.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

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

Godfroid, A. (2019). Investigating instructed second language acquisition using L2 learners' eye-tracking data. In R. P. Leow (Ed.), *The Routledge handbook of second language research in classroom learning* (pp. 44–57). Routledge. https://doi.org/10.4324/9781315165080-4

Godfroid, A., & Hui, B. (2025). Eye-tracking research in instructed second language acquisition. Language Teaching. *Advance Online Publication*. https://doi.org/10.1017/S0261444825000102

Gokturk, N., & Chukharev-Hudilainen, E. (2023). Strategy use in a spoken dialog system–delivered paired discussion task: A stimulated recall study. *Language Testing*, *40*(3), 630–657. https://doi.org/10.1177/02655322231152620

Kang, H., Kweon, S. O., & Choi, S. (2022). Using eye-tracking to examine the role of first and second language glosses. *Language Teaching Research*, *26*(6), 1252–1273. https://doi.org/10.1177/1362168820928567

Koval, N. G. (2019). Testing the deficient processing account of the spacing effect in second language vocabulary learning: Evidence from eye tracking. *Applied Psycholinguistics*, *40*(5), 1103–1139. https://doi.org/10.1017/S0142716419000158

He, X. S., & Loewen, S. (2022). Stimulating learner engagement in app-based L2 vocabulary self-study: Goals and feedback for effective L2 pedagogy. *System*, *105*, 102719. https://doi.org/10.1016/j.system.2021.102719

Hui, B., Rudzewitz, B., & Meurers, D. (2023). Learning processes in interactive CALL systems: Linking automatic feedback, system logs, and learning outcomes. *Language Learning & Technology*, *27*(1), 1–23. https://hdl.handle.net/10125/73527

Hui, B. & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*(5), 1089–1115. https://doi.org/10.1017/S0142716420000193

Hwang, H. B., Coss, M. D., Loewen, S., & Tagarelli, K. M. (2024). Acceptance and engagement patterns of mobile-assisted language learning among non-conventional adult L2 learners: A survival analysis. *Studies in Second Language Acquisition*, *46*(4), 1–27. https://doi.org/10.1017/S0272263124000354

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Lee, J. (2019). Task complexity, cognitive load, and L1 speech. *Applied linguistics*, *40*(3), 506–539. https://doi.org/10.1093/applin/amx054

Lee, A. H., & Lyster, R. (2016). Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language learning*, *66*(4), 809–833. https://doi.org/10.1111/lang.12167

Leow, R. (1997). Attention, awareness, and foreign language behavior. *Language Learning*, *47*(3), 467–505. https://doi.org/10.1111/0023-8333.00017

Leow, R. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. Routledge. https://doi.org/10.4324/9781315887074

Leow, R. P., Thinglum, A., & Leow, S. A. (2022). WCF processing in the L2 curriculum: A look at type of WCF, type of linguistic item, and L2 performance. *Studies in Second Language Learning and Teaching*, *12*(4), 651–673. https://doi.org/10.14746/ssllt.2022.12.4.6

Li, C., Dewaele, J. M., Pawlak, M., & Kruk, M. (2022). Classroom environment and willingness to communicate in English: The mediating role of emotions experienced by university students in China. *Language Teaching Research*, Advance online publication. https://doi.org/10.1177/13621688221111623

Li, C., Wei, L., & Lu, X. (2024). Task complexity and L2 writing performance of young learners: Contributions of cognitive and affective factors. *The Modern Language Journal.* Advance online publication. https://doi.org/10.1111/modl.12954

Lively, S. E., Pisoni, D. B., van Summers, W., & Bernacki, R. H. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *Journal of the Acoustical Society of America*, *93*(5), 2962–2973. https://doi.org/10.1121/1.405815

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(S1), 185–207. https://doi.org/10.1111/lang.12117

Loerts, H., Loewi, W., & Seton, B. (2020). *Essential statistics for applied linguistics: Using R or JASP*. Red Globe Press.

