Experimental and quasi-experimental designs

John Rogers & Andrea Révész

Introduction

Researchers within the field of applied linguistics have long used experiments to investigate cause-effect relationships regarding the use and learning of second languages (L2s). In experimental research, one or more variables are altered and the effects of this change on another variable are examined. This change or experimental manipulation is usually referred to as the treatment. Researchers typically draw upon either experimental or quasi-experimental research designs to determine whether there is a causal relationship between the treatment and the outcome. This chapter outlines key features and provides examples of common experimental and quasi-experimental research designs. We also make recommendations for how experimental designs might best be applied and utilized within applied linguistics research.

Experimental and quasi-experimental research

Experimental and quasi-experimental research designs examine whether there is a causal relationship between independent and dependent variables. Simply defined, the *independent variable* is the variable of influence and the *dependent variable* is the variable that is being influenced (Loewen & Plonsky, 2016). In other words, the independent variable is expected to bring about some variation or change in the dependent variable. For example, in a study examining the impact of oral corrective feedback on grammatical development, corrective feedback will serve as the independent variable and grammatical development as the dependent variable. Moderating variables are another type of variable that are often of interest in experimental and quasi-experimental research. Moderating variables are defined as variables that modify the relationship between an independent variable and a dependent variable. If the previous study of corrective feedback also investigates how working memory may influence the extent to which learners benefit from feedback (e.g., Révész, 2012), working memory will function as a moderating variable in the design.

Non-experimental designs can also be used to investigate cause-effect relationships between independent and dependent variables, but there are a number of defining features that mark true experimental research. True experiments involve the manipulation of one or more independent variables, and the dependent variables are carefully measured, typically in the form of pre- and post-testing. True experiments also include a control group and an

experimental group. The control only takes part in the pre- and post-testing, whereas the experimental group receives the experimental treatment in addition to completing the pre- and post-testing. Finally, true experiments are characterized by random assignment, that is, participants are randomly placed into the control and the experimental condition following a chance procedure (Gravetter & Lorzano, 2018; Hatch & Lazaraton, 1991; Kirk, 2009; Loewen & Plonsky, 2016; Nunan, 1992).

The main feature that distinguishes non-experiments from true experiments is the lack of random assignment. Quasi-experiments are a subtype of non-experiments, which attempt to mimic randomized, true experiments in rigor and experimental structure but lack random assignment (Cook & Wong, 2008; Kirk, 2009). Quasi-experimental studies do not require a true control group either but may include a comparison group. A comparison group is an additional experimental group that receives a different experimental treatment. Non-experiments may also take the form of pre-experimental designs. Pre-experimental designs use neither a control nor a comparison group (Nunan, 1992). As such, experimental and quasi-experimental designs allow researchers to draw more unambiguous conclusions as to the causal relationship between two variables (Marsden & Torgerson, 2012).

The quality of experimental research is usually considered in terms of its reliability and validity. Reliability refers to the extent to which a measurement or an experimental procedure elicits consistent interpretations about the construct that it sets out to measure (Norris & Ortega, 2003). The reliability of an experimental study may suffer due to various sources of random error in measurement, including issues to do with the context, data collection procedures, characteristics of the instruments, analytical procedures, and participant idiosyncrasies (Norris & Ortega, 2003). Reliability is considered a prerequisite for validity but does not guarantee it. Validity refers to the soundness of a study (Loewen & Plonsky, 2016), that is, the degree to which the results of a study accurately answer the question that it set out to answer (Gravetter & Forzano, 2018; Révész, 2012b). Any aspect of the experiment that raises doubts as to whether the results have led to accurate and meaningful interpretations threatens the validity of the research. There are many types of validity that a researcher may wish to take into consideration when designing a research project (see Loewen & Plonsky, 2016; Mackey & Gass, 2016, and Shadish, Cook & Campbell, 2002 for an overview), two of which, internal and external validity, are of particular relevance in this chapter.

Internal validity relates to the design of the study and captures the extent to which the manipulations in the independent variable(s) (e.g., the presence/absence of treatment, different types of treatments) are responsible for the observed changes in the dependent variable. A study can claim internal validity if the results can only be explained by the

independent variable, whereas a study lacks internal validity if the results may have been influenced by factors other than the independent variable. Any extraneous factor that may allow for alternative explanation poses a threat to internal validity. Threats to the internal validity of a study may be external, such as a coincidental outside event that influences the results, or internal, including factors to do with the soundness of the research design and procedures (Campbell, 1957; McLeary, McDowell, & Bartos, 2017; Shadish et al., 2002). Steps to help ensure internal validity include careful sampling, thorough piloting of instruments and procedures, adherence to the experimental procedure, and accurate data analysis (Loewen & Plonsky, 2016; Mackey & Gass, 2015; Shadish et al., 2002).

