CrossMark

Annual Review of Applied Linguistics (2014), **34**, 65–101.
© Cambridge University Press, 2014, 0267-1905/14 $16.00
doi: 10.1017/S0267190514000105

Quantitative Approaches for Analyzing Longitudinal Data in Second Language Research

Khaled Barkaoui

This article discusses methods used in second language (L2) research to analyze quantitative longitudinal data. Longitudinal studies are experimental and nonexperimental studies that collect repeated measures of the same variable(s) from the same participant(s) at two or more time points. Three challenging areas in longitudinal L2 research are first discussed: study design, measurement, and data analysis and modeling. Next, various traditional and recent quantitative approaches for analyzing longitudinal data are discussed, including difference or gain scores, repeated measures univariate and multivariate analysis of variance (RM ANOVA, MANOVA), multilevel modeling (MLM), autoregressive models and latent growth curve modeling (LGCM) within the structural equation modeling (SEM) framework, item response theory (IRT), single-case research designs and time series analysis (TSA), and event history analysis (EHA). Longitudinal L2 studies published in the last 10 years are reviewed to identify trends and patterns in the use of different quantitative approaches for analyzing longitudinal L2 data, describe best data analysis practices in such research, and provide recommendations for future longitudinal L2 studies. It is argued that, when feasible and appropriate, recent approaches (e.g., MLM, LGCM) have several conceptual, methodological, and practical advantages and can stimulate the development and empirical examination of more complex questions and models concerning L2 development over time than is possible with traditional techniques.

In the first and, so far, only thorough discussion of practices and challenges in the longitudinal study of second language (L2) learning, Ortega and Iberri-Shea (2005) stressed that because L2 learning takes a long time, we need to study it over time in order to understand it (cf. Ortega & Byrnes, 2008a). Ortega and Iberri-Shea noted that many questions concerning L2 learning require a longitudinal approach, such as how long it takes L2 learners to attain proficiency in L2, when and how L2 ability develops over time, the pace and pattern of L2 development over time, and the nature and durability of the effects of L2 instructional practices on L2 learning. However, Ortega and Iberri-Shea also noted a dearth of both longitudinal L2 studies and methodological discussions about longitudinal research in applied linguistics, compared to cross-sectional research. Consequently, researchers interested in conducting longitudinal

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

research on L2 learning can find little guidance concerning the "problems about the development of L2 competencies" that L2 researchers have investigated longitudinally and the "unique challenges and strengths" that the longitudinal investigation of L2 learning entails (pp. 26–27).

Ortega and Iberri-Shea (2005) identified and reviewed 38 longitudinal L2 studies published between 2002 and 2004, highlighting in particular the design choices of these studies, the array of questions they addressed, the types of data they collected (i.e., quantitative or qualitative), and their implications for future longitudinal L2 studies. This article continues the discussion started by Ortega and Iberri-Shea of "the unique challenges and strengths" that the longitudinal study of L2 learning entails, focusing primarily on practices and issues in longitudinal L2 data analysis (p. 27). Discussing longitudinal data analysis, however, requires first defining longitudinal research.

DEFINING LONGITUDINAL RESEARCH

Ortega and Iberri-Shea (2005) listed four conditions that make a study longitudinal: study length, the presence of multiwave data collection, the conceptual focus on capturing change by design, and the focus on establishing antecedent-consequent relationships through prolonged tracking of the phenomenon in its context rather than through experimental controls or comparisons. Of these four conditions, there seems to be an agreement in the longitudinal research literature on the second and third conditions. The first condition is difficult to define, while the last requirement may not apply to experimental studies that aim to examine the effects of a particular intervention on individual change over time. Consequently, in this article longitudinal research is defined in relation to the second and third conditions only to refer to studies in which observations or measures are collected repeatedly from the same participant(s) on the same outcome variable(s)[1] over time, thus allowing researchers to describe intraindividual (and interindividual) changes over time and possibly to explain such change in terms of certain other variables (Bijeleveld et al., 1998; Menard, 2002, 2008b; Ruspini, 2002; Taris, 2000; Twisk, 2003). Such designs are sometimes referred to as *repeated measures designs*. However, the term *repeated measures* is also used to describe designs that are not longitudinal, such as designs in which each participant is observed under different conditions and/or measured on more than one outcome variable at one point in time (Bijeleveld et al., 1998).[2] In contrast, experimental studies with observations of the same participants at two or more time points (e.g., pretest and posttest) are considered longitudinal (Gustafsson, 2010).

An often-cited advantage of longitudinal research is that it provides a better basis for making claims about causal relations between variables than cross-sectional research because longitudinal data can provide information about the temporal order of the designated cause and effect variables (Bijeleveld et al., 1998; Gustafsson, 2010; Ruspini, 2002; Taris, 2000).[3] However, the chief advantage of longitudinal research is that it allows the investigation and explanation of change over time (Baltes & Nesselroade, 1979; Bijeleveld et al., 1998; Gustafsson, 2010; Preacher, Wichman, MacCallum, & Briggs, 2008; Ross, 2005;

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Ruspini, 2002; Singer & Willett, 2003; Twisk, 2003). Baltes and Nesselroade (1979) listed five rationales (and advantages) for longitudinal research compared to cross-sectional research: examining intraindividual change (i.e., whether and how individuals change over time); investigating interindividual similarities or differences in intraindividual change (i.e., whether individuals change in the same or different ways); analyzing interrelationships in behavioral change (i.e., whether certain changes are correlated with each other); analyzing causes or predictors of intraindividual change over time; and analyzing causes or predictors of interindividual similarities or differences in intraindividual change.

Despite these advantages, there are fewer longitudinal studies than cross-sectional studies, mainly because of the high cost and complexity of gathering and analyzing longitudinal data, compared to cross-sectional data (Ruspini, 2002). In fact, as Singer and Willett (2003) noted, until the 1970s, many authors (e.g., Cronbach & Furby, 1970) insisted that "researchers should not even attempt to measure change because it could not be done well" (p. 3). However, recent developments in theory, research design, measurement theory and procedures, and data analysis methods have enabled researchers to ask and address important questions about change over time (Collins, 2006; Gustafsson, 2010; Hartmann, 2005; Menard, 2002, 2008b; Singer & Willett, 2003). These developments have not solved all problems; researchers investigating L2 learning longitudinally still have to be aware of and deal with several design, measurement, and data analysis challenges unique to longitudinal research. Some of these issues are discussed in the following sections because they can significantly impact decisions concerning longitudinal data analysis and interpretation.

SELECTION OF STUDIES FOR REVIEW

To identify publications relevant to this review, the longitudinal research literature (e.g., Bijeleveld et al., 1998; Gustafsson, 2010; Ho, O'Farrell, Hoong, & You, 2006; Laursen, T. D. Little, & Card, 2012; Menard, 2002; Newsom, Jones, & Hofer, 2012; Ruspini, 2002) was first reviewed to identify key approaches and techniques for analyzing quantitative longitudinal data. Based on this review, several approaches were identified, including traditional techniques, such as difference or gain scores, the paired t-test, repeated measures univariate and multivariate analysis of variance (RM ANOVA, MANOVA), analysis of covariance (ANCOVA), and regression analysis, as well as newer, advanced approaches, such as multilevel modeling (MLM), autoregressive models and latent growth curve modeling (LGCM) within the structural equation modeling (SEM) framework, single-case research designs and time series analysis (TSA), item response theory (IRT), and event history analysis (EHA). These approaches and techniques served as a set of themes or categories for locating, reviewing, organizing, and discussing relevant longitudinal L2 studies.

Electronic searches of more than 30 peer-reviewed applied linguistics journals were then conducted using the preceding list of statistical techniques in combination with terms such as *second language*, *additional language*, and *foreign language* as key words in order to identify experimental and observational (i.e., nonexperimental) L2 studies that quantitatively measured the same dependent

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

variable(s) on the same participant(s) on two or more occasions, regardless of study length. For the traditional techniques, the search was limited to the period January 2008 to June 2013. Because it was expected that few studies had used the newer analytic approaches, the search for such studies covered the period January 2003 to June 2013. Studies identified through online searches were then reviewed to determine if they satisfy the preceding definition of longitudinal studies. This led to the exclusion of some studies that measured and examined the relationships between different dependent variables over time, such as Dale, Harlaar, and Plomin (2012) who examined the longitudinal relationships between genetic and environmental factors at time 1, on one hand, and L2 achievement at time 2, on the other. Additionally, for studies using recent analytic approaches, only studies with at least three measurement occasions were included (see justification in the next section). The final set of studies included in the review consisted of 36 studies. Given the focus of this article, each study was examined in terms of the statistical techniques used to analyze longitudinal data as well as any design and measurement considerations and issues that have a bearing on the use of these techniques. Additionally, because of length restrictions, only some of these studies are described in detail in this article . The next two sections discuss several design and measurement considerations and issues that have direct implications for the statistical analysis of quantitative longitudinal L2 data.

DESIGN CONSIDERATIONS AND ISSUES

Research designed to capture change and development over time involves a number of decisions regarding which variables and measures to include, the selection of participants, the duration of the study, the definition and measurement of time, the timing, number, and spacing of measurement occasions, and whether to measure all participants at the same time or at different times (Collins, 2006; de Vaus, 2001; Hertzog & Nesselroade, 2003; Hofer, Thorvaldsson, & Piccinin, 2012; Ortega & Iberri-Shea, 2005; Rees & Klapper, 2008; Skiba, Dittmar, & Bressem, 2008; Taris, 2000). First, in terms of sample size, both the number of individuals and the number of repeated observations per individual need to be considered carefully since they play an important role in statistical power and analysis (Curran, Obeidat, & Losardo, 2010; Shaw & Liang, 2012). Second, because "time is the fundamental predictor in every study of change," researchers need to carefully define and measure time "reliably and validly in a sensible metric" (Singer & Willett, 2003, p. 10; cf. Ortega & Byrnes, 2008b; Ortega & Iberri-Shea, 2005). As Ortega and Byrnes (2008b) cautioned, some conceptualizations that rely on "biological and institutional time scales" (p. 37) may not be appropriate when examining L2 development over time (cf. Ortega & Iberri-Shea, 2005). However, the notion of time is not explicitly discussed or problematized in longitudinal L2 studies; these studies seem to continue to rely solely on biological and institutional time scales.

A third consideration is the temporal design (Collins, 2006) of the study; that is, the duration of the study as well as the timing (when), number (how many), frequency (how often), duration (for how long), and spacing of

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

measurement occasions (how long between occasions) (Collins, 2006; de Vaus, 2001; Hofer et al., 2012; Ortega & Byrnes, 2008b; Ortega & Iberri-Shea, 2005; Rees & Klapper, 2008; Singer & Willett, 2003). Decisions about the temporal aspects of longitudinal studies should be made primarily on the basis of a theory of the phenomenon under study and how it changes over time as well as the study research questions, rather than on the basis of methodology, statistics, or logistics (Collins, 2006; de Vaus, 2001; Ortega & Byrnes, 2008b; Ortega & Iberri-Shea, 2005; Singer & Willett, 2003; Skiba et al., 2008). Different phenomena and different populations (e.g., children vs. adults) may require different levels of intensity, frequency, and spacing of measurement occasions (Ortega & Iberri-Shea, 2005; Skiba et al., 2008). In the absence of a good understanding of the temporal course of a variable, more observations spaced close together, particularly around (i.e., before, during, and after) important transition points, are better than fewer in order to obtain a reasonable view of the shape and nature of individual growth trajectories[4] (Collins, 2006; Ortega & Iberri-Shea, 2005; Singer & Willett, 2003). However, more frequent measurement can lead to participant fatigue, attrition, and practice effects, as well as change the phenomenon being studied (Collins, 2006; de Vaus, 2001).

It is worth noting here that, because of their versatility and flexibility, newer statistical approaches allow researchers to make decisions about the temporal design of their longitudinal studies that are more consistent with their theories of change (Collins, 2006). For example, researchers have traditionally used equally spaced occasions for all participants, mainly because traditional statistical methods (e.g., RM ANOVA) require evenly spaced observations occurring at the same time for all participants (Collins, 2006). However, newer statistical approaches (e.g., MLM) allow variation across and within individuals in when, how frequently, and how far apart measures are collected (Collins, 2006; Singer & Willett, 2003). This is more consistent with theory since more frequent observations need to be collected during periods of rapid nonlinear change and fewer observations, spaced further apart, need to be collected during periods of little change (Collins, 2006). These new techniques are also more powerful when examining data from studies with more than two measurement occasions. This is important because there is a consensus in the literature that addressing questions about change over time requires three or more repeated measures of the same variable for each participant (Luke, 2008; Newsom, 2012; Preacher et al., 2008; Rogosa, 1995; Ross, 2005; Singer & Willett, 2003; Willett, 1988). Increasing the number of repeated measures can improve the precision and reliability of estimates of growth; provide more details about individual growth trajectories and individual differences in growth over time; and allow the examination of the effects of various individual and contextual factors on individual growth (Preacher et al., 2008; Singer & Willett, 2003).