Makowski, D., Ben-Shachar, M., Lüdecke, D. (2019). BayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of memory and language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, *22*(1), 6–27. https://doi.org/10.1037/met0000086

Öztürk, G. (2023). The relationship between reading and listening anxieties in EFL classrooms: Exploring the mediating effect of foreign language classroom anxiety. *Annual Review of Applied Linguistics*, 1–15. https://doi.org/10.1017/S0267190523000107

Pellicer-Sánchez, A., Tragant, E., Conklin, K., Rodgers, M., Serrano, R., & Llanes, A. (2020). Young learners' processing of multimodal input and its impact on reading comprehension: An eye-tracking study. *Studies in Second Language Acquisition*, *42*(3), 577–598. https://doi.org/10.1017/S0272263120000091

Pinheiro, J., & Bates, D. (2024). nlme: Linear and nonlinear mixed effects models. R package version 3.1-166, https://CRAN.R-project.org/package=nlme

Puimège, E., Montero Perez, M., & Peters, E. (2023). Promoting L2 acquisition of multiword units through textually enhanced audiovisual input: An eye-tracking study. *Second Language Research*, *39*(2), 471–492. https://doi.org/10.1177/02676583211049741

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr., C. (2012). *Psychology of reading*. Psychology Press.

Révész, A., Kourtali, N. E., & Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning*, *67*(1), 208–241. https://doi.org/10.1111/lang.12205

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition*, *38*(4), 703–737. https://doi.org/10.1017/S0272263115000339

Rosa, E. M., & Leow, R. P. (2004). Awareness, different learning conditions, and second language development. *Applied Psycholinguistics*, *25*(2), 269–292. https://doi.org/10.1017/S0142716404001134

Rosseel Y (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*(4), 422–445. https://doi.org/10.1037/1082-989X.7.4.422

Skehan, P. (2014). *Processing perspectives on task performance*. John Benjamins.

Sparks, R. L., & Alamer, A. (2022). Long-term impacts of L1 language skills on L2 anxiety: The mediating role of language aptitude and L2 achievement. *Language Teaching Research*, 1–22. https://doi.org/10.1177/13621688221104392

Tan, L. C., & Yap, M. J. (2016). Are individual differences in masked repetition and semantic priming reliable? *Visual Cognition*, *24*(2), 182–200. https://doi.org/10.1080/13506285.2016.1214201

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05

Teimouri, Y., Plonsky, L., & Tabandeh, F. (2022). L2 grit: Passion and perseverance for second-language learning. *Language Teaching Research*, *26*(5), 893–918. https://doi.org/10.1177/1362168820921895

Tsang, A. & Dewaele, J. (2023). The relationships between young FL learners' classroom emotions (anxiety, boredom, & enjoyment), engagement, and FL proficiency. *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2022-0077

Tykto, T., Panza, N., & Hui, B., (2024). The effect of multimodal input on L2 learners' reading comprehension: A pre-registered, eye tracking study. *Open Science Framework*. https://osf.io/hq57n/?view_only=ed878de940e44cadade885f21efeea45

Vasylets, O., Gilabert, R., & Manchón, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning*, *67*(2), 394–430. https://doi.org/10.1111/lang.12228

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P., Paananen, T., & Gelman, A. (2024). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.8.0, https://mc-stan.org/loo/

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P. C. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC (with discussion). *Bayesian Data Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Venables, W.N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Vuorre, M., & Bolger, N. (2018). Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience. *Behavior Research Methods*, *50*, 2125–2143. https://doi.org/10.3758/s13428-017-0980-9

Woll, N. (2018). Investigating dimensions of metalinguistic awareness: What think-aloud protocols revealed about the cognitive processes involved in positive transfer from L2 to L3. *Language Awareness*, *27*(1-2), 167–185. https://doi.org/10.1080/09658416.2018.1432057

Xu, T. S., Zhang, L. J., & Gaffney, J. S. (2022). Examining the relative effects of task complexity and cognitive demands on students; writing in a second language. *Studies in Second Language Acquisition*, *44*(2), 483–506. https://doi.org/10.1017/S0272263121000310

Yamagata, S., Nakata, T., & Rogers, J. (2022). Effects of distributed practice on the acquisition of verb-noun collocations. *Studies in Second Language Acquisition*, *45*(2), 1–27. https://doi.org/10.1017/S0272263122000225

Zalbidea, J. (2021). On the scope of output in SLA: Task modality, salience, L2 grammar noticing, and development. *Studies in Second Language Acquisition*, *43*(1), 50–82. https://doi.org/10.1017/S0272263120000261

Cite this article: Jia, R., & Hui, B. (2025). Modeling relationships between learning conditions, processes, and outcomes: An introduction to mediation analysis in SLA research. *Studies in Second Language Acquisition,* 1–36. https://doi.org/10.1017/S0272263125100867