External validity refers to the degree to which the results of a particular study hold true outside of the particular study, that is, the extent to which the results are generalizable. The generalizability of a study can be considered from various perspectives: whether the results are generalizable from the research participants to the wider population, from one research study to another; and from the research study to a real-world situation. External validity should not be assumed and is best controlled through replication (see Marsden, this volume; also McLeary et al., 2017; Porte, 2012; Shadish et al., 2002).

It is widely acknowledged that in experimental research there is a constant tension between internal and external validity (Chaudron, 2003; Hulstijn, 1997). For example, psycholinguistic studies typically involve tightly controlled experimental conditions to eliminate or minimize the effects of potential confounding variables (Hulstijn et al., 2014). However, as a result of emphasizing control, the experimental conditions may become so artificial and unnatural that they no longer resemble how language is used and learned in the real world, thus reducing external validity. Despite this tension, all experimental studies should strive to maximize both internal and external validity through striking a balance between sound study design and generalizability (Gravetter & Lorzano, 2018).

Common research designs in experimental and quasi-experimental research

When deciding upon an experimental design, there are a number of questions that researchers need to consider to ensure that the internal and external validity of the study are optimized. These include reflecting on the type of variables studied, the number of independent variables investigated, the absence or presence of pretesting, the number of treatment sessions required, and the size and nature of the sample to be selected. Each design option has its pros and cons, thus researchers inevitably need to make compromises in the decision-making process (Gass & Mackey, 2016). In the sections to follow, we introduce five common research designs used within experimental and quasi-experimental research, highlighting their advantages and limitations with a view to helping researchers select designs that are best suited to address

their research questions, while also taking into account constraints related to practicality and feasibility.

Pretest-posttest design

The pretest-posttest control group design is probably the most common experimental research design (Cook & Wong, 2008). In this design, the experimental group takes part in some type of treatment or intervention (marked by X in Table 1), which can consist of single or multiple training sessions. The design also includes a pretest and a posttest, in which both the experimental and control groups participate. The purpose of the pretest is to ensure the comparability of the two groups prior to the treatment; whereas the posttest allows the researchers to determine the immediate effects of the treatment on the outcome variable(s). In addition to the pre-test and immediate post-test, a delayed post-test or posttests are often included to examine the effects of the treatment over the longer term. The inclusion of the control group enables researchers to determine whether any observed changes from the pretest to the posttest in the experimental group are the result of the experimental treatment or can be attributed to other influences such as testing effects or maturation. As both experimental and the control group take the tests at the same time, time-related confounds are minimized (Gravetter & Forzano, 2018).

Table 1. Pre-test / Post-test control group design

Experimental Group	O X O
Control Group	O O

There are several considerations when designing the testing sessions. Regarding timing, it has been recommended that the pretest is administered a minimum of one week prior to the treatment session (Hulstijn, 2003) to decrease the likelihood that the effects of the treatment are confounded by testing effects that may arise from completing the pretest. The immediate post-test is typically administered immediately following the treatment phase of the experiment. The timing of the delayed post-tests varies; delayed posttests can be administered one week, one month, or even several months following treatment. In terms of content and procedures, each testing session (pretest, posttest, and delayed posttest) should be comparable. Within a testing session, single or multiple outcome measures may be employed. While single outcome measures are more practical to administer, the use of multiple outcomes measures, if carefully selected, are likely to provide a fuller picture of second

language development (e.g., Webb, 2005). An example of a study employing a pretest-posttest design is provided in the example below.

> Example 1. A pretest-posttest design
> Experiment: Peters and Webb (2018, Experiment 1) utilized an experimental pretest-posttest design to examine the effect of TV viewing on the incidental learning of L2 vocabulary.
> Independent variable: viewing versus not viewing L2 television
> Dependent variable: form recognition and meaning recall of L2 vocabulary
> Design: The participants, Dutch learners of L2 English, were randomly assigned to either a true control group (n=27) or an experimental group (n=36). The experiment consisted of three sessions: a pre-testing session (one week prior to treatment), the treatment session, and a post-testing session (administered one week following treatment). The control group only took part in the testing sessions. The experimental group, in addition to completing the pretest and posttest, participated in a treatment, as part of which they viewed a TV programme.