MEASUREMENT CONSIDERATIONS AND ISSUES

In addition to concerns about the validity, reliability, and bias of each measure used in longitudinal research, researchers face several other challenges

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

when measuring constructs and variables repeatedly over time, including the reliability of change, practice effects, measurement invariance, and regression to the mean, to name a few (Bontempo, Grouzet, & Hofer, 2012; de Vaus, 2001; Duncan, Duncan, & Strycker, 2006; Menard, 2002, 2008b; Taris, 2008). Reliability in longitudinal research concerns both the reliability of each measure and the reliability of estimates of change over time. One major challenge in longitudinal research is that the reliability and validity of any measure may change over time even when administered to the same individuals under the same conditions (Bates & Novosad, 2005; de Vaus, 2001; Ruspini, 2002; Singer & Willett, 2003; Taris, 2008). Consequently, researchers need to examine and establish measurement reliability and validity at each measurement occasion, rather than just at the beginning of the study as is typically done in longitudinal studies (Taris, 20008). The reliability of change concerns the question of how to disentangle true change in the construct being measured from unreliability in observed measures (Hartmann, 2005; Ruspini, 2002; Taris, 2000). The reliability of change estimates can be improved by increasing the number of measurement occasions for each participant, using multiple measures for each dependent variable, and improving the reliability of the outcome measures (Hartmann, 2005; Hertzog & Nesselroade, 2003; Selig & T. D. Little, 2012; Singer & Willett, 2003). Modern statistical approaches, such as SEM and IRT, can also separate change due to measurement unreliability from meaningful change.

A key validity question in the measurement of change is whether observed change over time is due to true change in the construct being measured or to change in the measurement of the construct (Bontempo et al., 2012; Llosa, 2012; Menard, 2008b; Ortega & Iberri-Shea, 2005; Pitts, Prost, & Winters, 2005; Selig & T. D. Little, 2012; Singer & Willett, 2003; Taris, 2008; van Gelderen, Schoonen, Stoel, de Glopper, & Hulstijn, 2007). If observed change is due mainly to change in measurement procedures, the validity of any claims about stability or change over time is undermined (Hartmann, 2005; Llosa, 2012; Singer & Willett, 2003; Taris, 2008; van Gelderen et al., 2007; Willett, 1988). For example, when different tasks are employed on different occasions, it is difficult to disentangle time-induced and task-induced variability (Ortega & Iberri-Shea, 2005). One strategy to avoid this problem is to use the same instrument repeatedly over time. However, this can lead to practice effects; that is, changes in observed performance that are due to prior experiences with the measurement procedures, rather than to changes in the construct being measured (Embretson & Poggio, 2012; Hertzog & Nesselroade, 2003; Hofer et al., 2012; Ortega & Iberri-Shea, 2005).

The study of change rests on the assumption of measurement invariance or equivalence over time (Hartmann, 2005; Hertzog & Nesselroade, 2003; Hofer et al., 2012; T. D. Little, 2013; Llosa, 2012; Selig & T. D. Little, 2012; Taris, 2008; van Gelderen et al., 2007). Measurement variance or nonequivalence occurs when measures do not reflect the same construct across occasions (or groups) because of practice effects and/or changes in some or all aspects of the measurement procedures over time, such as changes in item and task characteristics and difficulty, item interpretation, administration and scoring conditions, or rater severity (Bontempo et al., 2012; de Vaus, 2001; Embretson & Poggio, 2012; Hartmann, 2005). For example, in some studies, researchers may need to

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

develop multiple age- and/or grade-appropriate measures (e.g., tasks, tests) of the same construct with different characteristics, content, and/or levels of difficulty to be administered to participants at different ages or stages of development (Edwards & Wirth, 2009; Embretson & Poggio, 2012; Gustafsson, 2010; T. D. Little, 2013; Ortega & Iberri-Shea, 2005). While such an approach can create measures that can capture a wide range of inter- and intraindividual differences, avoid floor and ceiling effects, and enhance participant affective, cognitive, and linguistic investment in the tasks and tests used, it can also make comparing measures over time problematic (Gustafsson, 2010; Ortega & Iberri-Shea, 2005). SEM and IRT provide powerful tools to empirically examine the assumption of measurement invariance over time (Embretson & Poggio, 2012; Gustafsson, 2010; Hartmann, 2005; Lissitz, Doran, Schafer, & Willhoft, 2006; T. D. Little, 2013; Llosa, 2012; McCall & Hauser, 2007; Millsap & Cham, 2012; Taris, 2000, 2008; van Gelderen et al., 2007).

The threat of variance in measurement validity and reliability over time is not typically discussed or examined in longitudinal L2 studies, but recent studies conducted within the SEM framework tend to examine measurement validity, reliability, and invariance before analyzing and modeling L2 longitudinal data (e.g., Llosa, 2012; Matsumura, 2003; Schoonen, van Gelderen, Stoel, Hulstijn, & de Glopper, 2011; van Gelderen et al., 2007). Matsumura (2003) and Llosa (2012) provide detailed accounts of why and how to evaluate measurement longitudinal invariance within the SEM framework. Other longitudinal L2 studies have used IRT models to address the issue of measurement invariance over time (e.g., Kieffer & Lesaux, 2012b; Lundstrom & Baker, 2009; Revesz, 2012). Lundstrom and Baker (2009), for example, in a study that compared pretest and posttest essay scores for the same 91 participants who received two types of peer feedback, analyzed essay raw scores using an IRT model to obtain an ability estimate, adjusted for differences in rater severity, for each participant for each occasion and then used these IRT estimates, rather than raw scores, as dependent variables in all statistical comparisons across groups and occasions. Kieffer and Lesaux (2012b), in a study that involved administering different tests of the same construct (morphological awareness) to the same participants at different grades, used IRT scaling methods to link and place scores from the different tests on a common continuous scale that has the same meaning across occasions before examining changes in morphological awareness over time.

A final measurement issue is the phenomenon of regression to the mean; that is, "the tendency for extreme scores to move closer to the mean of the distribution at subsequent measurements" (T. D. Little, 2013, p. 43; cf. Bijeleveld et al., 1998; de Vaus, 2001; Taris, 2000, 2008). Regression to the mean is caused by measurement unreliability and floor and ceiling effects, which can lead to a decrease of the variance of participants' scores over time, thus distorting the findings and leading to the erroneous conclusion that observed change is due to true change or to the effect of an intervention (in experimental studies) when the change is simply an artifact of statistical regression toward the (group) mean (Bijeleveld et al., 1998; de Vaus, 2001; Hertzog & Nesselroade, 2003; T. D. Little, 2013; Taris, 2000, 2008). Several strategies can be used to avoid regression to the mean, such as including more than two measurement

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

occasions, excluding participants with extreme scores at time 1, using measures that avoid ceiling and floor effects, using multiple measures of the construct of interest, and using statistical techniques such as SEM to model error-free latent variables (Bijeleveld et al., 1998; T. D. Little, 2013; Taris, 2000, 2008; Willett, 1988). However, despite calls for the use of multiple measures of L2 proficiency (e.g., Rees & Klapper, 2008), most longitudinal L2 studies tend to use only one measure of the dependent variable, except for some studies conducted within an SEM framework.

DATA ANALYSIS CONSIDERATIONS AND ISSUES

Statistical analysis is essential in quantitative longitudinal research in order to discern and understand the shape and direction of change as well as to identify the causes and consequences of change (Preacher et al., 2008). In the past, the range of analytic techniques available to researchers was limited to difference scores and analysis of variance. However, technological and statistical advances over the last four decades have made available a wide array of statistical methods for analyzing longitudinal data (Gustafsson, 2010; Ho et al., 2006). One limitation of L2 longitudinal studies noted by Ortega and Iberri-Shea (2005) was that these studies often use traditional statistical techniques that are employed in cross-sectional research and do not benefit from recent advances in longitudinal quantitative analysis. To address this drawback, Ortega and Iberri-Shea urged researchers interested in studying L2 development over time to train themselves in the use of statistical methods for longitudinal data analysis. This article discusses the potentials and limitations of some of these methods. It is hoped that an awareness of the possibilities offered by different analytic techniques can help L2 researchers design their longitudinal studies appropriately (cf. Gustafsson, 2010).

Selecting an appropriate statistical technique to analyze longitudinal data depends on several criteria, including the research purpose and questions of the study (e.g., to describe or to explain change), number of variables, the design of the study, sample size, the number of repeated measures, characteristics of the data collection procedures, and characteristics of the data (e.g., continuous or categorical, normality; Bijeleveld et al., 1998; Menard, 2002, 2008b). Most importantly, as Collins (2006) emphasized, one needs to adopt a statistical model that corresponds as much as possible to the theoretical model of change and research questions of the study; a mismatch of theoretical and statistical models can "result in the addressing of irrelevant or even meaningless questions" (p. 509). Another consideration when selecting statistical techniques that is unique to longitudinal studies concerns the number of participants and measurement occasions: Some techniques require large numbers of participants and/or occasions, while others do not (Bijeleveld et al., 1998; Menard, 2002, 2008b). For example, TSA is appropriate for studies with a few cases and many occasions, while LGCM and MLM are feasible for studies with a few occasions and many cases (Menard, 2002).

Before discussing the array of analytic methods that have been used in quantitative longitudinal L2 studies, the issue of data missingness needs to

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

be discussed. There are two types of data missingness in longitudinal research: missing occasions and missing values (Bijeleveld et al., 1998). Missing occasions occur when participants drop out of the study after participating in early occasions (i.e., participant attrition), while missing values occur when participants participate in all occasions but do not respond to all the questions on a test or a questionnaire, for example (Bijeleveld et al., 1998). Attrition and missing values can reduce sample size, thus reducing statistical power and restricting the range of analysis possible with the data; they can also negatively impact estimates of descriptive statistics, trends, and relationships among variables (de Vaus, 2001; Feng, Cong, & Silverstein, 2012; Menard, 2002; Rees & Klapper, 2008; Ross, 2005). For example, if those who leave the study are different on some relevant characteristics from those who stay and/or if participants in one group in a multigroup study drop out at a higher rate than those in other groups, then attrition is nonrandom and can undermine the study's internal validity because the changing composition of the sample can confound the measurement of the dependent variable (Bijeleveld et al., 1998; de Vaus, 2001; Menard, 2002; Rees & Klapper, 2008; Ruspini, 2002; Taris, 2000). How to deal with missing data depends primarily on the magnitude, pattern, and reasons of missingness (Bijeleveld et al., 1998; de Vaus, 2001; Feng et al., 2012; T. D. Little, 2013; R. J. A. Little & Rubin, 1987; Ross, 2005; Taris, 2000; Twisk, 2003; van Gelderen et al., 2007).

As Rees and Klapper (2008) noted, there has been little or no discussion of issues of data missingness in the L2 literature. Additionally, most longitudinal L2 studies, particularly those using traditional analytic techniques, rarely, if ever, discuss issues of missing data. Typically, these studies report how many participants dropped out or are excluded from the study, but they do not indicate who those participants are, how they compare to those who remained in the study, why they dropped out, or whether and how attrition and data missingness affect the findings and conclusions of the study. The most common approach to dealing with missing data in these studies is excluding participants with any missing data using listwise[5] or pairwise deletion (e.g., Bae & Lee, 2012). This method is simple but assumes that the data are missing completely at random (MCAR), a strong and untenable assumption in most longitudinal studies (Bijeleveld et al., 1998; Ross, 2005; Taris, 2000; van Gelderen et al., 2007). Recently, several authors have questioned this approach to dealing with missing data in longitudinal L2 studies (e.g., Ross, 2005; Schoonen et al., 2011; van Gelderen et al., 2007). These authors conducted their longitudinal L2 studies within an SEM framework, which allows the explicit examination of data missingness and the use of multiple imputations,[6] which assumes that data are missing at random (MAR, a weaker assumption than MCAR), to deal with missing data before modeling longitudinal L2 data. These authors provide useful discussions of issues in, and appropriate methods for, dealing with missing data in longitudinal L2 research.

DIFFERENCE SCORES

The simplest method to measure change is the difference score (i.e., gain or change score), computed by subtracting time 1 scores from time 2

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

scores for each individual. Statistical analyses are then conducted using these change scores as the dependent variable. Though intuitively simple, difference scores have often been criticized as being intrinsically unreliable. Specifically, difference scores tend to be much less reliable than each of the two measures on which they are based (Allison, 1990; Newsom, 2012; Taris, 2008). The unreliability of change scores has led many to question their use (e.g., Cronbach & Furby, 1970; Lord, 1963). Cronbach and Furby (1970), for example, strongly discouraged the use of change scores, urging researchers "who ask questions regarding gain scores" to "frame their questions in other ways" (p. 80). Over time this advice has become "conventional wisdom" and many researchers have become hesitant to use change scores (Allison, 1990; Taris, 2008; Zumbo, 1999). As Zumbo (1999) put it:

> In the nearly 30 years since Cronbach and Furby's paper, the change score has become the most maligned index of change. Cronbach and Furby created such doubts about the difference score that granting agencies, journal editors, and even committees of graduate students' theses have been known to deplore the use of change scores. (p. 271)

Many authors have recently questioned the claimed deficiencies of the change score, arguing that these scores can be quite reliable under certain conditions and can yield powerful tests of causal hypotheses despite their unreliability (e.g., Allison, 1990; Hertzog & Nesselroade, 2003; Mellenbergh, 1999; Newsom, 2012; Rogosa, 1995; Taris, 2000, 2008; Willett, 1988; Zumbo, 1999). Willett (1988), for example, demonstrated that change scores are not necessarily unreliable; that when there is true interindividual differences in growth to be detected, their reliability could be very high; and that they certainly could be a valid measure of individual growth. These authors also provide detailed guidelines concerning whether and when to use difference scores when analyzing longitudinal data (e.g., Zumbo, 1999). However, the calls for rehabilitating gain scores do not mean using these scores indiscriminately. Rather, these calls urge researchers not to always discard gain scores and to be aware of situations when they are useful, without ignoring the fact that their reliability and validity can vary depending on various factors (Zumbo, 1999).

Only one L2 study was identified that used gain scores. Green (2005) compared the International English Language Testing System (IELTS) writing scores of more than 15,000 test takers who took the IELTS twice as well as the scores of 476 learners who responded to the IELTS writing section before and after a period of L2 instruction. In both cases, Green computed a gain score for each test taker by subtracting test 1 scores from test 2 scores. These gain scores were then compared across test-taker subgroups and used as the outcome variable in regression analyses to estimate the contribution of various student background factors (e.g., age, gender, years of English study, scores on other sections of the IELTS) to variability in gain scores (cf. Elder & O'Loughlin, 2003). Green found considerable individual variation in the rate of score gain and that relatively rapid gains appear more likely for those with initial scores at the lower end of the IELTS scale.

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Another approach is to use IRT models to estimate gain scores, thus removing the effects of measurement unreliability on gain estimates. Revesz (2012), for example, in a study examining whether working memory moderates the effects of recasts on improvement in 90 L2 learners' knowledge of one usage of the English past progressive construction (measured using three types of outcome measures), first submitted participants' pretest, posttest, and delayed test scores on each outcome measure to an IRT analysis to estimate gains and then examined the correlations between working memory measures and IRT-estimated gain scores. Revesz found that recasts generated the greatest gains on an oral production test and the least gains on a grammaticality judgment test and that working memory was associated with the extent of development achieved by learners who received recasts.

TRADITIONAL APPROACHES

Traditional approaches include the dependent t-test, RM ANOVA, MANOVA, ANCOVA, and regression analysis. In a study with one group and two measurement occasions (e.g., pretest and posttest), the key question is whether the dependent variable significantly changes over time or whether the intervention causes a change from the pretest to the posttest (Twisk, 2003; Zumbo, 1999). This question is addressed using the dependent or paired-samples t-test with time as the independent factor. Storch (2009), for example, used several dependent t-tests to examine changes in various linguistic and content characteristics of essays written by 25 L2 learners before and after one semester of study in an English-medium university. Storch found that some aspects of language showed some improvement (e.g., level of informality, structure and development of ideas), while others (e.g., fluency, linguistic accuracy, complexity) showed no improvement over time.

When measures are collected three or more times from the same participants, RM ANOVA is used. RM ANOVA involves the comparison of means at two or more time points, assessing whether the dependent variable has changed significantly over time (Newsom, 2012; Weinfurt, 2000). When two or more groups and two or more measurement occasions are included, mixed-design RM ANOVA can be used to address questions about overall group effects (i.e., Is there a difference between the groups, independent of time?), overall time effects (i.e., Is there a change over time in the dependent variable, independent of the different groups?), and group by time interaction effects (i.e., Is there a difference between groups in the magnitude or direction of change over time?; Bijeleveld et al., 1998; Newsom, 2012; Taris, 2000; Twisk, 2003; Weinfurt, 2000). If the RM ANOVA result is significant and if there are more than two measurement occasions, post hoc contrasts can be used to compare mean scores across occasions and/or to examine the shape (e.g., linear, quadratic, cubic) of the relationship between the dependent variable and time (i.e., trend analysis; Bijeleveld et al., 1998; Newsom, 2012; Taris, 2000; Twisk, 2003; Weinfurt, 2000). When there are two or more groups, researchers can examine whether there is a difference in the shape of

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

the relationship between the dependent variable and time across groups (Twisk, 2003).

Five recent studies were identified that used mixed-design RM ANOVA to analyze longitudinal data from quasi-experimental studies that aimed to examine the effects of different types of feedback on L2 learning (Bitchener & Knoch, 2008, 2010a, 2010b; Lyster & Izquierdo, 2009; Yang & Lyster, 2010). Each of these studies involved at least three measurement occasions (i.e., pretest, immediate posttest, and delayed posttest); Bitchener and Knoch (2010a) included five measurement occasions (i.e., three delayed posttests). Bitchener and Knoch (2008) examined the effects of one learner factor, residence status, on the dependent variable as well. Sample sizes ranged between 25 and 144 learners, while the length of the studies ranged from 2 weeks to 10 months. The five studies used post hoc pairwise comparisons to compare mean scores across groups (e.g., intervention vs. control groups) and across pairs of measurement occasions (e.g., pretest vs. immediate posttest, pretest vs. delayed posttest) to examine the significance and magnitude of time, group, and time by group effects on the dependent variables. None of these studies used trend analysis to examine the form of change over time. All five studies justified the use of RM ANOVA by the finding that there were no significant differences between groups on the pretest. Some of these studies included more than one dependent variable, such as measures of oral and written production (e.g., Lyster & Izquierdo, 2009; Yang & Lyster, 2010), but used RM ANOVA to compare scores on each dependent variable separately across groups and occasions. None of these studies discussed issues of data missingness although some of them seem to have suffered a large attrition rate (e.g., Bitchener & Knoch, 2010a). Additionally, although they all reported checking the data for assumptions required by standard ANOVA (e.g., normality, homogeneity of variance), none discussed whether a key assumption in RM ANOVA, sphericity, was met.[7]

Another approach to analyzing RM data (e.g., if the sphericity assumption is not met) is MANOVA, which is usually used to examine the effects of one or more independent variables on two or more dependent variables (Twisk, 2003; Weinfurt, 2000). While RM ANOVA views measurement occasion as a within-subjects independent variable with as many levels as occasions, MANOVA treats each measure at each occasion as a separate dependent variable (Bijeleveld et al., 1998; Twisk, 2003; Weinfurt, 2000). Three studies were identified that used MANOVA to analyze data collected at two or more time points on more than one dependent variable from the same participants (Bae & Lee, 2012; Vandergrift & Tafaghodtari, 2010; Yasuda, 2011). Yasuda (2011) seems to have used MANOVA because the study included multiple dependent variables, rather than because each variable was measured more than once. Vandergrift and Tafaghodtari (2010) used RM MANOVA to assess change in five indicators of metacognitive knowledge over time (before, during, and after an intervention) among 106 learners of L2 French. Bae and Lee (2012) provide a detailed account of the use of MANOVA to analyze longitudinal data. Bae and Lee measured the English writing abilities of 42 children participating in an L2 program three times (at 9-month intervals) to examine L2 writing development over time in terms of various indicators (e.g., grammar, coherence, spelling). Because the study

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

included multiple dependent variables that were measured multiple times, Bae and Lee used MANOVA with RM, or double multivariate MANOVA, to compare participants' scores on the various writing indicators across the three measurement occasions. Trend analyses were also conducted to identify the shape of change in writing indicators over time.

An alternative to RM ANOVA when analyzing data from intact (i.e., nonrandomized) groups in pretest-posttest designs and when the groups differ significantly on the pretest is ANCOVA. ANCOVA uses statistical control to equate the groups at pretest when comparing them at posttest (Bijeleveld et al., 1998; Newsom, 2012). ANCOVA, thus, addresses a different question about change than RM ANOVA. Specifically, while RM ANOVA asks whether the average change between two occasions is significant, ANCOVA asks whether there would be a significant difference between the groups on the posttest, if the groups were equivalent on the pretest (Bijeleveld et al., 1998; Newsom, 2012; Twisk, 2008; Weinfurt, 2000). Consequently, RM ANOVA and ANCOVA can lead to different and sometimes contradictory findings concerning the same RM data (Bijeleveld et al., 1998; Weinfurt, 2000).

Four studies were identified that used ANCOVA to analyze longitudinal data from quasi-experimental studies to examine the effects of different types of input (Gilmore, 2011), feedback (Sato & Lyster, 2012), approaches to teaching (Vandergrift & Tafaghodtari, 2010), and learning contexts (Serrano, Llanes, & Tragant, 2011) on various L2 learning outcomes. The length of these studies varied between a few weeks and 10 months and sample sizes between 62 and 147 participants. Each of these studies included more than one dependent variable measured before and after the intervention (i.e., pretest-posttest design). Because the groups differed significantly in terms of their pretest mean scores, ANCOVA was used to compare posttest results across groups after statistically controlling for pretest differences. Therefore, these studies did not examine change over time, although the authors did discuss their findings in relation to L2 development, progress, or growth (e.g., Gilmore, 2011; Serrano et al., 2011; Sato & Lyster, 2012). Additionally, as Weinfurt (2000) warned, the use of ANCOVA when groups differ significantly on the pretest can lead to wrong conclusions about change. Weinfurt instead recommended computing a gain score for each individual and then performing a one-way ANOVA on the mean gain scores for the groups in order to answer the more relevant question: "Did the groups undergo a very different mean change over time?" (p. 340). According to Weinfurt, ANCOVA should be used only when there are no significant differences between groups on the pretest mean scores, either as a result of random assignment of individuals to treatment or for other reasons (cf. Twisk, 2008; Zumbo, 1999).

Finally, after establishing that there are significant changes in a dependent variable over time (e.g., based on RM ANOVA results), regression analysis can be used to investigate the causes of change, which groups are more likely to change, and/or the relationship between the amount of change and a number of other variables (Newsom, 2012; Twisk, 2003; Zumbo, 1999). As noted earlier, Green (2005) computed a difference score for each participant and then used these difference scores as an outcome variable in a regression analysis with individual and contextual factors as predictors in order to explain individual

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

differences in the degree of change between two measurement occasions (cf. Elder & O'Loughlin, 2003).[8]

The traditional methods to analyzing longitudinal data previously described have several limitations. First, comparing pairs of time points (e.g., time 1 vs. time 2, time 2 vs. time 3) can reveal whether there is across-time change or not, but it does not inform us about the rate or shape of change (Taris, 2000). In fact, the use of traditional techniques in studies with two or more measurement occasions often reduces the longitudinal problem to a cross-sectional one (Ortega & Iberri-Shea, 2005; Twisk, 2003). As Ortega and Iberri-Shea (2005) cautioned, the use of traditional techniques in longitudinal L2 studies emphasizes group averages and comparisons, rather than longitudinal individual change. As a result, "often these multi-wave between-group comparisons are conceptually cross-sectional and only indirectly longitudinal" (p. 40; cf. Twisk, 2003, 2008). Second, traditional methods provide information about average (i.e., group-level) change over time, but cannot provide information about variability in growth trajectories across individuals (Newsom, 2012; Weinfurt, 2000). As Weinfurt (2000) warned, traditional methods assume that all members of a group have the same pattern of mean performance across time; any individual deviations from that pattern are attributed to error (cf. Usborne, Caouette, Qumaaluk, & Taylor, 2009; Van Geert & van Dijk, 2002). Traditional statistical techniques, therefore, encourage a view of intraindividual variability as noise or error to be removed and ignored when analyzing language development (Van Geert & van Dijk, 2002). Third, only dichotomous or categorical time-invariant variables, but not continuous or time-varying variables, can be included as between-subjects factors in RM ANOVA and MANOVA (Twisk, 2003).[9] Fourth, traditional methods are limited in their lack of control for measurement error and the ability to include intervening (i.e., mediator) variables (Ho et al., 2006). Practically, traditional methods do not permit individual variation in the intervals between measures and require balanced data; that is, each participant has to be measured on each variable an equal number of times and at the same point in time (Bijeleveld et al., 1998; Ho et al., 2006; Hox, 2002; Twisk, 2003). This often leads to listwise or pairwise deletion to create a complete data set before data analysis. The methods discussed in the following sections address most of these limitations.