Despite its utility and practicality, there are some limitations to the pretest-posttest design. A main issue is that the pretest may sensitize participants to the focus of the experiment, and this, in turn, may influence the results. To give an example, if participants notice that the pretest assesses their vocabulary knowledge, they might be inspired to pay more attention to vocabulary during the treatment. One way to control for this possibility is to include distractor items in the tests. This, however, has the obvious practical disadvantage of prolonging the length of the testing sessions. Another potential threat to the validity of this design is that participants in the control and experimental groups may communicate about the study outside the experiment, which might also contaminate the findings. Finally, a pretest-posttest design can only provide a limited picture of the L2 learning process. Longitudinal designs, such as the time series design, are more suitable to capture the effects of longer-term treatments on L2 development.

Time-series design

A time-series design is an example of longitudinal design in which researchers collect samples of language on a regular basis over a set period (Kirk, 2009; Mellow, Reeder, & Foster, 1996). By collecting data on multiple occasions, time-series designs can allow insight into the time course of language development, including changes that may be immediate, gradual, delayed, incubated, or residual (Mellow et al., 1996; Mellow, 2012) as well as the permanency of any effects resulting from a treatment. A time series design is characterized by multiple observations both before and after the treatment. The number of pre-treatment

and post-treatment observations can vary, and there is no need to have the same number of observations pre- and post-treatment (Kirk, 2009). The treatment may entail a single or multiple treatment sessions. Whether involving a single or multiple trainings, the treatment can vary in length, from including brief to extended sessions. Table 2 provides an illustration of a time series design, with a single treatment and eight observations, four before the treatment and four after the treatment.

Table 2. Time series design with a single treatment

Experimental Group	O1 O2 O3 O4 X O5 O6 O7 O8
Control / Comparison Group	O1 O2 O3 O4 O5 O6 O7 O8

An example time series design by Ishida (2004), is described below:

> Example 2: A time series design
> Experiment: Ishida (2004) utilized a time series design to investigate the impact of recasting on development in the use of the Japanese te-i-(ru) construction.
> Independent variable: presence versus absence of recasting
> Dependent variable: accuracy in the use of the Japanese te-i-(ru) construction, as reflected in accuracy rates during oral performance
> Design: The participants were four learners of L2 Japanese, who took part in eight 30-min one-on-one conversation sessions. The first two sessions served as the pretest, the middle four as the treatment, and the last two as the posttest. Two participants also participated in a delayed posttest seven weeks after the last posttest. The treatment involved providing recasts in response to errors in the use of the Japanese -te i- (ru) construction.

The use of multiple pre- and post-tests in time-series designs is instrumental in increasing the internal validity of the findings. The multiple pretests enable researchers to test whether there are any trends in the data before the treatment session. If trends are observed prior to the treatment, this indicates that the posttest scores might be influenced by factors other than the treatment, such as testing effects, fatigue and maturation (Gravetter & Forzano, 2018). Similarly, the multiple posttests make it possible to obtain a richer account of L2 development than a single posttest would allow for. It is possible, for instance, that a treatment only has a temporary effect that fades over time (Mackey & Gass, 2016), which can only be captured if multiple posttests are included in the design.

Time-series designs, however, fare less well in terms of external validity. Due to the larger number of observations and the richer analysis of language development they make possible,

time-series designs usually include a smaller number of participants than quantitative designs with fewer observational points. This inevitably has a negative impact on the generalizability of the findings to the wider population.

Latin square design

A Latin square design is frequently used within experiments that utilize multiple data collection instruments. This design can be traced back to Fischer (1925) and gets its name from an ancient puzzle that was concerned with the number of ways that Latin letters can be arranged in a square matrix so that each letter appears once in each row and once in each column (Kirk, 2009). A Latin square is a table made with the same number of rows and columns that can be used to counterbalance data collection instruments and to help control against test- and task-order effects (see Richardson, 2018 for a recent review). Simply put, in a Latin square design, the ordering of instruments (e.g., tests or tasks) are different for various participants or groups of participants. For instance, Lambert, Kormos and Minn (2017) used a Latin square design to investigate the effects of task repetition on L2 oral fluency. Participants carried out four different tasks, three monologue tasks and an opinion dialogue task. To make sure that the order of the tasks does not influence the results, the participants were randomly assigned to four groups. Each group completed the four tasks in a different order following a Latin square design, as shown in Table 3. Latin squares are also commonly employed when multiple versions of tests are included in a study. For example, to avoid practice effects, studies with pretest-posttest-delayed posttest designs often use three versions of all testing instruments, and these are typically administered in a Latin square design across participants in the testing sessions. Of course, besides counterbalancing instruments, Latin square designs can be applied in studies whose primary goal is to examine task or test order effects.