MULTILEVEL MODELING

MLM is a family of statistical models for analyzing data with nested or hierarchical structure (Barkaoui, 2013; Bijeleveld et al., 1998; Hox, 2002; Kozaki & Ross, 2011; Luke, 2008; Steele, 2008). Nested data means that observations at lower levels are nested within units at higher levels, such as students nested within classrooms and classrooms nested within schools. Longitudinal designs also give rise to nested data, where observations gathered on an individual over time are nested within the person (Bijeleveld et al., 1998; Henry & Slater, 2008; Hox, 2002; Steele, 2008). MLM distinguishes between two levels of analysis in longitudinal data: level 1 observations nested in level 2 units. As Singer and Willett (2003) explained, the goal of level 1 analysis is to describe the shape

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

of each individual's growth trajectory in terms of two parameters, initial status (i.e., intercept) and rate of change over time (i.e., slope), while the goal of level 2 analysis is to detect variability in change across individuals and to determine the relationship between one or more independent variables and the shape of each individual's growth trajectory (cf. Luke, 2008; Ross, 2005; Steele, 2008).

In MLM, each individual is allowed to have his or her own growth trajectory, and individual trajectories may vary across individuals in terms of shape (e.g., linear, nonlinear), intercept, and/or slope (Luke, 2008; Steele, 2008). If interindividual variation in rates or patterns of change around the average trajectory is significant, individual variables (e.g., gender, first language [L1]) can be used to explain interindividual differences in intraindividual change at level 2 (Henry & Slater, 2008; Hox, 2002; Luke, 2008). The level 2 models, thus, examine the factors influencing the rate and shape of change in the outcome variable over time. A third level could be added if individuals are nested within a higher level unit (e.g., classrooms) in order to examine the effects of contextual factors on growth trajectories (Kozaki & Ross, 2011). MLM, thus, can address several questions about change over time such as: What is the form and structure of change in a dependent variable (e.g., L2 proficiency) over time? Is individual change linear, nonlinear, continuous, or discontinuous over time? What is the relationship between initial level and rate of change? Do individuals and groups vary significantly in their initial level and rate and shape of change over time? What individual and contextual factors explain differences in individual initial level and rate of change over time? (Bijeleveld et al., 1998; Luke, 2008; Singer & Willett, 2003).

MLM has several conceptual and practical advantages when examining longitudinal data (Henry & Slater, 2008; Hertzog & Nesselroade, 2003; Hox, 2002; Kozaki & Ross, 2011; Luke, 2008; Preacher et al., 2008; Shaw & Liang, 2012; Singer & Willett, 2003; Steele, 2008; Usborne et al., 2009). First, because it puts emphasis on individual growth, MLM allows the description and estimation not only of the average trajectory for a group, but also interindividual variation (in terms of direction, shape, and rate) in growth trajectories around this mean. Second, MLM allows examining the extent to which interindividual variation in trajectories is influenced by various individual and contextual factors. Third, MLM can assess whether within-person changes in one variable (e.g., motivation) are associated with within-person changes in another variable (e.g., L2 proficiency) and whether this association varies by group. Practically, MLM can handle both short and long time series and unbalanced data sets with varying numbers of participants across levels and groups and varying numbers, timing, and spacing of measurement occasions across individuals, which allows the use of all available data, including data from individuals with missing data on some occasions, assuming data are MAR. Finally, MLM can accommodate time-varying and time-invariant predictors of change, handle multivariate and categorical outcome variables, and accommodate more than two hierarchical levels (e.g., time nested within students nested within classrooms nested within schools).

Seven studies were identified that used MLM to analyze longitudinal L2 data to address various questions concerning L2 development over time. Two of

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

these studies (Kozaki & Ross, 2011; Usborne et al., 2009) were observational, while the other five (Gutierrez-Clellen, Simon-Cereijido, & Sweet, 2012; Kieffer & Lesaux, 2012b; Kruk & Reynolds, 2012; Shneyderman & Abella, 2009; Tong, Irby, Lara-Alecio, Yoon, & Mathes, 2010) aimed to examine the effects of particular interventions on L2 learning. These studies involved three to five measurement occasions and included variables at two or three levels; level 3 variables were either classroom or school characteristics. Sample sizes ranged between 44 and 1,682 learners, while the length of the studies ranged from 8 months to 4 years. All six studies provide accessible discussions of the application of MLM to the analysis of longitudinal L2 data, but only two, one observational and one experimental, are described here.

Kozaki and Ross (2011) examined the effects of both individual and contextual factors and their interactions on individual growth in L2 proficiency. The study included 1,682 Japanese learners (15% females) clustered in 72 L2 classes in four streams. The outcome variable was L2 proficiency measured using learners' scores on the Test of English for International Communication (TOEIC)Bridge administered on three occasions over 2 years. Independent variables at the individual level included gender, two aspiration variables (aspiration to professional pursuit [APP] and orientation to the social norm [OSM]), and two L2 learning orientations (intrinsic learning orientation [ILO] and extrinsic learning orientation [ELO]) measured using a self-assessment survey in Japanese. Independent variables at the class level included stream, two normative aspiration measures (perceived peers' normative APP and normative OSM), and two normative L2 learning orientation measures (normative ILO and normative ELO). A three-level MLM model was used to examine the variation in individual learners' L2 growth trajectories (i.e., change in L2 proficiency scores) over 2 years (level 1) and whether and how variation in these trajectories is related to individual variables (level 2) and classroom factors (level 3). The moderating[10] effects of class-level factors on the relationships between individual level variables and L2 growth were also examined. Kozaki and Ross found that there was significant variation in growth trajectories across individuals and classes: Gender and APP were significantly related to variation in growth trajectories across individuals; the association between APP and L2 growth was moderated by the class-level factor, perceived normative APP; and normative OSM had a negative effect on the growth trajectories of individuals.

Kruk and Reynolds (2012) investigated the effects of exposure to French on growth in English phonological awareness, decoding, and reading comprehension among 44 English-speaking at-risk and not-at-risk readers attending French immersion and English programs over 3 years (Grades 1 to 3). Reading instruction in French immersion was provided in both French and English, while the English program provided instruction in English only. The outcome variables included measures of phonological awareness, decoding, and reading comprehension, which were assessed five times: first in the spring of Grade 1 and then every 6 months until the spring of Grade 3. The predictors included school program (French immersion or English), reading-risk status (at-risk or not-at-risk), and time (number of months prior to the end of the final testing occasion). Covariates included pretest measures of IQ, phonological awareness, decoding,

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

and reading comprehension. Two-level MLM (occasions nested within learner) was used to examine (a) the shape of change in the outcomes over time and (b) which factors predict change. Kruk and Reynolds found that the participants experienced significant growth on all outcome measures, with children in immersion achieving higher final scores (in Grade 3) in decoding and phonological awareness and experiencing more rapid growth in decoding. At-risk readers in French immersion experienced faster growth and higher final scores in reading comprehension.

MLM has its limitations too. First, MLM assumes that the repeatedly measured variable is free of measurement error (Hox, 2002; Preacher et al., 2008). Second, though it can examine moderating (i.e., interaction) effects, currently MLM cannot examine mediating (i.e., intervening) effects (Ho et al., 2006). Third, MLM treats growth as the ultimate and only outcome and does not allow using parameters of change as predictors of other variables or processes (Henry & Slater, 2008; Hox, 2002; Preacher et al., 2008; Stoolmiller, 2008). Finally, while it can examine the impact of time-varying predictors on change in a dependent variable, currently MLM cannot examine two or more growth processes simultaneously or assess the relationships between two or more change processes (Henry & Slater, 2008; Ho et al., 2006).

STRUCTURAL EQUATION MODELING APPROACHES

SEM is a collection of statistical techniques to examine and represent the nature and strength of the relationships between one or more independent variables and one or more dependent variables (Ho et al., 2006; T. D. Little, 2013; Matsumura, 2003; Schoonen et al., 2011; Ullman & Bentler, 2004; Winke, this volume). The variables can be continuous or discrete, latent or measured.[11] The main purpose of SEM is to test specific hypotheses concerning the relationships among a number of observed and unobserved variables in experimental and nonexperimental studies (Bijeleveld et al., 1998; Matsumura, 2003; Schoonen et al., 2011). SEM involves two main components: a measurement model that specifies and evaluates the relationships between the observed variables (indicators) and the hypothetical constructs or latent variables they measure, and a structural model that specifies and evaluates the hypothesized relationships among the latent variables themselves (Bijeleveld et al., 1998; T. D. Little, 2013; Matsumura, 2003; Schoonen et al., 2011; Ullman & Bentler, 2004; Winke, this volume). There are two main approaches to examining and modeling longitudinal data within the SEM framework: autoregressive or (Markov) simplex models and LGCMs (Bollen & Curran, 2004; Bollen & Zimmer, 2010; Curran & Bollen, 2001). Both approaches benefit from the many advantages of SEM, particularly the ability to use multiple indicators of each latent variable; estimate and correct for time-specific measurement error; examine measurement invariance over time; use variables simultaneously as independent and dependent variables; assess relationships at the level of the construct, rather than at the level of observed variables; and model moderating and mediating effects (Curran, 2003; Schoonen et al., 2011; Selig & T. D. Little, 2012; Ullman & Bentler, 2004).

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

AUTOREGRESSIVE SEM MODELS

In an autoregressive or (Markov) simplex model with one latent variable, the value of the variable at each time point is modeled as being influenced primarily by the value of that same variable at the preceding time point (Bollen & Curran, 2004; Bollen & Zimmer, 2010; Curran & Bollen, 2001; Gustafsson, 2010; T. D. Little, 2013; Selig & T. D. Little, 2012). This model, thus, takes into account the time dependence between measurements. As Selig and T. D. Little (2012) explained, by examining individuals' standings on the latent variable across adjacent occasions, the autoregressive model describes the stability of the construct from one occasion to the next. Autoregressive effects will be large, indicating a stable construct, when individuals' relative standings on the construct do not change over time. Autoregressive effects will be small, indicating construct change, when individuals' relative standings change substantially. By allowing status at time 1 to predict status at time 2, status at time 2 to predict status at time 3, and so on, the autoregressive model answers the question: "Does an individual who scores high at the beginning of time remain high on all future occasions?" (Singer & Willett, 2003, p. 280). When a study includes more than one latent variable, cross-lagged effects can be examined as well in an autoregressive cross-lagged model. Cross-lagged effects refer to the unique effect of a variable (e.g., *X*) measured at an earlier occasion (e.g., time 1) on another variable (e.g., *Y*) measured at a later occasion (e.g., time 2), controlling for the autoregressive effects of that variable at time 1 on itself at time 2 (Curran & Bollen, 2001; Gustafsson, 2010; T. D. Little, 2013; Selig & T. D. Little, 2012). The substantive interpretation of the cross-lagged parameter is that an earlier measure of *X* predicts a later measure of *Y* above and beyond the autoregressive prediction of *Y* from itself at an earlier occasion (Bollen & Curran, 2004; Curran & Bollen, 2001). The autoregressive SEM model was first used in L2 research in the early 1990s by Verhoeven (1994).

The autoregressive model has several advantages. First, this model is especially useful for identifying the relations between variables across time and for examining the direction of causation (e.g., whether X1 predicts Y2 and Y1 predicts X2) and reciprocal effects over time, such as the effect of motivation at time 1 on achievement at time 2 and the effects of the latter on motivation at time 3 (Gustafsson, 2010; Hertzog & Nesselroade, 2003; Selig & T. D. Little, 2012). As Selig and T. D. Little (2012) noted, when there is neither strong theory for the change process nor a strong expectation for the trajectory of change over time or when the research question is about the pattern of influences, rather than the direction of change, the autoregressive model can be particularly useful. Second, this model allows estimating the stability of individual differences over time (Bast & Reitsma, 1997; Selig & T. D. Little, 2012). Third, the simplex model is particularly well suited to longitudinal series in which there is occasion-to-occasion transmission (e.g., values of measures at time 3 depend on values at time 2, which in turn depend on values at time 1) (Bast & Reitsma, 1997). Finally, T. D. Little (2013) noted that the simplex model is most useful when change is gradual and consistent over time; that is, when individuals change at a steady rate, with minimal outside influences on the rate of change.

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Four studies were identified that used the autoregressive cross-lagged model to analyze longitudinal L2 data (Matsumura, 2003; Schoonen et al., 2011; van Gelderen et al., 2007; Verhoeven & Leeuwe, 2012). Each of these studies involved three measurement occasions, but they varied in terms of length between 6 months and 6 years and in terms of sample size between 137 and 1,700 learners. Matsumura (2003) provides one of the most detailed and accessible discussions and illustrations of the use of the autoregressive SEM model for examining longitudinal L2 data. Matsumura examined change over time in pragmatic use of English among 137 Japanese English as a foreign language (EFL) learners as a function of their varying levels of English proficiency and amount of exposure to English. Questionnaire responses and scores from the Test of English as a Foreign Language (TOEFL) were collected three times at 3-month intervals from each participant. Matsumura described in detail the process of developing and testing various models within the SEM framework in order to examine measurement validity, reliability, and invariance, and to address the research questions of the study. Matsumura specified, assessed, and compared four SEM models in order to identify the model that best described the relationships among L2 pragmatic development, L2 proficiency, and L2 exposure. All four models included autoregressive effects that represent the relationships between the same latent variables (e.g., L2 pragmatic competence, L2 proficiency) over time. Cross-lagged effects were also included in some models (e.g., effects of amount of L2 exposure at time 1 on L2 proficiency at time 2). The models also included correlated measurement errors to control for potential sources of covariance among measures such as the potential of practice or memory effects. Matsumura found that L2 pragmatic competence had a significant and persistent effect on itself over the duration of the study and that the amount of L2 exposure had greater potential to account for L2 pragmatic development than did level of L2 proficiency.

Schoonen et al. (2011) provide another detailed and accessible account of how the autoregressive model can be used to examine longitudinal L2 data. Schoonen et al. examined the relationships among writers' linguistic knowledge, linguistic processing efficiency, and metacognitive knowledge, on one hand, and writing proficiency, on the other hand, over time, in L1 (Dutch) and L2 (English) among 389 students in order to determine whether these relationships change over time and whether they are fundamentally different for L1 and L2 writing contexts. Participants completed three writing tasks in L1 and three parallel writing tasks in L2 each year in three consecutive years (Grades 8 through 10). Furthermore, data were collected each year about students' L1 and L2 linguistic knowledge, metacognitive knowledge, and efficiency in processing L1 and L2 linguistic knowledge. Various autoregressive cross-lagged models were specified and evaluated in order to investigate whether L1 and L2 writing proficiency improves over time; the extent to which initial levels of writing proficiency are predictive of later levels of proficiency; the relative contributions of linguistic knowledge, processing efficiency, and metacognitive knowledge to L1 and L2 writing development; the developmental differences between L1 and L2 writing with respect to the roles of linguistic knowledge, processing efficiency, and metacognitive knowledge; and the relationship between L1 and L2 writing proficiency over time. Schoonen et al. found that students' L2 writing proficiency

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

improved to a greater extent than did their L1 writing proficiency over the 2 years; L1 writing proficiency was stable with individual differences remaining largely the same over time (i.e., high autoregressive correlations over time); students differed in the pace of their L2 writing development from time 1 to time 2 (i.e., low autoregressive correlation between two occasions); individual differences in L2 writing proficiency did not change significantly from time 2 to time 3 (i.e., high autoregressive correlation); there was a strong relationship between L1 and L2 writing proficiency at time 1, with metacognitive knowledge and general fluency mediating this relation; this relationship did not change substantially over time; and all the subskill variables in the study were correlated with writing proficiency in L1 and L2.

A major limitation of the autoregressive model is that it does not allow the examination of the shape and rate of growth over time, individual differences in growth trajectories over time, the effects of individual and contextual factors on the shape and rate of growth, or the relationships between two or more growth processes (Collins, 2006; Duncan et al., 2006; Gustafsson, 2010; Hertzog & Nesselroade, 2003; Preacher et al., 2008; Selig & T. D. Little, 2012). As Singer and Willett (2003) cautioned, the autoregressive model does not address questions about individual change over time; rather, it assesses the stability of rank order of individuals on the outcome over time. Consequently, this model may not capture change that does not affect the rank ordering of individuals (Selig & T. D. Little, 2012). The autoregressive model is also limited in that it focuses on change between two adjacent time points, even when there are multiple measurement occasions; provides the same autoregressive and cross-lagged effects for all individuals in the sample; and considers change over time in terms of each variable depending on its immediately prior value but not its values for earlier time points (Curran & Bollen, 2001; Gustafsson, 2010). Because of these limitations, the autoregressive model is a poor choice when the goal is to examine the form, precedents, and consequences of individual growth trajectories (Curran & Bollen, 2001; Preacher et al., 2008; Selig & T. D. Little, 2012; Singer & Willett, 2003).

LATENT GROWTH CURVE MODELING

Like MLM, LGCM allows researchers to model and estimate a mean growth trajectory for a group on a dependent variable, variability in individual trajectories around the mean trajectory, and the extent to which specific time-invariant and time-varying variables predict individual variability in growth trajectories (Curran et al., 2010; Duncan et al., 2006; Henry & Slater, 2008; Jones, 2012; Kieffer, 2012; T. D. Little, 2013; Preacher et al., 2008; Ross, 2005; Stoolmiller, 2008; Ullman & Bentler, 2004). The process, like that in MLM, can be thought of as consisting of two stages. First, parameters of individual growth trajectories (i.e., intercepts and slopes) and differences in these trajectories, if any, are estimated and described. Second, predictor variables are used to explain variability in individual growth trajectories (Duncan et al., 2006; Preacher et al., 2008; Ullman & Bentler, 2004).

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

However, unlike MLM, LGCM posits the existence of continuous underlying (i.e., latent) individual growth trajectories that give rise to the repeated measures; the observed repeated measures are considered (fallible) indicators of these latent trajectories (Bollen & Curran, 2006; Duncan et al., 2006; Henry & Slater, 2008; Hertzog & Nesselroade, 2003; Jones, 2012; Preacher et al., 2008). As Bollen and Curran (2006) explained, the interest in LGCM is not specifically in the observed repeated measures, but in the unobserved (latent) trajectory factors that lead to (and are inferred from) the observed repeated measures. For any dependent variable, two latent factors are estimated that jointly define the underlying trajectory in LGCM: an intercept factor that represents the mean and variance of the collection of individual initial levels on the dependent variable and a slope factor that represents the mean and variance of the collection of the individual trajectory slopes or growth rates over time (Bollen & Curran, 2006; Curran et al., 2010; Duncan et al., 2006; Hertzog & Nesselroade, 2003; Ho et al., 2006; Jones, 2012; Preacher et al., 2008). To use SEM terminology, a simple LGCM with one latent variable is a two-factor multiple-indicator model; the two factors being the intercept and slope and the multiple indicators being the repeated measures of the latent variable for every individual over all time points (Bollen & Curran, 2006).

Because of their similarities, LGCM can address all the questions that MLM can address concerning the characteristics (i.e., direction, rate, and shape) of the mean trajectory of the entire group, individual differences in growth trajectories, the contribution of various variables to variability in individual growth trajectories, and the relationship between initial level and rate of change (Bollen & Curran, 2006; Grimm & Ram, 2012; Jones, 2012; Kieffer, 2012; Preacher et al., 2008; Ross, 2005; Stoolmiller, 2008; Ullman & Bentler, 2004). Additionally, LGCM can be used to examine more complex models and questions about change than MLM, such as whether change parameters (e.g., change rate) predict other outcome variables and whether change over time in one variable is related to change over time in another variable (Preacher et al., 2008; Ross, 2005). LGCM, therefore, allows examining not only antecedents (or predictors) of change, but also the consequences or effects of change on other variables and change processes (Kieffer, 2012; T. D. Little, 2013; Preacher et al., 2008; Ross, 2005). LGCM also permits modeling change in more than one latent or measured variable across time simultaneously to determine the extent to which their development is interrelated, specifying and testing different types of growth trajectories (e.g., linear, nonlinear), and examining moderating and mediating effects on individual and group growth trajectories (Curran et al., 2010; Duncan et al., 2006; Jones, 2012; Kieffer, 2012; T. D. Little, 2013; Preacher et al., 2008; Ross, 2005).

LGCM provides a more powerful tool for the construction, estimation, and testing of complex growth models than any other single statistical model for longitudinal data (Duncan et al., 2006; Preacher et al., 2008). It has all the advantages of SEM and almost the same advantages as MLM. Consequently, compared to other approaches, LGCM can be used to examine more complex relationships between preexisting predictors, growth, and future outcomes of growth, including testing for mediation or indirect effects that flow from preexisting predictors to growth and then to some distal outcome (Duncan et al., 2006; Kieffer, 2012;

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Ross, 2005; Singer & Willett, 2003; Stoolmiller, 2008). The flexibility and versatility of LGCM allow for greater correspondence between the statistical model and the theory under investigation (Duncan et al., 2006; Preacher et al., 2008). However, because it takes a multivariate approach to RM within subjects, unequal number and irregular spacing of measurement occasions across individuals can be problematic and may require complicated procedures to deal with in LGCM, especially in small samples (Duncan et al., 2006; Hox, 2002; Preacher et al., 2008; Stoolmiller, 2008). It is worth noting also that the autoregressive and LGCM models have recently been combined into the autoregressive latent trajectory model, resulting in a flexible and hybrid model that can be used to examine stability and change simultaneously (Bollen & Curran, 2004; Bollen & Zimmer, 2010; Curran & Bollen, 2001). Additionally, as a result of recent developments in software and estimation techniques, the few differences between MLM and LGCM are rapidly disappearing (Curran, 2003; Preacher et al., 2008).

Five studies were identified that used LGCM to examine L2 development over time (Kieffer, 2011, 2012; Kieffer & Lesaux, 2012a; Lervag & Aukrust, 2010; Ross, 2005). All five studies are complex and include a large amount of information about the theory guiding the study, study design, and analysis approach. These studies involved three to seven measurement occasions; varied in length from 18 months to 9 years; and included 77 to 9,189 learners. Ross (2005) was perhaps the first to use LGCM to research L2 development over time, and he provides an accessible discussion and a concrete example of the use of LGCM to analyze longitudinal L2 data. Ross examined the effects of the use of formative assessment and conventional summative assessment practices on language learning growth in an 8-year longitudinal study. Participants were 2,215 Japanese EFL learners who were assessed either with mainly conventional end-of-term summative assessments and tests or with formative assessments throughout a four-semester EFL program. Grade point averages (GPA) were obtained for each semester, and TOEFL reading and listening scores were collected three times for each student. Various LGCM models were specified and tested in order to examine growth in GPA and L2 proficiency over time; whether and how changes in achievement over time influence L2 proficiency growth; differences in achievement and L2 proficiency growth across groups; and growth in L2 listening and reading abilities over time. For example, the growth trajectories for achievement and L2 proficiency were modeled in parallel to examine the relationship between initial individual differences in achievement (achievement intercept) and initial L2 proficiency (proficiency intercept) and the effect of initial L2 proficiency level on the trajectory of changes in achievement over time. As Ross noted, LGCM provided a powerful tool to examine the effect of formative assessment on achievement growth and the influence of achievement growth on L2 proficiency growth. Ross found that there were differences between the summative and formative cohorts in the way changes in achievement influence growth in L2 proficiency: The effect of formative assessment was substantive for listening but small for reading; both groups attained the same level of L2 reading proficiency by the end of the program; and a gap in L2 listening proficiency growth was slowly but consistently widening in favor of the formative cohort.

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Lervag and Aukrust (2010) used LGCM to examine the form and rate of growth of reading comprehension and the role of decoding and vocabulary skills in the growth of reading comprehension among approximately 300 L1 and L2 learners of Norwegian. Participants were tested in terms of their reading comprehension on four occasions over a period of 18 months (i.e., every 6 months). Measures of word decoding, vocabulary breadth, vocabulary definitions, nonverbal abilities, and maternal education were collected at time 1. Several LGCM models were used to compare the starting level (i.e., intercept) and the growth rates (i.e., slopes) of reading comprehension in the L1 and L2 learners. Mothers' educational level, nonverbal abilities, and vocabulary and decoding skills were included as predictors of the initial levels and rate of growth of reading comprehension over time. The findings indicated that the L1 learners had better initial reading comprehension skills and faster growth in these skills over time; beginning reading comprehension skills were predicted by vocabulary and decoding skills in both L1 and L2 learners, but only vocabulary predicted the subsequent growth of reading comprehension skills; and vocabulary seemed to be a stronger predictor of growth in reading comprehension among the L2 learners than among the L1 learners.