Table 3. Example of a Latin square design (Lambert et al., 2017)

Groups	Task order			
1	Instruction monologue	Narration monologue	Opinion monologue	Opinion dialogue
2	Narration monologue	Opinion monologue	Opinion dialogue	Instruction monologue
3	Opinion monologue	Opinion dialogue	Instruction monologue	Narration monologue
4	Opinion dialogue	Instruction monologue	Narration monologue	Opinion monologue

Repeated measures design

Repeated measures designs, also known as within-participants designs, are characterized by a single group of participants who take part in all the different treatment conditions and/or are measured at multiple times (Abbuhl & Mackey, 2017; Gravetter & Forzano, 2018). In a within-participants design, the participant is subjected to all levels of the independent variable. This design derives its name from the fact that the design involves 'repeated' measurements of the same participant. Within-participants designs differ from between-participants designs, where the treatment conditions are assigned to different groups of participants, that is, different participants are tested on the various levels of the independent variable. Lambert et al.'s (2017) study that was presented earlier also constitutes an example of a repeated measures design as all participants completed all four tasks. Another example of a study adopting a repeated measures design, Rogers and Cheung (2018), is given below.

> Example 3. A pretest-posttest within-participants design
>
> Rogers and Cheung (2018) investigated the impact of spacing on L2 vocabulary learning in an authentic classroom setting.
>
> Independent variable: temporal spacing of treatment sessions (1 day versus 8 days)
>
> Dependent variable: learning of English adjectives, measured by performance on a multiple-choice picture identification task
>
> Design: The participants were Cantonese primary school students of L2 English in four different intact classes. They were taught half of the target vocabulary items under spaced-short conditions (1 day between treatment sessions) and half of the items under spaced-long conditions (8 days between treatment sessions). The items were counterbalanced across the two treatment conditions. All participants took part in the pretest and posttest as well as the treatment.

In this study, rather than assigning each of the four participating classes to a different experimental condition, the researchers manipulated the independent variable within participants, that is, each class studied half of the target items under one experimental condition and the other half under another experimental condition.

There are several advantages and disadvantages associated with the use of repeated measures designs. This type of design is advantageous in that it helps control for potential confounds, such as class effects and individual differences between learners, which might arise from the lack of randomized assignment in quasi-experimental research or low group sizes in true experimental studies. Given that different measurements come from the same individuals, groups equivalence can automatically be assumed. An additional benefit of repeated measures designs is that fewer participants are needed to attain sufficient power, as compared to between-participants designs. A disadvantage is that repeated measures designs may be affected by order effects, that is, the results might, at least in part, be attributed to the order in

which the different types of treatment conditions are administered rather than the difference in the conditions themselves. For example, results may deteriorate due to fatigue and boredom or improve as a result of more practice and task familiarity. Such order effects may be reduced by counterbalancing treatment conditions across participants (Rogers, 2017), for example, through adopting a Latin-square design.

Factorial design

Factorial designs include more than one independent variable, that is, factorial designs are employed to investigate the effects of two or more independent variables on the dependent variable. The independent variables in a factorial design are also referred to as factors. Factorial designs allow researchers to examine not only the impact of each independent variable separately but also the combined effects of the independent variables on the dependent variable. The separate effects of the independent variables are described as main effects and their combined effects are referred to as interaction effects. In factorial designs, a notation system is used to denote the number of levels associated with each independent variable. For instance, in a 2 x 3 design, there are two independent variables or factors: the first factor has two levels and the second factor has three. Factorial designs can include between-participants or within-participants factors only or can combine between- and within-participants factors. Factorial designs that include both between-participants and within-participants factors are usually described as mixed factorial designs.

Zalbidea (2018) provides a recent example of a study utilizing a factorial design. The researcher employed a mixed 2 x 2 factorial design to examine the impact of task complexity and modality on L2 performance. The two independent variables were task complexity, a within-participants factor, and modality, a between-participants variable. As shown in Table 4, each of the two independent factors had two levels (task complexity: simple versus complex; modality: written versus spoken). Task complexity was counterbalanced across participants to avoid order effects. Through adopting a factorial design, Zalbidea was not only able to examine the impact of modality and task complexity independently but also tease out how these independent factors interacted in influencing task performance.