TIME SERIES APPROACHES

Time series research refers to both time series designs (TSDs) and TSA (Barlow, Nock, & Hersen, 2009; Krishef, 1991). TSDs are designs that involve frequent observation or measurement of a dependent variable for an individual or a group of individuals over time with time being the major independent variable, but other independent variables may be included as well (Barlow et al., 2009; Krishef, 1991; Tabachnick & Fidell, 2007). As Menard (2002) noted, "technically, whenever we have data on the same variable for the same case for two or more [time points], we have a time series, but the term time series analysis is usually reserved to describe a family of methods for analyzing relatively long time series for a single case at a time" (p. 67; cf. Menard, 2008a; Tabachnick & Fidell, 2007). Once a time series design has been established and data have been collected, TSA is concerned with the statistical analysis of the data that has been obtained (Krishef, 1991).

Time series designs are one type of single-case research designs. Single-case (or *N*-of-1 or single-subject) research designs are a diverse set of designs in which the phenomenon of interest is studied using a single participant or a small group of participants (Ishida, 2004; Mellow, 2013; Mellow, Reeder, & Forster, 1996; Nock, Michel, & Photos, 2007).[12] These designs can be nonexperimental (e.g., case study) or experimental. Single-case experimental designs are designs in which an individual is observed multiple times before, during, and after an intervention in order to establish the effects of the intervention on a single individual (Barlow et al., 2009; Ishida, 2004; Krishef, 1991; McCormick, 1995; Mellow, 2013; Mellow et al., 1996). These designs differ from group experimental designs in several key characteristics: (a) only a single person is studied rather than a group of people; (b) measures of the dependent variable are gathered

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

frequently and over time before, throughout, and after the intervention to obtain multiple observations for each condition and capture change as it is occurring; and (c) control procedures, rather than control groups, are used (Barlow et al., 2009; Ishida, 2004; Krishef, 1991; McCormick, 1995; Mellow, 2013; Mellow et al., 1996; Nock et al., 2007). Control procedures involve observing the participant during an initial (baseline) phase in which all conditions are carefully controlled so that they differ from the intervention phase in only one way: Intervention procedures are absent (McCormick, 1995). The effectiveness of the intervention is tested by evaluating the presence and degree of change in the dependent variable that occurs with the introduction of the intervention (Nock et al., 2007). Single-case experimental studies can include more than one dependent variable and more than one participant.

Typically, measures of the dependent variable from single-case experimental designs are plotted against time in a line graph; visual inspection and comparison of data points in the baseline phase to data points during and/or after the intervention are then used to determine if a change has occurred as well as the trend, magnitude, latency, and reliability of the change[13] (Busk & Marascuilo, 1992; Ishida, 2004; Krishef, 1991; McCormick, 1995; Mellow, 2013; Mellow et al., 1996; Nock et al., 2007). However, visual inspection has been criticized as being subjective and not sufficiently sensitive or reliable (Barlow et al., 2009; Busk & Marascuilo, 1992; Krishef, 1991; McCormick, 1995; Nock et al., 2007). Consequently, some authors have recommended using statistical analyses instead of, or to supplement, visual inspection of data plots. The use of statistics in single-case experimental designs remains controversial, however (Barlow et al., 2009; Busk & Marascuilo, 1992; Kamil, 1995; Krishef, 1991; Nock et al., 2007).

Data from single-case designs have two characteristics that can be problematic for statistical analysis: (a) serial dependency or autocorrelation, that is, correlation of observations of the same variable over time, and (b) small number of cases and/or observations (Busk & Marascuilo, 1992; Kamil, 1995; Krishef, 1991; Nock et al., 2007). Autocorrelation may or may not be present in a single-case experiment (Barlow et al., 2009; Kamil, 1995). When it is not present, most parametric tests (e.g., RM ANOVA) can be used (Barlow et al., 2009; Kamil, 1995). When the data are significantly autocorrelated, nonparametric tests can be used instead, because parametric statistics assume that observations are independent (Barlow et al., 2009; Kamil, 1995; Nock et al., 2007). Another alternative is to use TSA, which allows the examination of intervention impact and the magnitude and trend (e.g., linear, quadratic) of change in a dependent variable as well as comparing trends across conditions or phases, while accounting for autocorrelation (Barlow et al., 2009; Kamil, 1995; Nock et al., 2007; Ruspini, 2002; Tabachnick & Fidell, 2007; Yanovitzky & VanLear, 2008). As Ruspini (2002) explained, TSA is based on "the assumption that the data consist of a systematic pattern and random noise that makes identifying the pattern difficult"; most TSA techniques involve "some form of filtering out noise in order to make the pattern clearer" (p. 109).[14] TSA, however, requires a very large number of observations (50 or more) for unambiguous model identification and assumes that the series is equally spaced at meaningful intervals (Barlow et al., 2009; Busk & Marascuilo,

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

1992; Menard, 2008a; Tabachnick & Fidell, 2007). TSA is highly technical and discussion of the topic is beyond the scope of this article.

Single-case experimental designs are powerful, flexible, and efficient tools for evaluating the effectiveness of interventions compared to large-sample designs (Ishida, 2004; Mellow, 2013; Mellow et al., 1996; Nock et al., 2007). The multiple assessments used in these designs allow the evaluation of how and why individuals change over time and whether, when and how an intervention affects individual growth trajectories (Ishida, 2004; Nock et al., 2007). As Mellow (2013) has argued, these designs allow applied linguists to investigate the nature of language use or acquisition over time, variability in language development and use over time, and the effect of a specific variable or intervention on language use or acquisition (cf. Ishida, 2004; Van Geert & van Dijk, 2002). However, as Ortega and Iberri-Shea (2005) noted, single-case research designs remain underused in L2 research.

Five studies were identified that used a single-case research design to examine L2 development over time (Crossley, Salsbury, & McNamara, 2009, 2010; Ishida, 2004; Mellow, 2008; Salsbury, Crossley, & McNamara, 2011). These studies involved one to six participants who were observed on 6 to 15 different occasions. Ishida (2004) used a single-case experimental design with four concurrent replications (i.e., $N = 4$) to examine the effects of intensive recasting on accuracy, variability over time, and durability in L2 learners' use of the Japanese aspectual form *-te i-(ru)*. Each learner participated in eight 30-minute, one-on-one conversation sessions: Two sessions established a baseline (i.e., pretests); four sessions involved receiving intensive recasts (i.e., treatment period); and two sessions served as posttests (i.e., eight observation occasions). Two learners also participated in another 30-minute session that served as a delayed posttest after 7 weeks. In addition to graphs, several statistical analyses were employed to interpret the findings. RM ANOVA was used to compare mean accuracy in the pretest, intervention, and posttest phases. The Pearson product-moment correlation coefficient was used to investigate the degree of relationship between the total number of recasts and any increases in accuracy. Ishida found that the overall accuracy rate increased significantly after the recasting treatment began, and the high accuracy rate was retained both in the immediate and delayed posttest sessions. Moreover, the accuracy increase significantly correlated with the number of recasts provided to the learners during the treatment period.

In three related studies, Crossley et al. (2009, 2010) and Salsbury et al. (2011) examined changes in vocabulary knowledge and use (e.g., depth of word knowledge, word frequency use) in the spontaneous speech data of six adult English as a second language (ESL) learners in an intensive English program. Each participant was interviewed every 2 weeks for a total of 18 sessions over a 1-year period. Speech samples from six sessions for each participant were then analyzed in terms of various lexical features. RM ANOVAs, with each of the measures of lexical features as a dependent variable and the temporal interval between observations as the independent variable, were conducted to examine whether length of time studying English is related to development in different aspects of vocabulary knowledge and use. Findings from the three studies indicated that, as learners spent time studying the English language, their lexical diversity

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

increased and their productive vocabularies became more abstract, less context dependent, and more tightly associated over time.

EVENT HISTORY ANALYSIS

EHA, also known as survival or duration analysis, is a set of techniques to address questions concerning whether and when events (e.g., attaining a particular level of L2 proficiency, achieving a particular language milestone) occur, the extent to which individuals vary in the timing of event occurrence, and how variation in the timing of event occurrence is related to specific individual and contextual factors (Bornstein, Hah, Haynes, Manian, & Tamis-LeMonda, 2005; Ruspini, 2002; Singer & Willett, 2003; Steele, 2008; Taris, 2000). Data for EHA consist of information about the timing (e.g., dates) of changes in participants' states or conditions as well as the length of the time intervals between entry to and exit from specific states (Ruspini, 2002; Taris, 2000). That is, while other approaches treat time as a discrete variable (i.e., data are collected at particular points in time only, such as every week), in EHA, time is treated as a continuous variable and the focus is on the occurrence of particular events (e.g., transition from one discrete state to another; Ruspini, 2002; Singer & Willett, 2003). Data for EHA about the occurrence of events can be collected retrospectively or prospectively.

Recently, Bloomfield, Ross, Masters, Gynther, and O'Connell (2014) used EHA (and LGCM) to examine factors affecting the rate of attrition of L2 skills over time. Specifically, they conducted EHA on the listening, reading, and speaking test scores of a sample of more than 800 L2 learners who were tested two to seven times over 6 years in order to estimate the effects of amount of time between tests, participation in formal language courses, and initial test scores on the occurrence of score loss (i.e., obtaining lower scores on later tests compared to earlier tests) over time. The findings indicated that speaking was the most vulnerable skill, with a large proportion of participants showing loss in their test scores when the time between tests exceeded 2 years. Reading and listening scores were more stable, with a large proportion of participants maintaining their scores even when the time between tests exceeded 3 years. Rate of loss did not significantly differ based on initial test score or participation in formal language courses (see also Ross, 2014).[15]

ITEM RESPONSE THEORY

Finally, when examining change over time, IRT models can be used to identify variation over time that is due to changes in learners' abilities and separate it from changes due to changes in measurement characteristics, such as changes in item difficulty and rater severity over time (e.g., Bornstein et al., 2005; Endler & Bond, 2006; Fischer, 1976; Roberts & Ma, 2006; Wang, Wilson, & Adams, 1998; Wolfe & Chiu, 1999a, 1999b). Fischer (1976), for example, developed an IRT model for measuring individual and group change in the context of experimental pretest-posttest designs. The model distinguishes between two parts of change over time: change due to the effects of the intervention and change due to

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

other factors. Similarly, Wolfe and Chiu (1999a, 1999b) describe a multistep IRT approach to disentangling changes in person ability estimates over time from changes in item interpretation and difficulties in both experimental and nonexperimental research. However, no L2 studies were identified that used these approaches to examine L2 development over time.

CONCLUDING REMARKS

As the preceding discussion shows, there is currently a wide array of methods available for analyzing quantitative longitudinal L2 data. Traditional methods tend to focus on interindividual variation and to emphasize the establishment of average change patterns that apply to all individuals, while newer approaches focus on intraindividual variation and emphasize understanding change within the individual before establishing average patterns (Collins, 2006). Traditional techniques are thus limited in their ability to describe and explain individual change over time (Curran et al., 2010; Preacher et al., 2008). Advanced methods, particularly MLM and LGCM, provide powerful tools to examine both intraindividual change and interindividual variation in change and to better understand and model hypotheses concerning growth, individual variability in growth, and predictors of growth (Collins, 2006; Henry & Slater, 2008).

One major concern regarding the use of MLM and LGCM to analyze longitudinal L2 data relates to sample size. Modeling change over time in both approaches requires at least three repeated measures per individual. Generally, LGCM is best suited for modeling trends measured over a limited number (i.e., 3–6) of occasions per individual in large samples to obtain stable parameter estimates (Curran et al., 2010; Duncan et al., 2006; Preacher et al., 2008). Although new developments in LGCM allow the use of relatively small samples (e.g., $N = 60$, according to Tabachnick & Fidell, 2007), the more complex the model and the larger the number of variables, the larger the sample size required for fitting the model. As for MLM, Curran et al. (2010) noted that MLM has been successfully used with samples as small as $N = 22$ (e.g., Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991), although sample sizes approaching at least 100 are often preferred (cf. Maas & Hox, 2005).[16] When sample size is small, traditional techniques should be used. Additionally, traditional methods might be the only option when data requirements for more sophisticated analyses are not met, researchers lack the knowledge or resources to tackle more complicated analyses, or the simplicity of the hypothesis of interest makes the effort of using more complex analysis approaches not worth the trouble (Newsom, 2012).

In concluding this article, it is important to emphasize the central roles of theory, research design, and proper measurement in longitudinal L2 research. As Hertzog and Nesselroade (2003) cautioned, there is a risk that an emphasis on statistical advances can divert attention away from the problem of matching theoretical conceptions about change with design, measurement, and analysis decisions when studying change over time. No statistical technique or model can determine change and causal relationships apart from strong theory, solid research design, and valid measures of outcome variables (Embretson & Poggio, 2012; Selig & T. D. Little, 2012). The integration of theory, design, measurement,

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

and statistical analysis in longitudinal research, thus, becomes essential. Study design and data collection and analysis must be preceded by considering "the nature of the change phenomenon that is to be observed" (Collins, 2006, p. 507), and decisions concerning statistical analyses must be guided by substantive theory and an understanding of the underlying assumptions and limitations of statistical techniques (Bollen & Curran, 2006; Collins, 1991, 2006; Duncan et al., 2006; Hertzog & Nesselroade, 2003; Preacher et al., 2008). As Collins (1991, 2006) has repeatedly emphasized, ideal longitudinal research is "characterized by the seamless integration of at least three elements: (a) a well-articulated theoretical model of change observed using (b) a temporal design that affords a clear and detailed view of the process, with the resulting data analyzed by means of (c) a statistical model that is an operationalization of the theoretical model" (Collins, 2006, p. 506; cf. Hertzog & Nesselroade, 2003; Preacher et al., 2008).

Because they focus on individual change over time, newer statistical models provide a better match for current theories of change (Collins, 2006). As shown earlier, advanced longitudinal data analytic techniques greatly expanded the range of questions, hypotheses, and models that researchers can examine concerning L2 development over time and enable them to examine old questions in more appropriate ways. These methods also have several conceptual and practical advantages and fewer assumptions that are more realistic and can be empirically evaluated, compared to traditional techniques. Of course, these methods cannot compensate for deficiencies in theory, study design, or measurement procedures. However, when feasible and appropriate, these methods are more consistent with theory and can stimulate the development and empirical examination of more complex questions and models concerning L2 development and individual variability in L2 development over time than is possible with traditional techniques. It is hoped that this review will encourage more L2 researchers to use these methods more frequently in their longitudinal studies.

ACKNOWLEDGMENTS

Preparation of this review was supported by a Minor Research and Development Grant from the Faculty of Education at York University. I would like to thank Ali Hadidi for helping with locating relevant longitudinal L2 studies for this review and an anonymous *Annual Review of Applied Linguistics* reviewer for their comments on an earlier version of the article.

NOTES

1 *Outcome* and *dependent variable* are used interchangeably in this article. *Predictor* and *independent variable* are also used interchangeably.

2 The term *repeated* is sometimes used for both, but in this review, *repeated* refers only to cases where the same dependent variable is measured repeatedly and does not include cases where multiple (and different) dependent variables are measured for the same individual at one point in time.

3 Three conditions are required for interpreting the association between two variables in causal terms: "covariation (i.e., statistically significant association between the two variables), nonspuriousness (i.e., the association between the two variables must not be due to the effects of other variables), and temporal order (i.e., the cause must precede its effect in time)" (Taris, 2000, p. 3).

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

4 The word *growth* is used in longitudinal research (e.g., growth modeling) and in this article to refer to any kind of change over time including increasing, decreasing, and complex, nonlinear change patterns (Jones, 2012).
5 Listwise or casewise deletion involves excluding any participant whose data are incomplete.
6 Multiple imputations consist in replacing each missing value with a set of plausible values that are more or less suitable estimates based on nonmissing data to create multiple complete data sets which are then analyzed as if there were no missing data (Bijeleveld et al., 1998; Feng et al., 2012; Taris, 2000).
7 Sphericity refers to the assumption that correlations between all pairs of scores across occasions are equal, irrespective of the time interval between occasions (Tanguma, 1999; Twisk, 2003; Weinfurt, 2000). Violation of this assumption leads to spurious findings (Tanguma, 1999; Weinfurt, 2000).
8 When the outcome variable is categorical, the assumptions of parametric tests are not met, and/or the sample is small, nonparametric statistical tests (e.g., Wilcoxon signed rank sum test, the Friedman test) could be used instead to analyze longitudinal data (see Bijeleveld et al., 1998; Newsom, 2012; Twisk, 2003).
9 Time-varying predictors are variables whose value changes over time such as age, motivation, and L2 proficiency. In contrast, time-invariant predictors are variables that are constant across time such as student L1 and gender.
10 Kozaki and Ross (2011) refer to mediating effects throughout the paper, when what they examined were *moderating* effects. Currently, mediating effects cannot be examined in MLM, but moderating effects (referred to as cross-level interactions in MLM) can.
11 A latent variable is an unobserved construct or factor that is assessed indirectly through one or more measured (observed) variables (Ullman & Bentler, 2004).
12 While the discussion here focuses on single case studies, group averages can also be used in TSA to avoid some of the pitfalls of a series made up of scores with no variation around them.
13 Trend refers to the direction of change (e.g., increase or decrease across phases); latency refers to how long it takes to see a change in response; and reliability concerns whether the changed responses are consistently and dependently displayed, and whether they are maintained over time (McCormick, 1995; Nock et al., 2007).
14 TSA is widely used in economics and policy studies.
15 Another relevant study not included in this review, Bornstein et al. (2005), used EHA to examine children's early L1 development.
16 Huttenlocher et al. (1991) included 22 participants with eight measurement occasions for each participant.

ANNOTATED BIBLIOGRAPHY

Laursen, B., Little, T. D., & Card, N. A. (Eds.). (2012). *Handbook of developmental research methods*. New York, NY: Guilford.

The 41 chapters in this volume cover a wide range of considerations and issues in longitudinal research, including research design, measurement, data collection strategies, and various traditional and modern techniques for longitudinal data analysis, which are described in accessible, nontechnical language and illustrated using real examples. This is a valuable resource for anyone studying change over time.

Menard, S. (Ed.). (2008). *Handbook of longitudinal research: Design, measurement, and analysis*. New York, NY: Academic Press.

This comprehensive handbook discusses various conceptual, methodological, and practical considerations and issues concerning research design, measurement,

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

data missingness and attrition, experimental research, and quantitative and qualitative data analysis in longitudinal research. This is a valuable resource for both novice and experienced researchers engaged in longitudinal research.

Neuman, S. B., & McCormick, S. (Eds.). (1995). *Single-subject experimental research: Applications for literacy*. Newark, DL: International Reading Association.

The eight chapters in this volume provide guidelines for designing, implementing, and reporting single-case studies, describe various single-case research designs, and discuss the collection and analysis of qualitative and quantitative data in single-case research. Several examples of single-case literacy studies are described.

Newsom, J. T., Jones, R. N., & Hofer, S. M. (Eds.). (2012). *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences*. New York, NY: Routledge.

This 11-chapter volume provides accessible discussions and examples of various approaches to quantitative longitudinal data analysis, including ANOVA, MLM, SEM, and EHA. Each chapter discusses the conceptual, methodological, and statistical aspects of each data analysis approach and relevant software for data analysis, and provides a list of recommended readings and examples of studies that used the approach under focus. Other chapters in the volume discuss practices, considerations, and issues concerning study design, sampling, measurement, missing data, mediation analysis, and the analysis of categorical data in longitudinal research.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.

Singer and Willett provide a comprehensive discussion of various considerations and issues in the measurement and modeling of change and growth over time. The first part of the book provides an accessible discussion of the conceptual, statistical, and practical aspects of MLM and LGCM in the context of analyzing longitudinal data to study change over time. The second part focuses on EHA. Both parts include numerous step-by-step examples of longitudinal data analysis.

Weinfurt, K. P. (2000). Repeated measures analyses: ANOVA, ANCOVA, and HLM. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 317–361). Washington, DC: American Psychology Association.

This chapter provides very accessible discussion, comparisons, and examples concerning the use of gain scores, RM ANOVA, MANOVA, ANCOVA, and MLM for longitudinal data analysis. The chapter also provides guidelines concerning the use of RM ANOVA, MANOVA, ANCOVA, and ANOVA on gain scores in experimental and quasi-experimental studies with two measurement occasions (i.e., pretest and posttest).

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 269–304). Bingley, UK: JAI Press.

This chapter provides a detailed and accessible discussion of the supposed unreliability and invalidity of gain scores; challenges these criticisms; and provides advice and detailed guidelines concerning whether, when, and how to use gain scores when analyzing data from two occasions as well as how to address the limitations of gain scores. The chapter also discusses and makes recommendations concerning the

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

use of RM ANOVA, ANCOVA, and ANOVA on gain scores in experimental and quasi-experimental studies with two measurement occasions.

REFERENCES

Allison, P. A. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114.

Bae, J., & Lee, Y. (2012). Evaluating the development of children's writing ability in an EFL context. *Language Assessment Quarterly*, *9*, 348–374.

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1–39). New York, NY: Academic Press.

Barkaoui, K. (2013). An introduction to multilevel modeling in language assessment research. *Language Assessment Quarterly*, *10*, 241–273.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behaviour changes* (3rd ed.). New York, NY: Pearson.

Bast, J., & Reitsma, P. (1997). Matthew effects in reading: A comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research*, *32*, 135–167.

Bates, J. E., & Novosad, C. (2005). Measurement of individual difference constructs in child development, or taking aim at moving targets. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 103–122). Malden, MA: Blackwell.

Bijeleveld, G. C. J. H., van der Kamp, L. J. T., Mooijaart, A., van der Kloot, W. A., van der Leeden, R., & ven der Burg, E. (1998). *Longitudinal data analysis: Designs, models, and methods*. Thousand Oaks, CA: Sage.

Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. *Language Teaching Research*, *12*, 409–431.

Bitchener, J., & Knoch, U. (2010a). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics*, *31*, 193–214.

Bitchener, J., & Knoch, U. (2010b). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing*, *19*, 207–217.

Bloomfield, A., Ross, S., Masters, M., Gynther, K., & O'Connell, S. (2014). How does foreign language proficiency change over time? Results of data mining official test records. In J. Connor-Linton & L. W. Amoroso (Eds.), *Measured language: Quantitative studies of acquisition, assessment, and variation* (pp. 199–212). Washington, DC: Georgetown University Press.

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods and Research*, *32*, 336–383.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley-Interscience.

Bollen, K. A., & Zimmer, C. (2010). An overview of the autoregressive latent trajectory (ALT) model. In K. van Montfort, J. H. L. Oud, & A. Satorra (Eds.), *Longitudinal research with latent variables* (pp. 153–176). Amsterdam, The Netherlands: Springer.

Bontempo, D. E., Grouzet, F. M. E., & Hofer, S. M. (2012). Measurement issues in the analysis of within-person change. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences* (pp. 97–142). New York, NY: Routledge.

Bornstein, M. H., Hah, C., Haynes, O. M., Manian, N., & Tamis-LeMonda, C. S. (2005). New research methods in developmental science: Applications and illustrations. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 509–533). Malden, MA: Blackwell.

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 159–185). Hillsdale, NJ: Erlbaum.

Collins, L. M. (1991). Measurement in longitudinal research. In L. A. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 137- 148). Washington, DC: American Psychological Association.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*, 505–528.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*, 68–80.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*, 307–334.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*, 573–605.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*, 529–569.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 107–135). Washington, DC: American Psychological Association.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*, 121–136.

Dale, P. S., Harlaar, N., & Plomin, R. (2012). Nature and nurture in school-based second language achievement. *Language Learning*, *62*, 28–48.

de Vaus, D. (2001). *Research design in social research*. Los Angeles, CA: Sage.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). New York, NY: Psychology Press.

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, *6*, 74–96.

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. In *IELTS Research Reports* (Vol. 4, pp. 207–254). Manchester, UK: IELTS Australia and British Council.

Embretson, S. E., & Poggio, J. (2012). The impact of scaling and measurement methods on individual differences in growth. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 82–108). New York, NY: Guilford.

Endler, L. C., & Bond, T. G. (2006). Tracking cognitive development with the Rasch model. In X. Liu & W. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 74–109). Maple Grove, MN: JAM Press.

Feng, D., Cong, Z., & Silverstein, M. (2012). Missing data and attrition. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences* (pp. 71–96). New York, NY: Routledge.

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. M. DeGruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). New York, NY: Wiley.

Gilmore, A. (2011). "I prefer not text": Developing Japanese learners' communicative competence with authentic materials. *Language Learning*, *61*, 786–819.

Green, A. (2005). EAP study recommendations and score gains on the IELTS academic writing test. *Assessing Writing*, *10*, 44–60.

Grimm, K. J., & Ram, N. (2012). Growth curve modeling from a structural equation modeling perspective. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 411–431). New York, NY: Guilford.