Table 4. Example of mixed 2 x 2 factorial design, based on Zalbidea (2018)

	Modality			
	Written modality (N=16)		Spoken modality (N=16)	
Order of task performance	Complex Task	Simple Task	Complex task	Simple Task
	Simple task	Complex task	Simple Task	Complex task

Considerations when designing an experiment

A full discussion of all the decisions to be made when designing an experimental or quasi-experimental study is beyond the scope of the current chapter (see Mackey & Gass, 2016, for a fuller discussion). However, there are several key considerations that we would like to highlight here.

Assignment of participants to experimental conditions

Randomized experimental designs are considered the gold standard for research investigating causal relationships (Cook & Wong, 2008). As such, randomized assignment is preferred over non-randomization in that it eliminates systematic differences that may preexist among groups (Kirk, 2009; Plonsky, 2017). It is not surprising, therefore, that in some research domains non-randomized designs have systematically been shown to result in smaller effect sizes than experimental research, presumably due to extraneous factors that are less closely controlled in the absence of randomization (e.g., Bloom, Michalopoulos, & Hill, 2005). However, in applied linguistics research, random assignment is not always possible due to reasons of practicality and/or ethical concerns. Further, randomization might not be appropriate given the objectives of the research. For instance, instructed second language acquisition researchers often wish to trial instructional interventions in authentic learning environments involving the use of intact classes. Clearly, the lack of random assignment in such cases may open the door for potential confounds that can limit the internal validity of the study. However, the resulting threats to internal validity may be offset by the enhanced ecological validity afforded by conducting research in a context that closely resembles natural classroom environments to which the results are meant to be generalized (Mackey, 2017). To conclude, when deciding whether to randomize or non-randomize participant assignment,

researchers need to carefully consider the objectives of the study, while taking account of potential practical constraints and ethical issues.

Control or comparison group

Another key consideration is whether to include a true control group or a comparison group in quasi-experimental research. While the use of a control group is generally recommended, it is often not possible to include a true control group in quasi-experimental research for practical or ethical reasons (e.g., Mackey & Gass, 2016; Plonsky, 2017). It is also worth noting that, in some circumstances, the inclusion of a comparison group might, in fact, be the preferred option. For instance, as mentioned above, when researchers investigate the impact of a particular instructional intervention, they may decide that intact classes constitute the most ecologically valid setting for the research to take place (see Mackey & Gass, 2016; Mackey, 2017; Plonsky, 2017 for discussions). In this case, a comparison group, engaged in normal classroom instruction, may serve as the best baseline to the experimental condition. Using a comparison rather than a control group might also offer advantages in some experimental contexts. For instance, Hamrick and Sachs (2017) have argued that the use of a trained control (i.e., comparison) group rather than a true control group may help control for hidden bias among participants in experimental SLA research utilizing artificial language systems.

Controlling for extraneous variables

The hallmark of experimental and quasi-experimental is using strict experimental control to maintain the internal validity of the findings. As such, researchers should take care to control for extraneous variables, and document how they have done so when reporting their research. Researchers can help guard the internal validity of their research design in several ways. Some key methods include employing random assignment to avoid selection bias, using a control and/or a comparison group to control for the effects of testing, using multiple pre- and post-tests to asses pre-existing trends and gain a fuller picture of longer-term treatment effects, establishing that test versions designed to be parallel are indeed comparable, piloting instruments and procedures, and reducing test- and task-order effects.

Reporting

Finally, it is also worth considering what details to include when writing up an experimental study. A general rule of thumb is that the description of the methodology should be sufficiently detailed to enable replication. To achieve this, it is essential to include details about the sampling procedures, the sample, number and timing of the treatment and testing sessions (both duration and amount of time between sessions), the instruments used in the treatment and testing sessions, and the steps and procedures followed. It is also important to

highlight how potential extraneous variables were controlled for. Although the importance of detailed reporting is widely acknowledged in the field of applied linguistics, crucial methodological details are often left unaccounted for in published research. For example, published research studies often do not include information about the number and length of treatment sessions and the amount of time separating them. Given that the frequency and duration of treatment sessions and the interval between them has been shown to influence learning and retention (Rogers, 2017), it is recommended that researchers include such details when writing up reports on experimental research.