Gustafsson, J. E. (2010). Longitudinal designs. In B. P. M. Creemers, L. Kyrikides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 77–101). New York, NY: Routledge.

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Gutierrez-Clellen, V., Simon-Cereijido, G., & Sweet, M. (2012). Predictors of second language acquisition in Latino children with specific language impairment. *American Journal of Speech-Language Pathology*, *21*, 64–77.

Hartmann, D. P. (2005). Assessing growth in longitudinal investigations: Selected measurement and design issues. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 319–339). Malden, MA: Blackwell.

Henry, K. L., & Slater, M. D. (2008). Assessing change and intraindividual variation: Longitudinal multilevel and structural equation modeling. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The Sage sourcebook of advanced data analysis methods for communication research* (pp. 55–87). Los Angeles, CA: Sage.

Hertzog, C., & Nesselroade, J. R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging*, *18*, 639–657.

Ho, H. Z., O'Farrell, L., Hoong, S., & You, S. (2006). Developmental research: Theory, method, design, and statistical analysis. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 207–225). Mahwah, NJ: Erlbaum.

Hofer, S. M., Thorvaldsson, V., & Piccinin, A. M. (2012). Foundational issues of design and measurement in developmental research. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 3–16). New York, NY: Guilford.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*, 236–248.

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*, 311–394.

Jones, R. N. (2012). Latent growth curve models. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences* (pp. 271–290). New York, NY: Routledge.

Kamil, M. L. (1995). Statistical analysis procedures for single-subject designs. In S. B. Neuman & S. McCormick (Eds.), *Single-subject experimental research: Applications for literacy* (pp. 84–103). Newark, DL: International Reading Association.

Kieffer, M. (2011). Converting trajectories: Reading growth in language minority learners and their classmates, kindergarten to grade 8. *American Educational Research Journal*, *48*, 1187–1225.

Kieffer, M. (2012). Early oral language and later reading development in Spanish-speaking English language learners: Evidence from a nine-year longitudinal study. *Journal of Applied Developmental Psychology*, *33*, 146–157.

Kieffer, M. J., & Lesaux, N. K. (2012a). Development of morphological awareness and vocabulary knowledge in Spanish-speaking language minority learners: A parallel process latent growth curve model. *Applied Psycholinguistics*, *33*, 23–54.

Kieffer, M. J., & Lesaux, N. K. (2012b). Effects of academic language instruction on relational and syntactic aspects of morphological awareness for sixth graders from linguistically diverse backgrounds. *The Elementary School Journal*, *112*, 519–545.

Kozaki, Y., & Ross, S. J. (2011). Contextual dynamics in foreign language learning motivation. *Language Learning*, *61*, 1328–1354.

Krishef, C. H. (1991). *Fundamental approaches to single subject design and analysis*. Malabar, FL: Krieger.

Kruk, R. S., & Reynolds, K. A. A. (2012). French immersion experience and reading skill development in at-risk readers. *Journal of Child Language*, *39*, 580–610.

Laursen, B., Little, T. D., & Card, N. A. (Eds.). (2012). *Handbook of developmental research methods*. New York, NY: Guilford.

Lervag, A., & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry*, *51*, 612–620.

Lissitz, R. W., Doran, H., Schafer, W. D., & Willhoft, J. (2006). Growth modeling, value added modeling, and linking: An introduction. In R. W. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 1–46). Maple Grove, MN: JAM Press.

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Little, R. J. A., & Rubin, R. D. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.
Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford.
Llosa, L. (2012). Assessing English language learners' progress: Longitudinal invariance of a standard-based classroom assessment of English proficiency. *Language Assessment Quarterly*, *9*, 331–347.
Lord, F. M. (1963). Elementary models for measuring change. In W. C. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison, WI: University of Wisconsin Press.
Luke, D. A. (2008). Multilevel growth curve analysis for quantitative outcomes. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 545–564). New York, NY: Academic Press.
Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*, 30–43.
Lyster, R., & Izquierdo, J. (2009). Prompts versus recasts in dyadic interaction. *Language Learning*, *59*, 453–498.
Maas, C. J., & Hox, J. J. (2005). Sufficient sample size for multilevel modeling. *Methodology*, *1*, 86–92.
Matsumura, S. (2003). Modeling the relationships among interlanguage pragmatic development, L2 proficiency, and exposure to L2. *Applied Linguistics*, *24*, 465–491.
McCall, M., & Hauser, C. (2007). Item response theory and longitudinal modeling: The real world is less complicated than we fear. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in schools* (pp. 143–174). Maple Grove, MN: JAM Press.
McCormick, S. (1995). What is single-subject experimental research? In S. B. Neuman & S. McCormick (Eds.), *Single-subject experimental research: Applications for literacy* (pp. 1–31). Newark, DL: International Reading Association.
Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement*, *23*, 87–89.
Mellow, J. D. (2008). The emergence of complex syntax: A longitudinal study of the ESL development of dependence resolution. *Lingua*, *118*, 499–521.
Mellow, J. D. (2013). Time series. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Malden, MA: Blackwell.
Mellow, J. D., Reeder, K., & Forster, E. (1996). Using time-series research designs to investigate the effects of instructing on SLA. *Studies in Second Language Acquisition*, *18*, 325–350.
Menard, S. (2002). *Longitudinal research* (2nd ed.). Thousand Oaks, CA: Sage.
Menard, S. (2008a). A brief introduction to time series analysis. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 579–600). New York, NY: Academic Press.
Menard, S. (2008b). Introduction: Longitudinal research design and analysis. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 3–12). New York, NY: Academic Press.
Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 109–126). New York, NY: Guilford.
Newsom, J. T. (2012). Basic longitudinal analysis approaches for continuous and categorical variables. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences* (pp. 143–179). New York, NY: Routledge.
Newsom, J. T., Jones, R. N., & Hofer, S. M. (Eds.). (2012). *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences*. New York, NY: Routledge.
Nock, M. K., Michel, B. D., & Photos, V. I. (2007). Single-case research designs. In D. McKay (Ed.), *Handbook of research methods in abnormal and clinical psychology* (pp. 337–350). Thousand Oaks, CA: Sage.
Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26–45.

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Ortega, L., & Byrnes, H. (2008a). The longitudinal study of advanced L2 capacities: An introduction. In L. Ortega & H Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 3–20). New York, NY: Routledge.

Ortega, L., & Byrnes, H. (2008b). Theorizing advancedness, setting up the longitudinal research agenda. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 281–300). New York, NY: Routledge.

Pitts, S. C., Prost, J. H., & Winters, J. J. (2005). Quasi-experimental designs in developmental research: Design and analysis considerations. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 80–100). Malden, MA: Blackwell.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Los Angeles, CA: Sage.

Rees, J., & Klapper, J. (2008). Issues in the quantitative longitudinal measurement of second language progress in the study abroad context. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 89–105). New York, NY: Routledge.

Revesz, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning*, *62*, 93–132.

Roberts, J. S., & Ma, Q. (2006). IRT models for the assessment of change across repeated measurements. In R. W. Lissitz, (Ed.), *Longitudinal and value added models of student performance* (pp. 100–129). Maple Grove, MN: JAM Press.

Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottma (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Erlbaum.

Ross, S. J. (2005). The impact of assessment method on foreign language proficiency growth. *Applied Linguistics*, *26*, 317–342.

Ross, S. (2014). Subjectivity and efficiency in language assessment: Explorations of a compensatory rating approach. In J. Connor-Linton & L. W. Amoroso (Eds.), *Measured language: Quantitative studies of acquisition, assessment, and variation* (pp. 51–62). Washington, DC: Georgetown University Press.

Ruspini, E. (2002). *Introduction to longitudinal research*. London, UK: Routledge.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*, 343–360.

Sato, M., & Lyster, R. (2012). Peer interaction and corrective feedback for accuracy and fluency development: Monitoring, practice, and proceduralization. *Studies in Second Language Acquisition*, *34*, 591–626.

Schoonen, R., van Gelderen, A., Stoel, R. D., Hulstijn, J., & de Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning*, *61*, 31–79.

Selig, J. P., & Little, T. D. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 265–278). New York, NY: Guilford.

Serrano, R., Llanes, À., & Tragant, E. (2011). Analyzing the effect of context of second language learning: Domestic intensive and semi-intensive courses vs. study abroad in Europe. *System*, *39*, 133–143.

Shaw, B. A., & Liang, J. (2012). Growth models with multiple regression. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences* (pp. 217–242). New York, NY: Routledge.

Shneyderman, A., & Abella, R. (2009). The effects of the extended foreign language programs on Spanish-language proficiency and academic achievement in English. *Bilingual Research Journal*, *32*, 241–259.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.

Skiba, R., Dittmar, N., & Bressem, J. (2008). Planning, collecting, exploring, and archiving longitudinal L2 data. In L. Ortega & H Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 21–88). New York, NY: Routledge.

Steele, F. (2008). Multilevel models for longitudinal data. *Journal of Royal Statistical Society*, *171*, 5–19.

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

Stoolmiller, M. (2008). Latent growth curve models. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 523–544). New York, NY: Academic Press.
Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing*, *18*, 103–118.
Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Pearson.
Tanguma, J. (1999). Analyzing repeated measures designs using univariate and multivariate methods: A primer. In B. Thompson (Ed.), *Advances in social science methodology: Vol. 5* (pp. 233–250). Bingley, UK: JAI Press.
Taris, T. W. (2000). *A primer in longitudinal data analysis*. Thousand Oaks, CA: Sage.
Taris, T. W. (2008). Reliability issues in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 139–152). New York, NY: Academic Press.
Tong, F., Irby, B. J., Lara-Alecio, R., Yoon, M., & Mathes, P. G. (2010). Hispanic English learners' responses to longitudinal English instructional intervention and the effect of gender: A multilevel analysis. *Elementary School Journal*, *110*, 542–566.
Twisk, J. W. R. (2003). *Applied longitudinal data analysis for epidemiology: A practical guide*. Cambridge, UK: Cambridge University Press.
Twisk, J. W. R. (2008). Causal inference in longitudinal experimental research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 279–293). New York, NY: Academic Press.
Ullman, J. B., & Bentler, P. M. (2004). Structural equation modeling. In M. Hardy & A. Bryman (Eds.), *The handbook of data analysis* (pp. 431–458). Los Angeles, CA: Sage.
Usborne, E., Caouette, J., Qumaaluk, Q., & Taylor, D. M. (2009). Bilingual education in an Aboriginal context: Examining the transfer of language skills from Inuktitut to English or French. *International Journal of Bilingual Education and Bilingualism*, *12*, 667–684.
Vandergrift, L., & Tafaghodtari, M. H. (2010). Teaching L2 learners how to listen does make a difference: An empirical study. *Language Learning*, *60*, 470–497.
Van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development*, *25*, 340–374.
van Gelderen, A., Schoonen, R., Stoel, R., de Glopper, K., & Hulstijn, J. (2007). Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology*, *99*, 477–491.
Verhoeven, L. (1994). Transfer in bilingual development: The linguistic interdependence hypothesis revisited. *Language Learning*, *44*, 381–415.
Verhoeven, L., & Leeuwe, J. (2012). The simple view of second language reading throughout the primary grades. *Reading and Writing*, *25*, 1805–1818.
Wang, W., Wilson, M., & Adams, R. J. (1998). Measuring individual differences in change with multidimensional Rasch models. *Journal of Outcome Measurement*, *2*, 240–265.
Weinfurt, K. P. (2000). Repeated measures analyses: ANOVA, ANCOVA, and HLM. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 317–361). Washington, DC: American Psychology Association.
Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.
Wolfe, E. W., & Chiu, C. W. T. (1999a). Measuring change across multiple occasions using the Rasch rating scale model. *Journal of Outcome Measurement*, *3*, 360–381.
Wolfe, E. W., & Chiu, C. W. T. (1999b). Measuring pretest-posttest change with a Rasch rating scale model. *Journal of Outcome Measurement*, *3*, 134–161.
Yang, Y., & Lyster, R. (2010). Effects of form-focused practice and feedback on Chinese EFL learners' acquisition of regular and irregular past tense forms. *Studies in Second Language Acquisition*, *32*, 235–263.
Yanovitzky, I., & VanLear, A. (2008). Time series analysis: Traditional and contemporary approaches. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The Sage sourcebook of*

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press

advanced data analysis methods for communication research (pp. 89–124). Los Angeles, CA: Sage.

Yasuda, S. (2011). Genre-based tasks in foreign language writing: Developing writers' genre awareness, linguistic knowledge, and writing competence.*Journal of Second Language Writing*, *20*, 111–133.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 269–304). Bingley, UK: JAI Press.

https://doi.org/10.1017/S0267190514000105 Published online by Cambridge University Press