Conclusion

This chapter has reviewed basic concepts in experimental and quasi-experimental research and outlined a number of experimental designs that are commonly used in the field of applied linguistics. It is hoped that the descriptions, discussion, and examples provided here will help applied linguists to deepen their understanding of the tools and methods available to them, with the ultimate goal of enhancing the quality of future experimental and quasi-experimental research in the area.

References

Abbuhl, R., & Mackey, A. (2017). Second language acquisition research methods. In K. King & N.H. Hornburger (Eds.), *Research methods in language and education* (3rd ed., pp. 183–193). New York: Springer.

Bloom, H.S, Michalopoulos, C., & Hill, C. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In H.S. Bloom (Ed.), *Learning more from social experiments* (pp. 173–235). New York: Russell Sage Foundation.

Campbell, D. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297–312.

Chaudron, C. (2003). Data collection in SLA research. In C.J. Doughty & M.H. Long (Eds.), *The handbook of second language acquisition* (pp. 762–828). Malden, MA: Wiley-Blackwell.

Cook, T., & Wong, V. (2008). Better quasi-experimental practice. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The Sage handbook of social research methods* (pp. 134–164). London: Sage.

Fischer, R. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Gravetter, F., & Forzano, L. (2018). *Research methods for the behavioral sciences*. Boston: Cengage.

Hamrick, P., & Sachs, R. (2018). Establishing evidence of learning in experiments in artificial linguistic systems. *Studies in Second Language Acquisition*, *40*(1), 153–169.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston: Heinle & Heinle.

Hulstijn, J.H. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition*, *19*(2), 131–143.

Hulstijn, J.H. (2003). Incidental and intentional learning. In C.J. Doughty & M.H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Malden, MA: Blackwell.

Hulstijn, J.H., Young, R. F., Ortega, L., Bigelow, M., DeKeyser, R., Ellis, N.C., Talmy, S. (2014). Bridging the gap: Cognitive and Social Approaches to Research in Second Language Learning and Teaching. *Studies in Second Language Acquisition*, *36*(3), 361–421.

Ishida, M. (2004). Effects of recasts on the acquisition of the aspectual form–*te i-(ru)* by learners of Japanese as a foreign language. *Language Learning, 54*, 311-94.

Kirk, R. (2009). Experimental design. In R.E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 23–45). London: Sage.

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196.

Loewen, S., & Plonsky, L. (2016). *An A-Z of applied linguistics research methods*. London: Palgrave.

Mackey, A., & Gass, S. (2016). *Second language research: Methodology and design*. New York: Routledge.

Mackey, A. (2017). Classroom-based research. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 541–561). New York: Routledge.

Marsden, E., & Torgerson, C.J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, *38*, 583–616.

McLeary, R., McDowell, D., & Bartos, B. (2017). *Design and analysis of time series experiments*. Oxford: Oxford University Press.

Mellow, J. (2012). Time Series. In C.A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–5). West Sussex, U.K.: Wiley-Blackwell.

Mellow, J.D., Reeder, K., & Forster, E. (1996). Using the time-series design to investigate the effects of pedagogic intervention on SLA. *Studies in Second Language Acquisition, 18,* 325–350.

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M. Long (Eds.), Handbook of second language acquisition (pp. 717–761). Malden, MA: Wiley Blackwell.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition, 40,* 551-577.

Plonsky, L. (2017). Quantitative research methods. In S. Loewen & M. Sato (Eds.), *Routledge handbook of instructed second language acquisition* (pp. 505–521). New York: Routledge.

Porte, G. (Ed.) (2012). Replication research in applied linguistics. Cambridge: Cambridge University Press.

Révész, A. (2012a). Working Memory and the Observed Effectiveness of Recasts on Different L2 Outcome Measures. *Language Learning*, *62*(1), 93–132.

Révész, A. (2012b). Coding second language data validly and reliably. In A. Mackey & S.M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 203–221). Oxford: Wiley-Blackwell.

Richardson, J.T.E. (2018). The use of Latin-square designs in educational and psychological research. *Educational Research Review*, *4*, 84–97.

Rogers, J. (2017). The Spacing Effect and its Relevance to Second Language Acquisition. *Applied Linguistics*, *38*(6), 906–911.

Rogers J., & Cheung, A. (in press). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research* [online first 15 Oct 2018].

Shadish, W., Cook, T. ., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52.

Zalbidea, J. (2017). One task fits all? The roles of task complexity, modality, and working memory capacity in L2 performance. *The Modern Language Journal*, *101*(2), 335–352.